05 - Anomaly Detection

SYS 4582/6018 | Spring 2019

05-anomaly.pdf

Contents

1 Anomaly Detection Intro

1.1 Required R Packages

We will be using the R packages of:

- `tidyverse` for data manipulation and visualization
- `mclust` for model-based clustering

1.2 Anomaly Detection

Anomaly Detection: The identification of unusual observations. Statistically, this means finding observations that come from a different distribution that the *normal* or *usual* observations.

1. **Goodness of Fit (GOF)**
   - Tests if data conform to a given distribution (or distributional family)
2. **Two-Sample Tests (A/B Testing)**
   - Tests if two datasets come from the same distribution
   - Often simplified to test if one group has a larger mean than the other
3. **Outlier Detection**
   - Tests if a single observation or small set of observations come from the same distribution as the rest of the data
4. **Hotspot Detection**
   - Identification of regions that have *unusually* high density
5. **Outbreak Detection**
   - A sequential method that repeatedly tests for a change in an event (i.e., point process) distribution
   - Focus on determining *if* a change occurred, and then determining *when* it occurred
   - For outbreak detection, the changes of interest are those that conform to an expected *outbreak pattern*

1.3 Example #1: Benford's Distribution

State/Territory	Real or Faked Area (km^2)	
Afghanistan	645,807	796,467
Albania	28,748	9,943
Algeria	2,381,741	3,168,262
American Samoa	197	301
Andorra	464	577
Anguilla	96	82
Antigua and Barbuda	442	949
Argentina	2,777,409	4,021,545
Armenia	29,743	54,159
Aruba	193	367
Australia	7,682,557	6,563,132
Austria	83,858	64,154
Azerbaijan	86,530	71,661
Bahamas	13,962	9,125
Bahrain	694	755
Bangladesh	142,615	347,722
Barbados	431	818
Belgium	30,518	47,123
Belize	22,965	20,648
Benin	112,620	97,768
...	...	...

Table from Fewster (2009) A Simple Explanation of Benford's Law, *The American Statistician*, 63, 1, pp 26–32

- Someone that fakes numbers, say on a financial statement, may be tempted to use a random number generator

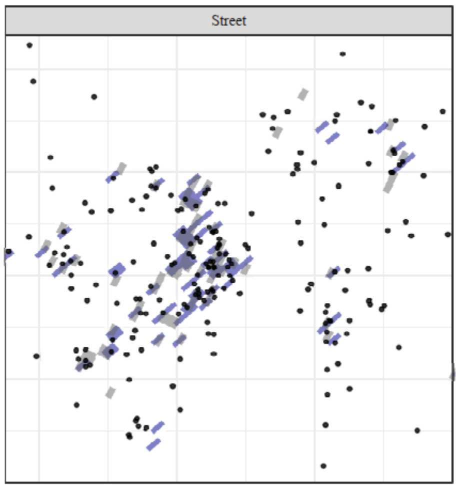
  - But watch out for Benford's Law
- Note on terminology:
  - Law = probability distribution
  - Anomalous Numbers = Random numbers (no known relationship)
- Benford's PMF:

$$\Pr(\text{first digit} = x) = \log_{10}\left(1 + \frac{1}{x}\right) \quad \text{for } x = 1, 2, \ldots, 9$$

1.4 Example #2: Crime Hotspots

In 2017, the National Institute of Justice (NIJ) held a Crime Forecasting Challenge

- Predict the crime hotspots for 4 crime types (burglary, street crime, motor vehicle theft, all types) for 5 forecasting windows (1 week ahead, ..., 3 months ahead)



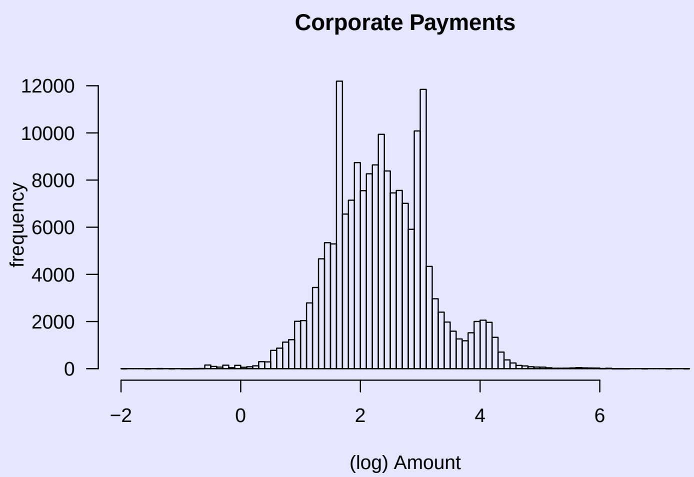
2 Goodness of Fit Testing

Tests if data conform to a given distribution (or distributional family).

Your Turn #1

Mark Nigrini provides a dataset of the 2010 payments from a division of a West Coast utility company https://www.nigrini.com/BenfordsLaw/CorporatePaymentsData.xlsx.

```
## Download xlsx file to local machine, then load into R:
## Notice that the data is in the "Data" sheet
library(readxl)
corp_payments = read_excel("../topics/anomaly/data/CorporatePaymentsData.xlsx",
                           sheet="Data") %>%
  filter(Amount > 0)   # only consider positive payments
```




```r
#-- Extract the first digit
first = stringr::str_extract(corp_payments$Amount, "[1-9]") %>% factor(levels=1:9)
```

Does the first digit of the `Amount` column conform to Benford's distribution?

2.1 GOF Hypothesis Test

- Let $D = (X_1, X_2, \ldots, X_n)$ be the observed random variables (i.e., the data).
- Let $\mathcal{H}_0$ be the *null* hypothesis
- Choose a *test statistic* $T = T(X_1, \ldots, X_n)$ that is a function of the observed data
  - T is a *random variable*; it has a distribution.
  - Let $t = T(x_1, \ldots, x_n)$ be the *observed* value of the test statistic
  - Common to adjust the test statistic so that *extreme* means large values of T
- The p-value is *the probability that chance alone would produce a test statistic as extreme as the observed test statistic if the null hypothesis is true*
  - E.g., p-value $= \Pr(T \geq t | \mathcal{H}_0)$
- Think of T or the p-value as the *evidence against* the null hypothesis
  - Its common to set a threshold (e.g., p-value $\leq .05$) and *reject* the null hypothesis when this threshold is crossed.
  - This is a form of *outlier detection*. Reject null if t_{obs} is an *outlier*; that is t_{obs} is from a different distribution that what is specified in $\mathcal{H}_0$.
- To calculate a p-value, we need to know/estimate the *distribution* of $T|\mathcal{H}_0$!
  - Even if we don't know the distribution of T under the null, we can often approximate it using simulation (Monte Carlo)

2.1.1 Example: one sample t-test

- In 2012, the Obama administration issued new rules on the fuel efficiency requirements for new cars and trucks by 2025.
  - The fleetwise fuel efficiency requirement is 54.5 mpg
- Suppose a car maker in 2025 designed a car to get an average fuel efficiency of 54.5 mpg.
  - Also, they think the fuel efficiency will be Normally

- The government officials randomly tested $n = 9$ cars. They got a sample mean of $\bar{x} = 53.0$ and sample standard deviation of $s = 2.5$.

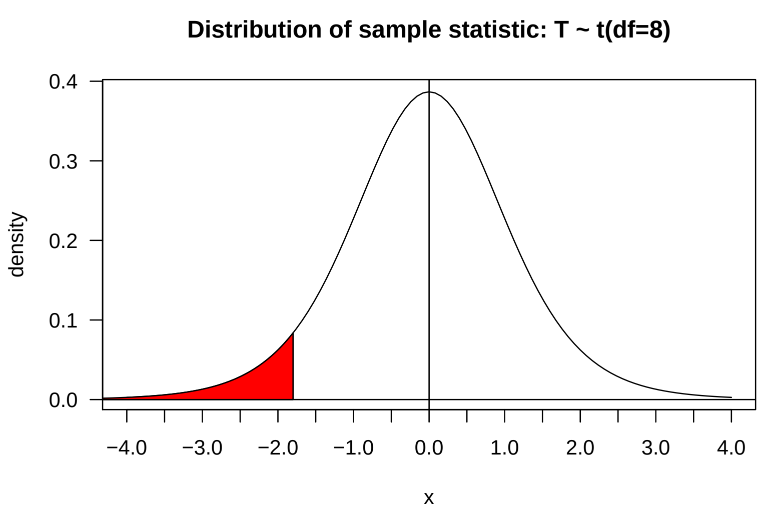
Your Turn #2

Is this enough evidence to conclude the car manufacturer failed to meet the requirements? Or can the results be attributed to chance fluctuation?

- Null Hypothesis:
  - the mpg come from a Normal distribution
  - mean of $\mu_0 = 54.5$
  - independent
  - standard deviation, σ is unknown
  - $X|\mathcal{H}_0 \overset{\text{iid}}{\sim} \mathcal{N}(\mu = 54.5, \sigma)$
- Let the test statistic be

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{53.0 - 54.5}{2.5/\sqrt{9}} = -1.80$$

- Under the null, T has a t-distribution with $df = n - 1$
  - Shoutout to William Sealy Gosset, a.k.a *Student*



- $\Pr(T \leq -1.8|\mathcal{H}_0) =$`pt((53.0-54.5)/(2.5/3), df=8)` $= 0.055$

2.1.2 Alternative Hypotheses

- The choice of test statistic depends on the expected deviations from the null
  - That is, we can come up with *better* test statistic if we know what sort of deviations from $\mathcal{H}_0$ are expected.
  - *better* meaning more *power* to correctly reject the null

2.2 Test Statistics

- Going back to the original question about the diamond prices conforming to Benford's distribution, we can state the *null hypothesis* formally:

$$\mathcal{H}_0 : X \overset{\text{iid}}{\sim} Benf$$

- X is the *first* digit(s)

- $Benf$ stands for Benford's distribution for the first digit(s). This has pmf:

$$f(x) = \log_{10}\left(1 + \frac{1}{x}\right)$$

  - Note: there are no parameters to estimate!

- A generic *alternative hypothesis* is:

$$\mathcal{H}_1 : X \not\sim Benf$$

- R code for a Benford pmf

```
#-- pmf for Benford's distribution
dbenford <- function(x) log10(1 + 1/x)

#-- first digit
dbenford(1:9)
#> [1] 0.30103 0.17609 0.12494 0.09691 0.07918 0.06695 0.05799 0.05115 0.04576

#-- first two digits
expand.grid(first=1:9, second=0:9) %>%
  mutate(two = paste0(first, second) %>% as.integer) %>%
  mutate(f = dbenford(two)) %>%
  select(first, second, f) %>%
  spread(second, f) %>%
  knitr::kable(digits=3)
```

first	0	1	2	3	4	5	6	7	8	9
1	0.041	0.038	0.035	0.032	0.030	0.028	0.026	0.025	0.023	0.022
2	0.021	0.020	0.019	0.018	0.018	0.017	0.016	0.016	0.015	0.015
3	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011
4	0.011	0.010	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009
5	0.009	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.007	0.007
6	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.006	0.006	0.006
7	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.005
8	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
9	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004

2.2.1 χ^2 Test Statistic

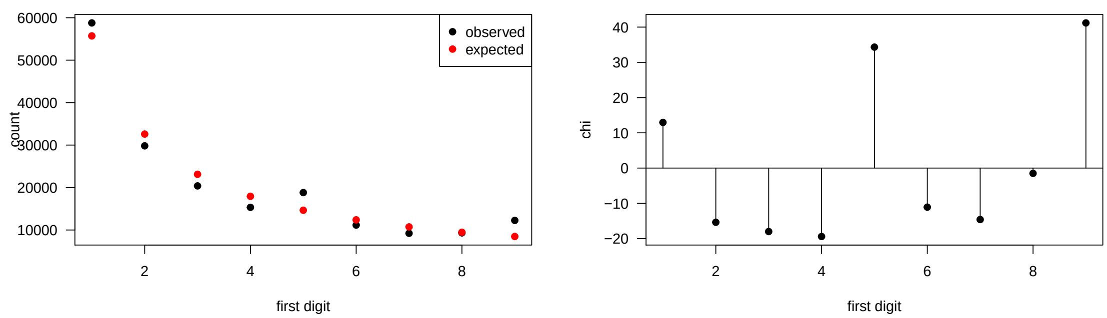
- The Pearson's χ^2 (chi-squared) test statistic is commonly used in goodness-of-fit testing.
- It requires the data to be *discrete* or *categorical*
  - Continuous data can be *binned*

$$\chi^2 = \sum_{j=1}^{J} \frac{(y_j - E_j)^2}{E_j} = \sum_{j=1}^{J} \frac{(y_j - np_j)^2}{np_j}$$

- J: number of categories or possible values
- y_j: *observed* count in category j
- E_j: *expected* count (i.e., under $\mathcal{H}_0$) in category j
- Let $n = \sum_{j=1}^{J} y_j$ be the total number of observations

- p_j: the proportion of events under the null (i.e., $\Pr(X = j|\mathcal{H}_0)$)
- Asymptotically, χ^2 converges to a chi-squared distribution with $J - 1$ degrees of freedom
- R code for corporate payments data

```
#-- Get counts
Y = table(first) %>% as.integer   # ensure first is factor with properly ordered
n = length(first)                 # number of observations

#-- chi-squared
n = length(first)     # number of observations
E = n*dbenford(1:9)   # expected count vector
chi = (Y-E)/sqrt(E)   # vector of deviations
(chisq = sum(chi^2)) # chi-squared test statistic
#> [1] 4317
```



- Note: there is a build-in R function `chisq.test()` which does these calculations

```
chisq.test(Y, p=dbenford(1:9))$statistic
#> X-squared
#>      4317
```

2.2.2 Likelihood Ratio Test Statistic

- When an *alternative hypothesis* can be specified, the log-likelihood ratio test statistic is commonly used in goodness-of-fit testing.
- The general binary hypothesis formulation is:
  - $\mathcal{H}_0 : X \sim f_0(X)$
  - $\mathcal{H}_1 : X \sim f_1(X)$
- The *likelihood ratio* is:

$$\begin{aligned} LR &= \frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)} \\ &= \prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)} \qquad \text{if } X\text{'s are iid} \end{aligned}$$

- The *log-likelihood ratio*, when the observations are iid, becomes:

$$\begin{aligned} llr &= \sum_{i=1}^{n} \log \frac{f_1(X_i)}{f_0(X_i)} \\ &= \sum_{i=1}^{n} \log f_1(X_i) - \sum_{i=1}^{n} \log f_0(X_i) \end{aligned}$$

- The hypotheses for the diamonds data:
  - $\mathcal{H}_0 : X \overset{\text{iid}}{\sim} Benf$
  - $\mathcal{H}_1 : X \overset{\text{iid}}{\sim} Cat(p_1, p_2, ..., p_9)$ where $\{p_k\}$ do *not* match Benford's probabilities.
- There are many reasonable choices for setting $\mathcal{H}_1$ parameters $(p_1, \ldots, p_9)$
  - *Discrete Uniform*: $p_1 = \ldots = p_9 = 1/9$
  - *MLE*: $\hat{p}_k = y_j/n$

Your Turn #3

Write out the log-likelihood for the two alternatives.

Using MLE, the llr = 2031.83.

- Note: $2 \times llr$ has an asymptotic chi-squared distribution (same as the chi-squared test statistic).

2.3 Testing

- The two-test statistics, χ^2 and llr, provide evidence against $\mathcal{H}_0$.
- But how do we know if these values are *unusually* large? Perhaps by chance alone (i.e. the data sample we observed) the values are as large as they are.
- We can answer this with a solid probabilistic statement if we knew the *distribution of the test statistic under the null hypothesis*
  - We don't often know this exactly, but there are often good approximations that hold as the sample size grows (asymptotically).
- There are two primary options:
  1. Use an asymptotic distribution (e.g., the chi-squared distribution)
  2. Use Monte Carlo simulation

2.3.1 Monte Carlo Simulation

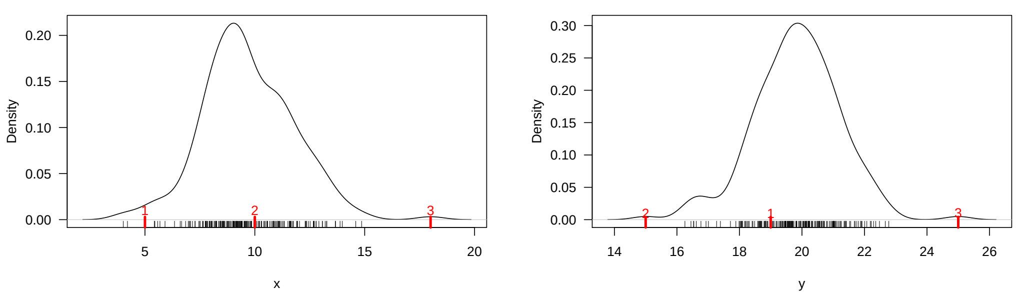
- If we can sample from the null hypothesis, then it becomes straightforward to estimate the distribution of *any* test statistic and consequently, p-values.
- **Monte Carlo based GOF Test**
  1. Calculate the test statistic for the original observation, t
  2. Generate M samples from the null distribution
     - $\{Y_1, \ldots, Y_M\}$
  3. For each sample, calculate the test statistic T^*
     - $\{T_1^*, \ldots, T_M^*\}$

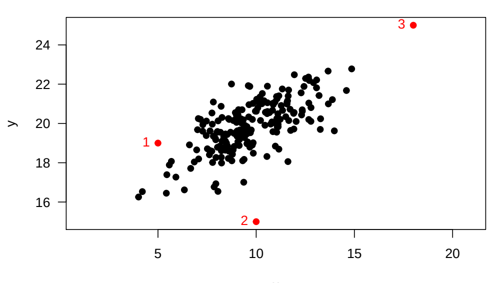
4. p-value = $\frac{1+\text{number of } T^*\text{'s greater than or equal to } t}{M+1}$

```
#-- Monte Carlo based p-value
n = length(x)
#> Error in eval(expr, envir, enclos): object 'x' not found

M = 1000                           # number of simulations
stat.chisq = numeric(M)            # initialize statistic
for(m in 1:M){
  #- generate observation under the null of Benford
  y.sim = rmultinom(1, size=n, prob=dbenford(1:9))
  #- calculate test statistic
  stat.chisq[m] = chisq.test(y.sim, p=dbenford(1:9))$statistic
}

#- calculate p-values
(1 + sum(stat.chisq > chisq)) / (M+1)   # chi-square p-value
#> [1] 0.000999
```

3 Outlier Detection

- Outlier Detection tests if a single observation or small set of observations come from the same distribution as the rest of the data
- Are any of the *red* points outliers?





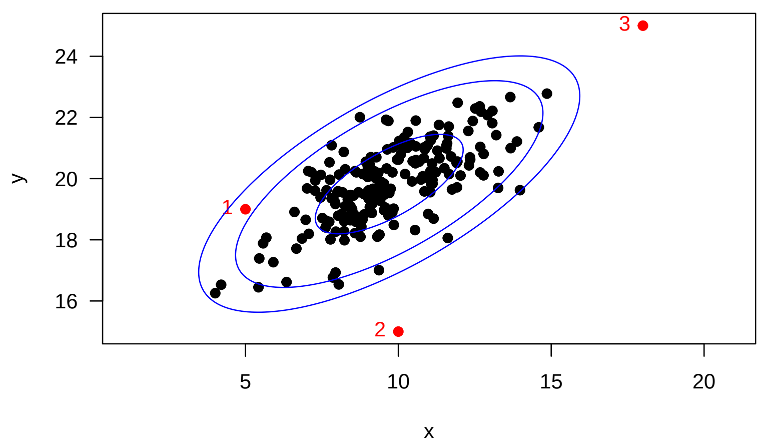
3.1 Distance based approach

- One approach, with strong connections to clustering, is to calculate the *distance* from an observation to the centroid
  - This assumes the "normal" observations are from a unimodal distribution
  - To allow for an ellipse shape (orientation) and different spreads in each dimension, use the squared *Mahalanobis Distance*

$$D_i^2 = (\mathbf{x_i} - \bar{\mathbf{x}})^\mathsf{T} \hat{\Sigma}^{-1} (\mathbf{x_i} - \bar{\mathbf{x}})$$

  - Note that $\bar{\mathbf{x}}$ and $\hat{\Sigma}$ are estimated from all of the data.
- Estimated Parameters:
  - $\bar{x} = (9.70, 19.82)$
  - $\hat{\Sigma} = (4.2171, 1.9482, 1.9482, 1.906)$

obs	x	y	Dsq
1	5	19	7.056
2	10	15	24.471
3	18	25	18.123



3.2 Likelihood Based Approach

- From the plots, it appears the a 2D Gaussian/Normal model could be a decent approximation to the distribution of the non-outlier observations
- We can use this to calculate the *log-likelihood* of observation i using the estimated parameters

$$
\begin{aligned}
\log L_i &= \mathcal{N}(\mathbf{x}_i; \mu = \bar{x}, \Sigma = \hat{\Sigma}) \\
&= -p \log 2\pi - \frac{1}{2} \log |\hat{\Sigma}| - \frac{1}{2}(\mathbf{x_i} - \bar{\mathbf{x}})^{\mathsf{T}} \hat{\Sigma}^{-1} (\mathbf{x_i} - \bar{\mathbf{x}}) \\
&= \frac{1}{2} D_i^2 + const
\end{aligned}
$$

- Robust estimation:
  - If indeed we have outliers, then these will be affecting out estimated parameters $\bar{x}$ and $\hat{\Sigma}$.
  - Robust estimation techniques can help limit the damage caused by the outliers
  - Another, more structured approach, is mixture models!

3.3 Mixture Model Approach

- We can go back to our mixture model formulation and propose that the outliers come from a different distribution than the normal observations
- In mixture formulation

$$f(x) = \pi h(x) + (1 - \pi) g(x)$$

  - $h(x)$ is the pdf for the outliers
  - $g(x)$ is the pdf for the normal observations

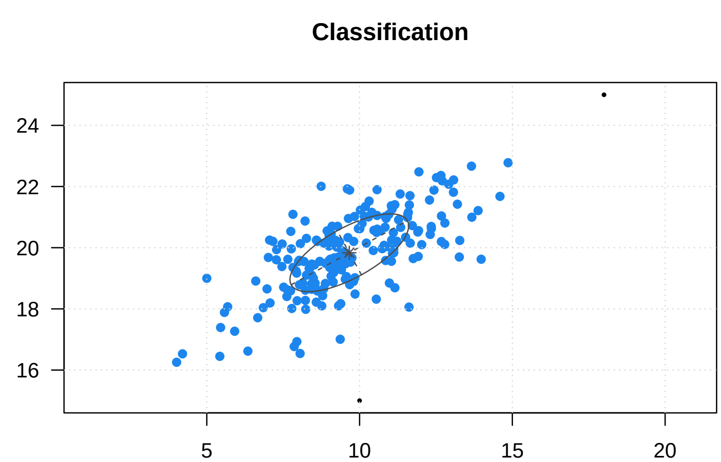
  - π is the prior probability that an observation will be an outlier
  - $E[\text{number of outliers}] = n\pi$

- There are several options for the outlier distribution $h(x)$
  - The uniform distribution on the *minimum bounding box* is one simple approach
    * $h(x) = 1/V$, where V is the volume of the bounding box
    * $V = \prod_{j=1}^{p} (max(x_j) - min(x_j))$

- The `Mclust()` function in the R package `mclust` permits this formulation using the `initialization=list(noise=TRUE)` argument

```
#-- Fit mixture model with uniform noise
library(mclust)
mc = Mclust(X, initialization=list(noise=TRUE))
summary(mc)
#> ----------------------------------------------------
#> Gaussian finite mixture model fitted by EM algorithm
#> ----------------------------------------------------
#>
#> Mclust EVE (ellipsoidal, equal volume and orientation) model with 1 component
#> and a noise term:
#>
#>  log.likelihood   n df   BIC   ICL
#>          -713.4 203  7 -1464 -1473
#>
#> Clustering table:
#>   1   0
#> 201   2
```

```
plot(mc, what="classification", asp=1, las=1)
grid()
```



4 Two-Sample Testing (A/B Testing)

4.1 A/B Testing

4.2 Two-Sample Goodness of Fit

5 Hotspot Detection

5.1 Mixture Model Formulation

6 Outbreak Detection