NOVOS RUMOS EM ESTATÍSTICA

Actas do IX Congresso Anual da Sociedade Portuguesa de Estatística

Editores
Lucília Carvalho
Fátima Brilhante
Fernando Rosado

Editores: Lucília Carvalho, Fátima Brilhante, Fernando Rosado
Título: *Novos Rumos em Estatística*
Editora: Sociedade Portuguesa de Estatística
ISBN: 972-98619-4-3

Design e produção gráfica da capa: Haderer & Müller.
Fotolitos da capa: Grafilis.
Impressão da capa: Minerva Comercial Sintrense.
Impressão: Reprografia do Instituto Superior de Agronomia
Tiragem: 250 exemplares
Depósito Legal nº: 183953/02

Comparação dos Parâmetros Direccionais de k Populações de Bingham

Adelaide Figueiredo
Faculdade de Economia do Porto e LIACC
Paulo Gomes
Instituto Nacional de Estatística

Resumo: Consideramos um teste proposto por Gomes e Figueiredo (1999), Figueiredo (2000) para a igualdade dos parâmetros direccionais de k populações de Bingham e estudamos a potência desse teste para duas populações de Bingham, em alguns casos. Trata-se de um teste muito importante no contexto da selecção a priori de variáveis, quando determinamos regras de classificação de uma nova variável a grupos previamente definidos de variáveis provenientes de populações de Bingham (ver Figueiredo, 2000).

Palavras–chave: Dados direccionais, Lei de Bingham, MANOVA, Selecção de variáveis

Abstract: We consider a test proposed by Gomes e Figueiredo (1999), Figueiredo (2000) for investigating whether the directional parameters of k Bingham distributions differ significantly and we study the power of the test in some cases for two Bingham populations. This test is important when we define classification rules to affect a new variable into one or more Bingham populations (see Figueiredo, 2000), a problem which gives us new highlights on the selection of variables.

1 Introdução

Consideremos o quadro de dados multivariado com n indivíduos caracterizados por p variáveis. Enquanto que na abordagem clássica, supõe-se que as variáveis estão fixas e os indivíduos são seleccionados aleatoriamente de uma população de indivíduos, nós consideramos uma abordagem em que os indivíduos estão fixos e as variáveis são seleccionadas aleatoriamente de uma população de variáveis.

No contexto do problema de selecção de variáveis são conhecidos contributos inovadores para este problema (Gomes, 1987, Gomes e Figueiredo, 1998, 1999, Figueiredo, 2000), entre os quais, destacamos a determinação de regras de classificação de uma nova variável a grupos previamente definidos de variáveis com distribuições de Bingham. Consideramos um caso particular da distribuição de Bingham, a distribuição Scheidegger-Watson bipolar definida na esfera n-unitária, $S_{n-1} = \{\mathbf{x} \in \Re^n : \|\mathbf{x}\| = 1\}$. A função densidade de probabilidade da

2 Teste para a igualdade dos parâmetros direccionais de k populações de Bingham

Consideremos k populações de Bingham $B_n(\mathbf{u}_i, \xi_i)$ com parâmetros ξ_i conhecidos e seja $X_i = [\mathbf{x}_{i1}|\mathbf{x}_{i2}|...|\mathbf{x}_{ip_i}]$ uma amostra aleatória de p_i variáveis da i-ésima população.

Para testar $H_0 : \mathbf{u}_1 = \mathbf{u}_2 = ... = \mathbf{u}_k = \mathbf{u}$, assumimos que $\mathbf{x}_{ij}$, $i = 1, ..., k$, $j = 1, ..., p_i$ são gerados a partir do seguinte modelo

$$1 - \langle \mathbf{x}_{ij}, \mathbf{u}_i \rangle^2 = e_{ij}^2, \tag{7}$$

onde e_{ij}^2 é uma variável aleatória tal que para ξ_i grande verifica (6).

Consideramos um teste baseado na decomposição da variabilidade total na variabilidade dentro-grupos e variabilidade entre-grupos, ou seja na identidade seguinte

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{p_i}\xi_i\left(1-\langle\widehat{\mathbf{u}},\mathbf{x}_{ij}\rangle^2\right) &= \sum_{i=1}^{k}\sum_{j=1}^{p_i}\xi_i\left(1-\langle\widehat{\mathbf{u}}_i,\mathbf{x}_{ij}\rangle^2\right)+\\ &+\sum_{i=1}^{k}\sum_{j=1}^{p_i}\xi_i\left(\langle\widehat{\mathbf{u}}_i,\mathbf{x}_{ij}\rangle^2-\langle\widehat{\mathbf{u}},\mathbf{x}_{ij}\rangle^2\right),\end{aligned} \tag{8}$$

onde o estimador de máxima verosimilhança $\widehat{\mathbf{u}}_i$ é o vector próprio associado ao maior valor próprio da matriz ${X_i}^t X_i$ e o estimador de máxima verosimilhança $\widehat{\mathbf{u}}$ é o vector próprio associado ao maior valor próprio de $\sum_{i=1}^{k} \xi_i {X_i}^t X_i$.

A identidade anterior é equivalente à seguinte

$$2\left(\sum_{i=1}^{k}\xi_i p_i - \lambda\right) = 2\left(\sum_{i=1}^{k}\xi_i p_i - \sum_{i=1}^{k}\lambda_i\right) + 2\left(\sum_{i=1}^{k}\lambda_i - \lambda\right), \tag{9}$$

onde λ representa o maior valor próprio de $\sum_{i=1}^{k} \xi_i {X_i}^t X_i$ e λ_i é o maior valor próprio da matriz $\xi_i {X_i}^t X_i$.

Vejamos que cada uma das variáveis aleatórias da identidade anterior toma valores não negativos.

Verifica-se que $\sum_{i=1}^{k} \lambda_i - \lambda \geq 0$, pois o maior valor próprio de uma soma de matrizes é menor ou igual que a soma dos maiores valores próprios das matrizes.

Atendendo a que a variável aleatória $\left(\sum_{i=1}^{k} \xi_i p_i - \lambda\right)$ pode ser escrita como $\sum_{i=1}^{k}\sum_{j=1}^{p_i} \xi_i \left(1 - \cos^2\theta_{ij}\right)$, onde θ_{ij} é o ângulo entre $\widehat{\mathbf{u}}$ e $\mathbf{x}_{ij}$, e $\cos^2\theta_{ij} \leq 1$, resulta que $\left(\sum_{i=1}^{k} \xi_i p_i - \lambda\right) \geq 0$.

3 Potência empírica do teste

Sejam $B_n(\mathbf{u}_1, \xi)$ e $B_n(\mathbf{u}_2, \xi)$ duas populações de Bingham com o mesmo parâmetro de concentração conhecido e pretendemos testar $H_0 : \mathbf{u}_1 = \mathbf{u}_2$. Vamos supôr para determinar a potência, amostras das populações com igual dimensão $p_1 = p_2 = p$ e os casos em que $n = 30$, $p = 30, 50$, $n = 50$, $p = 50, 100$ e $\xi = 10, 20$, para cada n.

Como só é conhecida a distribuição assintótica da estatística de teste sob H_0, determinamos os percentis de ordem 0.95 da estatística de teste com base em *10000* réplicas e supondo $\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{e}_n = (0, ..., 0, 1)$ - Tabela 1.

	n	30		50	
	p	30	50	50	100
ξ	10	2.245	2.329	2.386	2.877
	20	1.355	1.355	1.723	1.689
	30	1.359	1.363	1.276	1.274
	50	1.425	1.428	1.300	1.296

Tabela 1: Percentis de ordem 0.95 da estatística no caso de populações de Bingham com o mesmo parâmetro de concentração

Na Tabela 2 encontra-se a potência empírica do teste obtida a partir de *3000* réplicas da estatística sob a hipótese H_1: $\mathbf{u}_1 = \mathbf{e}_n$ e $\mathbf{u}_2 = \mathbf{u}$, onde $\mathbf{u}$ forma com $\mathbf{e}_n$ o ângulo $\theta = 18^o(18^o)90^o$. Ver Figura 1.

	n	30				50			
	p	30	50	30	50	50	100	50	100
θ	ξ	10	10	20	20	10	10	20	20
18°		0.076	0.105	0.431	0.732	0.062	0.083	0.183	0.447
36°		0.214	0.484	0.995	1	0.107	0.184	0.779	0.999
54°		0.487	0.864	1	1	0.166	0.394	0.993	1
72°		0.706	0.964	1	1	0.221	0.569	1	1
90°		0.477	0.875	1	1	0.070	0.158	1	1

Tabela 2: Potência empírica do teste no caso de populações de Bingham com o mesmo parâmetro de concentração

Podemos concluir que para $n = 30$ e $n = 50$, a potência é elevada a partir de um valor de ξ relativamente baixo. Podemos observar que fixando n, p e ξ (não suficientemente pequeno para cada n), a potência empírica aumenta à medida que aumenta o ângulo θ. Para θ, p e n fixados, a potência empírica aumenta quando aumenta o parâmetro de concentração ξ. Para cada n, ξ e θ, a potência empírica cresce à medida que aumenta a dimensão das amostras. Para o mesmo

n	ξ_1	ξ_2	p/θ	18°	36°	54°	72°	90°
30	10	20	30	0.088	0.308	0.702	0.917	0.940
			50	0.106	0.566	0.942	0.996	0.998
	10	30	30	0.091	0.308	0.708	0.924	0.958
			50	0.130	0.566	0.953	0.997	0.998
	20	30	30	0.576	1	1	1	1
			50	0.887	1	1	1	1
50	10	20	50	0.077	0.163	0.325	0.473	0.398
			100	0.092	0.315	0.648	0.842	0.832
	10	30	50	0.077	0.143	0.315	0.507	0.536
			100	0.084	0.271	0.635	0.865	0.902
	20	30	50	0.442	0.971	1	1	1
			100	0.512	1	1	1	1

Tabela 4: Potência empírica do teste no caso de de populações de Bingham com parâmetros de concentração diferentes

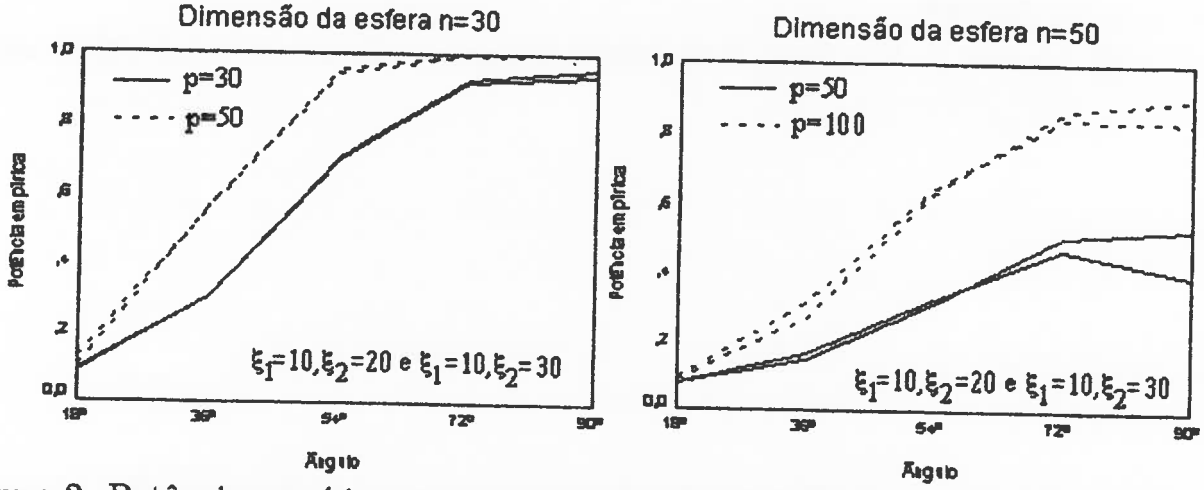

Figura 2: Potência empírica no caso de populações de Bingham com parâmetros de concentração diferentes