Legendre Polynomials for Uniform Mixture Detection

Rajan Shankar
School of Mathematics and Statistics, University of Sydney
Supervised by Michael Stewart

Contents

Uniform mixture detection problem

Uniform mixture detection problem

What is a mixture model?

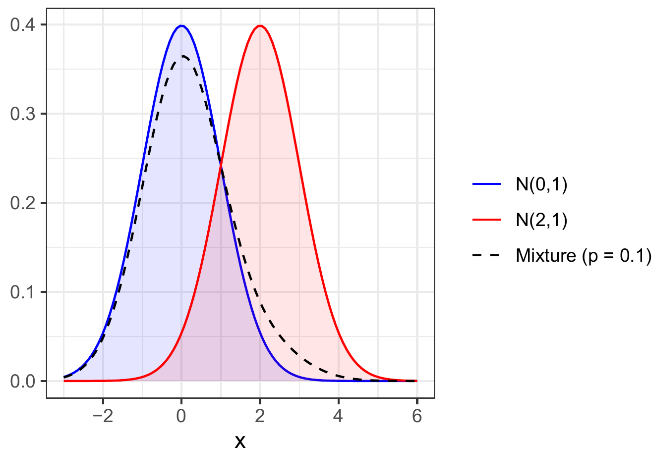
- Probability density function (pdf):

$$(1-p)f(x) + pg(x)$$

- Cumulative distribution function (CDF):

$$(1-p)F(x) + pG(x)$$

- Example:



What is a uniform mixture model?

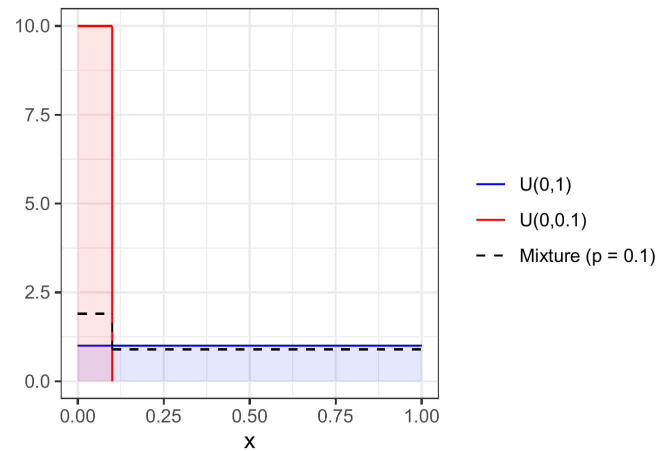
- Probability density function (pdf):

$$1-p+pg(x)$$

- Cumulative distribution function (CDF):

$$(1-p)x + pG(x)$$

- Example:



Uniform mixture detection problem

As a model for p-values

- If the null hypothesis of a test is *actually* true, then we expect the p-value to follow a $U(0, 1)$ distribution
- Now, consider a scenario where we need to conduct many identical tests
- What if, in some small proportion of cases, the null hypothesis is *actually* false?
- We can model this scenario with a uniform mixture model!

Uniform mixture detection problem

Motivational example

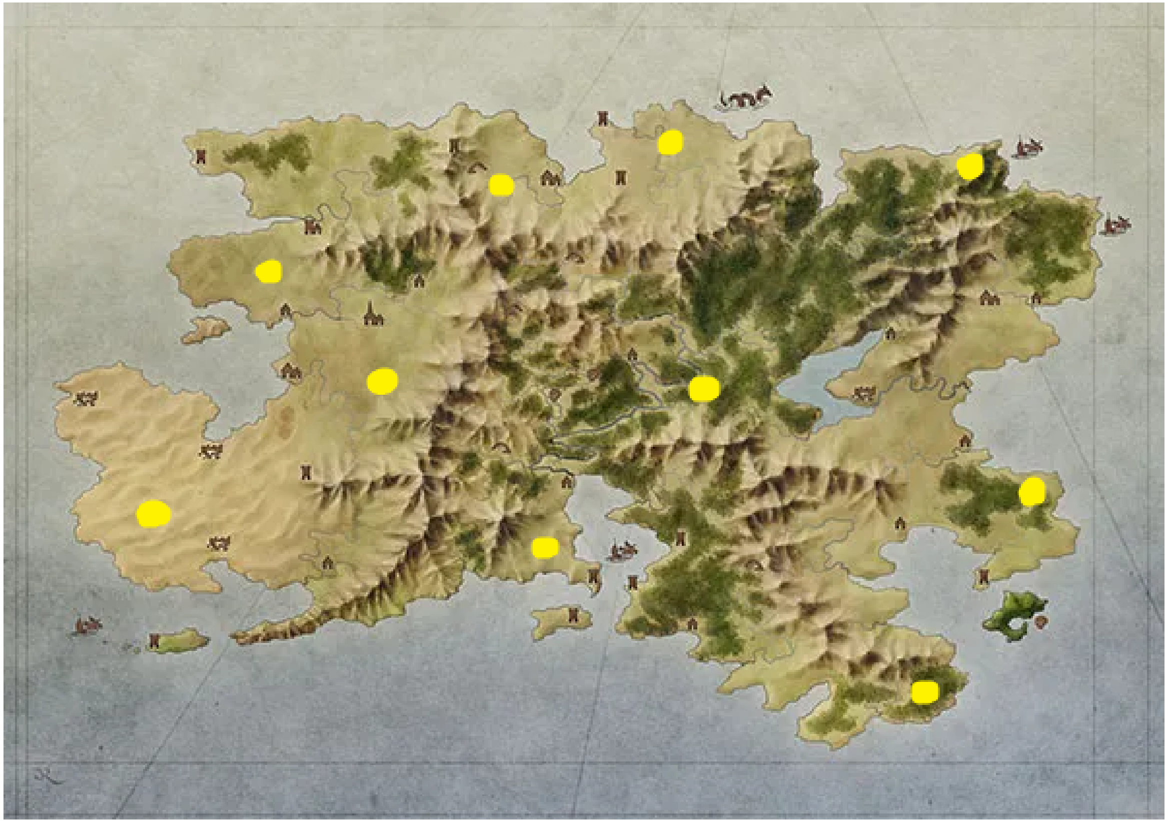
- Donoho and Jin, 2004: Higher criticism for detecting sparse heterogenous mixtures
- Bioweapon
- Causes an increase of some chemical in the blood
- Blood test returns a p-value

Uniform mixture detection problem

Motivational example

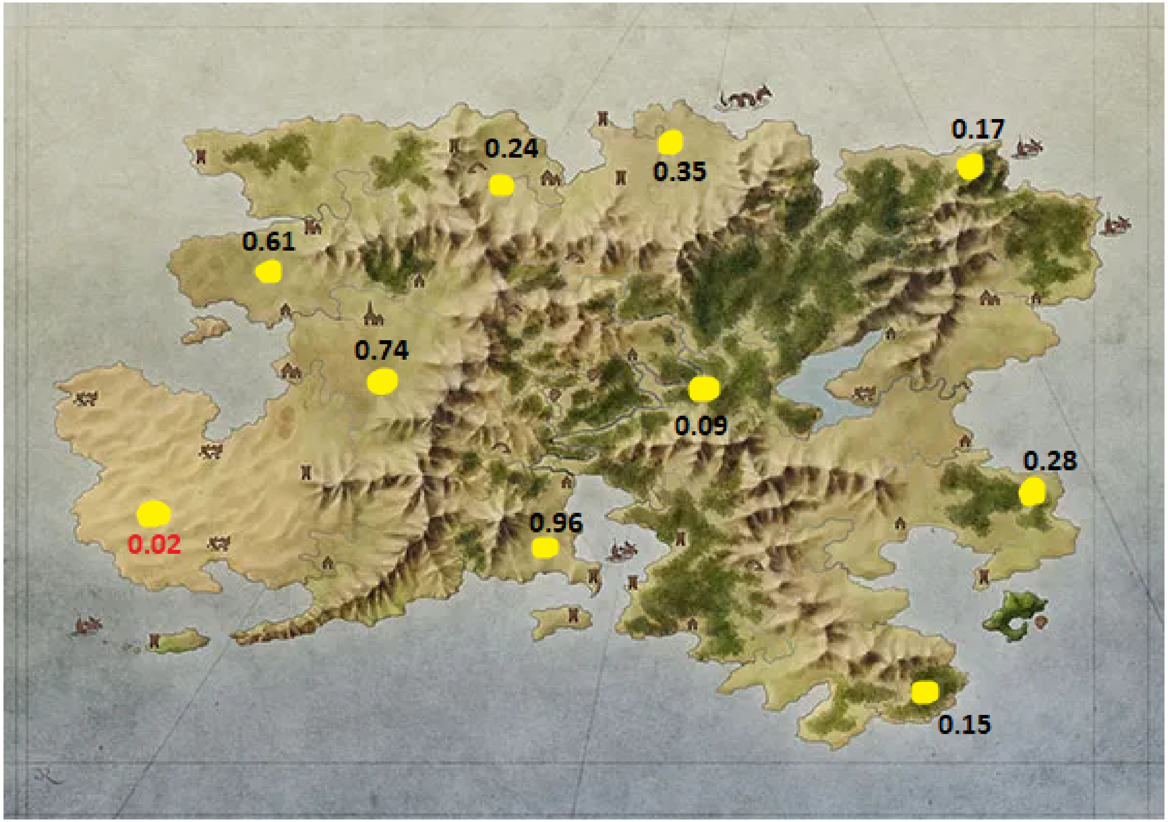
Uniform mixture detection problem

Motivational example



Uniform mixture detection problem

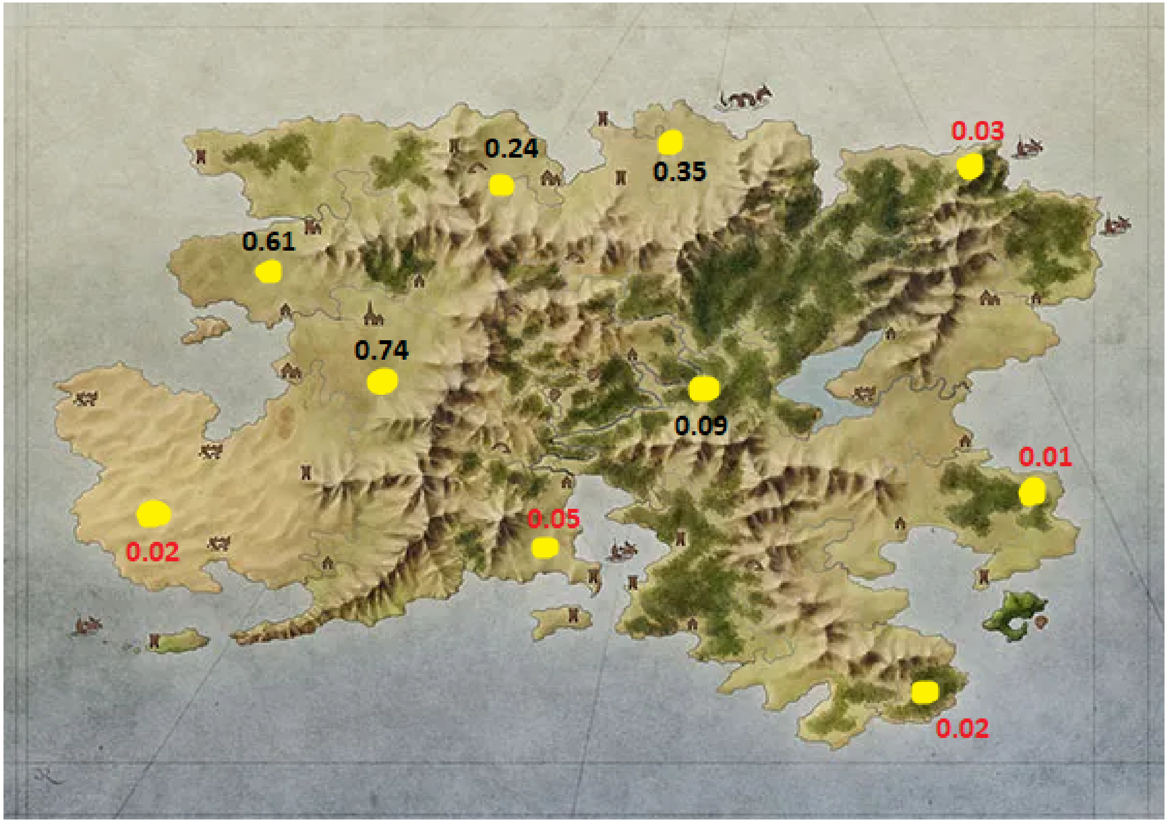
Motivational example



- Is there evidence that the bioweapon has affected some of the population?

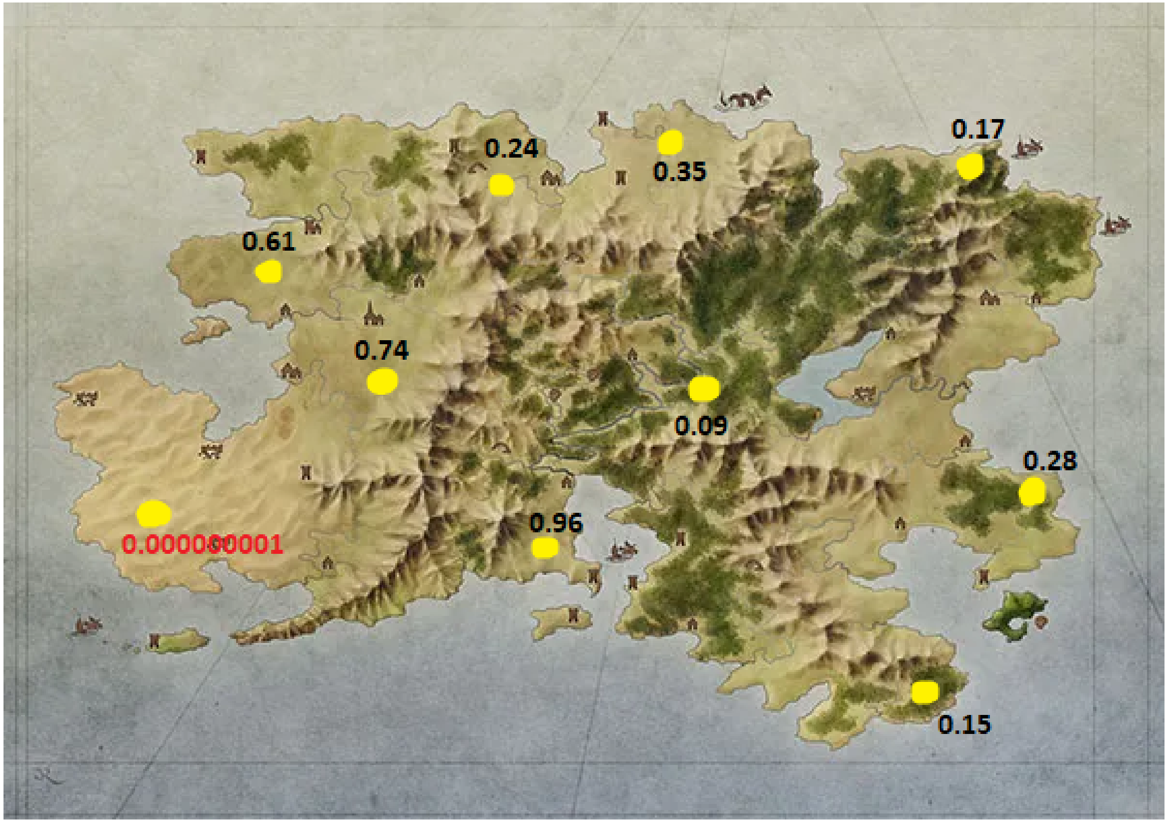
Uniform mixture detection problem

Motivational example



- Is there evidence that the bioweapon has affected some of the population?

Motivational example



- Is there evidence that the bioweapon has affected some of the population?

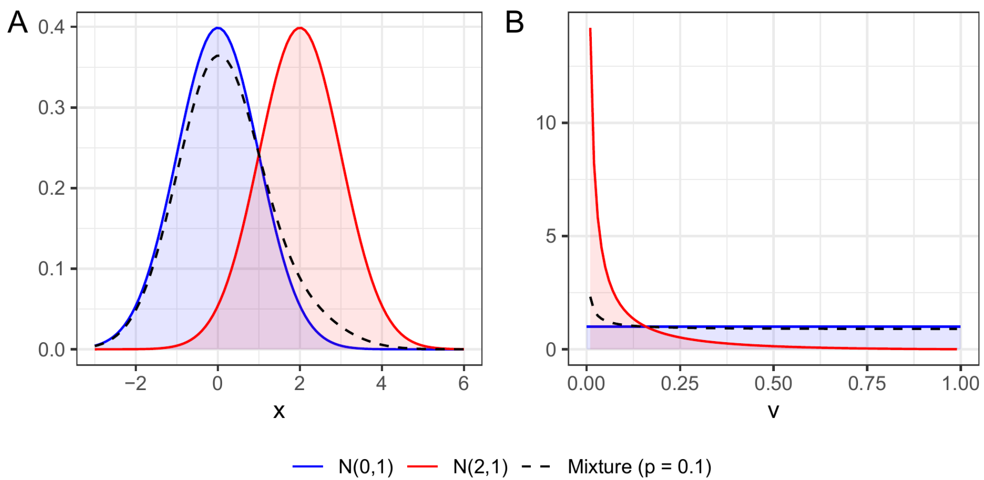
Uniform mixture detection problem

Transforming to a uniform mixture model

- We can transform a mixture model into a *uniform* mixture model
- Let the CDF of X be $(1-p)F(x) + pG(x)$ and let $V = 1 - F(X)$. Then, the CDF of V is:

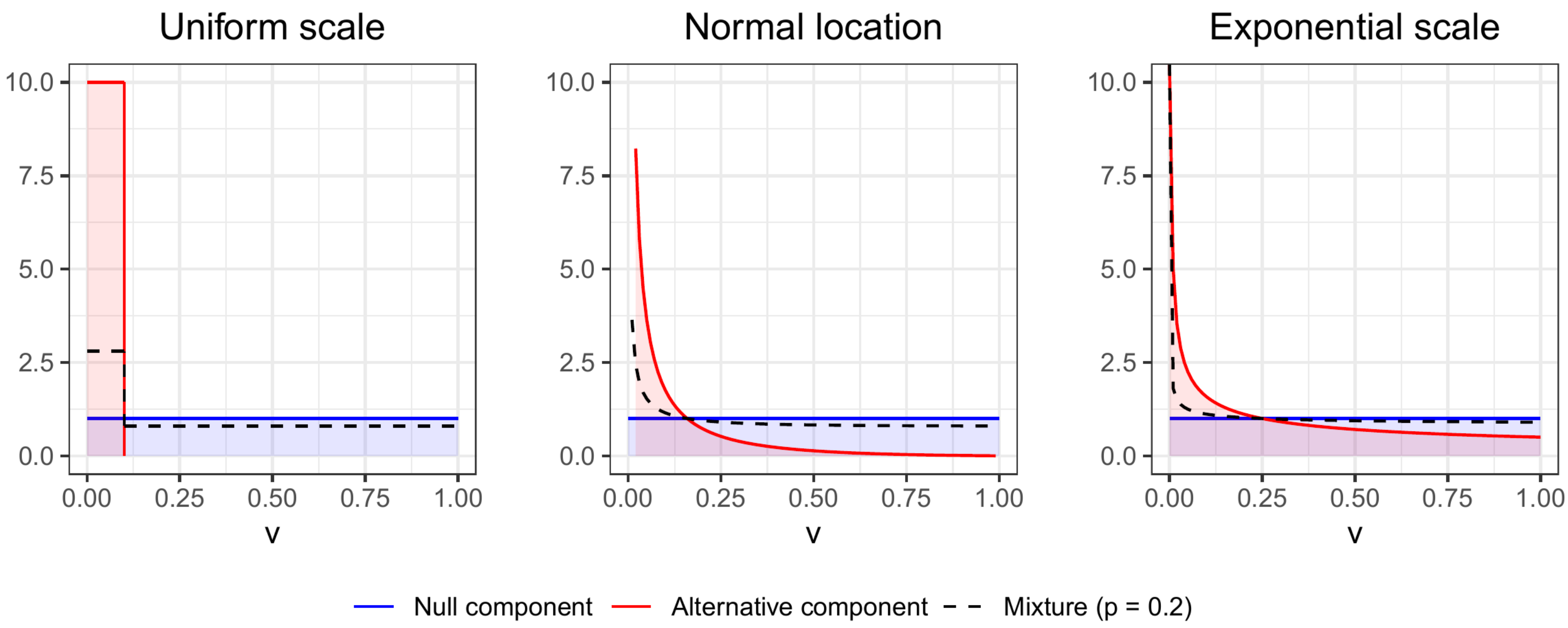
$$(1-p)v + p\left(1 - G\left[F^{-1}(1-v)\right]\right)$$

- Differentiating this gives a density of the form $1 - p + p(\ldots)$
- Example:



Uniform mixture detection problem

Common uniform mixture models



Legendre polynomials

Legendre polynomials

Definition

- The k^{th} Legendre polynomial is the coefficient of t^k in the power series expansion of the generating function given by:

$$\frac{1}{\sqrt{1-2xt+t^2}} = \sum_{n=0}^{\infty} P_n(x)t^n$$

Shifted Legendre polynomials

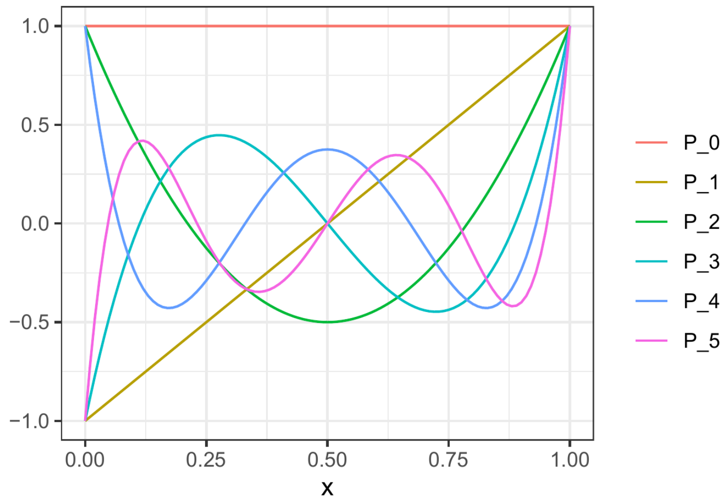
- P_k is defined on the interval $[-1, 1]$
- Uniform mixture models are defined on the interval $[0, 1]$
- Apply the following shift:

$$\tilde{P}_k(x) := P_k(2x-1)$$

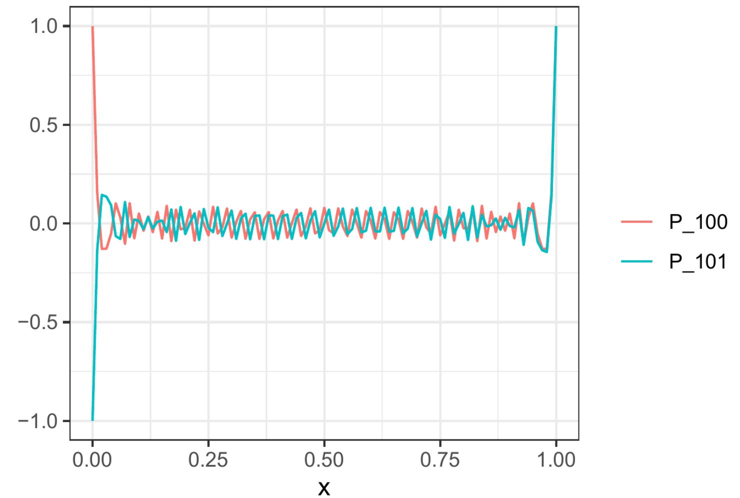
Legendre polynomials

Visualisation

- First few polynomials



- Higher-degree polynomials



Legendre polynomials

Orthogonal system

- Let $\tilde{P}_k$ and $\tilde{P}_j$ be shifted Legendre polynomials of degree k and j respectively. Then:

$$\langle \tilde{P}_k, \tilde{P}_j \rangle := \int_0^1 \tilde{P}_k(x)\tilde{P}_j(x)\,dx = \begin{cases} 0 & \text{if } k \neq j \\ \frac{1}{2k+1} & \text{if } k = j \end{cases}$$

Uncorrelated property

- If $X \sim U(0,1)$, then $\tilde{P}_k(X)$ and $\tilde{P}_j(X)$ are uncorrelated for $k \neq j$. Proof:

$$E\left[\tilde{P}_k(X)\tilde{P}_j(X)\right] = \int_0^1 \tilde{P}_k(x)\tilde{P}_j(x)\,dx = 0$$

$$E\left[\tilde{P}_k(X)\right] = \int_0^1 \tilde{P}_k(x)\,dx = \int_0^1 \tilde{P}_k(x)\tilde{P}_0(x)\,dx = 0$$

$$\mathrm{Cov}\left[\tilde{P}_k(X), \tilde{P}_j(X)\right] = \underbrace{E\left[\tilde{P}_k(X)\tilde{P}_j(X)\right]}_{=0} - \underbrace{E\left[\tilde{P}_k(X)\right]}_{=0}\underbrace{E\left[\tilde{P}_j(X)\right]}_{=0} = 0$$

Legendre polynomials

Use in uniform mixture detection

- Model data via the density $1 - p + pg(x)$
- Hypotheses are $H_0 : p = 0, H_1 : p \neq 0$
- If H_0 is actually true, i.e. $X \sim U(0,1)$, then we expect $\tilde{P}_k(X) = 0$ for all $k \geq 1$
- If H_0 is actually false, then we might expect $\tilde{P}_k(X)$ to be some other value
- Think of each polynomial as measuring the deviation from H_0 in a certain 'abstract' direction

$S_{n,k}$ statistic

$S_{n,k}$ statistic

Definition

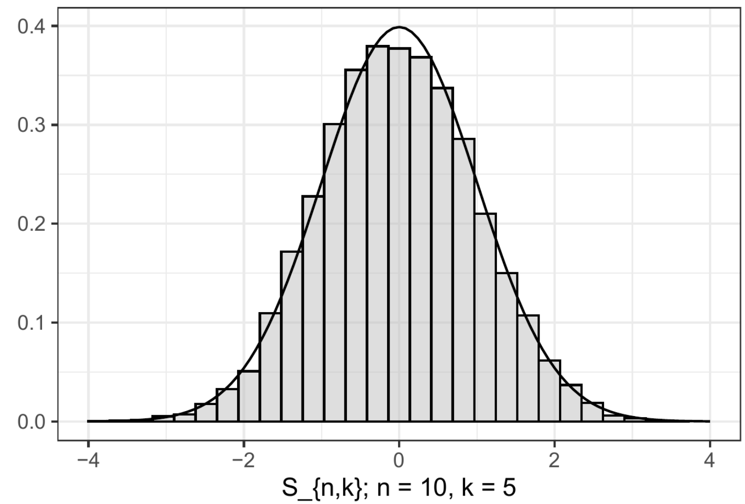
- Let $x_i, i = 1, \ldots, n$ be a sample of data. Then:

$$S_{n,k} := \sqrt{\frac{2k+1}{n}} \sum_{i=1}^{n} \tilde{P}_k(x_i)$$

- Simply a sum of $\tilde{P}_k$ over the sample, standardised to have mean 0 and variance 1
- Interpreted as the amount in which the data supports the abstract direction that $\tilde{P}_k$ is measuring

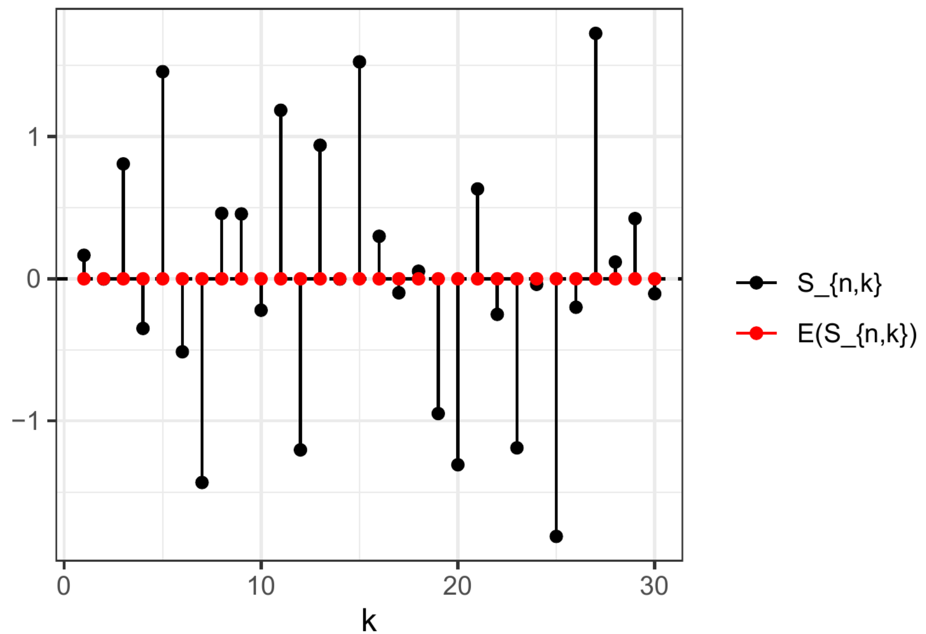
Asymptotic normality

- Central limit theorem kicks in quickly



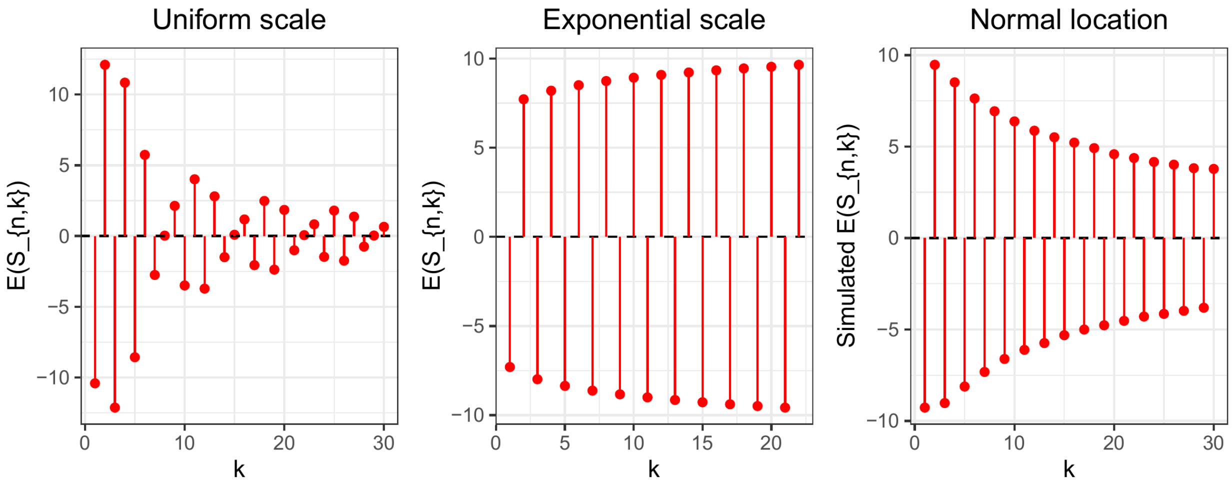
Spectrum diagrams

- Plot of $S_{n,k}$ vs k under a specific mixture model
- Under $H_0 : p = 0$, we expect $S_{n,k} = 0$ for all $k \geq 1$:



$S_{n,k}$ statistic

- Under $H_1 : p \neq 0$, we expect $S_{n,k}$ to follow the patterns below for our three different mixture models:



- These plots can be used to inspire test statistics
- It is surprising that we are able to theoretically derive $E(S_{n,k})$ for the uniform scale and exponential scale mixture models
- Example (uniform scale):

$$E(S_{n,k}) = \sqrt{n(2k+1)} \cdot p \sum_{\ell=0}^{k} (-1)^{k+\ell} \binom{k}{\ell} \binom{k+\ell}{\ell} \frac{\theta^\ell}{\ell+1}$$

Constructing test statistics

Constructing test statistics

Statistical power

- One of the main themes in mathematical statistics is to analyse the notion of power
- Power depends on the statistical test and the alternative hypothesis
- It is defined as:

$$P(\text{reject } H_0 \mid H_1 \text{ is true})$$

- We like tests that have high power across common alternative hypotheses

Constructing test statistics

The test statistics

- Our spectrum diagrams inspire our construction of test statistics
- Sum of squares:

$$\sum_{k=1}^{K} S_{n,k}^2$$

- Threshold sum of squares:

$$\sum_{k=1}^{K} S_{n,k}^2 \cdot 1_{\{|S_{n,k}| \geq c\}}$$

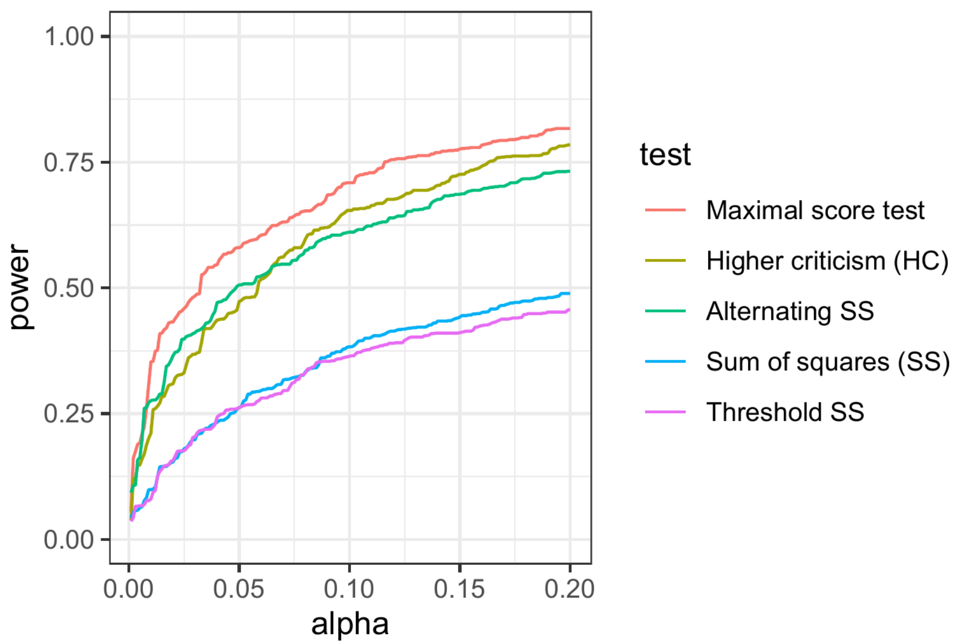
- Alternating sum of squares:

$$\sum_{k=1}^{K} \max\{0, (-1)^k \cdot S_{n,k}\}^2$$

Constructing test statistics

Simple power analysis

- A ROC curve is a plot of power vs significance level
- We simulate power under the normal location mixture model for a variety of tests:



Takeaways

Takeaways

What we looked at

- Uniform mixture detection problem
- Legendre polynomials
- $S_{n,k}$ spectrum diagrams and constructing test statistics

Questions my research looks at

- What does the spectrum diagram look like under different mixture models?
- How does the performance of tests based on the Legendre polynomials compare to other tests in the literature?
- What rate should K grow at as the sample size n increases?

End