A Robust Maximal F-ratio Statistic to Detect Clusters Structure

**L. A. García-Escudero, A. Gordaliza,
A. Mayo-Iscar and C. Matrán**

Departamento de Estadística e Investigación Operativa, Universidad de Valladolid
Valladolid, Spain

Cinquième Rencontre de Statistiques Mathématiques
BORDEAUX-SANTANDER-TOULOUSE-VALLADOLID
BOSANTOUVAL2009

Bordeaux, June 5th, 2009

Outline

1. Introduction
2. The Maximal F-ratio
3. A Trimmed version of the Maximal F-ratio
4. Asymptotics for the Trimmed Maximal F-ratio
5. Hypothesis Testing
6. An Example in Genetics

Outline

1. Introduction
2. The Maximal F-ratio
3. A Trimmed version of the Maximal F-ratio
4. Asymptotics for the Trimmed Maximal F-ratio
5. Hypothesis Testing
6. An Example in Genetics

F-ratio statistics

- F-ratio is a widely used tool in Statistical Data Analysis

$$F = \frac{Between\ \ Groups\ \ Variance}{Within\ \ Groups\ \ Variance}$$

$$F = \frac{B_k}{W_k} = \frac{\sum_{j=1}^{k} n_j \left\|\overline{y}_j - \overline{y}\right\|^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left\|y_{i,j} - \overline{y}_j\right\|^2}$$

- Supervised (known group ownerships):
  - ANOVA
    Testing the existence of differences between the groups means.
  - Linear Discriminant Analysis
    Derivation of the canonical variates.
- Unsupervised (unknown group ownerships):
  - Maximal F-Ratio in Cluster Analysis
    Testing the number of clusters.

Outline

1. Introduction
2. The Maximal F-ratio
3. A Trimmed version of the Maximal F-ratio
4. Asymptotics for the Trimmed Maximal F-ratio
5. Hypothesis Testing
6. An Example in Genetics

$k-$means (I)

- McQueen (1967), Hartigan and Wong (1979).
- Well-known and widely used (non-hierarchical) clustering method.
- $x_1, x_2, ..., x_n$: p-variate observations.
- $C_1, ..., C_k$: partition into k groups. $m_1, ..., m_k$: groups sample means.
- k-means $m_1^n, ..., m_k^n$: Solution of the Minimization problem

$$\min_{C_1,\dots,C_k} \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - m_j \right\|^2 = \min_{m_1,\dots,m_k} \sum_{i=1}^{n} \inf_{1 \le j \le k} \left\| x_i - m_j \right\|^2$$

- Within-groups sum of squares:

$$W_k^n = \sum_{i=1}^{n} \inf_{1 \le j \le k} \left\| x_i - m_j^n \right\|^2$$

$k-$means (II)

- $x_1, x_2, ..., x_n$: p-variate observations
- $C_1, ..., C_k$: partition into k groups. $m_1, ..., m_k$: groups sample means.
- m^n: overall mean
- $n_1, ..., n_k$: groups sizes
- k-means $m_1^n, ..., m_k^n$: Solution of Maximization problem

$$\max_{C_1,\ldots,C_k} \sum_{j=1}^{k} n_j \|m_j - m^n\|^2$$

- Between-groups sum of squares:

$$B_k^n = \sum_{j=1}^{k} n_j \|m_j^n - m^n\|^2$$

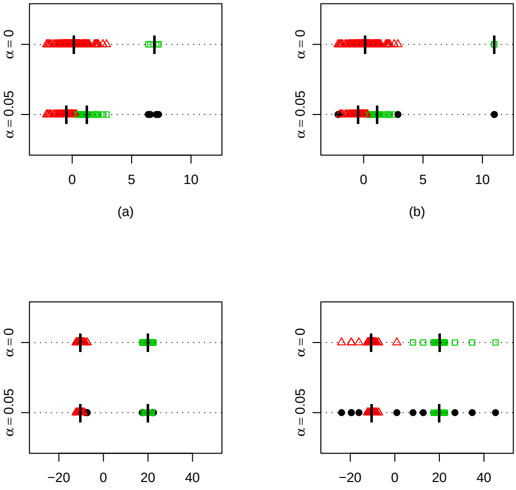
Maximal $F-$Ratio Statistic (I)

$$R_k^n = \frac{B_k^n}{W_k^n}$$

- Maximal $F-$Ratio Statistic to test two clusters structure:
  - Engelman and Hartigan (1969) proposed a test for "clusters structure" for univariate data and $k = 2$.
  - The test divides the sample into two subsets maximizing the likelihood ratio that the two subsets are sampled from two normals with different means, against the null hypothesis that the means are equal.
  - The test reject H_0 when the 2-means centers are "separated" enough.
- Maximal $F-$Ratio Statistic to check k clusters structure:
  - Calinski and Harabasz (1974): Extension to check "$k - 1$ or less groups" against "at least k groups".
    - It is not straightforward: We need to specify the distribution in H_0.
    - "Pseudo $F-$test": Descriptive Criterion for choosing k.
    - Good results in Milligan and Cooper (1985)'s study among 30 procedures for determining the number of clusters.
  - Hartigan (1978) proved the asymptotic normality for the k-mean centers and the Maximal F-ratio in the univariate case.

Maximal $F-$Ratio Statistic (II)

- Extension to the multivariate setting
  - Lee (1979) extended the Maximal $F-$Ratio principle to test clusters structure in the multivariate case.
  - Pollard (1982) extended Hartigan's results to the multivariate case by using Empirical Process Theory.
- Main drawbacks:
  - Hartigan and Pollard's results need some moment conditions on the underlying distribution to be applied.
  - Lack of Robustness: $k-$means and Maximal F-ratio can be severely affected by a small amount of anomalous observations



(a) (b) (c) (d)

(a) 100 observations: 95% $N(0,1)$ and 5% outliers.
(b) 100 observations: 99% $N(0,1)$ and 1% remote outliers.
(c) 200 observations drawn from a normal mixture.
(d) 200 observations: 95% from a normal mixture and 5% background noise.

Outline

1. Introduction
2. The Maximal F-ratio
3. A Trimmed version of the Maximal F-ratio
4. Asymptotics for the Trimmed Maximal F-ratio
5. Hypothesis Testing
6. An Example in Genetics

Robust Maximal F-ratio based on Trimmed $k-$means

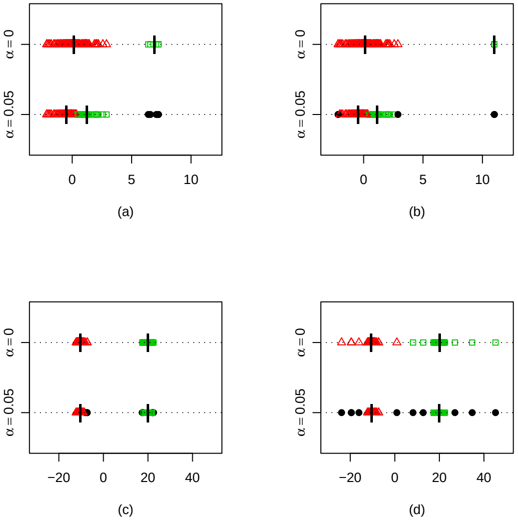
- Trimming data in clustering problems is not an easy task.
- Gordaliza (1991) introduced "impartial" trimming.
- Cuesta-Albertos et al (1997) extended the idea to clustering.
- Trimmed $k-$means: $m_1^n, ..., m_k^n$ solution of the Minimization problem:

$$\min_{\mathbf{Y}} \min_{m_1,\dots,m_k} \sum_{x_i \in \mathbf{Y}} \inf_{1 \le j \le k} \|x_i - m_j\|^2 ,$$

  - $\alpha \in (0,1)$ is the trimming proportion.
  - $\mathbf{Y}$ ranges over all subsets $\mathbf{Y} \subset \{x_1, ..., x_n\}$ with $[n(1-\alpha)]$ elements.
  - The trimmed k-means $m_1^n, ..., m_k^n$ induce a partition of the non-trimmed observations onto k clusters $C_1^n \cup ... \cup C_k^n$.
  - The clusters are balls having the same radius.

- Trimmed Maximal F-Ratio:

$$R_k^n(\alpha) := \frac{B_k^n(\alpha)}{W_k^n(\alpha)} := \frac{\sum_{j=1}^k n_j \left\| m_j^n - m^n \right\|^2}{\sum_{j=1}^k \sum_{x_i \in C_j^n} \left\| x_i - m_j^n \right\|^2}$$



(a) 100 observations: 95% $N(0,1)$ and 5% outliers.
(b) 100 observations: 99% $N(0,1)$ and 1% remote outliers.
(c) 200 observations drawn from a normal mixture.
(d) 200 observations: 95% from a normal mixture and 5% background noise.

Outline

1. Introduction
2. The Maximal F-ratio
3. A Trimmed version of the Maximal F-ratio
4. Asymptotics for the Trimmed Maximal F-ratio
5. Hypothesis Testing
6. An Example in Genetics

Population version

- Population version of Trimmed $k-$means. $m_1^0, \dots, m_k^0$ solution of:

$$\min_{A_0 \,:\, P(A_0)=1-\alpha} \ \min_{m_1^0,\dots,m_k^0} \int_{A_0} \inf_{1\le j\le k} \left\| x - m_j^0 \right\|^2 dP(x)$$

- Population version of Trimmed Maximal $F-$ratio:

$$R_k^0(\alpha) := \frac{B_k^0(\alpha)}{W_k^0(\alpha)} := \frac{\sum_{j=1}^k P(C_j^0) \left\| m_j^0 - m^0 \right\|^2}{\sum_{j=1}^k \int_{C_j^0} \left\| x - m_j^0 \right\|^2 dP(x)}.$$

  - $C_1^0, \dots, C_k^0$: partition of the non trimmed area into k disjoint groups.
  - $m_1^0, \dots, m_k^0$: groups sample means (Trimmed $k-$means).
  - m^0: overall trimmed mean.

Asymptotic properties

- Theorem: Consistency
  Let $\alpha \in (0,1)$ and P be a continuous probability distribution and assume that there exists a unique α-trimmed k-mean. Then

$$R_k^n(\alpha) \to R_k^0(\alpha) \; P\text{−a.e.}$$

- Theorem: Asymptotic Normality
  If P has bounded density not identically null on the boundary of the optimal trimming set and there exists a unique α-trimmed k-mean, then:

$$\sqrt{n}(R_k^n(\alpha) - R_k^0(\alpha)) \xrightarrow{\mathcal{L}} N(0, LVL')$$

  - Proof based on Empirical Processes Theory.
  - No moment conditions on P are needed.

Asymptotic means and variances of $R_k^n(\alpha)$

$k = 2$

α:	.0001	.001	.01	.05	.1	.2
Asymp. mean	1.756	1.776	1.877	2.097	2.263	2.483
Asymp. var.	6.515	6.485	7.605	12.841	19.462	34.475

$k = 3$

α:	.0001	.001	.01	.05	.1	.2
Asymp. mean	4.273	4.347	4.698	5.412	5.921	6.568
Asymp. var.	74.775	75.119	79.328	95.600	115.570	160.341

Table: Asymptotic means and variances of the maximal F-ratio for the $N(0,1)$ distribution, $k = 2$ and $k = 3$ and different trimming levels.

Hartigan (1978)'s values for the untrimmed ($\alpha = 0$) case, when $k = 2$:

- Asymptotic mean: $2/(\pi - 2) = 1.752$
- Asymptotic variance:$8/\pi \cdot (1 - 3/\pi)/(1 - 2/\pi)^4 = 6.582$

Outline

1. Introduction
2. The Maximal F-ratio
3. A Trimmed version of the Maximal F-ratio
4. Asymptotics for the Trimmed Maximal F-ratio
5. Hypothesis Testing
6. An Example in Genetics

Hypothesis Testing

- Asymptotic results can be used to test $H_0 : k = 1$ against $H_1 : k > 1$.
- Reject H_0 when $R_k^n(\alpha)$ is greater than the (asymptotic) critical value.
- The test depends on the assumed model for H_0.
- Large sample size n will be required because of the asymptotic character of the results.
- We conduct a simulation study in the univariate normal case to analyze:
  - The empirical power of the test.
  - The gain in robustness provided by the trimming.
  - The behavior for finite sample sizes.

Empirical powers

		$\varepsilon = .0$			$\varepsilon = .01$		
	α:	0	.01	.05	0	.01	.05
$D = 0$	$n = 500$	.08	.08	.05	.87	.33	.02
	$n = 1000$	.06	.06	.04	.95	.26	.01
	$n = 10000$	.05	.05	.04	1.00	.03	.00
$D = 2$	$n = 500$	.94	.90	.76	.41	.60	.66
	$n = 1000$	1.00	.99	.94	.42	.62	.89
	$n = 10000$	1.00	1.00	1.00	.47	.63	1.00
$D = 3$	$n = 500$	1.00	1.00	1.00	.06	.67	1.00
	$n = 1000$	1.00	1.00	1.00	.01	.72	1.00
	$n = 10000$	1.00	1.00	1.00	.00	.98	1.00

Table: Empirical powers based on 5000 random samples.
Significance level: 0.05.
$f_{D,\varepsilon}(x) = (1-\varepsilon)f_D(x) + \varepsilon\varphi(x-20)$
$f_D(x) = 0.5\varphi(x) + 0.5\varphi(x-D)$ with $\varphi(\cdot)$ being the density of $N(0,1)$.
Powers obtained using the asymptotic distributions of R_2^n and $R_2^n(\alpha)$.

Outline

1 Introduction

2 The Maximal F-ratio

3 A Trimmed version of the Maximal F-ratio

4 Asymptotics for the Trimmed Maximal F-ratio

5 Hypothesis Testing

6 An Example in Genetics

Ordering and selecting genes

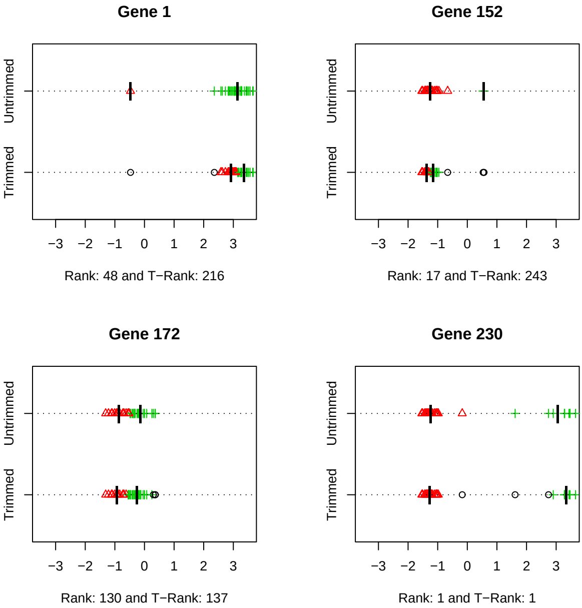
- Characteristics of genetical studies:
  - Genetical studies usually involve a large number of genes.
  - The presence of noise is frequent (genomic and proteomic data are frequently affected by measurement errors)
- Selection of genes:
  - We look for informative genes exhibiting differences between the expression levels across the individuals in the study.
  - When group ownerships are known, ordering genes according to their F-Ratios can be used as a first screening phase in gene selection (Dudoit et al. 2002).
  - When group ownerships are unknown, we can resort to Maximal $F-$Ratios to order genes and select a reduced set of informative genes to apply clustering methods.
  - The presence of noise suggests the use of Robust Maximal $F-$Ratios to order and select genes.

Golub's leukemia dataset (I)

- Example based on the training set of the famous Golub's leukemia dataset (Golub et al 1999).
- We study the data subset *leukemia* which accompanies the contributed package supclust in the R-project repository.(Dettling and Bühlmann 2004)
- Study of gene expression in two types of acute leukemias:
  - Acute lymphoblastic leukemia (ALL).
  - Acute myeloid leukemia (AML).
- The data set consists of 250 gene expressions measured on 38 individuals.
- We will assume that no leukemia type classification is available (unsupervised study).

Golub's leukemia dataset (II)

- We obtain an "importance" index of genes according to the values of the statistics
  - Maximal $F-$Ratio Statistic: R_2^n (Rank)
  - Trimmed Maximal $F-$Ratio Statistic: $R_2^n(.08)$ (T-Rank).
- The trimming size $\alpha = .08$ is fixed to avoid the influence of the 3 most outlying observations.
- Genes with smallest values in Rank and T-Rank are marked as the most "interesting" ones.


Gene 1
Untrimmed
Trimmed
−3 −2 −1 0 1 2 3
Rank: 48 and T−Rank: 216
Gene 152
Untrimmed
Trimmed
−3 −2 −1 0 1 2 3
Rank: 17 and T−Rank: 243
Gene 172
Untrimmed
Trimmed
−3 −2 −1 0 1 2 3
Rank: 130 and T−Rank: 137
Gene 230
Untrimmed
Trimmed
−3 −2 −1 0 1 2 3
Rank: 1 and T−Rank: 1

Ordering subsets of p genes

- The statistic $R_2^n(\alpha)$ can be computed from p-dimensional samples.
- The trimmed maximal F-ratio can be used for ranking subsets of p variables which jointly serve to detect interesting clusters structures.
- This is a computationally hard problem if the number of genes J is large as long as $\binom{J}{p}$ subsets need to be explored.

García-Escudero, L. A., Gordaliza A., Mayo-Iscar, A. and Matrán, C. (2009). A Robust Maximal F-ratio Statistic to Detect Clusters Structure. *Communications in Statistics - Theory and Methods*, **38**, 682-694.

Thank you very much

Merci beaucoup