MINISTÉRIO DO PLANEJAMENTO E ORÇAMENTO
FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA-IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034 - 7175

R. bras. Estat., Rio de Janeiro, v.53, n.199/200, p.1- 91, jan./dez.,1992

REVISTA BRASILEIRA DE ESTATÍSTICA

Órgão oficial do IBGE

Publicação semestral,editada pelo IBGE, que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.
Temas, abordando aspectos do desenvolvimento metodológico, serão aceitos desde que relevantes para os órgãos produtores de estatísticas.
Os originais para publicação deverão ser submetidos em três vias (que não serão devolvidas) para:

Djalma G. Pessoa
Editor-Responsável - RBEs

ENCE
Rua André Cavalcante, 106 - Bairro de Fátima
20231-050 - Rio de Janeiro - RJ

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.
Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Capa
Pedro Paulo Machado

Revista brasileira de estatística / Fundação Instituto Brasileiro de Geografia e Estatística - v.1, n.1 (jan./mar. 1940)- . — Rio de Janeiro: IBGE, 1940-

v.

Trimestral (1940-1986), semestral (1987-)
Órgão oficial do IBGE.
Continuação de: Revista de economia e estatística.
Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v.50 (1980-1989)

ISSN 0034-7175 - Revista brasileira de estatística

1 - Estatística - Periódicos. I. IBGE

IBGE. CDDI. Departamento de Documentação e Biblioteca CDU 31 (05)
RJ-IBGE/88-05 Rev. PERIÓDICO

Impresso no Brasil/Printed in Brazil

SUMÁRIO

ARTIGOS

COMPARAÇÃO DE MODELOS DE SOBREVIVÊNCIA APLICADOS A UM ESTUDO DE LEUCEMIA EM CRIANÇAS

Enrico A. Colosimo*
Maria Lourdes G. Nogueira*
Nádia R. M. Rocha*
Marcos B. Viana**

1 INTRODUÇÃO

Análise de sobrevivência é uma das áreas da estatística que mais tem crescido nos últimos 20 anos. Este crescimento pode ser observado tanto no desenvolvimento e aprimoramento de métodos estatísticos, quanto no número de aplicações. Uma evidência quantitativa do crescimento de aplicações de análise de sobrevivência em medicina pode ser vista em Baillar III & Mosteller (1992, Capítulo 3). Segundo estes autores, nos artigos do conceituado periódico *The New England Journal of Medicine*, o uso de métodos de análise de sobrevivência cresceu de 11% em 1979 para 32% em 1989. Dentre os vários métodos estatísticos considerados, estes foram os que mais cresceram no período avaliado.

Os conjuntos de dados de sobrevivência têm como variável resposta o tempo até a ocorrência de um evento de interesse. Este tempo é denominado **tempo de falha**,

* Departamento de Estatística, ICEx – UFMG

** Departamento de Pediatria, Faculdade de Medicina – UFMG

podendo ser o tempo até a morte do paciente, bem como até a cura ou recidiva de uma doença. Em estudos de câncer, por exemplo, é usual registrar a entrada do paciente no estudo, a remissão (após o tratamento, o paciente fica livre dos sintomas da doença), a recorrência da doença (recidiva) e a morte do paciente.

A principal característica de dados de sobrevivência é a presença de **censura**, que é a observação parcial da resposta. Isto é, por alguma razão, o acompanhamento do paciente foi interrrompido. Seja porque o paciente mudou de cidade, o estudo foi interrompido para a análise dos dados ou o paciente morreu de causa diferente da estudada. Em medicina, a censura do tipo aleatória é a mais usual (Kalbfleisch & Prentice, 1980). Este tipo de censura é caracterizado pelo fato dos pacientes entrarem no estudo em tempos diferentes. Sem a presença de censura, as técnicas estatísticas clássicas, como análise de regressão e planejamento de experimentos, poderiam ser usadas na análise deste tipo de dados. No entanto, são os métodos de análise de sobrevivência que possibilitam incorporar na análise estatística a informação contida nos dados censurados.

O termo análise de sobrevivência se refere basicamente a situações médicas envolvendo dados censurados. Entretanto, condições similares ocorrem em outras áreas onde se usam as mesmas técnicas de análise de dados. Em engenharia, são comuns os estudos onde produtos ou equipamentos são colocados sob teste para se estimar características relacionadas aos seus tempos de vida. Exemplos podem ser vistos em Nelson (1990). Os engenheiros denominam esta área de confiabilidade. O mesmo acontece em ciências sociais, onde várias situações de interesse são estudadas envolvendo dados censurados (Allison, 1984). O crescimento observado no número de aplicações em medicina também é observado nestas outras áreas. Mesmo sendo comuns as técnicas de análise para dados censurados em todas estas áreas, neste artigo será utilizado o termo análise de sobrevivência. Isto porque estas técnicas serão aplicadas em um estudo médico.

Geralmente, em medicina, queremos identificar fatores de prognóstico para uma certa doença (ou terapia) ou estudar um tratamento de interesse enquanto controlamos outros fatores. Existem duas classes de modelos para analisar dados de sobrevivência: os modelos paramétricos, que na literatura são denominados modelos de tempo de vida acelerada, e os modelos semiparamétricos, chamados modelos de riscos proporcionais. O objetivo deste artigo é comparar estas duas classes de modelos usando como ilustração um conjunto de dados de leucemia em crianças. O artigo está assim dividido: a Seção 2 apresenta as duas classes de modelos de sobrevivência; o conjunto de dados é apresentado na Seção 3, e na Seção 4 as duas classes de modelos são aplicadas aos dados.

2 MODELOS DE ANÁLISE DE SOBREVIVÊNCIA

2.1 Modelos de Tempo de Vida Acelerada

Os modelos de tempo de vida acelerada utilizam famílias paramétricas de distribuições para o tempo de sobrevivência T. Esta classe de modelos é definida considerando o seguinte modelo loglinear em T:

$$[Y = \log T = X'\beta + \sigma W,]$$

onde X é a matriz de delineamento, β é o vetor de parâmetros, σ é mais um parâmetro a ser estimado e W é um vetor de erros com distribuição especificada. Cabe observar que Y é associado com as covariáveis X através de um modelo linear e, portanto, o papel das covariáveis é acelerar (ou desacelerar) o tempo de sobrevivência T. Os parâmetros β e σ no modelo acima são estimados pelo método de máxima verossimilhança (Cox & Hinkley, 1974). Os tempos de falha contribuem com a função de densidade para a função de verossimilhança e os tempos de censura com a função de sobrevivência, $P(T > t)$. Um modelo freqüentemente usado assume distribuição Weibull para o tempo de sobrevivência T, o que implica na distribuição dos valores extremos para o vetor de erros W (Kalbfleisch & Prentice, 1980; Colosimo & Garcia, 1991). Outros modelos de interesse utilizam as distribuições exponencial, lognormal e gama.

2.2 Modelos de Riscos Proporcionais

Os modelos de riscos proporcionais pertencem a uma classe de modelos com a propriedade de que diferentes indivíduos têm funções de risco proporcionais. A função de risco é a taxa de falha do indivíduo no tempo t condicional a sua sobrevivência até este tempo. Este modelo foi proposto por Cox (1972) e especifica que a função de risco pode ser escrita como:

$$[\lambda(t) = \lambda_0(t)\, g(X, \beta),]$$

onde $\lambda_0(t)$ é uma função de risco padrão não–especificada. A parte paramétrica do modelo é geralmente tomada como $g(X, \beta) = exp(X'\beta)$. Os dois componentes do

modelo têm um efeito multiplicativo na função de risco $\lambda(t)$. Devido à presença do componente não-paramétrico ($\lambda_0(t)$) no modelo, o método de máxima verossimilhança não pode ser usado para estimar o vetor de parâmetros β. Cox (1975) propôs o método de máxima verossimilhança parcial. Para grandes amostras, os estimadores de β obtidos têm as mesmas propriedades que os estimadores de máxima verossimilhança usuais, isto é, eles são consistentes e assintoticamente normais (Tsiatis, 1981; Andersen & Gill, 1982).

2.3 Modelo Paramétrico versus Modelo Semiparamétrico

As duas classes de modelos respondem às perguntas feitas em estudos de sobrevivência. Entretanto, existem vantagens e desvantagens no uso de cada uma delas. Os modelos paramétricos pressupõem a especificação de uma distribuição para o vetor de erros. Podemos então usar o método de máxima verossimilhança usual para fazer inferências sobre os parâmetros do modelo. Outra vantagem desta classe de modelos é a fácil interpretação dos coeficientes estimados, pois a sua forma é a mesma do clássico modelo de regressão linear. Por outro lado, se a distribuição para o vetor de erros for incorretamente especificada, as estimativas de máxima verossimilhança serão viciadas.

O modelo de riscos proporcionais se comporta de forma diferente. Devido à presença do componente não-paramétrico, esta classe de modelos é extremamente flexível. Por exemplo, o modelo paramétrico com distribuição Weibull para os tempos de sobrevivência é um caso particular dos modelos de riscos proporcionais. Outra vantagem deste modelo é poder incluir covariáveis dependentes do tempo. Entretanto, o método de máxima verossimilhança parcial é menos eficiente que o de máxima verossimilhança padrão. Esta perda de eficiência é devida ao fato de se descartar parte da informação contida na amostra ao usar a verossimilhança parcial.

Pode-se notar que a escolha de um destes modelos é um impasse entre possíveis vícios ao usar os modelos paramétricos e perda de precisão no caso dos modelos de riscos proporcionais. Em medicina, os modelos de riscos proporcionais estão ganhando cada vez mais popularidade. Isto é justificado porque muitos dos estudos médicos são altamente sujeitos a vícios, pois são observacionais e envolvem um grande número de covariáveis. Por outro lado, em engenharia os modelos paramétricos praticamente dominam. Os estudos nesta área são geralmente controlados e as questões de interesse são mais facilmente respondidas pelos modelos paramétricos devido à sua simplicidade.

3 DESCRIÇÃO DOS DADOS DE LEUCEMIA

Existem vários exemplos de aplicação dos modelos de análise de sobrevivência. Na área médica, eles são muito utilizados na identificação de fatores de prognóstico para uma doença, bem como na comparação de tratamentos. Em oncologia, por exemplo, qualquer nova terapêutica ou droga para o combate ao câncer requer um estudo, onde a resposta de interesse é geralmente o tempo de sobrevida ou de remissão dos pacientes.

O conjunto de dados que motivou este artigo foi obtido de um estudo de leucemia em crianças, desenvolvido pelo Grupo Cooperativo Mineiro para Tratamento de Leucemias Agudas. Uma descrição deste conjunto de dados será apresentada a seguir.

A leucemia aguda é a neoplasia de maior incidência na população com menos de 15 anos de idade. Calcula-se que nesta faixa etária a incidência anual gire em torno de 5 a 6 casos novos por 100 mil crianças, sendo a grande maioria dos casos de Leucemia Linfoblástica Aguda (LLA).

Apesar do progresso alcançado no tratamento, em particular, da leucemia linfoblástica, as leucemias agudas continuam sendo a causa mais comum de morte por neoplasia. O objetivo do tratamento médico de uma criança com LLA é obter longos períodos de sobrevida livre da doença, o que, muitas vezes, significa sua "cura". Os avanços terapêuticos obtidos nos últimos 25 anos têm sido grandes na LLA. Na década de 60, menos de 1% das crianças com LLA sobreviviam mais de 5 anos após o diagnóstico. Atualmente, com a intensificação da quimioterapia para os grupos com prognóstico mais desfavorável, 60 a 70% do total de crianças, com diagnóstico de LLA, são sobreviventes de longo prazo e encontram-se provavelmente "curadas". Nos grupos de melhor prognóstico, as proporções de "cura" já se situam no patamar de 90%.

Com objetivo de entender melhor quais fatores afetam o tempo de sobrevivência de uma criança brasileira com leucemia linfoblástica aguda, um grupo de 128 crianças, com idade inferior a 15 anos, foi acompanhado no período de 1988 a 1992, em alguns hospitais de Belo Horizonte. A variável resposta de interesse é o tempo a partir da remissão (ausência da doença) até a recidiva ou morte (o que ocorrer primeiro). Das 128 crianças, 120 entraram em remissão e são elas que formam o conjunto de dados em estudo.

Dois fatores de prognóstico para a LLA, leucometria inicial e idade, são conhecidos na literatura médica desde a década de 70. Para cada criança, além destes dois fatores, foram medidas ao diagnóstico uma série de covariáveis. Uma análise estatística inicial usando basicamente técnicas não-paramétricas de análise de sobrevivência, tais como o estimador de Kaplan-Meier e o teste logrank (Lee, 1992), identificou as covariáveis que não afetam o tempo de falha. Estas covariáveis foram descartadas. Restaram as covariáveis descritas a seguir, cada uma delas com duas categorias, para serem incluídas nos modelos. Os respectivos pontos de cortes encontram-se entre parênteses.

LEUINI: leucometria inicial (75000 leucócitos/mm^3);

IDADE: idade em meses (96);

ZPESO: valor do peso padronizado pela idade e sexo (−2);

PAS: porcentagem de linfoblastos medulares que reagiram positivamente ao ácido periódico de Schiff (5%);

VAC: porcentagem de vacúolos no citoplasma dos linfoblastos (15%).

Maiores informações sobre este estudo podem ser encontradas em Viana (1993) e Viana et al. (1994).

Uma análise estatística destes dados usando as duas classes de modelos será apresentada a seguir, possibilitando a comparação dos mesmos.

4 AJUSTE DOS MODELOS

Existe uma série de pacotes computacionais que ajustam as duas classes de modelos de sobrevivência. O ajuste dos modelos paramétricos pode ser feito no SAS (procedimento lifereg) ou no SYSTAT (módulo survival), enquanto o ajuste do modelo de riscos proporcionais pode ser feito no SAS (procedimento phreg), COXSURV, EGRET, SYSTAT (módulo survival) ou BMDP (programa 2L). O SAS apresenta uma vantagem sobre o SYSTAT no ajuste dos modelos paramétricos: pode-se ajustar o modelo paramétrico gama generalizado. O ajuste do modelo de riscos proporcionais foi feito usando o COXSURV (Campos-Filho & Franco, 1990) pela facilidade deste pacote na manipulação dos dados.

Inicialmente, foi ajustado o modelo paramétrico gama e pelo teste da razão de verossimilhança foi avaliado se os modelos mais comuns, Weibull e lognormal, se ajustavam adequadamente ao conjunto de dados em estudo. Todos estes modelos probabilísticos são casos particulares do modelo gama generalizado. O modelo

lognormal foi o que melhor se ajustou aos dados (valor p = 0,62 para o teste da razão de verossimilhança). Este modelo confirma resultados anteriormente obtidos envolvendo dados de sobrevivência de leucemia (Feinleib & Macmahon, 1960). A Tabela 1 mostra os resultados obtidos para as duas classes de modelos.

Tabela 1
Resultados dos ajustes dos modelos lognormal e de riscos proporcionais ao conjunto de dados de leucemia

COVA-	MODELO PARAMÉTRICO			MODELO SEMI-PARAMÉTRICO		
RIÁVEL	Estimativa	Erro Padrão	Valor p	Estimativa	Erro Padrão	Valor p
LEUINI	$-1,183$	0,342	0,001	$-1,120$	0,355	0,002
IDADE	$-0,628$	0,323	0,052	$-0,648$	0,336	0,053
ZPESO	1,499	0,490	0,002	1,977	0,451	0,000
PAS	1,082	0,385	0,005	1,300	0,449	0,004
VAC	$-1,025$	0,376	0,006	$-1,271$	0,397	0,001
Escala	1,138	0,134				

A Tabela 1 mostra que os resultados obtidos pelos dois modelos ajustados são bastante similares, mesmo sendo os dois modelos de classes diferentes. Em particular, as estimativas dos coeficientes associados às covariáveis são bem próximas. De uma forma geral, isto confirma a adequação do modelo lognormal a este conjunto de dados. A maior precisão da análise usando o modelo paramétrico pode ser observada pela comparação das estimativas dos erros padrão dos coeficientes estimados. Estas estimativas são ligeiramente menores para o modelo paramétrico.

Estes resultados mostram uma pequena vantagem do modelo paramétrico lognormal sobre o modelo de riscos proporcionais. Juntando a isto a sua maior facilidade na interpretação dos coeficientes estimados, sugere um uso mais freqüente de modelos paramétricos em estudos médicos. Esta opinião também é compartilhada em um artigo recente por Wei (1992). Entretanto o uso de modelos paramétricos deve ser feito com cuidado, sempre acompanhado de uma análise detalhada da adequação do modelo proposto. Modelos incorretamente especificados para os tempos de falhas acarretam em vícios nos resultados. Por exemplo, se a distribuição especificada for a de Weibull, que é muito usada em análise de sobrevivência, **idade** e **zpeso** passam a não ser mais importantes para explicar o tempo da remissão até a recidiva ou morte (ambas com valor $p > 0,10$). Isto não está de acordo com os resultados já conhecidos na literatura nem é confirmado pela análise usando o modelo de riscos proporcionais.

BIBLIOGRAFIA

ALLISON, P. D. *Event history analysis sage.* London: [s.n.], 1984. (University Paper, 46).

ANDERSEN, P. K., GILL, R. Cox's regression model for counting processes: a large sample study. *Annals of Statistics.* Hayward, v. 10, p. 1100–1200, 1982.

BAILAR III, J. C, MOSTELLER, F. *Medical uses of statistics.* 2. ed. Boston: NEJM Book, 1992.

CAMPOS–FILHO, N. FRANCO, E. L. F. Microcomputer–assisted multivariate survival analysis using Cox's proportional hazards regression model. *Computer. Methods and Programs in Biomedicine,* v. 31, p. 81–87, 1990.

COLOSIMO, E. A., GARCIA, N. C. Testing treatment differences in censored survival data: a small sample study. *Brazilian Journal of Probability and Statistics.* v. 5, p. 135–148, 1991.

COX, D. R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B,* London, v. 34, p. 187–220, 1972.

———. Partial likelihood. *Bioometrika,* v. 62, p. 269–276, 1975.

———, HINKLEY, D. V. *Theoretical statistics.* London: Chapman and Hall, 1974.

FEINLEIB, M., MAcMAHON, B. Variation in the duration of survival of patients with chronic leukemia. *Blood,* Orlando, v. 17, p. 332–349, 1960.

KALBFLEISCH, J. D., PRENTICE, R. L. *The statistical analysis of failure time data.* New York: J. Wiley, 1980.

LEE, E. T. *Statistical methods for survival data analysis.* New York: J. Wiley, 1992.

NELSON, W. *Accelerated life testing:* statistical models, data analysis and test plans. New York: J. Wiley, 1990.

TSIATIS, A. A. A large sample study of Cox's regression model. *Annals of Statistiscs,* Hayward, v. 9, p. 93–108, 1981.

VIANA, M. B. *O estado nutricional como fator de prognóstico na leucemia linfoblástica da criança:* uma análise multivariada. São Paulo, 1993. Tese (Doutoramento em Pediatria) – Escola Paulista de Medicina, 1993.

———. et al. Malnutrition as a prognostic factor in lymphoblastic leukaemia: a multivariate analysis. *Archives of Disease in Childhood.* London, v. 71, p. 304–310, 1994.

WEI, L. J. The accelerated failure time model: a useful alternative to the Cox regression. *Statistics in Medicine,* Sussex, v. 11, p. 1871–1879, 1992.

RESUMO

Em análise de sobrevivência existem dois problemas típicos: (1) identificação de fatores de prognóstico associados à ocorrência de um evento de interesse; (2) comparação de tratamentos, controlando outros fatores que podem afetar a resposta. Para analisar dados de sobrevivência são frequentemente utilizadas duas classes de modelos: os modelos paramétricos, também denominados na literatura de modelos de tempo de vida acelerada e os modelos semiparamétricos, também conhecidos por modelos de riscos proporcionais.

Neste trabalho as duas classes de modelos são comparadas, apresentando suas vantagens e desvantagens na análise de dados de sobrevivência. Uma aplicação é feita utilizando os dados de um estudo de leucemia em crianças.

ABSTRACT

Two typical problems in survival analysis are: (1) identification of prognostic factors associated with the occurence of a particular event of interest; (2) comparison of treatment effects, while adjusting for other covariates that might affect the response. Two classes of models can be used for those purposes: parametric models, also known as accelerated lifetime models and semi–parametric models, also denominated by proportional hazards models.

In this paper the two classes of models are compared, specially showing theirs advantages and problems in the analysis of a survival data set. As an application, these classes of models are used in a data set from a study of pediatric leukemia.

MUESTREO SISTEMÁTICO EN BASES DE DATOS DEL DBASE III+

Armando H. Seuc*

1 INTRODUCCIÓN

La importancia de los procedimentos muestrales en las investigaciones de todo tipo son ampliamente reconocidas. De hecho, la inmensa mayoria de las investigaciones en todos los campos utilizan una muestra como instrumento para estudiar la población de interés. En particular el muestreo sistemático es especialmente útil cuando se cuenta con la población ordenada de alguna manera, ya que es muy fácil de realizar y se ha demostrado que en general tiene una eficacia similar a la del muestreo simple aleatorio (Cochran (1963)). Cuando la población está constituida por los records de una base de datos, concretamente del DBASEIII+, puede aprovecharse el orden en el que los records aparecen en la base (dado por el número del record) para hacer una selección sistemática de la muestra. En este trabajo describimos un procedimiento sencillo e interactivo para ejecutar dicha tarea.

* Laboratorio de Metodologia de la Investigación – Instituto Nacional Endocrinologia.

R. bras. Estat., Rio de Janeiro, 53(199/200): 15–20, jan./dez. 1992.

2 PROCEDIMIENTO

2.1 Muestreo Sistemático

La idea del muestreo sistemático es seleccionar para la muestra el p–ésimo sujeto de cada q sujetos, partiendo de que los N sujetos en la población están ordenados de alguna manera. La determinación del número p se hace mediante una tabla de números aleatorios; la única restricción es que p debe ser un número entre 1 y q. La determinación del valor de q se hace de forma que el tamaño de muestra resultante n sea (aproximadamente) el preestablecido n', este último calculado según los objetivos y características de la investigación (Cochran (1963)), Armitage (1987), Machin & Campbell (1987).

A manera de ejemplo, supongamos que de una población de diabéticos de tamaño $N = 1000$, registrada en un listado consecutivo (según el orden del número de historia clínica), se precisa seleccionar una muestra de aproximadamente 330 pacientes. Para determinar q, hacemos

$$q = N/n = 1000/330 = 3.03$$

es decir, q es aproximadamente igual a 3.

Para la determinación del valor de p, en cualquier tabla de números aleatorios seleccionamos un número entre 1 y 3, digamos el 2. El procedimiento de selección consistirá entonces en seleccionar el segundo sujeto de cada 3; es decir, se incluirán en la muestra el paciente nº 2, el nº 5, el nº 8, etc. Formalmente, las unidades en la muestra serán aquellas en las posiciones

$$p + q * i, \quad \text{para} \quad i = 0, 1, \cdots \tag{1}$$

donde i se incrementa sucesivamente hasta que se agoten las unidades en la relación ordenada.

Generalizando el enfoque anterior, la simplificación de la fracción n/N nos dirá exactamente cuántas unidades deberán seleccionarse de cada cuántas. Si denotamos r/q la simplificación de la fracción n/N, debemos entonces seleccionar r números aleatorios entre 1 y q; las unidades en la muestra serán entonces aquellas en las

posiciones

$$\begin{array}{l} p_1 + q * i, \\ p_2 + q * i, \\ \quad \cdots \\ p_j + q * i, \\ \quad \cdots \\ p_r + q * i, \end{array} \quad \text{para} \quad i = 0, 1, \cdots \qquad 1.1$$

donde los p_j son los números aleatorios seleccionados de la correspondiente tabla, y nuevamente i se incrementa sucesivamente hasta que se agoten las unidades en la relación ordenada.

2.2 Bases de Datos

Cada vez con mayor frecuencia se crean bases de datos computarizadas, las cuales almacenan grandes volúmenes de información acerca de cualesquiera unidades, ya sean éstas sujetos en general, pacientes, viviendas, animales, hospitales, etc. Al estar estas bases de datos computarizadas , la recuperación de información a partir de determinados criterios sobre las variables captadas resulta fácil y econômico.

Otra ventaja de tener estas bases de datos computarizadas es que permiten hacer una selección automatizada de muestras a partir de la población constituida por cada una de estas bases de datos. Esta ventaja desgraciadamente no ha sido suficientemente explotada; en muchas ocasiones, para seleccionar una muestra a partir de una población en una base de datos computarizada, lo que se hace es *imprimir la base de datos, y hacer la selección sobre el listado impreso resultante!!* A esta situación han contribuido varios factores: a) el DBASEIII+, uno de los sistemas de base de datos de mayor difusión, no cuenta con comandos ni funciones que permitan la selección de muestras a partir de ficheros .DBF mediante el procedimiento básico y elemental del simple aleatorio, y b) la poca generalización de procedimientos alternativos sencillos que permitan tal propósito.

2.3 Procedimiento

Cuando el muestreo sistemático es adecuado y conveniente, un procedimiento extremadamente sencillo e interactivo para seleccionar una muestra de la base de datos consiste en copiar hacia un fichero creado al efecto, aquellos records en las posiciones determinadas por la expresión (1) o de manera más general por la expresión (1.1).

En la base del procedimiento se encuentra la función matemática del DBASEIII+ "MOD" (módulo). La sintaxis de esta función es

$$MOD(r, s)$$

donde "r" y "s" son expresiones numéricas cualesquiera. Esta función retorna el residuo de dividir "r" entre "s"; por ejemplo,

$$MOD(5, 4) = 1$$

(Townsed (1986)).

Los pasos del procedimiento son: 1) Determinar los valores "p_j" y "q" a partir de los valores de N y n , según se describió en 2.1.

2) Estando en uso la base de datos con la población, la que llamaremos "FPOB.DBF", se ejecuta el comando

COPY TO FMUE FOR (MOD (RECNO ( ),q) = t_1 .or.
MOD (RECNO ( ), q) = t_2 .or. $\cdots$ .or.
MOD (RECNO ( ), q) = t_r

donde FMUE es el fichero que contendrá la muestra de records y t_j = modulo (p_j, q). Siguiendo el ejemplo, donde $j = 1$, $q = 3$ y $p_1 = 2$, entonces

t_1 = modulo (2, 3) = 2 y el comando seria

$$COPY \quad TO \quad FMUE \quad FOR \quad (MOD(RECNO(\,), 3) = 2).$$

La secuencia de pasos que acabamos de mostrar será la indicada siempre que la muestra sea un 50% o menos de la población, es decir, si $n/N \leq 0.50$. Si $n/N > 0.50$, es más cómodo "eliminar" records de la base de datos. En el ejemplo anterior, si en lugar de $n = 330$ se hubiera determinado que el tamaño de muestra necesario era $n = 670$ aproximadamente (es decir tenemos ahora que $n/N = 2/3$, el muestreo sistemático nos exigiria seleccionar dos sujetos de cada tres, lo que equivale a eliminar un sujeto de cada tres. En general, si se requiere eliminar r unidades de cada q , la secuencia de pasos seria:

1) Determinar los valores "p_j" y "q" a partir de los valores de N y n, según se describió anteriormente, con la variante de que ahora cada valor de p_j indica que la p_j-ésima unidad de cada q unidades, será eliminada.

2) Hacer copia del fichero "FPOB.DBF" , que es la base de datos que contiene la población. Este nuevo fichero, que llamaremos "FMUE.DBF" , se crea interactivamente con el comando "COPY TO FMUE" estando en uso el fichero "FPOB.DBF".

3) Estando en uso "FMUE.DBF", se ejecuta interactivamente el comando

DELETE FOR (MOD (RECNO ( ) , q) = t_1 .or $\cdots$ .or

MOD (RECNO ( ), q) = t_r) donde t_j = modulo (p_j, q)

4) Ejecutar el comando PACK.

3 CONCLUSIONES

El procedimiento descrito en este trabajo permite la selección de muestras de tamaño n de poblaciones de tamaño N para cualquiera sea la fracción n/N , cuando el método del muestreo sistemático es adecuado y conveniente.

En general, cuando la población objeto de estudio está ordenada de alguna forma (por ejemplo, el orden de los records en una base de datos computarizada, el cual puede corresponderse con el orden alfabético de los sujetos incluidos en dicha base) y no hay razones para pensar que exista una relación cíclica entre la (s) variables (s) de interés en la investigación y el orden de las unidades en la población, el muestreo sistemático es prácticamente equivalente al simple aleatorio y por lo tanto totalmente justificado (Cochran (1963)).

Consideramos por tanto que este procedimiento, debido a su sencillez, carácter interactivo y flexibilidad, merece tener una buena difusión entre estadísticos e investigadores en general.

REFERENCIAS

COCHRAN, W. G. *Sampling techniques.* New York: J. Wiley, 1963.

TOWNSEND, C. *Mastering DbaseIII Plus.* Sybex, 1986.

ARMITAGE, P., BERRY. G. *Statistical methods in medical research.* Oxford: Blackwell, 1987.

MACHIN, D., CAMPBEL, M. J. *Statistical tables for the design of clinical trials.* Oxford: Blackwell, 1987.

RESUMEN

Se describe una forma sencilla e interactiva de seleccionar una muestra de una población en un fichero del DBASEIII+, mediante muestreo sistemático.

ABSTRACT

A simple and interactive way of selecting a sample from a population in a DBASEIII+ data base file, using systematic sampling is described.

TESTES PARA HIPÓTESES NA FORMA DE DESIGUALDADES LINEARES E NÃO–LINEARES EM REGRESSÃO

Gilberto A. Paula*

1 INTRODUÇÃO

A teoria de testes do tipo *one–sided* em regressão linear é datada da década de 60 (vide Kudô, 1963). Todavia, a extensão do método para situações mais gerais, tais como hipóteses com desigualdades não–lineares em modelos não–lineares, é bastante recente (vide Kodde e Palm, 1986). Nos últimos 30 anos um grande número de trabalhos e pelo menos dois livros foram publicados na área de Inferência com Restrições de Ordem. Uma excelente referência é o livro de Robertson,Wright e Dykstra(1988).

O estudo da distribuição do teste da razão de verossimilhança para hipóteses com restrições de ordem em populações normais é talvez o principal tópico no assunto. Mostra–se que a distribuição nula do teste da razão de verossimilhança é uma mistura de Qui–quadrados (ou de distribuições do tipo F), com pesos que não dependem dos parâmetros sob a hipótese nula. Esses pesos não são em geral expressos em forma

* Instituto de Matemática e Estatística – USP.

fechada para cinco ou mais restrições. O uso de métodos de integração numérica (Bohrer e Chow, 1978), de aproximações para os pesos (Lee, Robertson e Wright, 1993) ou o estudo de situações de importância prática para as quais os pesos tomam formas simplificadas (Robertson, Wright e Dykstra, 1988), têm sido alguns dos tópicos de pesquisa no assunto.

A extensão dos resultados de k populações normais para a família exponencial, bem como para outras distribuições, por exemplo, a multinomial, tem sido também largamente estudada. Nesses casos, sob condições gerais de regularidade, o teste da razão de verossimilhança tem distribuição nula assintótica que é uma mistura de Qui–quadrados com pesos em forma similar aos do caso normal.

Embora os testes para hipóteses com restrições de ordem sejam em geral mais poderosos com relação a outros testes competitivos, a obtenção de formas fechadas para a função poder tem sido somente alcançada em situações particulares. O uso de métodos de Monte Carlo ou mesmo de aproximações em torno da hipótese nula são os procedimentos mais usuais para o cálculo do poder.

Alguns artigos com resultados relevantes relacionados com testes do tipo *one–sided* em regressão foram publicados nos últimos anos. Silvapulle(1991) tem investigado a aplicação de tais testes em modelos de regressão com log–verossimilhança côncava. Entre esses modelos estão os modelos lineares generalizados com ligação canônica e alguns modelos para análise de dados de sobrevivência. Wolak(1991) tem mostrado para uma classe geral de modelos de regressão em que a hipótese nula é expressa na forma de desigualdades não–lineares, que não necessariamente a situação menos favorável ocorre quando todas as restrições são satisfeitas em forma de igualdades. Esse resultado se verifica, em particular, para o caso usual de regressão normal linear com desigualdades lineares (vide Perlman, 1969). Wolak(1991) propôs um lema onde define um subconjunto da hipótese nula que contém a situação menos favorável. Em geral pode ser relativamente complexo obter tal subconjunto. Primeiro, pelo fato da distribuição nula assintótica, que continua sendo uma mistura de Qui–quadrados, ter uma dimensão que depende da região menos favorável. Segundo, devido à dificuldade de estimar os pesos, uma vez que esses dependem dos parâmetros da hipótese nula menos favorável. Torna–se assim importante a investigação de situações de interesse prático em que os pesos não dependam dos parâmetros sob a hipótese nula. Isso faria com que a situação menos favorável fosse obtida quando todas as desigualdades forem satisfeitas em forma de igualdades. Wolak(1989) sugere, alternativamente, a utilização de testes para hipóteses locais em que os resultados assintóticos se verificam sob condições gerais de regularidade.

Paula e Sen(1994a) verificaram para uma subclasse de modelos lineares genera-

lizados e para algumas hipóteses com restrições de ordem, que a distribuição nula assintótica da estatística da razão de verossimilhança não depende dos parâmetros sob a hipótese nula quando o estudo é balanceado. Esse resultado provavelmente pode ser estendido para outras classes de regressão.

Nas seções seguintes procuramos descrever brevemente os principais resultados relacionados com testes do tipo *one–sided* em regressão, bem como o que pode ser investigado no assunto. Na Seção 2 fazemos uma revisão para o modelo normal linear com alguns resultados recentes. A extensão para os modelos lineares generalizados é discutida na Seção 3. Um exemplo ilustrativo na área médica é apresentado na Seção 4. Finalmente, alguns possíveis tópicos de pesquisa são apresentados na última seção.

2 REGRESSÃO NORMAL LINEAR

Sejam $Y_1, \ldots, Y_n$ variáveis aleatórias independentes e normalmente distribuídas com variância constante σ^2 e médias $E(Y_i) = \mu_i, i = 1, \ldots, n$. Adicionalmente, suponha que μ_i está relacionado linearmente com $x_i = (x_{i1}, \ldots, x_{ip})^T$, isto é, $\mu(x_i) = \sum_{j=1}^{p} x_{ij}\beta_j$, onde $x_{i1}, \ldots, x_{ip}$ representam valores conhecidos de p variáveis explicativas.

As hipóteses usuais do tipo *one–sided* são definidas por

$$H_0 : C\beta = d$$
$$H_1 : C\beta \geq d,$$

com pelo menos uma desigualdade estrita em H_1. A matriz C tem dimensão $q \times p$ $(q \leq p)$, e é assumida de posto completo, $\beta = (\beta_1, \ldots, \beta_p)^T$ são parâmetros a serem estimados e d é um vetor $q-$dimensional de constantes.

A estimação de máxima verossimilhança sob a hipótese H_1 é discutida em McDonald e Diamond(1983). Kudô(1963) tem mostrado que a distribuição nula da estatística da razão de verossimilhança é uma mistura de Qui–quadrados

$$Pr\{LR_{01} \geq c\} = \sum_{\ell=0}^{q} Q(q, \ell; \Delta) Pr\{\chi^2_\ell \geq c\}, \qquad (1)$$

onde $c \geq 0$, χ^2_0 é a distribuição degenerada na origem, $\Delta = C(X^TX)^{-1}C^T$ e $Q(q, \ell; \Delta)$ denota a probabilidade do vetor $C\hat{\beta}^* - d$ ter exatamente ℓ valores positivos, com $\hat{\beta}^*$ sendo a estimativa de máxima verossimilhança sob a hipótese

H_1. Essas probabilidades são em geral muito complexas de serem obtidas para $q \geq 5$. Em Wolak(1987) há formas fechadas para $q \leq 4$. Bohrer e Chow(1978) apresentam um algoritmo para cálculo dessas probabilidades para $5 \leq q \leq 10$.

Como ilustração, suponha o caso de $q = 2$. Pelas fórmulas de Wolak(1987) essas probabilidades são expressas nas formas seguintes:

$$Q(2,0;\Delta) = \frac{1}{2}\pi^{-1}\cos^{-1}(\rho_{12}),\ Q(2,1;\Delta) = \frac{1}{2}$$

$$\text{e } Q(2,2;\Delta) = \frac{1}{2} - \frac{1}{2}\pi^{-1}\cos^{-1}(\rho_{12}),$$

onde $\rho_{12} = \delta_{12}/\sqrt{\delta_{11}\delta_{22}}$. Nesse caso, o valor crítico c_α é equivalente à solução em c dada por

$$\alpha = \sum_{\ell=0}^{2} Q(2,\ell;\Delta) Pr\{\chi^2_\ell \geq c\}.$$

De Gouriéroux, Holly e Monford(1982), a solução c_α pertence ao intervalo $[c_1\ c_2]$ onde c_j é definido por

$$\Phi_j^{-1}\left(1 - \frac{\alpha}{1-u}\right)\ (j = 1,2),$$

$u = Q(2,0;\Delta)$ e $\Phi_j(x) = Pr\{\chi^2_j \leq x\}$. Logo, o valor c_α pode ser obtido através de um método de busca no intervalo $[c_1\ c_2]$. Em Gouriéroux, Holly e Monford(1982) há uma tabela com alguns valores críticos para níveis de significância de 5% e 10%.

Outra situação de interesse prático seria a seguinte:

$$H_1 : C\beta \geq d$$
$$H_2 - H_1,$$

onde $H_2 : \beta \in \mathbb{R}^p$. Nesse caso, Perlman(1969) tem mostrado que a distribuição menos favorável sob H_1 é dada por

$$Pr\{LR_{12} \geq c\} = \sum_{\ell=0}^{q} Q(q,\ell;\Delta) Pr\{\chi^2_{q-\ell} \geq c\}, \tag{2}$$

que coincide com a distribuição dada em (1), com χ^2_ℓ sendo substituído por $\chi^2_{q-\ell}$.

A distribuição nula da estatística da razão de verossimilhança quando σ^2 é substituído pela estimativa usual s^2, é uma mistura de distribuições do tipo F, tanto em (1) quanto em (2) (vide Wolak, 1987). Simplificações e aproximações para as probabilidades $Q(q,\ell;\Delta)$ são em geral obtidas quando a matriz de correlação associada à matriz Δ não depende dos parâmetros sob a hipótese nula. Essa

propriedade se verifica, em particular, para o caso de k populações normais. Duas situações de interesse prático serão discutidas a seguir.

2.1 Ordem Simples

Sem perda de generalidade, suponha que $Y_{1j}, \ldots, Y_{kj}$ são variáveis aleatórias independentes, tais que $Y_{ij} \sim N(\mu_i, \sigma^2), i = 1, \ldots, k$ e $j = 1, \ldots, r_{ij}$. Uma situação de importância prática seria verificar se as k médias estão em ordem monotônica. Por exemplo, se μ_i denota o efeito de um tratamento segundo um nível de exposição i, e se é assumido que o efeito correspondente ao nível $i+1$ é pelo menos igual ao efeito sob o nível i, pode ser de interesse testar as seguintes hipóteses:

$$H_0 : \mu_1 = \ldots = \mu_k$$
$$H_1 : \mu_1 \leq \ldots \leq \mu_k,$$

com pelo menos uma desigualdade estrita em H_1. A distribuição nula da estatística da razão de verossimilhança, também conhecida como Qui-quadrado barra (vide Barlow e al., 1972), é dada por

$$Pr\{\bar{\chi}^2_{01} \geq c\} = \sum_{\ell=1}^{k} P_\omega(\ell, k) Pr\{\chi^2_{\ell-1} \geq c\}, \tag{3}$$

onde $\omega = (\omega_1, \ldots, \omega_k)$ denotam os pesos e nesse caso são definidos por $\omega_i = r_i$. Aqui a matriz Δ é de ordem $(k-1) \times k$, com coeficientes de correlação dados por $(\rho_{ii} = 1)$

$$\rho_{i(i+1)} = \rho_{(i+1)i} = -\left[\frac{r_i r_{i+2}}{(r_i + r_{i+1})(r_i + r_{i+2})}\right]^{1/2},$$

e $\rho_{ij} = 0$ para $\mid i - j \mid > 1$. Como pode-se notar, os níveis de probabilidade $P_\omega(\ell, k) = Q(k-1, \ell-1; \Delta)$ não dependem das médias $\mu_i's$ sob H_0. No entanto, a dificuldade de computação permanece, especialmente para $k \geq 5$.

Em particular para estudos balanceados, $r_1 = \ldots = r_k$, as probabilidades $P_\omega(\ell, k)$ são expressas numa forma recursiva simples (vide Robertson, Wright e Dykstra, 1988)

$$P_S(1, k) = \tfrac{1}{k},$$
$$P_S(k, k) = \tfrac{1}{k!}$$

e $P_S(\ell, k) = \frac{1}{k} P_S(\ell - 1, k - 1) + \frac{k-1}{k} P_S(\ell, k-1), \ell = 2, \ldots, k-1$, onde $P_S(\ell, k) = P_\omega(\ell, k)$. Siskind(1976) tem conjecturado que a distribuição nula da estatística $\bar{\chi}^2_{01}$ sob a hipótese de pesos iguais deve levar a aproximações razoáveis para o caso de pesos

desiguais quando os tamanhos amostrais não são muito diferentes. Robertson e Wright(1983) têm confirmado essa conjectura para hipóteses em ordem monotônica. Eles mostraram que a aproximação é adequada se a razão $r_{\max}/r_{\min}$ não exceder 3.5.

Recentemente, Lee, Robertson e Wright(1993) encontraram limites para (3) em duas situações de importância prática. Em particular, se os tamanhos amostrais estão em ordem monotônica, por exemplo, $r_1 \leq \ldots \leq r_k$, mostra-se que

$$\inf_{\omega} Pr\{\bar{\chi}^2_{01} \geq c\} = \sum_{\ell=1}^{k} A_k(\ell) Pr\{\chi^2_{\ell-1} \geq c\}, \tag{4}$$

onde $\inf_\omega$ denota o ínfimo com relação a todos os pesos possíveis $\omega_i = r_i$, tais que $\omega_1 \leq \ldots \leq \omega_k$, enquanto o termo $A_k(\ell)$ é o coeficiente de s^ℓ na expansão da função $sC(\frac{1}{2}s + k - 1 - \frac{1}{2}, k-1)$, com $C(a,m) = a(a-1)\ldots(a-m+1)/m!$ para todo real a e inteiro não negativo m, e $C(a,0) = 1$. De Robertson, Wright e Dykstra(1988), o supremo de (3) é dado por

$$\sup_{\omega} Pr\{\bar{\chi}^2_{01} \geq c\} = \sum_{\ell=1}^{k} \binom{k-1}{\ell-1} (\tfrac{1}{2})^{k-1} Pr\{\chi^2_{\ell-1} \geq c\}. \tag{5}$$

Esse tipo de delineamento seria conveniente para situações em que k tratamentos com graus de severidade em ordem monotônica são comparados. Nesse caso, é razoável que um número menor de indivíduos seja submetido à medida que o grau de severidade aumenta.

Lee, Robertson e Wright(1993) também encontraram limites para $Pr\{\bar{\chi}^2_{01} \geq c\}$ no caso em que $\omega_1 \leq \omega_i$ ou $\omega_1 \geq \omega_i$, $i = 2, \ldots, k$. O primeiro tipo de delineamento pode ser interessante quando um tratamento particular é muito mais suscetível a efeitos colaterais que os restantes.

Outra situação usual e de importância seria: $H_1 : \mu_1 \leq \ldots \leq \mu_k$ *versus* $H_2 - H_1$, onde $H_2 : \mu_i \in \mathbb{R}$. Nesse caso, a distribuição nula menos favorável da estatística Qui-quadrado barra é dada por

$$Pr\{\bar{\chi}^2_{12} \geq c\} = \sum_{\ell=1}^{k} P_\omega(\ell, k) Pr\{\chi^2_{k-\ell} \geq c\}. \tag{6}$$

Aqui $\inf_\omega Pr\{\bar{\chi}^2_{12} \geq c\} = \sup_\omega Pr\{\bar{\chi}^2_{01} \geq c\}$ e $\sup_\omega Pr\{\bar{\chi}^2_{12} \geq c\} = \inf_\omega Pr\{\bar{\chi}^2_{01} \geq c\}$ sob as duas hipóteses de pesos restritos discutidas acima.

2.2 Ordem Simples na Forma de Árvore

Suponha novamente $Y_{ij} \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, k$ e $j = 1, \ldots, r_{ij}$, com Y_{ij} sendo mutuamente independentes. Se as médias $\mu_1, \ldots, \mu_k$ denotam efeitos de k tratamentos, pode haver interesse particular em testar a hipótese de homogeneidade dos tratamentos contra a alternativa de que há pelo menos um melhor que um tratamento controle. Nesse caso, é assumido que o controle não é mais eficiente que os demais tratamentos. Se denotarmos por μ_1 o efeito correspondente ao controle, essa situação fica representada pelas hipóteses

$$H_0 : \mu_1 = \ldots = \mu_k$$

$$H_1 : \mu_1 \leq [\mu_2, \ldots, \mu_k],$$

onde $[\mu_2, \ldots, \mu_k]$ denota qualquer ordem e é assumido pelo menos uma desigualdade estrita em H_1. A distribuição nula da estatística Qui-quadrado barra tem a mesma forma dada em (3), contudo os coeficientes de correlação associados à matriz Δ são agora dados por $(\rho_{ii} = 1)$

$$\rho_{ij} = \rho_{ji} = \left[\frac{r_i r_j}{(r_1 + r_i)(r_1 + r_j)}\right]^{1/2},$$

para $1 \leq i \neq j \leq k$. Para pesos iguais há uma fórmula recursiva para as probabilidades $P_\omega(\ell, k)$ em Robertson, Wright e Dykstra(1988, p.83).

Similarmente ao caso anterior, mostra-se que para a situação $H_1 : \mu_1 \leq [\mu_2, \ldots, \mu_k]$ contra $H_2 - H_1$, onde $H_2 : \mu_i \in \mathbb{R}$, a distribuição nula menos favorável da estatística Qui-quadrado barra coincide com (6), com os mesmos coeficientes de correlação dados acima. Limites superiores e inferiores para $Pr\{\bar{\chi}^2_{12} \geq c\}$ sob restrições de ordem monotônica e ordem simples em forma de árvore para os pesos são dados em Lee, Robertson e Wright(1993) para os casos H_0 *versus* H_1 e H_1 *versus* $H_2 - H_1$.

Singh e Wright(1993) têm proposto aproximações para o caso de restrições de ordem simples do tipo *loop*, isto é, quando $H_1 : \mu_1 \leq \mu_j \leq \mu_k, j = 1, \ldots, k-1$. Esse tipo de hipótese seria de interesse quando vários tratamentos são comparados com um controle e não é possível testar todas as combinações de tratamentos. A hipótese de que os tratamentos são pelo menos iguais ao controle e que uma combinação particular é pelo menos tão eficiente quanto os tratamentos individuais poderia ser testada aqui.

3 MODELOS LINEARES GENERALIZADOS

Suponha $Y_1, \ldots, Y_n$ variáveis aleatórias independentes tais que Y_i pertence à família exponencial de distribuições. A média e a variância de Y_i são dadas por $\mathrm{E}(Y_i) = \mu_i$ e $\mathrm{Var}(Y_i) = \phi^{-1}V_i$, onde ϕ^{-1} denota o parâmetro de dispersão e $V_i = V(\mu_i)$ é a função de variância. Adicionando a suposição de que $g(\mu_i) = \eta_i$, onde $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, tem-se os Modelos Lineares Generalizados (vide McCullagh e Nelder, 1989).

Silvapulle(1991) tem estudado a distribuição nula assintótica da estatística da razão de verossimilhança para as seguintes hipóteses:

$$H_0 : h_1(\beta) = 0, h_2(\beta) = 0$$
$$H_1 : h_1(\beta) \neq 0, h_2(\beta) \geq 0,$$

onde $h(\beta)$ é um vetor $k - \text{dimensional}$, enquanto $h_1(\beta)$ e $h_2(\beta)$ são subvetores com dimensões q e $k - q$, respectivamente. Quando $q = 0$, há pelo menos uma desigualdade estrita em H_1. É assumido que $\Pi(\beta) = (\partial/\partial\beta)h(\beta)$ é contínua e de posto completo. Silvapulle(1991) propôs um lema bastante geral para a obtenção da distribuição nula assintótica da estatística da razão de verossimilhança no caso acima. No entanto, apenas enfocou a aplicação do mesmo em modelos de regressão para dados censurados, com distribuição normal, logística e do valor extremo. Em outro artigo, ainda não publicado, Silvapulle tem mostrado que o lema proposto pode também ser estendido para os MLGs com ligação canônica sob as condições (D) e (N) de regularidade dadas em Fahrmeir e Kaufmann(1985).

Com a suposição adicional de que a log-verossimilhança é côncava e usando os resultados de Kodde e Palm (1986, Caso II), mostra-se assintoticamente que a distribuição nula da estatística da razão de verossimilhança é uma mistura de Qui-quadrados

$$Pr\{LR_{01} \geq c\} = \sum_{\ell=0}^{k} Q(\ell, k; \Delta(\beta)) Pr\{\chi^2_\ell \geq c\},$$

onde $\Delta(\beta) = \Pi_2(\beta)(X^T V X)^{-1}\Pi_2(\beta)^T$, $\Pi_2(\beta) = (\partial/\partial\beta)h_2(\beta)$, $V = \mathrm{diag}\{V_1, \ldots, V_n\}$ e as demais quantidades são as mesmas da regressão linear. É importante lembrar que a concavidade da log-verossimilhança é garantida para MLGs com ligação canônica. Na prática a matriz $\Delta(\beta)$ pode ser estimada substituindo-se β pela estimativa irrestrita $\hat{\beta}$.

A simplificação das probabilidades $Q(\ell, k; \Delta(\beta))$ continua sendo um tópico relevante de pesquisa, especialmente para $k \geq 5$. Em particular, a obtenção de limites superiores e inferiores sob hipóteses de importância prática poderia ser investigado aqui.

O outro caso de hipóteses do tipo *one-sided* é o seguinte:

$$H_1 : h(\beta) \geq 0$$
$$H_2 - H_1,$$

onde $H_2 : \beta \in \mathbb{R}^p$. Wolak(1991) tem mostrado que devido à dependência de $\Delta(\beta)$ em β, a distribuição nula assintótica menos favorável da estatística da razão de verossimilhança pode não ser alcançada quando $h(\beta) = 0$.

Definindo o subconjunto de $h(\beta)$, $h_b(\beta) = (h_{b_1}(\beta), \ldots, h_{b_m}(\beta))^T$, onde $m \leq k$ e $b_j \in K = \{1, \ldots, k\}$, Wolak(1991) mostra para grandes amostras que a distribuição nula sob a hipótese $h_b(\beta) = 0$, tal que $h_{\bar{b}}(\beta) > 0$, é dada por

$$Pr\{LR_{12} \geq c\} = \sum_{\ell=0}^{m} Q(\ell, m; \Delta_b(\beta)) Pr\{\chi^2_{m-\ell} \geq c\},$$

onde $h_{\bar{b}}(\beta)$ é o complemento de $h_b(\beta)$, $\Delta_b(\beta) = \Pi_b(\beta)(X^T V X)^{-1}\Pi_b(\beta)^T$ e $\Pi_b(\beta) = (\partial/\partial\beta)h_b(\beta)$ tem dimensão $m \times p$. É possível mostrar apenas que a hipótese menos favorável ocorre quando $h_b(\beta) = 0$. O problema é definir esse subconjunto. Não obstante, Wolak(1991) tem encontrado para $c > 0$ os limites

$$\frac{1}{2}Pr\{\chi^2_1 \geq c\} \leq Pr\{LR_{12} \geq c\} \leq \frac{1}{2}Pr\{\chi^2_{k-1} \geq c\} + \frac{1}{2}Pr\{\chi^2_k \geq c\},$$

que em geral são conservadores, especialmente quando k cresce. Esses limites são os mesmos para $Pr\{LR_{01} \geq c\}$. Kodde e Palm(1986) apresentam os valores críticos superiores correspondentes a vários níveis de significância.

Um resultado importante que pode ser aplicado aqui (vide Perlman, 1969) é o fato da distribuição nula menos favorável ser alcançada quando $m = k$ nos casos em que $\Delta(\beta)$ não depende de β. Aplicações desse resultado são apresentadas em Paula e Sen(1994a,b).

3.1 Exemplo

Como ilustração, iremos utilizar os dados referentes a um estudo de caso–controle realizado no Setor de Anatomia e Patologia do Hospital Heliópolis, em São Paulo, no período de 1970 a 1982 (Paula e Tuder, 1986). Um total de 175 pacientes

com processo infeccioso pulmonar foi classificado segundo as seguintes variáveis: (i) $RESP$, tipo de tumor (1 maligno, 0 benigno); (ii) $IDADE$, idade em anos; (iii) $SEXO$ (1 masculino, 0 feminino); (iv) HL, intensidade da célula histiócitos–linfócitos (1 ausente, 2 discreta, 3 moderada, 4 intensa) e (v) FF, intensidade da célula fibrose–frouxa (1 ausente, 2 discreta, 3 moderada, 4 intensa). As informações referentes às variáveis $RESP, HL$ e FF foram obtidas após biópsia realizada na região pleural de cada paciente ou por autópsia no caso de óbito.

Considere, como exemplo, a aplicação do modelo logístico–linear apenas com os efeitos principais

$$\Pr\{RESP = 1 \mid \eta\} = \{1 + \exp(-\eta)\}^{-1},$$

onde $\eta = \beta_1 + \beta_2 * IDADE + \beta_3 * SEXO + \sum_{i=1}^{4} \beta_{4i} * HL_i + \sum_{i=1}^{4} \beta_{5i} * FF_i$, com HL_i e FF_i sendo variáveis binárias correspondentes aos níveis de HL e FF, respectivamente. É assumido que $\beta_{41} = \beta_{51} = 0$. Uma observação importante é que, devido ao fato da amostragem ter sido retrospectiva, o uso do modelo acima para previsões somente é válido corrigindo–se a estimativa da constante, β_1 (vide McCullagh e Nelder, 1989, p.113).

Do ponto de vista médico é muito razoável supor que a chance de tumor maligno no nível i de HL ou FF é pelo menos igual à chance no nível $i+1$. Assim, pode ser de interesse testar as seguintes hipóteses:

$$H_0 : \beta_{42} = \beta_{43} = \beta_{44} = 0$$
$$H_1 : \beta_{44} \leq \beta_{43} \leq \beta_{42} \leq 0$$

e

$$H_0 : \beta_{52} = \beta_{53} = \beta_{54} = 0$$
$$H_1 : \beta_{54} \leq \beta_{53} \leq \beta_{52} \leq 0,$$

com pelo menos uma desigualdade estrita em H_1. Isto é, verificar se há pelo menos dois grupos de intensidades em cada caso com chances diferentes de tumor maligno.

Como temos três restrições em H_1, as expressões para computação dos pesos $Q(3, \ell; \Delta)$, onde $\Delta = C(X^T V X)^{-1} C^T$ e C é uma matriz 3×9, são bem mais complexas que para o caso de duas restrições descrito na Seção 2. Wolak(1987) apresenta expressões em forma fechada para $Q(3, \ell; \Delta)$, que são reproduzidas abaixo

$$Q(3,0;\Delta) = \tfrac{1}{2} - Q(3,2;\Delta), \quad Q(3,1;\Delta) = \tfrac{1}{2} - Q(3,3;\Delta),$$
$$Q(3,2;\Delta) = \tfrac{1}{4}\pi^{-1}\{3\pi - \cos^{-1}(\rho_{12.3}) - \cos^{-1}(\rho_{13.2}) - \cos^{-1}(\rho_{23.1})\},$$
$$Q(3,3;\Delta) = \tfrac{1}{4}\pi^{-1}\{2\pi - \cos^{-1}(\rho_{12}) - \cos^{-1}(\rho_{13}) - \cos^{-1}(\rho_{23})\},$$

onde ρ_{ij} é o $ij-$ ésimo elemento da matriz de correlação associada à Δ e $\rho_{ij.k}$

denota o coeficiente de correlação parcial entre $C_i^T\hat{\beta}$ e $C_j^T\hat{\beta}$ dado $C_k^T\hat{\beta}$, sendo definido por

$$\rho_{ij.k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}},$$

onde C_i^T é a $i-$ ésima linha de C.

O desvio correspondente ao ajuste do modelo foi de $D(y;\hat{\mu}) = 157,40$ (166 graus de liberdade), indicando um ajuste adequado. O valor da estatística da razão de verossimilhança foi de $LR_{01} = 15.67$ para o caso HL e $LR_{01} = 3.57$ para o caso FF. Após computarmos $\Delta(\hat{\beta})$ e os respectivos pesos chegamos aos seguintes níveis descritivos:

$$P = 0.377 * Pr\{\chi_1^2 \geq 15.67\}$$
$$+0.375 * Pr\{\chi_2^2 \geq 15.67\} + 0.123 * Pr\{\chi_3^2 \geq 15.67\} \cong 0.0003$$

e

$$P = 0.438 * Pr\{\chi_1^2 \geq 3.57\}$$
$$+0.295 * Pr\{\chi_2^2 \geq 3.57\} + 0.062 * Pr\{\chi_3^2 \geq 3.57\} \cong 0.0946.$$

Portanto, podemos afirmar que há pelo menos dois tipos de intensidades para a célula HL com chances diferentes de tumor maligno. Está implícito que essas chances estão em ordem monotônica. Por outro lado, para o tipo de célula FF, não há evidências fortes de diferenças entre as chances de tumor maligno segundo o grau de intensidade. Os dados referentes a esse exemplo podem ser obtidos diretamente com o autor.

4 MODELOS MAIS GERAIS

Como foi mencionado anteriormente, a concavidade de alguns modelos paramétricos tem facilitado o desenvolvimento da teoria assintótica de testes do tipo *one-sided* em regressão. Em particular, para os modelos de análise de sobrevivência com distribuição do valor extremo, logística e normal, Silvapulle(1991) tem usado esse fato para estabelecer condições de regularidade sob as quais a distribuição nula assintótica da estatística da razão de verossimilhança é uma mistura de Qui-quadrados. Provavelmente é possível obter, nesse caso, algumas simplificações ou aproximações para $Pr\{LR_{01} \geq c\}$ e $Pr\{LR_{12} \geq c\}$ em certos estudos balanceados.

Aplicações dos resultados de Silvapulle(1991) em modelos mais gerais, tais como os modelos não-lineares de família exponencial (Cordeiro e Paula, 1989; Paula,1992) ou modelos de dispersão (Jorgensen, 1987) poderiam também ser investigadas.

AGRADECIMENTO

Este trabalho foi desenvolvido com suporte financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

BIBLIOGRAFIA

BARLOW, R. E. et al. *Statistical inference under order restrictions.* New York: J. Wiley, 1972.

BOHRER, R., CHOW W. Algorithm AS122. Weights for one-sided multivariate inference. *Applied Statistics*, v. 27, p. 100-104, 1978.

CORDEIRO, G. M., PAULA, G. A. Improved likelihood ratio statistics for exponential family nonlinear models. *Biometrika*, London, v. 76, p. 93-100, 1989.

FAHRMEIR, L., KAUFMANN, H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics.* Hayward, v. 13, p. 342-368, 1985.

GOURIÉROUX, C., HOLLY A., MONFORT, A. Likelihood ratio test, Wald test and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, Chicago, v. 50, p. 63-80, 1982.

KODDE, D. A., PALM, F. C. Wald criteria for jointly testing equality and inequality restrictions. *Econometrica*, Chicago, v. 54, p. 1243-1248, 1986.

KUDÔ, A. A multivariate analogue of one-sided test. *Biometrika*, London, v. 50, p. 403-418. 1963.

JORGENSEN, B. Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society B*, London, v. 49, p. 127-162. 1987.

LEE, C. C., ROBERTSON, WRIGHT, F. T. Bounds on distributions arising in order restricted inferences with restricted weights. *Biometrika*, London, v. 80, p. 405-416, 1993.

McCULLAGH, P., NELDER, J .A. *Generalized linear models.* 2. ed. London: Chapman and Hall, 1989.

McDONALD, J. M., DIAMOND, I. Fitting generalized linear models with linear inequality constraints. *GLIM Newsletter*, v. 6, p. 29-36, 1983.

PAULA, G. A. Bias correction for exponential family nonlinear models. *Journal of Statistical Computational and Simulation*, v. 40, p. 43-54, 1992.

———, SEN, P. K. *One–sided tests in generalized linear models with paralell regression lines*, [S.l.:s.n.], 1994a. (Relatório técnico, IMEUSP-9402).

———. Tests of ordered hypotheses in linkage in heredity. *Statistics and Probability Letters*, v. 20, 1994b.

———, TUDER, R. M. Utilização da regressão logística para aperfeiçoar o diagnóstico de processo infeccioso pulmonar. *Revista Ciência e Cultura*, São Paulo, v. 38, n. 6, p. 1046-1050, 1986.

PERLMAN, M. D. One-sided problems in multivariate analysis. *Annals of Mathematical Statistics*, Hayward, v. 40, p. 549-567, 1969.

ROBERTSON, T., WRIGHT, M. T. On approximation of the level probabilities and associated distributions in order restricted inference. *Biometrika*, London, v. 70, p. 597-606, 1983.

———, DYKSTRA, R. L. *Order restricted statistical inference*, New York: J. Wiley, 1988.

SILVAPULLE, M. J. On limited dependent variable models: maximum likelihood estimation and test of one-sided hypothesis. *Econometric Theory*, v. 7, p.385-395, 1991.

SINGH, B., WRIGHT, F. T. The level probabilities for a simple loop ordering. *Annals of the Institute of Statistical Mathematics*, v. 45, p. 279-292, 1993.

SISKIND, V. Approximate probability integrals and critical values for Bartholomew's test of ordered means. *Biometrika*, London, v. 63, p. 641-654, 1976.

WOLAK, F. A. An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American Statistical Association*, v. 82, p. 782-793, 1987.

———. Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models. *Econometric Theory*, v. 5, p. 1-35, 1989.

———. The local nature of hypothesis tests involving inequality constraints in nonlinear models. *Econometrica*, Chicago, v. 59, p. 981-995, 1991.

RESUMO

É comum encontrar problemas práticos na área de Estatística em que as hipóteses de interesse estão na forma de desigualdades lineares ou não-lineares. Por exemplo, numa regressão logística, pode haver interesse em testar a hipótese de homogeneidade dos riscos relativos associados a um fator particular de exposição contra a alternativa de que os mesmos estão em ordem monotônica. Ou, num modelo de análise de variância, em que um tratamento novo é comparado com um grupo de tratamentos tradicionais, pode ser de interesse testar a hipótese de homogeneidade dos tratamentos contra a alternativa de que o tratamento novo é mais eficiente que os demais tratamentos. Em ambos os casos os resultados assintóticos clássicos não se verificam mais para os testes Estatísticos. Apenas para ilustrar, o teste da razão de verossimilhança tem agora uma distribuição nula assintótica que é uma mistura de Qui-quadrados. Esse resultado vale também para as Estatísticas de Wald e de escore. Os principais resultados relacionados com o teste da razão de verossimilhança para hipóteses na forma de desigualdades lineares e não-lineares em regressão são revisados neste artigo. Inicialmente discutimos o modelo normal linear, com ênfase para o caso de k populações. Posteriormente discutimos a extensão dos resultados para os modelos lineares generalizados e outros modelos não-lineares. Um exemplo na área médica é apresentado como ilustração.

ABSTRACT

It is common to find practical problems in Statistics where the hypotheses of interest are defined in terms of linear or non-linear inequalities. For example, in a logistic regression, it could be of interest to test the hypothesis of homogeneity of the relative risks associated to a particular factor of exposure against the alternative that these risks are in a monotonic order. Or, in an analysis of variance model, where a new treatment is compared with a group of traditional treatments, it could be of interest to test the hypothesis of homogeneity of the treatments against the alternative that the new treatment is more efficient than the others. In both cases the classical asymptotic results do not hold for the statistical tests. Just for the sake of illustration, the Likelihood Ratio Test would then have as asymptotic null distribution a mixture of chi-squares distributions. This result also holds for the Walds and Scores statistics. In this paper we review the main results concerning the Likelihood Ratio Test for hypotheses in the form of linear and non-linear inequalities. We start discussing the linear normal model focusing on the k-populations case and then discuss an extension of the results for Generalized Linear Models and other non-linear models. As an illustration, we present an example in the medical field.

COMPARAÇÃO DE PODERES DOS TESTES DA RAZÃO DE VEROSSIMILHANÇA, DE WALD E ESCORE EM MODELOS LINEARES GENERALIZADOS

Gauss M. Cordeiro*
Denise A. Botter**
Silvia L. de P. Ferrari**

1 INTRODUÇÃO

As estatísticas da razão de verossimilhança, de Wald e escore são assintoticamente equivalentes no sentido de que têm a mesma distribuição assintótica não só sob a hipótese nula mas também sob uma seqüência de hipóteses alternativas de Pitman, convergindo para a hipótese nula à taxa $n^{-1/2}$, onde n é o tamanho da amostra. Um critério de escolha entre as três estatísticas pode se basear na comparação dos poderes dos testes a elas associados sob esta seqüência de alternativas até ordem $n^{-1/2}$. Peers (1971) desenvolveu expansões para as funções de poder dos três testes até ordem $n^{-1/2}$ sob alternativas de Pitman, no caso de hipóteses nulas simples. Estendendo os resultados de Peers para hipóteses nulas compostas, Hayakawa (1975) trabalhou com os testes da razão de verossimilhança e de Wald enquanto Harris e Peers (1980) trabalharam com o teste escore.

* Departamento de Estatística – UFPE.

** Departamento de Estatística – USP.

Consideremos um conjunto de variáveis aleatórias contínuas $Y_1, \ldots, Y_n$ independentes mas não necessariamente identicamente distribuídas. Denotemos por y_l o valor observado de Y_l e por $y = (y_1, \ldots, y_n)^T$ o vetor das n observações. Seja $l = l(\beta)$ a função de log-verossimilhança total dado y que depende de um parâmetro desconhecido $\beta = (\beta_1, \ldots, \beta_p)^T$ com p componentes. Sejam $U(\beta) = \partial l(\beta)/\partial\beta$ e $K = E\{U(\beta)U^T(\beta)\}$ a função escore total e a matriz de informação total de Fisher para β, respectivamente. Consideremos, também, a partição $\beta = (\beta_1^T, \beta_2^T)^T$, onde $\beta_1 = (\beta_1, \ldots, \beta_q)^T$ e $\beta_2 = (\beta_{q+1}, \ldots, \beta_p)^T$ com $q \leq p$. Queremos testar a hipótese nula $H_0 : \beta_1 = \beta_1^{(0)}$ contra a hipótese alternativa $H : \beta_1 \neq \beta_1^{(0)}$, onde $\beta_1^{(0)}$ é um vetor especificado de dimensão q.

Denotemos por $\hat{\beta}$ o estimador de máxima verossimilhança de β sob o modelo irrestrito e por $\tilde{\beta}_2$ o estimador de máxima verossimilhança de β_2 sob H_0. A partição de β induz às correspondentes partições $U^T = (U_1^T, U_2^T)$,

$$K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \quad \text{e} \quad K^{-1} = \begin{pmatrix} K^{11} & K^{12} \\ K^{21} & K^{22} \end{pmatrix},$$

onde K^{-1} representa a inversa da matriz K e K_{22}^{-1} representa a matriz de covariância assintótica de $\tilde{\beta}_2$. Definimos, também, as matrizes

$$A = \begin{pmatrix} 0 & 0 \\ 0 & K_{22}^{-1} \end{pmatrix}$$

e $M = K^{-1} - A$, ambas de dimensão $p \times p$. As funções avaliadas no ponto $\hat{\beta}$ serão escritas com um circunflexo e as avaliadas em $\tilde{\beta} = (\beta_1^{(0)T}, \tilde{\beta}_2^T)^T$ serão escritas com um til. As estatísticas da razão de verossimilhança (S_1), de Wald (S_2) e escore (S_3) utilizadas para testar H_0 são expressas como

$$S_1 = 2\{l(\hat{\beta}) - l(\tilde{\beta})\}\,, \tag{1.1}$$

$$S_2 = (\hat{\beta}_1 - \beta_1^{(0)})^T \hat{K}^{11^{-1}} (\hat{\beta}_1 - \beta_1^{(0)}) \tag{1.2}$$

e

$$S_3 = \tilde{U}_1^T \tilde{K}^{11} \tilde{U}_1\,, \tag{1.3}$$

onde $K^{11} = \{K_{11} - K_{12}K_{22}^{-1}K_{21}\}^{-1}$ é a matriz de covariância assintótica de $\hat{\beta}_1$. Para grandes amostras, a distribuição das três estatísticas sob H_0 é aproximadamente χ_q^2 e, para uma seqüência de alternativas de Pitman é aproximadamente $\chi'^2_{q,\lambda}$, isto é, qui-quadrado não-central com q graus de liberdade e um parâmetro de não-centralidade λ apropriado (Cox e Hinkley, 1974, Seção 9.3).

Consideremos a seqüência de alternativas de Pitman $H : \beta_1 = \beta_1^{(0)} + \xi$, onde $\xi = (\xi_1, \ldots, \xi_q)^T$ é um vetor cujos componentes são de ordem $n^{-1/2}$. Hayakawa (1975) e Harris e Peers (1980) mostraram que, para modelos contínuos e sob certas condições de regularidade, as funções de poder dos testes da razão de verossimilhança, de Wald e escore podem ser escritas, sob a seqüência de hipóteses alternativas definidas acima, respectivamente, como $\overline{P}_i = \overline{P}_i(x_\alpha) = 1 - P(S_i \leq x_\alpha)$, para $i = 1, 2, 3$, onde

$$P(S_i \leq x_\alpha) = G_{q,\lambda}(x_\alpha) + \sum_{j=0}^{3} b_{ij} G_{q+2j,\lambda}(x_\alpha) + o(n^{-1/2}) , \tag{1.4}$$

sendo $G_{r,\lambda}(x) = P(\chi'^2_{r,\lambda} \leq x)$. Aqui, x_α é tal que $G_{q,0}(x_\alpha) = 1 - \alpha$ e $\lambda = \delta^T K \delta$ com

$$\delta = \begin{pmatrix} I_q \\ -K_{22}^{-1} K_{21} \end{pmatrix} \xi,$$

onde I_q denota a matriz identidade de dimensão $q \times q$. As quantidades b_{ij}, de ordem $n^{-1/2}$, são funções de cumulantes conjuntos de derivadas de $l(\beta)$, dos vetores δ e ξ e das matrizes A e M. Todos os componentes dos termos b_{ij}, exceto ξ, devem ser avaliados sob H_0.

Na Seção 2, obtemos fórmulas matriciais simples para as quantidades b_{ij} para o teste de um subconjunto de parâmetros da parte sistemática de um modelo linear generalizado (MLG) (McCullagh e Nelder, 1989), supondo o parâmetro de dispersão conhecido. Na Seção 3, comparamos os poderes dos testes da razão de verossimilhança, de Wald e escore para alguns MLGs especiais e mostramos que, em alguns casos, podemos identificar regiões do espaço paramétrico onde, até ordem $n^{-1/2}$, um teste é mais poderoso que os demais. Na Seção 4, apresentamos um estudo de simulação e, na Seção 5, uma aplicação numérica.

2 PODERES DOS TESTES EM MLGS

Nesta seção, apresentamos expressões matriciais para as quantidades b_{ij} que determinam as expansões, até ordem $n^{-1/2}$, para os poderes dos testes da razão de verossimilhança, de Wald e escore sob alternativas de Pitman em MLGs.

Em MLGs as variáveis aleatórias $Y_1, \ldots, Y_n$ são supostas independentes com cada Y_l tendo densidade da forma

$$\pi(y; \theta_l, \phi) = exp[\phi\{y\theta_l - b(\theta_l)\} + a(y, \phi)] , \tag{2.1}$$

onde $a(.,.)$ e $b(.)$ são funções conhecidas. Assumimos que ϕ ($\phi > 0$) é conhecido e denominamos ϕ^{-1} de parâmetro de dispersão. Temos $E(y_l) = \mu_l = db(\theta_l)/d\theta_l$, $\text{var}(y_l) = \phi^{-1}V_l$, onde $V = V(\mu) = d\mu/d\theta$ é a função de variância e $\theta = \int V^{-1}d\mu = q(\mu)$, sendo $q(.)$ uma função biunívoca.

Um MLG é definido por (2.1) e pela parte sistemática $d(\mu_l) = \eta_l$ que relaciona a média $\mu = (\mu_1, \ldots, \mu_n)^T$ com o preditor linear $\eta = X\beta$, onde $\eta = (\eta_1, \ldots, \eta_n)^T$, X é a matriz modelo conhecida de dimensão $n \times p$ e de posto p ($p < n$) tal que $X^T = (x_1, \ldots, x_n)$ com $x_l = (x_{l1}, \ldots, x_{lp})^T$ e $\beta = (\beta_1, \ldots, \beta_p)^T$ é um conjunto de parâmetros desconhecidos a ser estimado. Assumimos que $d(.)$, denominada função de ligação, é uma função biunívoca.

Como discutido na Seção 1, consideramos a partição $\beta = (\beta_1^T, \beta_2^T)^T$ com $\beta_1 = (\beta_1, \ldots, \beta_q)^T$ e $\beta_2 = (\beta_{q+1}, \ldots, \beta_p)^T$. A matriz modelo particionada de maneira análoga é $X = (X_1, X_2)$. Estamos interessados em testar $H_0 : \beta_1 = \beta_1^{(0)}$ contra $H : \beta_1 \neq \beta_1^{(0)}$. A função escore total e a matriz de informação total de Fisher para β são, respectivamente, $U = \phi X^T W^{1/2} V^{-1/2}(y - \mu)$ e $K = \phi X^T W X$, onde $V = \text{diag}\{V_1, \ldots, V_n\}$ e $W = \text{diag}\{w_1, \ldots, w_n\}$ com $w_l = V_l^{-1}(d\mu_l/d\eta_l)^2$. Definimos $Z = \{z_{lm}\} = X(X^T W X)^{-1}X^T$ e $Z_2 = \{z_{2lm}\} = X_2(X_2^T W X_2)^{-1}X_2^T$ que, exceto pelo fator multiplicativo ϕ^{-1}, têm interpretações simples como as matrizes de covariância assintótica de $\hat{\eta} = X\hat{\beta}$ e $\tilde{\eta} = X_1\beta_1^{(0)} + X_2\tilde{\beta}_2$, respectivamente. As estatísticas S_1, S_2 and S_3 para testar H_0 vêm de (1.1)-(1.3) e são dadas por $S_1 = 2\{l(\hat{\beta}_1, \hat{\beta}_2) - l(\beta_1^{(0)}, \tilde{\beta}_2)\}$, $S_2 = \phi(\hat{\beta}_1 - \beta_1^{(0)})^T(\hat{R}^T\hat{W}\hat{R})(\hat{\beta}_1 - \beta_1^{(0)})$ e $S_3 = \tilde{s}^T\tilde{W}^{1/2}X_1(\tilde{R}^T\tilde{W}\tilde{R})^{-1}X_1^T\tilde{W}^{1/2}\tilde{s}$. Aqui $s = (s_1, \ldots, s_n)^T$, onde $s_l = \phi^{1/2}(y_l - \mu_l)V_l^{-1/2}$, é o vetor de resíduos de Pearson e $R = X_1 - X_2C$, onde $C = (X_2^T W X_2)^{-1}X_2^T W X_1$, representa uma matriz $n \times p$ cujas colunas são os vetores de coeficientes da regressão normal linear das colunas de X_1 sobre a matriz modelo X_2 com matriz de pesos W. Uma maneira muito simples de calcularmos S_3, como a diferença entre as somas de quadrados de resíduos de duas regressões normais lineares ponderadas, pode ser encontrada em Pregibon (1982). Já a estatística da razão de verossimilhança S_1 pode ser calculada como a diferença entre os desvios dos modelos definidos pelas matrizes X_2 e X (McCullagh e Nelder, 1989).

Os cumulantes conjuntos de derivadas da função de log-verossimilhança, necessários ao cálculo das quantidades b_{ij}, foram obtidos por Cordeiro (1983) e por Cordeiro, Ferrari e Paula (1993). Após a substituição destes cumulantes nas expressões dos b_{ij}'s (vide Hayakawa, 1975 e Harris e Peers, 1980) e um longo desenvolvimento algébrico, obtemos

$$b_{11} = \frac{\phi}{2}\{1^T(F + G)ET^2 1 - 1^T F T^3 1\} - \frac{1}{2}1^T F Z_{2d} T 1 ,$$

$$b_{12} = \frac{\phi}{6} 1^T (F - G) T^3 1 \ , \ b_{13} = 0 \ , \tag{2.2}$$

$$b_{21} = \frac{\phi}{2} \{1^T (F + G) E T^2 1 - 1^T F T^3 1\} - \frac{1}{2} \{2 1^T G (Z - Z_2)_d T 1 + 1^T F Z_d T 1\} \ ,$$

$$b_{22} = -\frac{\phi}{2} 1^T G T^3 1 + \frac{1}{2} 1^T (F + 2G)(Z - Z_2)_d T 1 \ , \ b_{23} = \frac{\phi}{6} 1^T (F + 2G) T^3 1 \ , \tag{2.3}$$

$$b_{31} = \frac{\phi}{2} \{1^T (F + G) E T^2 1 - 1^T F T^3 1\} - \frac{1}{2} \{1^T (F - G)(Z - Z_2)_d T 1 + 1^T F Z_{2d} T 1\} \ ,$$

$$b_{32} = \frac{1}{2} 1^T (F - G)(Z - Z_2)_d T 1 \ , \ b_{33} = \frac{\phi}{6} 1^T (F - G) T^3 1 \ . \tag{2.4}$$

onde, $T = \mathrm{diag}\{t_1, \ldots, t_n\}$ sendo t_l o l-ésimo elemento do vetor $t = (t_1, \ldots, t_n)^T = X\delta$, $T^i = \mathrm{diag}\{t_1^i, \ldots, t_n^i\}$ para $i = 2, 3$, $E = \mathrm{diag}\{e_1, \ldots, e_n\}$ sendo e_l o l-ésimo elemento do vetor $e = (e_1, \ldots, e_n)^T = X_1 \xi$, $Z_d = \mathrm{diag}\{z_{11}, \ldots, z_{nn}\}$ e $Z_{2d} = \mathrm{diag}\{z_{211}, \ldots, z_{2nn}\}$ são matrizes diagonais. Além disto, $F = \mathrm{diag}\{f_1, \ldots, f_n\}$ com $f = 1/V(d\mu/d\eta) d^2\mu/d\eta^2$, $G = \mathrm{diag}\{g_1, \ldots, g_n\}$ com $g = 1/V(d\mu/d\eta) d^2\mu/d\eta^2 - 1/V^2(dV/d\mu)(d\mu/d\eta)^3$ e 1 é o vetor de uns de dimensão n. O parâmetro de não-centralidade λ, que é $O(1)$, é dado por

$$\lambda = \phi 1^T T W T 1 \ . \tag{2.5}$$

Note-se que as quantidades b_{ij} são $O(n^{-1/2})$. Observamos que as fórmulas (2.2)-(2.5) são funções da matriz modelo X, das médias desconhecidas e do parâmetro de dispersão. Elas envolvem a função de ligação com suas primeira e segunda derivadas e a função de variância com sua primeira derivada. Todos os termos destas fórmulas podem ser calculados facilmente, uma vez que requerem apenas operações simples com matrizes. Além disto, as fórmulas (2.2)-(2.5) podem ser operadas analiticamente fornecendo expressões em forma fechada num grande número de situações em que a matriz de informação tem forma fechada. No entanto, como cada termo das quantidades b_{ij} depende da parametrização adotada, torna-se difícil interpretar as fórmulas (2.2)-(2.4).

Expansões até ordem $n^{-1/2}$ para os poderes dos testes da razão de verossimilhança, de Wald e escore na classe dos MLGs sob alternativas de Pitman seguem, então, de (1.4) com os termos b_{ij} dados por (2.2)-(2.4) e o parâmetro de não-centralidade λ dado por (2.5).

3 COMPARAÇÃO DE PODERES

O objetivo desta seção é comparar os poderes dos testes da razão de verossimilhança, de Wald e escore em MLGs, identificando regiões do espaço paramétrico onde um teste é mais poderoso que os demais. Uma vez que, até ordem $n^{-1/2}$, os três testes têm o mesmo tamanho e que, até primeira ordem, têm o mesmo poder, faz sentido compararmos seus poderes até ordem $n^{-1/2}$. Para tanto, denotemos por P_i, para $i = 1,2,3$, a soma dos termos de ordem um e $n^{-1/2}$ de $\overline{P}_i$ (vide Seção 1). É importante ressaltar aqui que todas as comparações de poderes são feitas até ordem $n^{-1/2}$ e que, portanto, quando afirmamos que um teste é mais poderoso que outro, esta afirmação é válida somente se termos de ordem inferior a $n^{-1/2}$ são ignorados.

Sejam $m(r,\lambda,x) = G_{r+6,\lambda}(x) - G_{r+4,\lambda}(x)$ e $n(r,\lambda,x) = G_{r+4,\lambda}(x) - G_{r+2,\lambda}(x)$, onde $G_{r,\lambda}$ foi definida na Seção 1. É fácil verificar que, fixados r e λ, $m(r,\lambda,x) < 0$ e $n(r,\lambda,x) < 0$, para todo valor de x. Das equações (1.4) e (2.2)-(2.4) segue que a diferença entre duas funções de poder quaisquer pode ser escrita como combinação linear das quantidades $m(r,\lambda,x)$ e $n(r,\lambda,x)$, ou seja,

$$P_1 - P_3 = k_1\, m(r,\lambda,x) + k_2\, n(r,\lambda,x)\ ,\ P_3 - P_2 = k_3\, m(r,\lambda,x) + k_4\, n(r,\lambda,x)$$

e

$$P_1 - P_2 = k_5\, m(r,\lambda,x) + k_6\, n(r,\lambda,x)\ , \tag{3.1}$$

onde

$$k_1 = \frac{\phi}{6} 1^T (F-G) T^3 1\ ,\ k_2 = \frac{1}{2} 1^T (F-G)(Z-Z_2)_d T 1\ ,\ k_3 = \frac{\phi}{2} 1^T G T^3 1\ ,$$

$$k_4 = \frac{3}{2} 1^T G (Z-Z_2)_d T 1\ ,\ k_5 = \frac{\phi}{6} 1^T (F+2G) T^3 1 \text{ e } k_6 = \frac{1}{2} 1^T (F+2G)(Z-Z_2)_d T 1\ .$$

Uma análise das equações (3.1) mostra que para os modelos com função de ligação canônica, como o modelo linear inverso para dados com distribuição exponencial, a matriz diagonal G é nula implicando em $k_3 = k_4 = 0$, ou seja, na igualdade entre P_2 e P_3. No entanto, se $k_3 \geq 0$ e $k_4 \geq 0$ com $k_3 + k_4 > 0$, observamos que $P_2 > P_3$; por outro lado, se $k_3 \leq 0$ e $k_4 \leq 0$ com $k_3 + k_4 < 0$ temos que $P_3 > P_2$. Já, $P_1 = P_2$ quando $k_5 = k_6 = 0$, isto é, quando $F = 2G$ ou $d^2\mu/d\eta^2 = 2/(3V)dV/d\mu(d\mu/d\eta)^2$. Os MLGs para os quais esta igualdade se verifica são aqueles cujas funções de ligação são definidas por $\eta = \int V^{-3/2} d\mu$, ou seja, onde as funções de log-verossimilhança são

localmente simétricas na vizinhança do verdadeiro valor do parâmetro β. Para o modelo gama, esta função é dada por $\eta = \mu^{-1/3}$. Por outro lado, se $k_5 \geq 0$ e $k_6 \geq 0$ com $k_5 + k_6 > 0$ constatamos que $P_2 > P_1$, sendo que este resultado se inverte para $k_5 \leq 0$ e $k_6 \leq 0$ com $k_5 + k_6 < 0$. Finalmente, $P_1 = P_3$ quando $k_1 = k_2 = 0$, ou seja, quando $F = G$. É obvio que esta igualdade se verifica somente para o modelo normal, independentemente da função de ligação adotada. Contudo, se $k_1 \geq 0$ e $k_2 \geq 0$ com $k_1 + k_2 > 0$ temos que $P_3 > P_1$, sendo que esta situação se inverte para $k_1 \leq 0$ e $k_2 \leq 0$ com $k_1 + k_2 < 0$. A igualdade $P_1 = P_2 = P_3$ se verifica somente para o modelo normal com ligação identidade.

Exemplo

Fazemos agora uma aplicação das equações (3.1). Consideremos um MLG para o qual expressamos as funções de ligação e de variância como

$$\eta = \begin{cases} \mu^{\gamma} & , \gamma \neq 0 \\ \log \mu & , \gamma = 0 \end{cases} \quad \text{e} \quad V = \mu^{\rho} ,$$

definidas somente para $\mu > 0$. Para η (com $\gamma \neq 0$) e V definidas acima temos

$$F - G = \frac{\rho}{\gamma^3} B , \quad G = \frac{(1-\gamma-\rho)}{\gamma^3} B , \quad F + 2G = \frac{3(1-\gamma) - 2\rho}{\gamma^3} B , \tag{3.2}$$

onde $B = \text{diag}\{b_1, \ldots, b_n\}$ com $b_l = \mu_l^{2-\rho-3\gamma}$. Substituindo (3.2) nas expressões dos k_i's em (3.1), encontramos

$$\begin{aligned} k_1 &= \frac{\phi}{6}\frac{\rho}{\gamma^3}\alpha_1 , & k_2 &= \frac{1}{2}\frac{\rho}{\gamma^3}\alpha_2 , & k_3 &= \frac{\phi}{2}\frac{(1-\gamma-\rho)}{\gamma^3}\alpha_1 , \\ k_4 &= \frac{3}{2}\frac{(1-\gamma-\rho)}{\gamma^3}\alpha_2 , & k_5 &= \frac{\phi}{6}\frac{\{3(1-\gamma)-2\rho\}}{\gamma^3}\alpha_1 , & k_6 &= \frac{1}{2}\frac{\{3(1-\gamma)-2\rho\}}{\gamma^3}\alpha_2 , \end{aligned} \tag{3.3}$$

onde $\alpha_1 = \sum b_l t_l^3$ e $\alpha_2 = \sum (z_{1l} - z_{2l}) b_l t_l$. É importante observar que para $0 < \rho < 1$, a distribuição resultante não faz parte da família exponencial (Jørgensen, 1987). Além disto, para outros valores de ρ, como por exemplo, $\rho = 1$ (modelo de Poisson), a distribuição resultante não é contínua. Como os resultados deste trabalho são válidos para MLGs contínuos, somente faz sentido compararmos poderes para certos valores de ρ. Em particular, estamos interessados em $\rho = 2$ e $\rho = 3$ que correspondem, respectivamente, aos modelos gama e normal inverso. Combinando (3.1) e (3.3), vêm, para $\gamma \neq 0$ e $\alpha_1 \leq 0$ e $\alpha_2 \leq 0$ com $\alpha_1 + \alpha_2 < 0$,

$$\text{sinal}\,(P_I - P_3) = \text{sinal}\,\frac{\rho}{\gamma} ,$$

$$\text{sinal}\,(P_3 - P_2) = \text{sinal}\,\frac{1-\gamma-\rho}{\gamma}\ ,$$

$$\text{sinal}\,(P_1 - P_2) = \text{sinal}\,\frac{3(1-\gamma)-2\rho}{\gamma}\ . \tag{3.4}$$

Por outro lado, para $\alpha_1 \geq 0$ e $\alpha_2 \geq 0$ com $\alpha_1 + \alpha_2 > 0$ os sinais das diferenças dos poderes em (3.4) se invertem.

Considerando, agora, $\gamma = 0$ ($\eta = \log \mu$) obtemos

$$F - G = \rho B\ , \quad G = (1-\rho)B\ , \quad F + 2G = (3-2\rho)B\ , \tag{3.5}$$

onde o elemento (l,l) da matriz B definida acima se reduz a $b_l = \mu_l^{2-\rho}$. A substituição de (3.5) em (3.1) conduz a

$$k_1 = \tfrac{\phi}{6}\rho\alpha_1\ , \qquad k_2 = \tfrac{1}{2}\rho\alpha_2\ , \qquad k_3 = \tfrac{\phi}{2}(1-\rho)\alpha_1\ ,$$

$$k_4 = \tfrac{3}{2}(1-\rho)\alpha_2\ , \quad k_5 = \tfrac{\phi}{6}(3-2\rho)\alpha_1\ , \quad k_6 = \tfrac{1}{2}(3-2\rho)\alpha_2\ ,$$

onde α_1 e α_2 foram definidos logo após (3.3).

Para $\alpha_1 \leq 0$ e $\alpha_2 \leq 0$ com $\alpha_1 + \alpha_2 < 0$, encontramos

$$\text{sinal}\,(P_1 - P_3) = \text{sinal}\,\rho\ ,$$

$$\text{sinal}\,(P_3 - P_2) = \text{sinal}\,(1-\rho)\ ,$$

$$\text{sinal}\,(P_1 - P_2) = \text{sinal}\,(3-2\rho)\ . \tag{3.6}$$

Novamente, para $\alpha_1 \geq 0$ e $\alpha_2 \geq 0$ com $\alpha_1 + \alpha_2 > 0$ os sinais das diferenças dos poderes em (3.6) se invertem.

Na Tabela 3.1, comparamos os poderes dos testes da razão de verossimilhança, de Wald e escore em função de valores de γ, α_1 e α_2 para $\rho = 2$ e $\rho = 3$. Esta tabela mostra que, uma vez conhecidos os sinais de α_1 e α_2, podemos construir intervalos de valores de γ onde um dos três testes é mais poderoso que os demais.

4 ESTUDO DE SIMULAÇÃO

Apresentamos nesta seção um estudo de simulação onde comparamos os desempenhos dos testes da razão de verossimilhança, de Wald e escore. Para este estudo, consideramos um vetor $y = (y_1, \ldots, y_n)^T$ de n observações amostradas de um modelo

gama simples ($X = 1$), onde assumimos ϕ conhecido e igual a 3 e adotamos a função de ligação identidade. A hipótese nula de interesse aqui é $H_0 : \mu = 3$. As estatísticas dos testes da razão de verossimilhança, de Wald e escore sob a hipótese H_0 são expressas, respectivamente, como

$$S_1 = 6n \left\{ \frac{\overline{y}}{3} - 1 - \log \overline{y} + \log 3 \right\} ,$$

$$S_2 = 3n \left(\frac{\overline{y} - 3}{\overline{y}} \right)^2 ,$$

$$S_3 = \frac{n}{3} (\overline{y} - 3)^2 \tag{4.1}$$

onde $\overline{y}$ é a média amostral.

Tabela 3.1
Comparações entre os poderes dos testes da razão de verossimilhança (P_1), de Wald (P_2) e (P_3)

		α_1 e α_2	
ρ	γ	$\alpha_1 \leq 0$ $\alpha_2 \leq 0$	$\alpha_1 \geq 0$ $\alpha_2 \geq 0$
		$\alpha_1 + \alpha_2 < 0$	$\alpha_1 + \alpha_2 > 0$
$\rho = 2$	$\gamma < -1$	$P_1 < P_3 < P_2$	$P_2 < P_3 < P_1$
(modelo gama)	$\gamma = -1$	$P_1 < P_2 = P_3$	$P_2 = P_3 < P_1$
	$-1 < \gamma < -1/3$	$P_1 < P_2 < P_3$	$P_3 < P_2 < P_1$
	$\gamma = -1/3$	$P_1 = P_2 < P_3$	$P_3 < P_1 = P_2$
	$-1/3 < \gamma < 0$	$P_2 < P_1 < P_3$	$P_3 < P_1 < P_2$
	$\gamma \geq 0$	$P_3 < P_1 < P_2$	$P_2 < P_1 < P_3$
$\rho = 3$	$\gamma < -2$	$P_1 < P_3 < P_2$	$P_2 < P_3 < P_1$
(modelo normal inverso)	$\gamma = -2$	$P_1 < P_2 = P_3$	$P_2 = P_3 < P_1$
	$-2 < \gamma < -1$	$P_1 < P_2 < P_3$	$P_3 < P_2 < P_1$
	$\gamma = -1$	$P_1 = P_2 < P_3$	$P_3 < P_1 = P_2$
	$-1 < \gamma < 0$	$P_2 < P_1 < P_3$	$P_3 < P_1 < P_2$
	$\gamma \geq 0$	$P_3 < P_1 < P_2$	$P_2 < P_1 < P_3$

Em primeiro lugar, simulamos as probabilidades de erro do tipo I dos três testes de hipóteses. Através da sub-rotina GGAMR do pacote computacional IMSL (1982), geramos sob H_0 50000 amostras de tamanho n ($n = 15, 30$ e 60) de observações da distribuição gama. Em seguida, calculamos as estatísticas S_1, S_2 e S_3, dadas em (4.1) e a proporção de vezes em que estas estatísticas excederam os valores críticos x_α obtidos da distribuição χ^2_1 para quatro níveis nominais α ($\alpha = 1,0; 2,5; 5,0$ e $10,0\%$). Essas taxas de rejeição estão apresentadas na Tabela 4.1.

Tabela 4.1

Valores simulados das taxas de rejeição (em porcentagem) das estatísticas da razão de verossimilhança (S_1), de Wald (S_2) e escore (S_3) sob a hipótese nula

n	Níveis nominais (%)	S_1	S_2	S_3
15	1,0	1,0	2,1	1,1
	2,5	2,5	3,7	2,5
	5,0	4,9	6,0	4,8
	10,0	10,0	10,6	9,8
30	1,0	1,0	1,6	1,0
	2,5	2,5	3,1	2,4
	5,0	4,9	5,5	4,8
	10,0	10,0	10,3	9,8
60	1,0	1,0	1,4	1,0
	2,5	2,6	2,9	2,5
	5,0	5,0	5,3	5,0
	10,0	9,9	10,1	9,8

Os resultados desta tabela mostram que, para todos os tamanhos de amostras considerados, as diferenças entre as taxas de rejeição de S_1 e S_3 são muito pequenas e

que estas taxas de rejeição estão muito próximas dos correspondentes níveis nominais. No entanto, as taxas de rejeição de S_2 só se aproximam das taxas de rejeição de S_1 e S_3 e, conseqüentemente, dos níveis nominais correspondentes, à medida que crescem os tamanhos das amostras.

Em segundo lugar, simulamos os valores dos poderes dos testes da razão de verossimilhança, de Wald e escore, respectivamente. Geramos 50000 amostras de tamanho n de observações da distribuição gama sob as alternativas contíguas H_n : $\mu = 3 + \xi$, onde $\xi = \delta/\sqrt{n}$ com δ assumindo os valores $-1,5$; $-0,5$; $0,5$ e $1,5$. Feito isto, calculamos os valores simulados dos poderes dos três testes, isto é, a proporção de vezes em que as estatísticas S_1, S_2 e S_3, dadas em (4.1), excederam os valores críticos x_α fixados anteriormente. As Tabelas 4.2-4.4 mostram esses valores simulados para amostras de tamanho $n = 15, 30$ e 60, respectivamente. Nestas tabelas apresentamos, também, os valores teóricos dos poderes dos três testes, calculados através das expansões assintóticas das funções de poder sob H_n até primeira ordem e até ordem $n^{-1/2}$ obtidas de:

$$
\begin{aligned}
\overline{P}_1(x_\alpha) &= G_{1,\lambda}(x_\alpha) + \tfrac{\delta^3}{27\sqrt{n}}\{2G_{1,\lambda}(x_\alpha) - 3G_{3,\lambda}(x_\alpha) + G_{5,\lambda}(x_\alpha)\} + o(n^{-1/2}) \ , \\
\overline{P}_2(x_\alpha) &= G_{1,\lambda}(x_\alpha) + \tfrac{\delta}{27\sqrt{n}}\left\{2\delta^2 G_{1,\lambda}(x_\alpha) - 9\left(\tfrac{\delta^2}{3} - 2\right) G_{3,\lambda}(x_\alpha)\right\} \\
&\quad + \tfrac{\delta}{27\sqrt{n}}\left\{9\left(\tfrac{\delta^2}{3} - 2\right) G_{5,\lambda}(x_\alpha) - 2\delta^2 G_{7,\lambda}(x_\alpha)\right\} + o(n^{-1/2}) \ , \qquad (4.2) \\
\overline{P}_3(x_\alpha) &= G_{1,\lambda}(x_\alpha) + \tfrac{\delta}{27\sqrt{n}}\left\{2\delta^2 G_{1,\lambda}(x_\alpha) - 9\left(\tfrac{\delta^2}{3} + 1\right) G_{3,\lambda}(x_\alpha)\right\} \\
&\quad + \tfrac{\delta}{27\sqrt{n}}\{9G_{5,\lambda}(x_\alpha) + \delta^2 G_{7,\lambda}(x_\alpha)\} + o(n^{-1/2}) \ ,
\end{aligned}
$$

onde o parâmetro de não-centralidade λ é dado por $\lambda = \delta^2/3$. É importante observar que na Tabela 4.2 o símbolo 0^- indica um valor negativo do poder calculado teoricamente até ordem $n^{-1/2}$.[1]

Tabela 4.2

Poderes (em porcentagem) dos testes da razão de verossimilhança, de Wald e escore, para $n = 15$

δ	Níveis nominais (%)	P_1^*			P_2^*		P_3^*	
		(0)	(1)	(2)	(1)	(2)	(1)	(2)
-1,5	1,0	4,4	4,5	4,4	10,5	11,9	1.5	1,5
	2,5	8,5	8,9	8,7	16,5	17,3	5,1	4,7
	5,0	13,9	14,7	14,4	22,8	23,2	10,6	9,7
	10,0	22,4	23,6	23,7	31,1	31,2	19,9	18,9
-0,5	1,0	1,3	1,3	1,3	2,8	4,0	0,6	0,7
	2,5	3,1	3,1	3,1	5,3	6,3	2,0	2,0
	5,0	6,0	6,0	5,9	8,5	9,4	4,7	4,6
	10,0	11,4	11,5	11,4	14,1	14,5	10,1	9,6
0,5	1,0	1,3	1,3	1,4	0⁻	1,2	2,1	2,2
	2,5	3,1	3,1	3,2	1,0	2,4	4,2	4,1
	5,0	6,0	5,9	5,9	3,4	4,5	7,2	7,1
	10,0	11,4	11,4	11,5	8,7	9,4	12,7	12,3
1,5	1,0	4,4	4,3	4,3	0⁻	0,7	7,3	7,2
	2,5	8,5	8,2	8,2	0,6	2,6	12,2	11,6
	5,0	13,9	13,2	13,4	5,1	6,3	17,3	16,7
	10,0	22,4	21,2	21,3	13,7	14,3	24,9	24,2

Nota: - P_1^*, P_2^* e P_3^* correspondem, respectivamente, aos poderes dos testes da razão de verossimilhança, de Wald e escore.

Nota: - Os números (0) e (1) indicam, respectivamente, os valores teóricos dos poderes, calculados até primeira ordem e até ordem $n^{-1/2}$.

Nota: - O número (2) indica os valores simulados dos poderes.

Tabela 4.3

Poderes (em porcentagem) dos testes da razão de verossimilhança, de Wald e escore, para $n = 30$

δ	Níveis nominais (%)	P_1^*			P_2^*		P_3^*	
		(0)	(1)	(2)	(1)	(2)	(1)	(2)
-1,5	1,0	4,4	4,5	4,6	8,7	9,7	2,3	2,4
	2,5	8,5	8,8	8,9	14,2	15,0	6,1	6,0
	5,0	13,9	14,4	14,7	20,2	20,4	16,6	16,9
	10,0	22,4	23,3	23,2	28,6	28,8	20,6	20,2
-0,5	1,0	1,3	1,3	1,3	2,4	2,9	0,8	0,9
	2,5	3,1	3,1	3,1	4,6	5,2	2,4	2,4
	5,0	6,0	6,0	6,0	7,8	8,3	5,1	5,0
	10,0	11,4	11,4	11,6	13,3	13,9	10,5	10,4
0,5	1,0	1,3	1,3	1,4	0,3	0,9	1,8	2,0
	2,5	3,1	3,1	3,2	1,6	2,3	3,9	3,9
	5,0	6,0	5,9	6,0	4,1	4,8	6,8	6,7
	10,0	11,4	11,4	11,4	9,5	10,0	12,3	12,2
1,5	1,0	4,4	4,3	4,3	0,1	1,2	6,5	6,5
	2,5	8,5	8,3	8,4	2,9	3,8	11,0	10,8
	5,0	13,9	13,4	13,5	7,7	8,4	16,3	16,0
	10,0	22,4	21,5	21,6	16,2	16,5	24,2	23,9

Nota: - P_1^*, P_2^* e P_3^* correspondem, respectivamente, aos poderes dos testes da razão de verossimilhança, de Wald e escore.

Nota: - Os números (0) e (1) indicam, respectivamente, os valores teóricos dos poderes, calculados até primeira ordem e até ordem $n^{-1/2}$.

Nota: - O número (2) indica os valores simulados dos poderes.

Tabela 4.4

Poderes (em porcentagem) dos testes da razão de verossimilhança, de Wald e escore, para $n = 60$

δ	Níveis nominais (%)	P_1^*			P_2^*		P_3^*	
		(0)	(1)	(2)	(1)	(2)	(1)	(2)
-1,5	1,0	4,4	4,4	4,5	7,5	8,1	2,9	3,0
	2,5	8,5	8,7	8,9	12,5	12,8	6,8	6,9
	5,0	13,9	14,3	14,4	18,3	18,7	12,3	12,2
	10,0	22,4	23,0	23,2	26,8	27,2	21,2	21,1
-0,5	1,0	1,3	1,3	1,4	2,1	2,4	0,9	1,0
	2,5	3,1	3,1	3,1	4,2	4,6	2,6	2,6
	5,0	6,0	6,0	6,1	7,2	7,7	5,3	5,4
	10,0	11,4	11,4	11,9	12,7	13,2	10,8	11,1
0,5	1,0	1,3	1,3	1,4	0,6	1,0	1,7	1,7
	2,5	3,1	3,1	3,0	2,0	2,4	3,6	3,6
	5,0	6,0	6,0	5,9	4,7	4,9	6,6	6,5
	10,0	11,4	11,4	11,6	10,1	10,4	12,0	12,0
1,5	1,0	4,4	4,3	4,4	1,3	1,9	5,9	5,7
	2,5	8,5	8,4	8,4	4,6	5,0	10,3	10,1
	5,0	13,9	13,6	13,7	9,5	9,8	15,6	15,6
	10,0	22,4	21,8	22,0	18,0	18,3	23,6	23,6

Nota: - P_1^*, P_2^* e P_3^* correspondem, respectivamente, aos poderes dos testes da razão de verossimilhança, de Wald e escore.

Nota: - Os números (0) e (1) indicam, respectivamente, os valores teóricos dos poderes, calculados até primeira ordem e até ordem $n^{-1/2}$.

Nota: - O número (2) indica os valores simulados dos poderes.

Uma análise destas tabelas mostra que, em quase todos os casos estudados, os valores simulados dos poderes estão mais próximos dos valores teóricos obtidos por expansões assintóticas sob H_n até ordem $n^{-1/2}$ do que dos valores teóricos calculados pelas mesmas expansões assintóticas até primeira ordem. Na comparação dos poderes simulados dos três testes encontramos as seguintes desigualdades: para valores negativos de δ, temos $P_3 < P_1 < P_2$ enquanto, para valores positivos de δ, temos esta desigualdade invertida, ou seja, $P_2 < P_1 < P_3$. Notemos que as desigualdades obtidas acima estão de acordo com as apresentadas na Tabela 3.1 (no modelo simples α_1 e α_2 apresentam o mesmo sinal que se iguala ao sinal de $\xi = \mu - \mu^{(0)}$) quando lá consideramos $\rho = 2$ e $\gamma = 1$.

5 UMA APLICAÇÃO

O objetivo desta seção é apresentar uma ilustração numérica através da qual comparamos os poderes dos testes da razão de verossimilhança, de Wald e escore. Consideremos os dados da Tabela 5.1 analisados por Cox e Snell (1981, págs. 148-150) relativos a tempos de sobrevivência y de 17 pacientes sofrendo de leucemia. Eles consideraram o modelo de regressão exponencial com estrutura log-linear $\log\mu_l = \alpha + \beta x_l$, para $l = 1, \ldots, 17$, onde μ_l é o tempo médio de sobrevivência desde o diagnóstico e x_l é o logaritmo na base 10 da contagem inicial de células brancas no sangue. McCullagh e Nelder (1989, pág. 464) consideraram a estatística da razão de verossimilhança S_1 para o teste de $H_0 : \beta = 0$ *versus* $H : \beta \neq 0$. O valor 6,826 para S_1 sugere que H_0 deve ser rejeitada ao nível de significância de 1%.

Tabela 5.1

Tempo de sobrevivência em semanas e $\log_{10}$ da contagem inicial de células brancas para 17 pacientes com leucemia

y	x	y	x	y	x
65	3,36	121	4,00	22	4,54
156	2,88	4	4,23	1	5,00
100	3,63	39	3,73	1	5,00
134	3,41	143	3,85	5	4,72
16	3,78	56	3,97	65	5,00
108	4,02	26	4,51		

O ajuste do modelo sob H_0 forneceu $\tilde{\alpha} = 4,135$ enquanto sob o modelo irrestrito encontramos $\hat{\alpha} = 8,477$ e $\hat{\beta} = -1,109$. Definindo $s = \sum x_l y_l - n\bar{x}\bar{y}$ e $s_a = \sum (x_l - \bar{x})^a$, para $a = 2, 3$, com $\bar{x} = \sum x_l/n$ e $\bar{y} = \sum y_l/n$, e lembrando que $\phi = 1$, $w_l = 1$ e $V_l = \mu_l^2$, obtemos dos resultados da Seção 2, $S_2 = \hat{\beta}^2 s_2$ e $S_3 = s^2(\bar{y}s_2)^{-1}$, respectivamente. Embora o valor 7,700 para S_2 sugira a rejeição de H_0 ao nível de significância de 1%, o valor 5,681 para S_3 sugere a não rejeição desta hipótese para este nível de significância.

Uma pergunta natural que surge agora é a seguinte: qual das três estatísticas deve ser utilizada? Um critério usual de escolha é utilizar a estatística do teste de maior poder. No entanto, como sabemos (vide Seção 3) que até primeira ordem os testes da razão de verossimilhança, de Wald e escore têm o mesmo poder, podemos comparar os poderes dos três testes até ordem $n^{-1/2}$. Estas comparações seguem da Tabela 3.1, pois o modelo exponencial com estrutura log-linear que estamos estudando é um caso particular do MLG considerado no exemplo da Seção 3 ($\gamma = 0$ e $\rho = 2$). Neste caso, para escolhermos o teste de de maior poder basta obtermos os sinais das quantidades α_1 e α_2. Mas, das expressões de α_1 e α_2 dadas logo após (3.3), temos que $\alpha_1 = \sum t_l^3$ e $\alpha_2 = \sum (z_{1l} - z_{2l}) t_l$, uma vez que $b_l = 1$, para $l = 1, \ldots, n$. Observando que $t_l = (x_l - \bar{x})\beta$ e que $z_{1l} - z_{2l} = (x_l - \bar{x})^2/s_2$, encontramos $\alpha_1 = \beta^3 s_3$ e $\alpha_2 = \beta s_3/s_2$, ou seja, $\alpha_1 = -0,097\beta^3$ e $\alpha_2 = -0,015\beta$. Logo, quando $\beta > 0$, temos $\alpha_1 < 0$ e $\alpha_2 < 0$ e, quando $\beta > 0$, temos $\alpha_1 > 0$ e $\alpha_2 > 0$. Segue, então, da Tabela 3.1, que para $\beta > 0$, o teste de maior poder é o de Wald enquanto para $\beta < 0$, o teste de maior poder é o escore. Portanto, uma vez que a estimativa de β é negativa, é razoável escolher a estatística escore para proceder o teste.

NOTAS

[1] Os valores das probabilidades $G_{r,\lambda}(x_\alpha) = P(X'^2_{r,\lambda} \leq x_\alpha)$ das fórmulas em (4.1) foram calculados através da sub-rotina MDCHN do pacote computacional IMSL (1982).

AGRADECIMENTOS

Os autores agradecem ao Editor Responsável e a um revisor anônimo pelas sugestões que ajudaram a melhorar a apresentação deste trabalho. Esta pesquisa foi parcialmente financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq.

BIBLIOGRAFIA

CORDEIRO, G. M. Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society B*, London, v. 45, p. 404-413, 1983.

———, FERRARI, S. L. P., PAULA, G. A. Improved score tests for generalized linear models. *Journal of the Royal Statistical Society B*, London, v. 55, p. 661-674, 1993.

COX, D. R., HINKLEY, D. V. *Theoretical statistics*. New York: J. Wiley, 1974.

———, SNELL, E. J. *Applied statistics*: principles and examples. London: Chapman and Hall, 1981.

HARRIS, P., PEERS, H. W. The local power of the efficient scores test statistic. *Biometrika*, London, v. 67, p. 525-529, 1980.

HAYAKAWA, T. The likelihood ratio criterion for a composite hypothesis under a local alternative. *Biometrika*, London, v. 62, p. 451-460, 1975.

IMSL LIBRARY *reference manual*. 9. ed. Houston: IMSL, 1982.

JORGENSEN, B. Exponential dispersion models. *Journal of the Royal Statistical Society B*, London, v. 49, p.127–162, 1987.

McCULLAGH, P., NELDER, J. A. *Generalized linear models*. London: Chapman and Hall, 1989.

PEERS, H. W. Likelihood ratio and associated test criteria. *Biometrika*, London, v. 58, p. 577-587, 1971.

PREGIBON, D. Score tests in GLIM with applications. In: GILCHRIST, R.,(Ed.). *GLIM 82*: Proceedings of the International Conference on Generalized Linear Models, 1., London. Berlin, Springer 1982. p. 87-97. (Lecture notes in statistics, 14).

RESUMO

Este trabalho apresenta fórmulas simples e de fácil aplicação para expansões, até ordem $n^{-1/2}$, onde n é o tamanho da amostra, dos poderes dos testes da razão de verossimilhança, de Wald e escore na classe dos modelos lineares generalizados, sob uma seqüência de hipóteses alternativas de Pitman. Os poderes dos três critérios são obtidos para o teste de um subconjunto de parâmetros da parte sistemática do modelo quando o parâmetro de dispersão é conhecido. As fórmulas derivadas apresentam vantagens em aplicações numéricas, uma vez que requerem apenas operações simples com matrizes. Elas são, também, suficientemente simples para serem utilizadas analiticamente, produzindo expressões em forma fechada quando aplicadas em modelos especiais cujas matrizes de informação de Fisher apresentam formas fechadas. Os poderes dos três testes são comparados sob condições específicas e, para alguns modelos especiais, pode-se identificar regiões do espaço paramétrico onde, até ordem $n^{-1/2}$, um teste é mais poderoso que os demais. Um estudo de simulação e uma aplicação numérica comparando os poderes dos três testes para dois modelos lineares generalizados especiais são apresentados.

ABSTRACT

This paper presents simple and easy-to-use formulas for expansions to the order of $n^{-1/2}$ – where n is the sample size – of the powers of the Likelihood Ratio, Walds and Scores tests in the class of Generalized Linear Models, under a sequence of Pitman alternative hypotheses. The powers of the three criteria are obtained for the test of a subset of parameters of the systematic part of the model when the dispersion parameter is known. The derived formulas are used to advantage in numerical applications since they only require simple matrices operations. For their simplicity they can be used analytically, yielding closed-form expressions in the special case of models having closed-form Fisher information matrices. The powers of the three tests are compared under specific conditions and for some special models it is possible to identify regions of the parameter space where, to the order , one of the tests is more powerful than the others. A simulation study and a numerical application comparing the powers of the three tests for two Generalized Linear Models are presented.

UTILIZAÇÃO DE TÉCNICAS DE ANÁLISE MULTIVARIADA PARA REDUÇÃO DE VARIÁVEIS NUM PROBLEMA DE CONTROLE ECOLÓGICO

Lucia Silva Kubrusly*

1 INTRODUÇÃO

Este trabalho originou–se em um projeto desenvolvido no departamento de Ecologia da UFRJ para avaliação e reparo dos danos causados pelos despejos, que durante quase 10 anos foram lançados no lago Batata, localizado no estado do Pará. Este rejeito era proveniente da lavagem de bauxita realizada no processo de beneficiamento do minério. O referido projeto foi iniciado em 1987, e sua base experimental consiste em coleta e análise de material da água e do sedimento, em diferentes pontos do lago, distribuídos entre as áreas impactada (mais próxima ao local de despejo), não impactada (distante dos despejos) e intermediária.

Os primeiros resultados da pesquisa indicaram que o lançamento do efluente no lago Batata deveria ser interrompido, o que ocorreu em dezembro de 1989 (veja Esteves, Bozelli, Roland (1990)). Os estudos estão orientados no sentido de rest-be-

* Laboratório Nacional de Computação Científica – LNCC/CNPq.

Título original: Data Editing Issues and Strategies at the Brazilian Central Statistical Office

R. bras. Estat., Rio de Janeiro, 53(199/200): 53–67, jan./dez. 1992.

lecer o equilíbrio ecológico do lago. Para tanto é necessário controlar determinadas variáveis ou indicadores ao longo do tempo. O trabalho de análise de material prossegue em cinco pontos de coleta abrangendo áreas impactada, não impactada e intermediária.

A fim de identificar semelhanças e diferenças entre os pontos de coleta, foi realizada uma análise de dados tendo por base um conjunto de onze variáveis observadas nos cinco pontos do lago. Foram utilizadas as técnicas de análise de grupamento e análise de componentes principais.

Pretendendo–se prosseguir por alguns anos o controle da qualidade da água, por meio de análise de material retirado nos cinco pontos de coleta, é interessante reduzir o número de variáveis observadas a fim de viabilizar a pesquisa por um longo período.

Neste trabalho o problema de redução de variáveis será tratado escolhendo–se alguns subconjuntos das onze variáveis originais com $n < 11$ variáveis. Para cada subconjunto serão aplicadas as técnicas de análise de grupamento e componentes principais. As soluções assim obtidas serão comparadas com a solução fornecida pelo conjunto original de onze variáveis. O subconjunto cuja solução for mais "próxima" da original será o escolhido.

2 A BASE DE DADOS E MÉTODOS DE ANÁLISE ESTATÍSTICA

A base de dados analisada neste trabalho se compõe de variáveis que medem concentrações de elementos encontrados no sedimento dos cinco pontos de coleta e uma variável morfológica. São elas as concentrações de alumínio (Al), bário (Ba), cobre (Cu), ferro (Fe), manganês (Mn), zinco (Zn), cálcio (Ca), magnésio (Mg), potássio (K) e sódio (Na). Como variável morfológica foi escolhida a profundidade (prof) devido a sua grande variação de acordo com as diferentes épocas do ano (enchente, cheia, vazante e seca). A amostra analisada foi escolhida tal que abrangesse todas as quatro estações da águas, por três anos seguidos.

Quanto aos métodos estatísticos utilizados, foram escolhidas as técnicas de análise de grupamento e análise de componentes principais, sempre no sentido de se obter estruturas de "semelhanças" e "diferenças" entre os pontos de coleta.

2.1 Análise de Grupamento

Segundo Lucas (1982), o problema de análise de grupamento pode ser colocado da seguinte forma:

Seja $X = \{x_1, \cdots, x_n\}$ um conjunto de características (variáveis) e

$E = \{e_1, \cdots, e_m\}$ o conjunto de elementos que se deseja grupar. Com base no conjunto X, determinar uma partição dos elementos de E em grupos g_i tal que:

se e_r e $e_s \in g_i \Longrightarrow e_r$ e e_s são semelhantes

se $e_r \in g_i$ e $e_s \in g_j \Longrightarrow e_r$ e e_s são distintos.

Em outras palavras, a análise de grupamento tem como característica detectar semelhanças e diferenças entre objetos, dado um certo conjunto de variáveis. De um modo geral, as semelhanças e diferenças detetadas são dadas pelas distâncias entre os objetos no espaço das variáveis (veja Anderberg, 1973).

2.2 Análise de Componentes Principais

O modelo de análise de componentes principais pode ser descrito da seguinte forma (veja Johnson, Wichern (1988)):

Seja $X = \{x_1, \cdots, x_n\}$ um conjunto de variáveis observadas sobre m indivíduos. As componentes principais C_p são definidas como:

$$C_p = \sum_{j=1}^{n} a_{pj} X_j$$

sujeito a:

$$\begin{gathered} var(C_p) = \quad \text{maxima} \\ \textstyle\sum_{j=1}^{n} a^2 pj = 1 \\ cor(C_p, C'_p) = 0 \quad \text{para} \quad p \neq p', \quad p = 1, \cdots, n \end{gathered}$$

Na solução do modelo acima, as componentes principais são obtidas a partir dos auto–valores e auto–vetores da matriz de correlação (ou covariância) de X, possibilitando a identificação de grupos de variáveis correlacionadas, e o aparecimento de variáveis isoladas. Portanto no presente trabalho a análise de componentes principais será utilizada para identificar "semelhanças" e "diferenças" apontadas pelas correlações entre variáveis.

Sendo o objetivo deste trabalho obter informações sobre "diferenças" e "semelhanças" entre os pontos de coleta, estes serão os objetos na análise de grupamento,

e serão as variáveis na análise de componentes principais. Para uma discussão sobre análise no espaço das variáveis e análise no espaço dos objetos veja Lebart, Morineau, Warwick (1984).

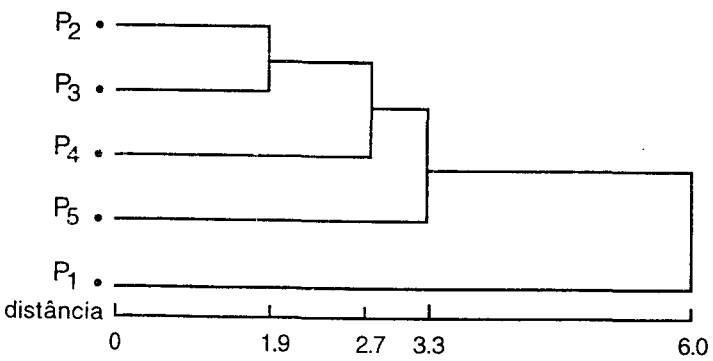
3 PADRÃO DE SEMELHANÇA FORNECIDO PELO CONJUNTO ORIGINAL DE VARIÁVEIS

3.1 Análise de Grupamento

A análise de grupamento foi realizada com auxílio do pacote BMDP–V.90, com utilização do método do centróide e distância euclidiana entre as variáveis padronizadas. Os objetos analisados se constituem dos pontos de coleta, aqui denominados P_1, P_2, P_3, P_4, P_5. O dendograma abaixo representa o resultado da análise de grupamento.

Figura 1

Análise de grupamento para o conjunto original



Para melhor entendimento desta solução é interessante observar que os pontos P_1, P_2 e P_3 estão em área considerada não impactada, P_5 está em área fortemente impactada, e P_4 está em área intermediária. Observa–se que com base no conjunto original de variáveis, os pontos P_2 e P_3 são os que mais se assemelham. Quanto aos demais, apresentam–se razoavelmente distintos. O ponto P_4 está mais "próximo" do grupo (P_2, P_3) do que P_5, e o ponto P_1 apresenta–se totalmente isolado.

Essa estrutura de semelhança/diferença pode ser explicada por P_2 e P_3 estarem

ambos na área não impactada e pertenceram ao corpo central do lago. A exclusão deste grupo dos pontos P_4 e P_5 pode ser explicada pelo grau de impacto causado pelo rejeito. Já o isolamento do ponto P_1 é explicado por razões alheias à presença do rejeito no lago. Este ponto está localizado numa região de forte estreitamento das margens, e por isso apresenta–se com características bastante diversas do restante do lago. Devido ao estreitamento, a movimentação da água é muito maior, e sendo assim, neste trecho, água e sedimento têm características mais próximas de um rio do que de um lago propriamente dito.

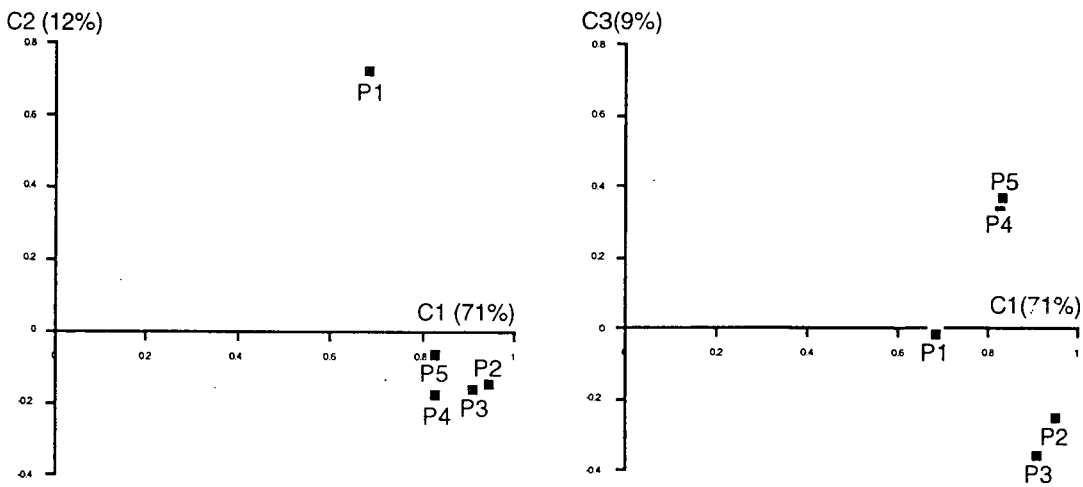
3.2 Análise de Componentes Principais

A análise de componentes principais foi realizada com auxílio do BMDP–V.90. Conforme já mencionado anteriormente, os pontos de coleta são as variáveis nesta análise.

A solução da análise de componentes principais para o conjunto original apresenta três componentes responsáveis por 92% da variância total. Os resultados estão representados na figura 2. Os eixos representam as componentes C_1, C_2 e C_3. As coordenadas dos pontos no gráfico são os pesos de cada variável (ponto de coleta) na composição da componente.

Figura 2

Análise de componentes principais para o conjunto original



É interessante observar que no plano $C_1 \times C_2$, responsável por 83% da variância total dos dados, o fenômeno mais nítido é o isolamento do ponto P_1. Os demais pontos se encontram agrupados neste plano, embora a maior proximidade dos pontos P_2 e P_3 indique uma semelhança mais acentuada entre os mesmos. Acrescentando–se a componente C_3 (responsável por 9% da variância total), aparece uma distinção

entre os pares (P_5, P_4), impactado, e (P_2, P_3), não impactado.

4 O PROBLEMA DE REDUÇÃO DE VARIÁVEIS

Diversos métodos de redução ou exclusão de variáveis têm sido utilizados (veja Jolliffe (1986) e Krzanowski (1988)) no sentido de excluir informações redundantes, ou, ao contrário, excluir variáveis isoladas em um certo conjunto, dependendo do objetivo da análise estatística realizada.

Neste trabalho, a abordagem adotada será um pouco diferente, pois o conjunto original é suposto padrão, e pretende–se obter um subconjunto que carregue as principais características deste. Mais especificamente, deseja–se escolher um subconjunto do conjunto original que reproduza a estrutura de semelhança e diferença já obtidas pela análise de grupamento e análise de componentes principais.

Como candidatos a substituir o conjunto original foram escolhidos quatro subconjuntos. Os dois primeiros foram escolhidos por critérios não estatísticos, e os outros dois foram escolhidos por critérios que serão apresentados a seguir. A escolha do melhor candidato a substituir o conjunto original terá como base o resultado que cada conjunto fornecerá, quando submetidos aos métodos de análise de grupamento e componentes principais.

A seguir estão as descrições dos quatro subconjuntos e os resultados das análises estatísticas para os mesmos

$S_1 = \{Al, Ba, Cu, Fe, Mn, Zn\}$; formado pelos elementos com as mais altas e as mais baixas concentrações.

$S_2 = \{Ca, Mg, K, Na, prof\}$; formado pelas variáveis que não estão em S_1.

$S_3 = \{Ba, Cu, Ca, Mg, K\}$.

O critério para escolha deste subconjunto foi baseado no número de diferenças apontadas nas médias de cada variável, entre todos os pares de pontos de coleta (veja (Ferreira, Esteves, Kubrusly, 1994)). Este procedimento pode ser descrito como:

Dados:

$n =$ número de variáveis originais;

p matrizes de dados com $X_1, \cdots, X_n$ variáveis, associadas aos pontos de coleta $P_1, \cdots, P_p$;

$m =$ número desejado de variáveis;

passo 0:

$q = 1$

passo 1:

tome X_q e faça teste de igualdade de médias para todos os pares de pontos de coleta

passo 2:

faça SCORE X_q = número de testes nos quais a hipótese de igualdade foi rejeitada

passo 3:

se $q = n$, escolha as m variáveis com maior SCORE e pare; senão faça $q = q + 1$ e volte para o passo 1.

O último conjunto definido foi $S_4 = \{Ba, Mn, Zn, Ca, Na\}$.

Para a escolha desse subconjunto, foi usado o procedimento descrito em Krzanowiski (1988) que pode ser resumido nos seguintes passos:

Dados:

n = número de variáveis originais;

matriz de dados com $X_1, \cdots, X_n$;

$f(x)$ = função objetivo para redução de variáveis;

m = número desejado de variáveis;

passo 0:

$q = n$

passo 1:

forme q subconjuntos com $q - 1$ variáveis

passo 2:

determine o valor de $f(x)$ para cada subconjunto e escolha aquele que otimiza f

passo 3:

faça $q = q - 1$

passo 4:

se $q = m$ pare; senão volte para o passo 1.

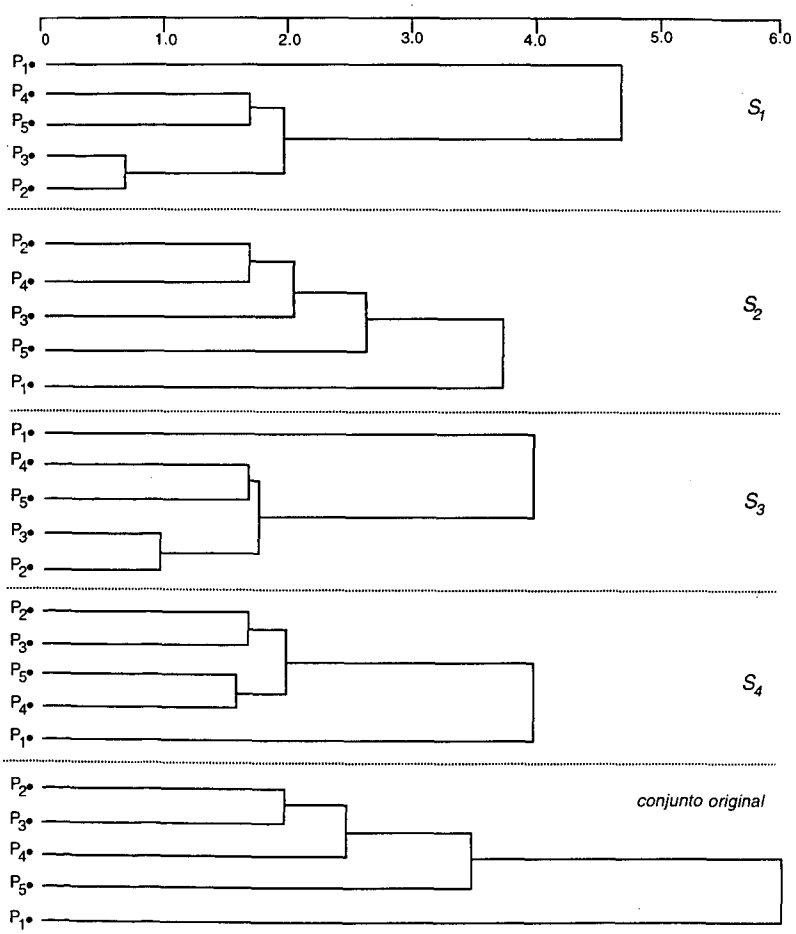
Neste trabalho $f(x)$ foi escolhida como o maior auto–valor da matriz de correlação das variáveis.

4.1 Escolha do substituto para o conjunto original com base nas soluções de análise de grupamento

A fim de proceder a escolha do substituto para o conjunto original, foram realizadas análises de grupamento para os quatro subconjuntos descritos acima. Nestas análises foram utilizados o mesmo pacote estatístico (BMDP-V.90), o mesmo método (centróide) e a mesma medida de distância (euclidiana). Será escolhido o subconjunto cuja solução da análise de grupamento for a mais próxima da solução obtida para o conjunto original. A seguir estão os resultados na figura 3.

Figura 3

Análise de Grupamento para os cinco conjuntos de variáveis



Na avaliação das diferentes soluções obtidas, os níveis de distância nos dendogramas não são diretamente comparáveis devido a diferença no número de variáveis dos diferentes conjuntos. Para efeito de comparação é necessário utilizar níveis de distância relativos. Assim, a cada nível de distância fornecido pela solução de S_i,

associa–se uma medida de distância relativa dada por

$$\frac{y_i = dS_i}{MdS_i,}$$

onde

dS_i é uma distância em S_i, e MdS_i é a distância máxima em S_i

Nas soluções apresentadas na figura 3, são destacados cinco níveis de distância, aos quais correspondem as seguintes formações de grupo:

nível 1 – (P_2, P_3)

nível 2 – (P_2, P_3, P_4)

nível 3 – (P_2, P_3, P_4, P_5)

nível 4 – (P_4, P_5)

nível final – $(P_1, P_2, P_3, P_4, P_5)$

Os níveis de distância n_1, n_2 e n_3 estão na solução do conjunto original. O nível n_4, está caracterizado em três das cinco soluções obtidas. O nível final n_f está necessariamente em todas as soluções.

Usando estes níveis de distância, e chamando o conjunto original de S_0, a medida de distância relativa descrita acima pode ser redefinida da seguinte maneira:

$$y_{ij} = \frac{n_j(S_i)}{n_f(S_i)} \quad \text{e} \quad y_{0j} = \frac{n_j(S_0)}{n_f(S_0)}$$

onde

$n_j(S_i)$ é a distância do nível j em S_i, e

$n_j(S_0)$ é a distância do nível j em S_0, $j = 1, \cdots, 4$.

A diferença entre a solução de um conjunto S_i e a solução do conjunto S_0 pode ser medida pela seguinte função:

$$f_j = \frac{1}{4}\sum_{j=1}^{4}(y_{ij} - y_{0j})^2.$$

Será escolhido como substituto do conjunto original S_0 aquele que apresentar o menor valor para f_i. Abaixo estão os valores desta função para $i = 1, 2, 3, 4$.

$$f_1 = 24.0 \times 10^{-3}$$
$$f_2 = 19.5 \times 10^{-3}$$
$$f_3 = 3.3 \times 10^{-3}$$
$$f_4 = 12.6 \times 10^{-3}$$

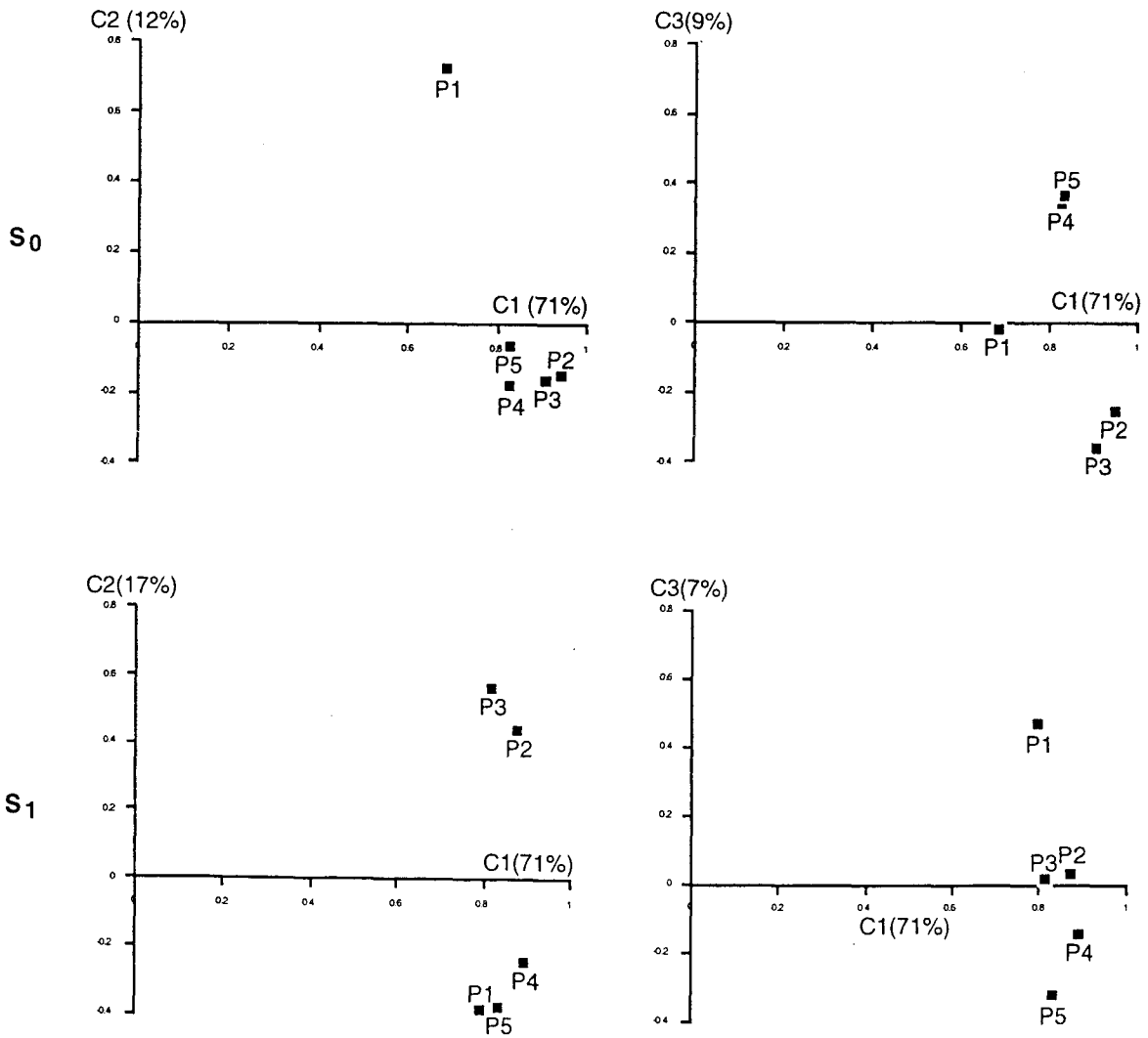
De acordo com o critério definido acima, o conjunto que melhor substitui o conjunto original na solução de análise de grupamento, é o conjunto S_3, formado pelas variáveis $\{Ba, Cu, Ca, Mg, K\}$.

4.2 Escolha do substituto para o conjunto original com base nas soluções de análise das componentes principais

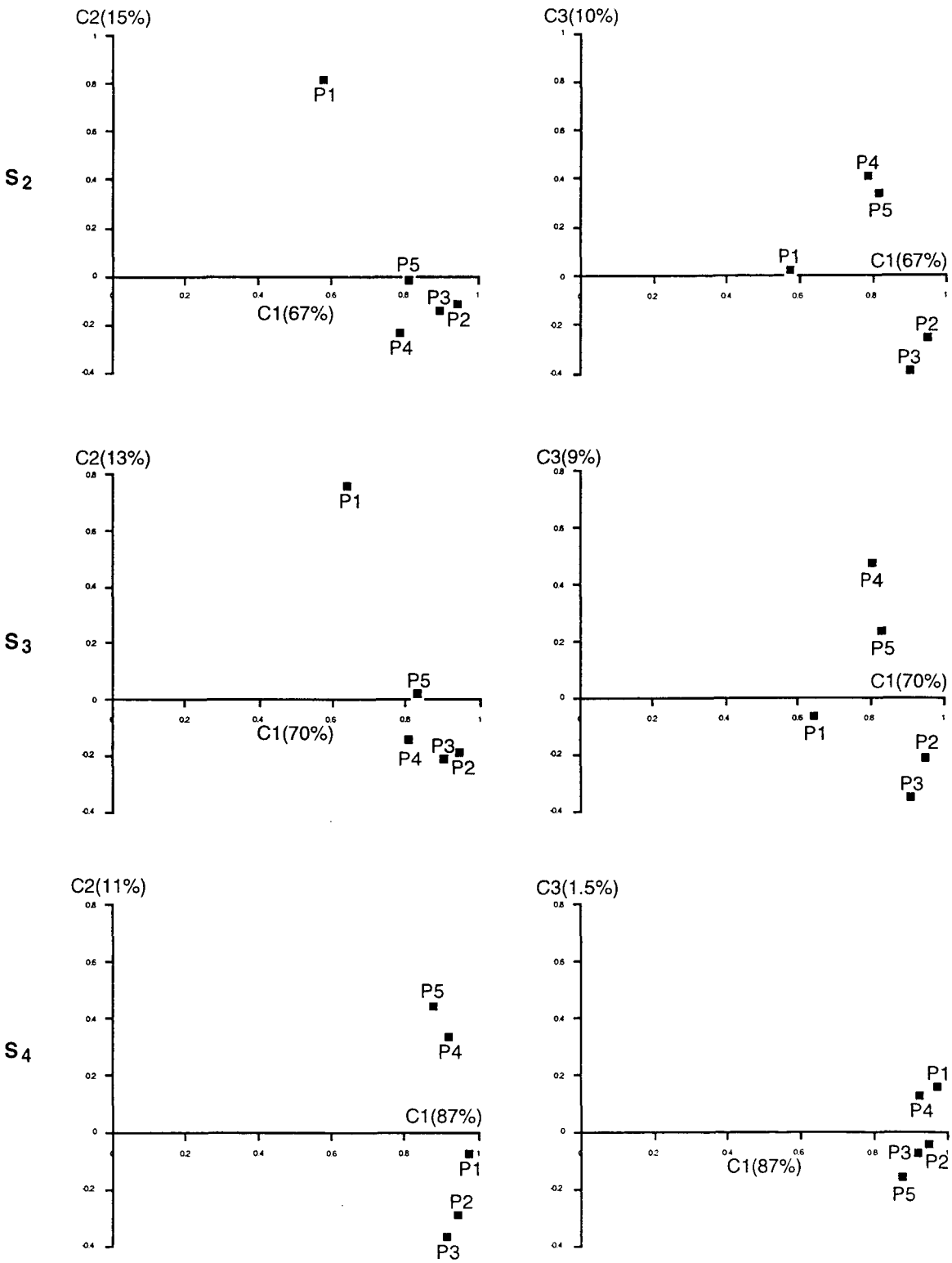
A escolha do conjunto que melhor substitui o conjunto original na análise de componentes principais, como na seção anterior, será decidida por uma avaliação das soluções fornecidas por S_i, $i = 1, \cdots, 4$. Os resultados destas análises estão representados na figura abaixo.

Figura 4

Análise de componentes principais para os cinco conjuntos de variáveis



continua ...


S2
C2(15%)
P1
P5
P3
P2
C1(67%)
P4
C3(10%)
P4
P5
P1
C1(67%)
P2
P3
S3
C2(13%)
P1
P5
C1(70%)
P4
P3
P2
C3(9%)
P4
P5
C1(70%)
P1
P2
P3
S4
C2(11%)
P5
P4
C1(87%)
P1
P2
P3
C3(1.5%)
P1
P4
P3
P2
C1(87%)
P5

A comparação da disposição dos pontos no espaço gerado por C_1, C_2, e C_3 não pode ser feita diretamente, pois C_1, C_2 e C_3 são diferentes para cada um dos cinco conjuntos. É necessário então se obter um único espaço no qual as diferentes soluções sejam representadas e comparadas.

Considerando as componentes C_1, C_2 e C_3 extraídas para os conjuntos S_i, pode–se escrever:

$$\begin{aligned} C_1^j &= \sum_{j=1}^{5} a_{1j}^i X_j \\ C_2^j &= \sum_{j=1}^{5} a_{2j}^i X_j \\ C_3^j &= \sum_{j=1}^{5} a_{3j}^i X_j \end{aligned}$$

Sendo assim, as soluções da análise de componentes principais podem ser representadas por

$$b^i = (\lambda_1^i a_{11}^i, \cdots, \lambda_1^i a_{15}^i, \lambda_2^i a_{21}^i, \cdots, \lambda_2^i a_{25}^i, \cdots, \lambda_3^i a_{31}^i, \cdots, \lambda_3^i a_{35}^i) \quad i = 0, 1, \cdots, 4$$

onde λ_p^i é a variância de C_p^i, e $i = 0$ representa a solução para o conjunto original.

Considere agora uma matriz de dados cujas colunas são os vetores $b^i \quad i = 0, 1, \cdots, 4$, e suponha que a partir desta matriz sejam extraídas as componentes principais G. As componentes assim obtidas formam um único espaço no qual os vetores b^i são representados, e eventuais semelhanças poderão ser observadas. Em outras palavras, a técnica da análise de componentes principais está sendo novamente usada para identificar semelhanças. Desta vez entre os vetores b^i, aos quais correspondem as soluções obtidas para os cinco conjuntos de variáveis. Esta abordagem é semelhante a adotada em Kubrusly e Roditi (1989).

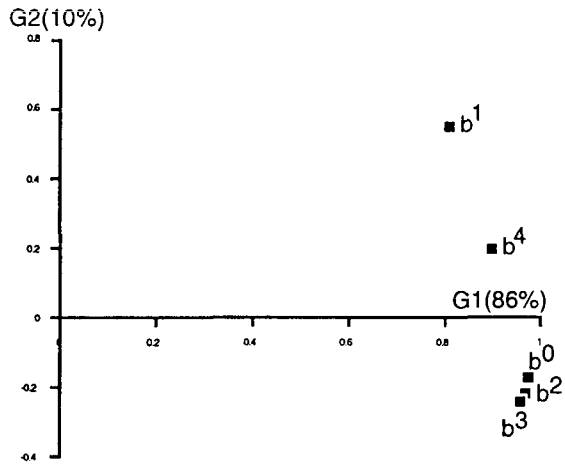
As componentes principais G podem ser escritas por:

$$G_p = \sum_{i=0}^{4} g_{ip} b^i, \quad p = 1, \cdots, 5$$

Realizando–se efetivamente esta análise, foram mantidas duas componentes G_1 e G_2, responsáveis por cerca de 93% da variância total. O resultado está representado na figura 5.

Figura 5

Representação dos vetores b_i no plano $G_1 \times G_2$



A fim de interpretar este resultado, é importante lembrar mais uma vez que os vetores b^i representam as componentes principais obtidas a partir de S_0 e S_i, e ainda que a semelhança entre estes vetores é traduzida por suas distâncias no plano $G_1 \times G_2$. Sendo assim, pode-se concluir que o conjunto S_2 (formado por $\{Ca, Mg, K, Na, prof\}$) é o que fornece solução da análise de componentes principais que mais se assemelha aquela obtida para o conjunto original. Portanto este é o subconjunto que está melhor indicado para substituir o original, quando o modelo estatístico utilizado for o de análise de componentes principais.

5 CONCLUSÃO

Neste trabalho foram apresentados dois critérios para comparação de diferentes soluções dos modelos de análise de grupamento e análise de componentes principais. Ambos os critérios foram utilizados para reduzir o nmero de variáveis presentes na análise.

O critério para comparação de soluções de análise de grupamento foi obtido por uma função de distância que mede a diferença entre pares de soluções. O critério para comparação de soluções de análise de componentes principais baseou–se em uma nova análise de componentes principais sobre os resultados já obtidos pela mesma técnica. Este critério fornece, simultaneamente, a representação das diferenças e semelhanças de todas as soluções testadas.

Na utilização dos dois critérios acima, houve uma diferença na indicação do melhor subconjunto para substituir o conjunto original. Para o modelo de análise de

grupamento, o conjunto escolhido foi S_3, cujo valor da função de distância f_i é quase quatro vezes menor que o segundo valor observado. Para o modelo de análise de componentes principais o subconjunto escolhido foi S_2. No entanto, pela figura 5, observa–se que S_3 poderia substituir o conjunto original quase tão bem quanto S_2. Neste caso parece razoável escolher S_3 para ambas as análises. Esta parece ser uma feliz coincidência, pois, dada as diferenças das duas técnicas de análise escolhidas, não há porque se esperar que o subconjunto que preserva semelhanças e diferenças na análise de grupamento seja o mesmo que preserva semelhanças e diferenças na análise de componentes principais. Isto porque, neste trabalho, esta ltima foi obtida da matriz de correlação dos dados, enquanto que a primeira teve por base a matriz de distância euclidiana.

No que diz respeito ao problema de redução de variáveis, se for acatada a proposta de se escolher S_3 para ambas as análises, ficam excluídas seis das onze variáveis originais. No entanto, considerando–se a escolha dos dois subconjuntos $S_3 = \{Ba, Cu, Ca, Mg, K\}$ e $S_2 = \{Ca, Mg, K, Na, prof\}$, nota–se a presença de três variáveis comuns. Portanto para a obtenção dos dois conjuntos é necessário a observação de apenas sete das onze variáveis originais, sendo, neste caso, excluídas as variáveis Al, Fe, Mn, Zn.

BIBLIOGRAFIA

ANDERBERG, M.R. *Cluster analysis for applications.* [S.l.]: Academic Press, 1973.

BIOMEDICAL Computer Programs – BMDP, Version 90.

ESTEVES, Bozelli Roland. Lago Batata: um laboratório de limnologia tropical. *Ciência Hoje*, São Paulo, v. 11, n. 64, 1990.

FERREIRA, C. L., ESTEVES, F. A., KUBRUSLY, L. S. Influência do rejeito de lavagem de bauxita sobre as concentrações de Al, Ba, Ca, Cu, K, Fe, Mg, Mn, Na, e Zn no sedimento de um lago amazônico. Apresentado no Congresso Brasileiro de Ecologia, Londrina, dez. 1994.

JOHNSON, R. A., WICHERN, D. W. *Applied multivariate statistical analysis.* [S.l.]: Prentice–Hall, 1988.

JOLLIFFE, I. T. *Principal component analysis.* Berlin: Springer–Verlag, 1986.

KRZANOWSKI, W. J. *Principles of multivariate analysis*: a user's perpective. Oxford: Oxford University Press, 1988.

KUBRUSLY, L. S., RODITI, D. Determinação de semelhanças regionais: uma metodologia utilizando análise de componentes principais, *Revista Brasileira de Estatística*, Rio de Janeiro, v. 50, n. 193, p. 87–99, jan./jun. 1989.

LEBART, L., MORINEAU, A., WARWICK, K. M. *Multivariate descriptive statistical analysis*. New York: J. Wiley, 1984

LUCAS, L. C. S. Análise de grupamento. *Revista Brasileira de Estatística*, Rio de Janeiro, v. 43, n. 172, p. 589–723, out./dez. 1982.

RESUMO

Em problemas de controle ambiental é freqüente a necessidade de repetição periódica de coleta e análise de dados. Neste caso, é interessante reduzir o número de variáveis envolvidas, para tornar o estudo das variações ambientais viável por um longo prazo. Neste trabalho a redução de variáveis é feita tendo como critério a comparação dos resultados das análises estatísticas realizadas no conjunto de variáveis originais, e em alguns subconjuntos do mesmo, candidatos a substituí–lo.

ABSTRACT

In solving ecological control problems it is quite often necessary to collect data and to process them periodically. In such a case, it is interesting to reduce the number of variables in order to extend the search to a longer period. In this paper a variable reduction is implemented by comparing two kinds of results. One of them comes from the statistical analysis performed on the original set S0 of variables. The other comes from the same statistical analysis performed on certain subsets of S0. Two criteria has been developed to select which of these subsets could replace the original set.

UM ALGORITMO PARA ANALISAR CADEIAS DE MARKOV ESTACIONÁRIAS E FINITAS

Marco Antônio Giacomelli*

1 INTRODUÇÃO

Uma Cadeia de Markov é um modelo probabilístico adequado para descrever uma seqüência de variáveis aleatórias observadas de acordo com um determinado parâmetro, sendo que este parâmetro geralmente é associado ao tempo decorrido.

Muitas são as aplicações de Cadeias de Markov, como por exemplo: modelagem para crescimento de populações, migrações de populações, sistemas físicos, modelos econômicos, jogos de azar, etc.

Para uma Cadeia de Markov podemos obter algumas informações e parâmetros, tais como: a existência ou não de uma distribuição de probabilidade limite, classificação de estados, os tempos médios de absorção e as probabilidades de absorção por uma subclasse fechada, além dos tempos médios de recorrência. São exatamente estas informações que visamos a obter através de um algoritmo.

Neste estudo foram consideradas apenas Cadeias estacionárias, finitas e com

* Departamento de Estatística UFRS

espaço de parâmetros discreto, uma vez que o algoritmo aqui utilizado é derivado do método desenvolvido por Isaacson e Madsen (1976) para Cadeias com estas restrições. Este método propõe um algoritmo para analisar a ergodicidade da Cadeia, denominado de método da soma das colunas, e outro algoritmo específico para a classificação de estados, cálculo dos tempos médios de absorção, das probabilidades de absorção e dos tempos médios de recorrência. Analisando estes dois algoritmos, tem–se a impressão de que são independentes um do outro. Contudo, a proposta deste artigo é derivar um único algoritmo, baseado nos dois propostos por Isaacson e Madsen (1976), tal que forneça a distribuição de probabilidade limite (se esta existir), a classificação de estados, os tempos médios de absorção, as probabilidades de absorção e os tempos médios de recorrência.

Um esboço geral do algoritmo proposto para a análise de uma Cadeia de Markov é apresentado no fluxograma 1. Nas próximas seções, cada etapa do fluxograma 1 será discutida em detalhes, propondo–se posteriormente um algoritmo para implementá–las. Por outro lado, uma visão detalhada do algoritmo para uma análise completa de uma Cadeia de Markov encontra–se no fluxograma 4.

Finalmente, cabe salientar que o algoritmo aqui proposto pode ser traduzido para uma linguagem de programação qualquer. Uma opção, por exemplo, é o "Quick Basic", que foi utilizado para implementar o fluxograma 4.

2 VERIFICANDO A ERGODICIDADE

Uma propriedade importante em uma Cadeia de Markov é a existência de ergodicidade. Esta propriedade pode ser formulada da seguinte maneira: considere $P = \{p_{ij}\}$ a matriz de transição e S o espaço de estados para uma Cadeia de Markov $\{x_m\}_{m \geq 0}$. Gostaríamos de saber se para $i \in S$

$$\lim_{m \to \infty} p_{ij}^m = \pi(j)$$

para todo $j \in S$, independentemente de i, e se

$$\sum_{j \in S} \pi(j) = 1$$

Colocando de outra maneira, queremos saber se para m suficientemente grande as linhas da matriz $P^m = \{p_{ij}^m\}$ convergirão para a distribuição de probabilidade limite (distribuição estacionária) $\pi = \{\pi(i)\}$. Se tal propriedade é verificada então

$\pi_m = \pi_0 \ P^m$ convergirá para π quando $m \to \infty$, independentemente de qual seja a distribuição inicial $\pi_0 = \{\pi_0(i)\}$, onde $\pi_m = \{\pi_m(i)\}$ é a distribuição de probabilidade da variável aleatória X_m observada no instante m.

Há quatro maneiras de se verificar a ergodicidade. A primeira delas é a menos eficiente, consistindo em obter as matrizes P^m e ao mesmo tempo observar se ocorre convergência das linhas destas para a distribuição limite. Este método claramente é ineficiente, uma vez que as linhas de P^m poderão convergir para m elevado e, por outro lado, se a cadeia não for ergódica,o processo nunca terminará devido a inexistência de um limite.

O segundo método tem como fundamento o seguinte teorema:

Teorema 3.1. Seja $\{x_m\}_{m\geq 0}$ uma Cadeia de Markov irredutível, recorrente positiva e aperiódica. Os limites $\pi(j)$ satisfazem as seguintes propriedades:

(i) $\lim_{m\to\infty} p_{ij}^m = \pi(j) = \frac{1}{\mu_j}$, onde μ_j é p tempo médio de recorrência do estado j;

(ii) $\pi(j) > 0, \quad \sum_{j\in S} \pi(j) = 1 \quad$ e $\quad \pi(j) = \sum_{k\in S} \pi(k) p_{kj}$;

(iii) a distribuição de probabilidade $\pi = \{\pi(i)\}$ que satisfaz (i) e (ii) existe e é única.

Prova: Isaacson e Madsen (1976).

O Teorema 3.1 afirma que uma vez satisfeitas suas condições então os limites $\pi(j)$ certamente existem e são equivalentes ao inverso dos tempos médios de recorrência. Um fato importante é que como estamos limitados a cadeias finitas, a condição de ser recorrente positiva sempre será verdadeira, de acordo com a proposição a seguir.

Proposição 3.1. Em uma Cadeia finita e irredutível, todos os estados são recorrentes positivos.

Prova: Isaacson e Madsen (1976).

No caso da cadeia não ser irredutível o Teorema 3.1 e a Proposição 3.1 são válidos desde que exista uma única subclasse fechada. Define–se uma subclasse fechada F como sendo um subconjunto do espaço de estados S tal que

$$\sum_{j\in F} p_{ij} = 1,$$

para qualquer $i \in F$. Assim, existirá uma distribuição limite $\pi = \{\pi(j)\}$ tal que $\pi(j) > 0$ para estados pertencentes a uma subclasse fechada e $\pi(j) = 0$ para estados transientes. Portanto, pelo segundo método deve–se classificar os estados, detectando–se as subclasses fechadas juntamente com os respectivos períodos. Para proceder–se na classificação de estados é preciso obter a matriz $P^{(n-1)^2}$, onde $n \geq 2$ denota a dimensão da matriz P. A razão disso será explicada adiante.

Outro método sugerido é oriundo das propriedades algébricas das matrizes de transição, o qual é derivado do seguinte teorema:

Teorema 3.2. Para uma Cadeia de Markov $\{x_m\}_{m\geq 0}$ finita com uma única subclasse fechada e aperiódica a respectiva matriz de transição $P = \{p_{ij}\}$ possui somente um único autovalor $\lambda_k = 1$, $k \in S$. Todos os outros autovalores são tais que $\|\lambda_j\| < 1$. Além disso, o vetor x, o qual é a solução de $xP = \lambda_k x$, é tal que suas componentes são todas não negativas.

Prova: Isaacson e Madsen (1976).

Do teorema anterior, pode–se deduzir que uma matriz de transição P com somente um único autovalor igual a 1 é irredutível e aperiódica ou possui somente uma subclasse fechada aperiódica. O autovetor x correspondente a $\lambda_k = 1$ equivale à distribuição limite π.

Este método aparentemente simples apresenta um inconveniente. O procedimento para a obtenção de autovalores é encontrar as raízes de um polinômio. Embora existam muitos algoritmos para tal finalidade, todos estão sujeitos a problemas de arredondamentos internos do computador, levando muitas vezes a resultados inconsistentes.

Para apresentar o quarto método, denominado de "método da soma das colunas", antes será preciso definir o coeficiente de ergodicidade.

Definição 3.1. Seja $P = \{p_{ij}\}$ a matriz de transição para uma Cadeia de Markov $\{X_m\}_{m\geq 0}$. O coeficiente de ergodicidade de P, denotado por $\alpha(P)$, é definido como:

$$\alpha(P) = \inf_{i<k} \sum_{j\in S} \min\{p_{ij}, p_{kj}\}, \quad i,k \in S$$

O coeficiente de ergodicidade possui a propriedade de ser limitado, ou seja, $0 \leq \alpha(P) \leq 1$. Um valor de $\alpha(P^m) = 0$, para algum $m \geq 1$, ocorre quando pelo menos um par de linhas (x,y) apresenta a seguinte propriedade: para todo $j \in S$, tem–se que

$$p_{xj}^m > 0,\ p_{yj}^m = 0 \quad \text{ou} \quad p_{xj}^m = 0\ p_{yj}^m > 0 \quad \text{ou} \quad p_{xj}^m = p_{yj}^m = 0$$

Essa característica implica em

$$\sum_{j\in S} \min\{p_{xj}^m, p_{yj}^m\} = 0$$

e conseqüentemente

$$\inf_{i<k} \sum_{j\in S} \min\{p_{ij}^m, p_{kj}^m\} = 0$$

A existência dessa propriedade em uma matriz de transição pode ser devido a uma das seguintes situações: ou a cadeia possui uma única subclasse fechada periódica ou possui mais de uma subclasse fechada. Com base nas afirmações antes colocadas poderíamos concluir que se $\alpha(P^m) = 0$, para algum $m \geq 1$, então certamente a cadeia não é ergódica. Entretanto, tal conclusão não é correta. Pode acontecer de $\alpha(P^k) = 0$ e $\alpha(P^m) > 0$ para $m > k$. Daí, o procedimento de concluir sobre a inexistência de ergodicidade para o primeiro valor de k tal que $\alpha(P^k) = 0$ é errôneo. Por outro lado, $\alpha(P^k) > 0$ implica em existência de ergodicidade.

Uma interpretação para o coeficiente de ergodicidade, quando este é maior que zero, é o grau de estabilidade das linhas da matriz de transição. Define–se estabilidade como a propriedade de uma matriz possuir todas as linhas iguais. No caso de $\alpha(P^m) = 1$, está–se diante de uma matriz a qual atingiu perfeita estabilidade de suas linhas. Em outro extremo, $\alpha(P^m)$ próximo de zero implica em uma matriz distante da estabilidade.

Considerando ainda o problema da ergodicidade, a solução para verificá–la pode ser o procedimento de calcular $\alpha(P^m)$ até que este seja maior que zero. Mas claramente este procedimento é ineficiente, pois se caso a cadeia não for ergódica o processo de calcular $\alpha(P^m)$ será infinito. Uma regra para decisão sobre a ergodicidade é dada pelo seguinte teorema:

Teorema 3.3. Considere $P = \{p_{ij}\}$ a matriz de transição para uma Cadeia de Markov $\{X_m\}_{m\geq 0}$. Então, $\alpha(P^k) > 0$ para algum $k \in \mathbb{N}^*$ se e somente se $\alpha(P^{n(n-1)/2}) > 0$, onde $n \geq 2$ é a cardinalidade do espaço de estados.

Prova: Isaacson e Madsen (1976).

De acordo com o teorema acima, o procedimento consiste em calcular $\alpha(P^{n(n-1)/2})$ e verificar se este é maior que zero, caso não for, então a cadeia não é ergódica.

Comparando com o método que se utiliza da classificação de estados é preferível utilizar o coeficiente de ergodicidade, pois $\frac{n(n-1)}{2} \leq (n-1)^2$, $n \geq 2$. Outra vantagem em utilizar–se o coeficiente de ergodicidade é a possibilidade de derivar um método ainda mais eficaz, denominado de método da soma das colunas.

De início surge imediatamente uma pergunta: existe uma maneira rápida e precisa para o cálculo de $\alpha(P^m)$? A solução para tal problema é substituir P por outra matriz $Z = \{z_{ij}\}$, denominada de matriz adjacente. A matriz Z consistirá de "0" e "1", ou seja, troca–se as entradas positivas de P por 1 e mantém–se as iguais a zero. Fica claro que Z não é uma matriz estocástica, mas mesmo assim ainda podemos

definir o coeficiente de ergodicidade como:

$$\alpha(Z^m) = \inf_{i<k} \sum_{j \in S} \min\{z_{ij}^m, z_{kj}^m\}$$

para $i, k \in S$ e $m \geq 1$. O coeficiente $\alpha(Z^m)$ terá um limite inferior igual a zero e um limite superior igual à dimensão n da matriz Z. As conclusões a respeito de $\alpha(Z^m)$ são análogas àquelas sobre $\alpha(P^m)$. Note que o coeficiente $\alpha(Z^m)$ não terá a mesma interpretação de $\alpha(P^m)$ quanto à estabilidade das linhas.

A principal vantagem em utilizar a matriz Z está em

$$\sum_{j \in S} \min\{z_{ij}, z_{kj}\} = \sum_{j \in S} z_{ij} z_{kj},$$

sendo muito mais rápido de se calcular no computador. Outra vantagem da matriz adjacente é o fato das entradas da matriz–produto Z^m resultarem sempre em números inteiros.

Uma propriedade do coeficiente de ergodicidade (ver Isaacson e Madsen (1976)) é que se $\alpha(P^k) > 0$ então $\alpha(P^m) > 0$ para $m > k$. Esta propriedade é bastante razoável, pois em uma Cadeia ergódica as linhas da matriz P^m convergem para a distribuição limite quando $m \to \infty$ e conseqüentemente $\alpha(P^m)$ convergirá monotonicamente para 1. Esse comportamento do coeficiente de ergodicidade também é válido para a matriz Z, ou seja, se $\alpha(Z^k) > 0$ então $\alpha(Z^m) > 0$ para $m > k$. Sendo assim, esta propriedade do coeficiente de ergodicidade permite–nos trabalhar com a seqüência de matrizes $Z, Z^2, Z^4, \cdots, Z^{2^k}$ ao invés da seqüência $Z, Z^2, Z^3, \cdots, Z^k$. Utilizando a primeira, o número de multiplicações de matrizes será menor.

O algoritmo do método da soma das colunas fundamenta–se no cálculo do coeficiente de ergodicidade. Para $2^k < \frac{n(n-1)}{2}$, verifica–se se alguma coluna de Z^{2^k} soma n ou $n-1$. Se a soma for n então $\alpha(Z^{2^k})$ certamente será maior que zero, levando–nos a concluir que a Cadeia é ergódica. Caso a soma for $n-1$, existirá uma possibilidade do coeficiente ser positivo, e portanto, calcula–se $\alpha(Z^{2^k})$. Quando $2^k \geq \frac{n(n-1)}{2}$, calcula–se $\alpha(Z^{2^k})$. Se $\alpha(Z^{2^k}) > 0$ concluimos que a Cadeia é ergódica, caso contrário a Cadeia será declarada como não ergódica.

3 CLASSIFICAÇÃO DE ESTADOS

Para obter os tempos médios de absorção, as probabilidades de absorção, os tempos médios de recorrência e a distribuição limite π necessitamos primeiro classificar estados. O primeiro passo é encontrar um estado recorrente e depois detectar todos os estados que se comunicam com ele. Continua–se este procedimento até que todas as subclasses fechadas tenham sido identificadas. Os estados que não forem classificados como recorrentes serão considerados transientes. A melhor alternativa para se encontrar um estado recorrente i é dada pelo seguinte teorema:

Teorema 4.1. Considere $\{x_m\}_{m\geq 0}$ uma Cadeia de Markov. O estado i é recorrente se a soma dos elementos da respectiva coluna na matriz Z^M, com $M=(n-1)^2$, for um máximo, isto é:

$$\sum_{j\in S} z_{ji}^M = \max_{k\in S} \sum_{j\in S} z_{jk}^M, \quad i,k \in S$$

Prova: Isaacson e Madsen (1976).

Obtido o estado recorrente i, a classe fechada $F(i)$ gerada por este é constituída como segue. Seja

$$K_0=\{j: z_{ij}^M = 1\}, \quad K_1=\{j: z_{ij}^{M+1}=1\}, \cdots, K_d=\{j: z_{ij}^{M+d}=1\}.$$

Caso $K_1 = K_0$, então a classe de equivalência gerada por i é aperiódica, permitindo–nos concluir que $K_0 = F(i)$. Caso $K_d = K_0$, para $d>1$, então

$$F(i) = \bigcup_{j=0}^{d-1} K_j.$$

O período de $F(i)$ é dado por d.

O próximo passo será encontrar os outros estados recorrentes, juntamente com suas respectivas classes de equivalência. Até aqui todos os estados de $F(i)$ foram classificados, podendo–se omiti–los do próximo passo. Além disso, os estados externos a $F(i)$, pelos quais $F(i)$ pode ser alcançada, devem ser transientes. Logo, estes estados também têm de ser desconsiderados. Devido a estas restrições, a matriz Z^M deverá ser alterada, substituindo–se por zeros todas as linhas e colunas correspondentes aos estados que possam alcançar $F(i)$. Para tal, escolhe–se um estado qualquer de cada classe k_j, pois como estas são classes de equivalência a

escolha de um estado dentro destas é arbitrária. Este fato é devido à propriedade das classes de equivalência, a qual diz que estados pertencentes a uma mesma classe possuem propriedades em comum. Por convenção será utilizado o primeiro elemento das classes de equivalência.

Vamos definir

$$D_0 = \{j: z^M_{Jk_0} = 1\}, \quad D_1 = \{j: z^M_{Jk_1} = 1\}, \cdots, D_{d-1} = \{j: z^M_{Jk_{d-1}} = 1\},$$

onde $k_0, k_1, \cdots, k_{d-1}$ são os primeiros elementos de $K_0, K_1, \cdots, K_{d-1}$, respectivamente. Agora obtemos

$$D = \bigcup_{v=0}^{d-1} D_v$$

descontando–se as interseções, pois as D_v nem sempre serão disjuntas. A classe D inclui todos os estados que intercomunicam–se com o estado recorrente i e todos os estados transientes que possam alcançá–lo. Agora é definida uma nova matriz $Z^* = \{z^*_{ij}\}$, tal que $z^*_{ij} = z^M_{ij}$ se ambos estados i e j pertencerem ao complementar de D e $z^*_{ij} = 0$. Caso contrário, para localizar outro estado recorrente i, com sua respectiva classe de equivalência, deve–se detectar qual coluna de Z^* tem uma soma máxima, ou seja,

$$\sum_{j \in S} z^*_{ji} = \max_{k \in S} \sum_{j \in S} z^*_{jk}.$$

O processo continua até a matriz Z^* possuir todas as entradas nulas e conseqüentemente

$$\max_{k \in S} \sum_{j \in S} z^*_{jk} = 0.$$

Neste ponto todas as subclasses fechadas foram identificadas. Os estados que não foram classificados como recorrentes serão declarados como transientes.

4 OBTENDO OS TEMPOS MÉDIOS DE ABSORÇÃO E AS PROBABILIDADES DE ABSORÇÃO

O tempo médio de absorção é interpretado como o tempo médio esperado até a Cadeia alcançar qualquer subclasse fechada, dado que este encontra–se em um estado transiente $j \in T$ onde T é a classe de estados transientes. Define–se este

parâmetro como

$$\eta_j = \sum_{m=1}^{\infty} m b_j^m,$$

sendo b_j^m a probabilidade do estado transiente j ser absorvido por qualquer subclasse fechada na m–ésima transição. O procedimento para a obtenção de η_j é feito através da matriz $N = (I - Q)^{-1}$, denominada de matriz fundamental. Observe que I é a matriz identidade e Q uma matriz subestocástica correspondendo às probabilidades de transição entre estados transientes. A matriz $(I - Q)$ sempre terá a inversa não singular. A soma de cada linha de N, correspondendo a um estado transiente j, fornecerá η_j.

A probabilidade de absorção de um estado transiente j por um estado recorrente i, pertencente à subclasse fechada $F(i)$, é denotada por α_{ji}. Tendo em vista este fato,

$$\sum_{i \in F(i)} \alpha_{ji}$$

fornece a probabilidade do estado j ser absorvido pela subclasse $F(i)$. Para o cálculo de α_{ji}, recorre–se ao seguinte resultado:

Proposição 5.1. A matriz $M = NR$ tem como entradas as probabilidades α_{ji}, onde R é a matriz das probabilidades de transição de estados transientes para recorrentes.

Prova: Isaacson e Madsen (1976).

Denotando–se por F a união de todas as subclasses fechadas, tem–se que

$$\sum_{k \in F} \alpha_{ji} = 1,$$

para todo j transiente. Portanto, concluimos que M é uma matriz estocástica. A partir deste fato também podemos concluir que um estado transiente tem probabilidade 1 de ser absorvido por uma subclasse fechada.

5 OBTENDO A DISTRIBUIÇÃO ESTACIONÁRIA E OS TEMPOS MÉDIOS DE RECORRÊNCIA

Tendo-se constatado a existência de ergodicidade, prossegue-se na consecução da distribuição limite π. Para encontrar π é necessário obter a solução do sistema de equações lineares $\pi P = \pi$. Uma vez que este sistema é homogêneo e indeterminado

é preciso introduzir uma restrição para resultar em uma solução única. A restrição

$$\sum_{k \in S} \pi(k) = 1,$$

substituída em qualquer uma das equações do sistema resultará em uma solução única.

Antes de proceder na solução do sistema será preciso considerar um problema o qual poderá vir a ocorrer. Para um estado transiente j, $\pi(j) = 0$, tendo-se que desconsiderar estes estados no sistema de equações. A razão disso é devido à possibilidade de uma solução na qual resulte $\pi(j) > 0$, onde na verdade $\pi(j)$ é exatamente igual a zero. Isto pode acontecer devido a arredondamentos internos do computador. Para evitar esse problema, constrói-se um sistema de $n - b$ equações e $n - b$ variáveis, sendo b o número de estados transientes. Com um sistema de equações reduzido será mais rápida e precisa sua solução. Aos estados transientes será atribuído $\pi(j) = 0$.

O tempo médio de recorrência de um estado recorrente i corresponde ao instante médio esperado em que o processo o alcançará pela segunda vez. Ele é denotado por μ_i e possui a importante propriedade

$$\mu_i = \frac{1}{\mu(i)}.$$

6 UM EXEMPLO

Considere uma Cadeia de Markov com $S = \{1, 2, 3, 4, 5, 6\}$ e matriz de transição dada por:

$$P = \begin{vmatrix} 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1/3 & 2/3 & 0 & 0 & 0 \\ 0 & 1/2 & 1/8 & 1/8 & 0 & 1/4 \\ 1/3 & 0 & 0 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1/4 & 1/2 & 0 & 0 & 1/4 \end{vmatrix}$$

A partir de agora iremos aplicar o fluxograma 4 para termos uma melhor compreensão da técnica.

A matriz adjacente resulta em

$$Z = \begin{vmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{vmatrix}$$

A dimensão da matriz P é (6×6), daí $n = 6$. Uma vez que as colunas de Z^{2^k} não somam n até $2^k \geq (n-1)^2$, isto é, $k = 5$, então procede-se na classificação de estados. Para $k = 5$, tem-se $M = 2^k = 32$ e

$$Z^M = \begin{vmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \end{vmatrix}$$

Procedendo-se na identificação de subclasses fechadas, iniciamos por detectar um estado recorrente.

O estado $i = 1$ é recorrente, pois $\sum_{j \in S} z_{j1}^M$ é um máximo. As classes de equivalência para este estado são:

$$K_0 = \{j : z_{1j}^M = 1\} = \{1, 4\}, \qquad K_1 = \{j : z_{1j}^{M+1} = 1\} = \{1, 4\}.$$

Como $Z^{M+1} = Z^M$ então $K_0 = K_1$ e, portanto, $d = 1$. Por conseguinte, a primeira classe fechada é dada por $F(1) = K_0 = \{1, 4\}$. Os estados externos a $F(1)$ e pertencentes a esta são detectados através de

$$D_0 = \{j : z_{j1}^M = 1\} = \{1, 2, 3, 4, 6\},$$

$$D_1 = \{j : z_{j1}^{M+1} = 1\} = \{1, 2, 3, 4, 6\},$$

Assim,

$$D = \{1, 2, 3, 4, 6\} \qquad D^c = \{5\}.$$

Prosseguindo,

$$Z^* = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Com base nessa nova matriz obtemos o máximo $\sum_{j\in S} z^*_{j5} = 1$, ou seja, $i = 5$ é um estado recorrente. Novamente constrói-se as classes:

$$K_0 = \{j: z^M_{5j} = 1\} = \{5\}, \quad K_1 = \{j: z^{M+1}_{5j} = 1\} = \{5\}$$

Logo, $K_1 = K_0$, ou seja, $d = 1$, $F(5) = \{5\}$, $D_0 = \{j: z^*_{j5} = 1\} = \{5\}$, $D_1 = \{j: z^M_{j5} = 1\} = \{5\}$, $D = \{5\}$ e $D^c = \{1, 2, 3, 4, 6\}$.

Transformando novamente Z^* temos

$$Z^* = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Assim, $\max_{i\in S} \sum_{j\in S} z^*_{ji} = 0$ e, portanto, não há mais estados recorrentes.

A classe de estados transientes é dada por $T = \{2, 3, 6\}$.

Para calcular os tempos médios de absorção e as probabilidades de absorção constrói-se as matrizes:

$$Q = \begin{vmatrix} 1/3 & 2/3 & 0 \\ 1/2 & 1/8 & 1/4 \\ 1/4 & 1/2 & 1/4 \end{vmatrix} \qquad R = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 1/8 & 0 \\ 0 & 0 & 0 \end{vmatrix},$$

$$(I - Q) = \begin{vmatrix} 2/3 & -2/3 & 0 \\ -1/2 & 7/8 & -1/4 \\ -1/4 & -1/2 & 3/4 \end{vmatrix} \quad \text{e} \quad N = \begin{vmatrix} 17/2 & 8 & 8/3 \\ 7 & 8 & 8/3 \\ 15/2 & 8 & 4 \end{vmatrix}$$

Utilizando as matrizes $M = NR$ e N, obtém-se:

Est.	Tempo md. absorc.	Prob. absorc. por $F(1)$	Prob. absorc. por $F(5)$
2	$\frac{115}{6}$	1	0
3	$\frac{53}{3}$	1	0
6	$\frac{39}{2}$	1	0

Cálculo do coeficiente de ergodicidade:

$$\begin{aligned} &\text{linhas } 1 \text{ e } 2: \quad \textstyle\sum_{j\in S} \min(z^{32}_{1j}, z^{32}_{2j}) = 2, \\ &\qquad\vdots \\ &\text{linhas } 5 \text{ e } 6: \quad \textstyle\sum_{j\in S} \min(z^{32}_{5j}, z^{32}_{6j}) = 0, \end{aligned}$$

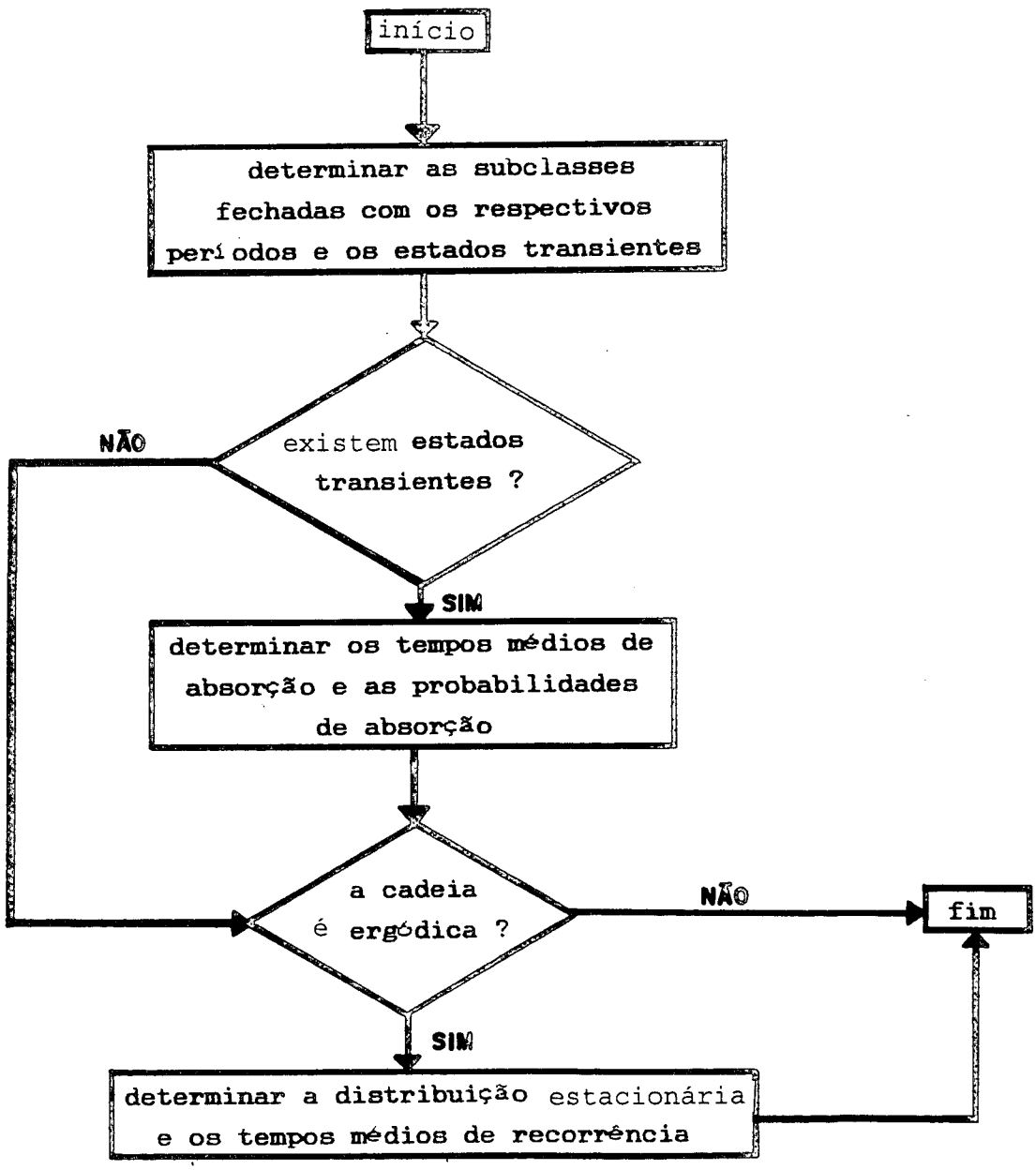
Portanto

$$\alpha(Z^{32}) = \inf_{i<k} \sum_{j \in S} \min(z_{ij}^{32}, z_{kj}^{32}) = 0$$

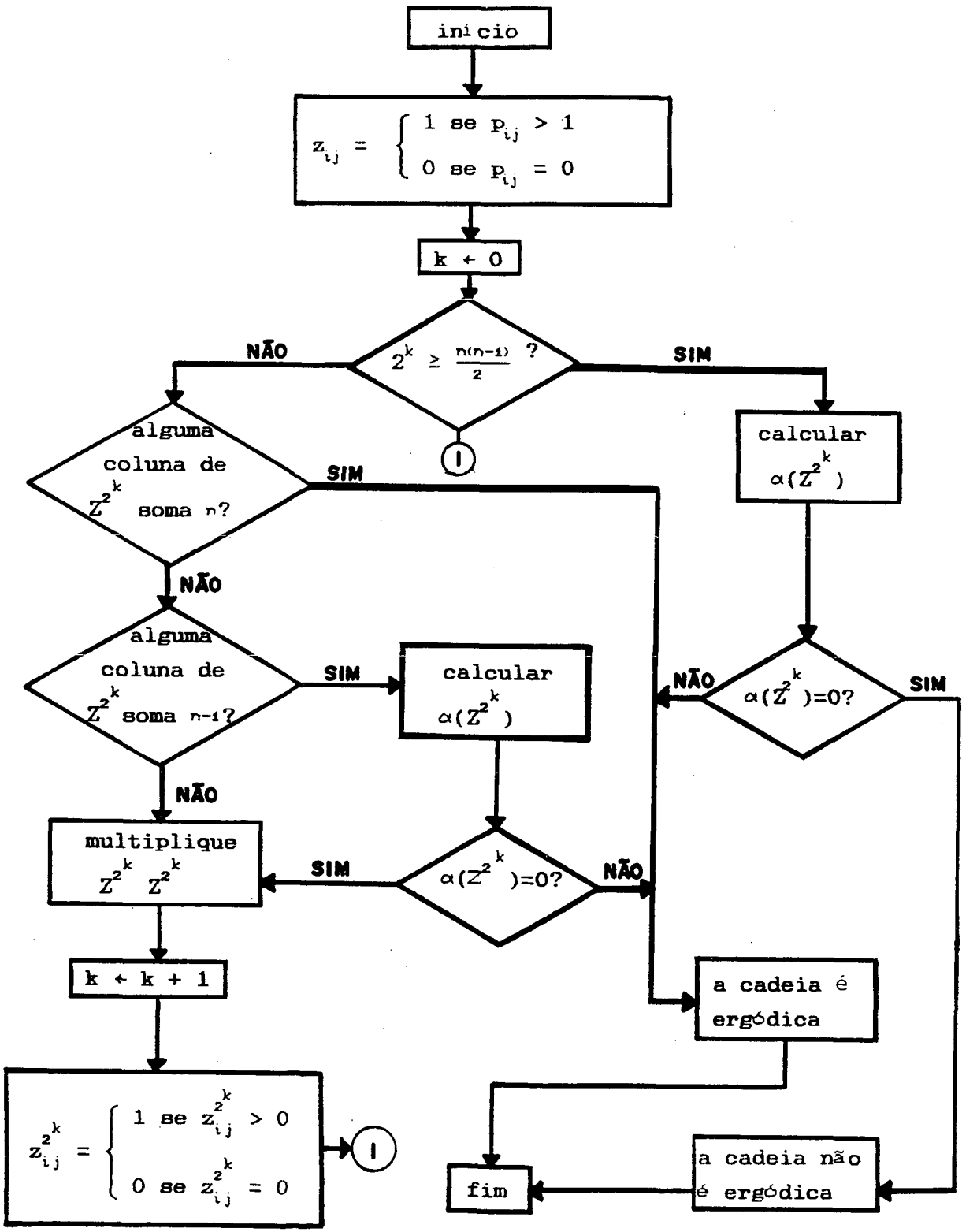
e conseqüentemente a cadeia não é ergódica.

7 APÊNDICE

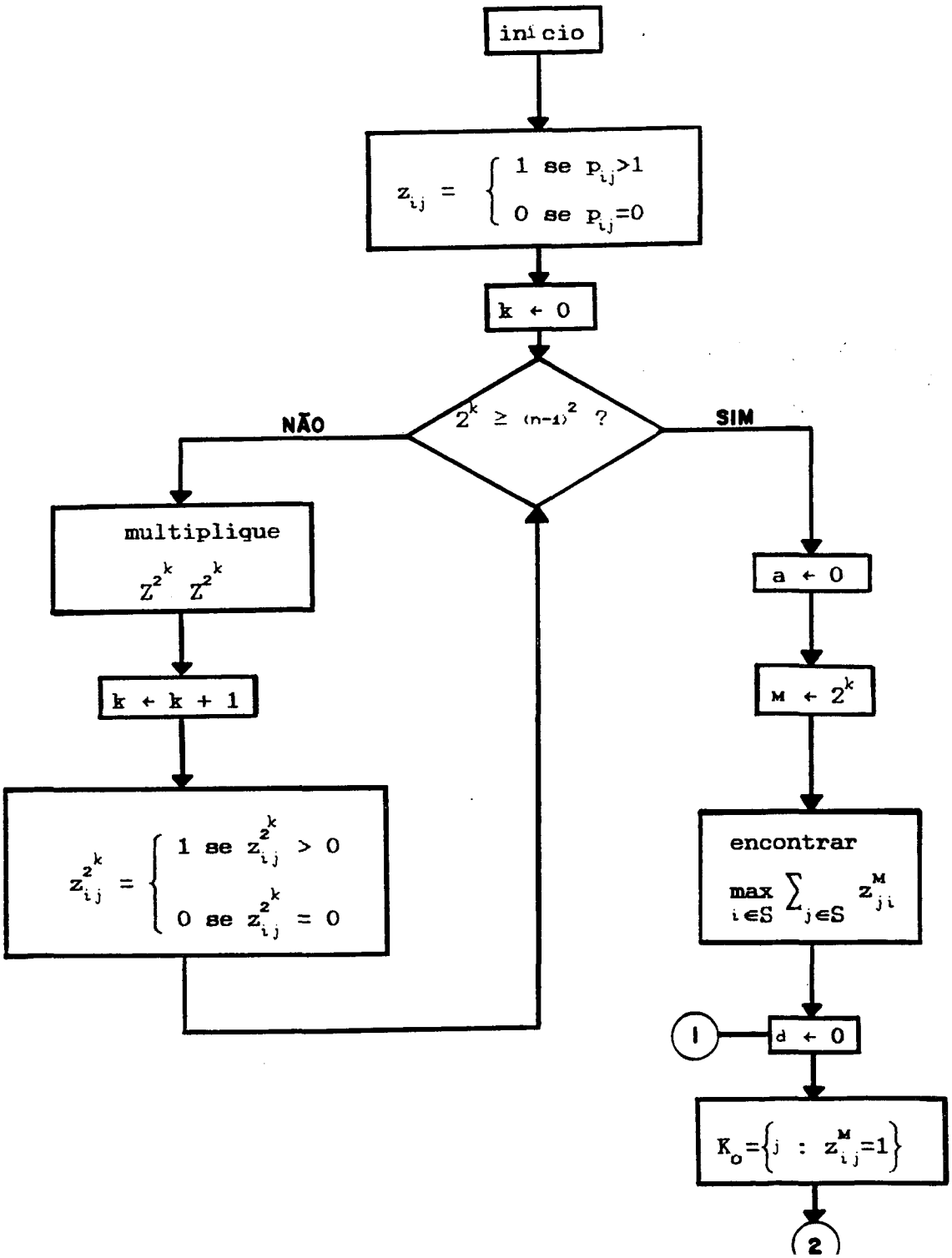
Fluxograma 1: uma visão geral da análise



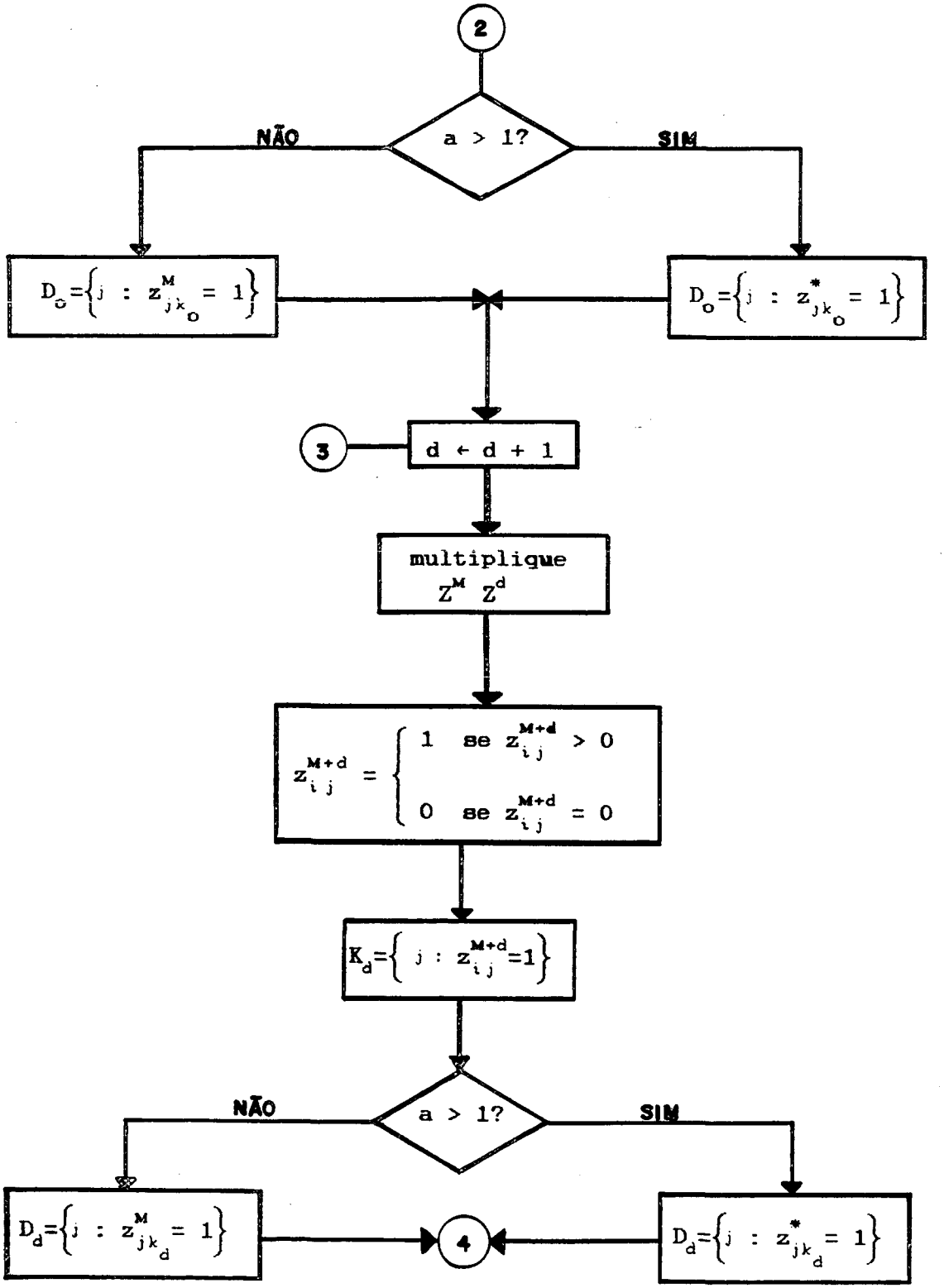
Fluxograma 2: o método da soma das colunas



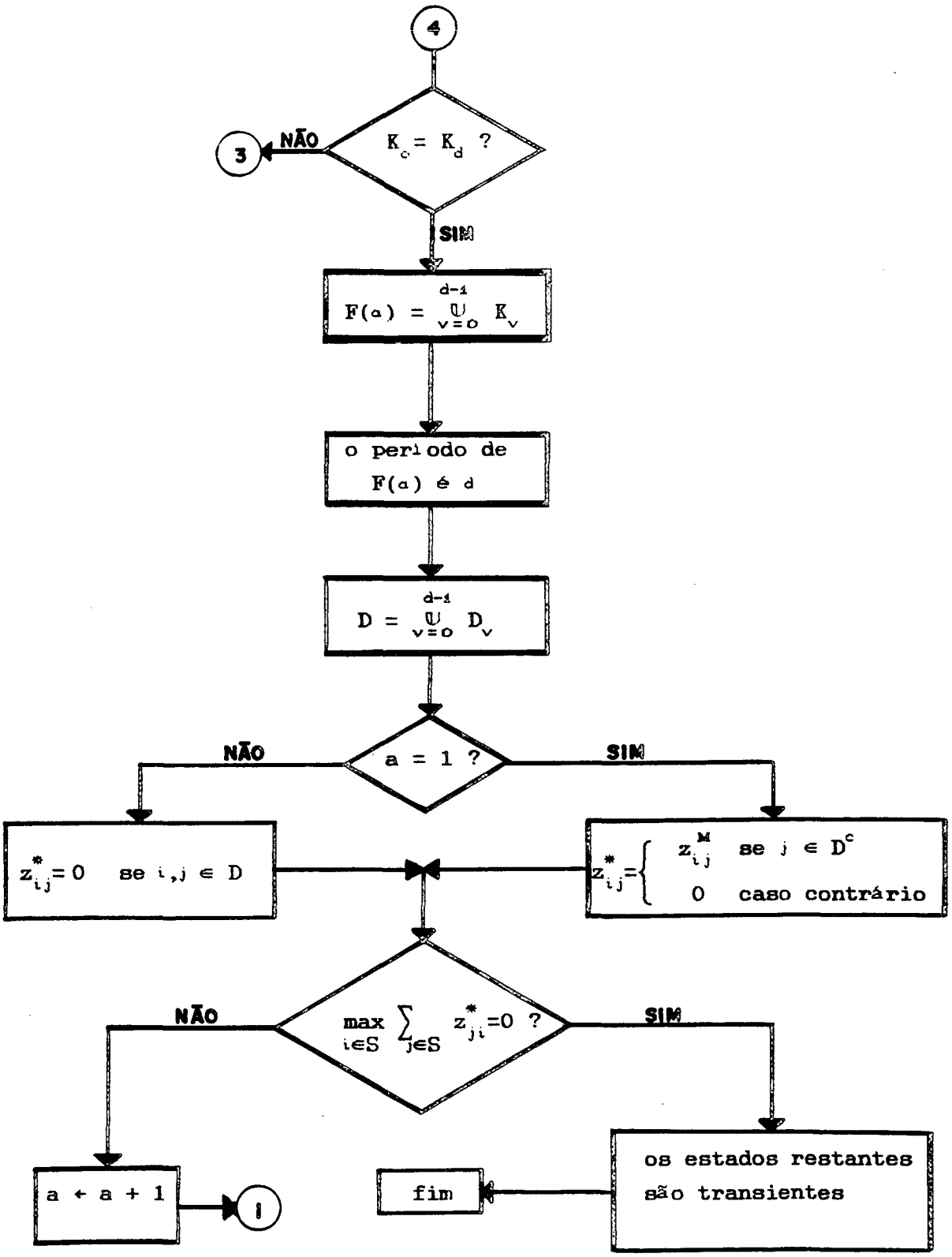
Fluxograma 3: classificação de estados



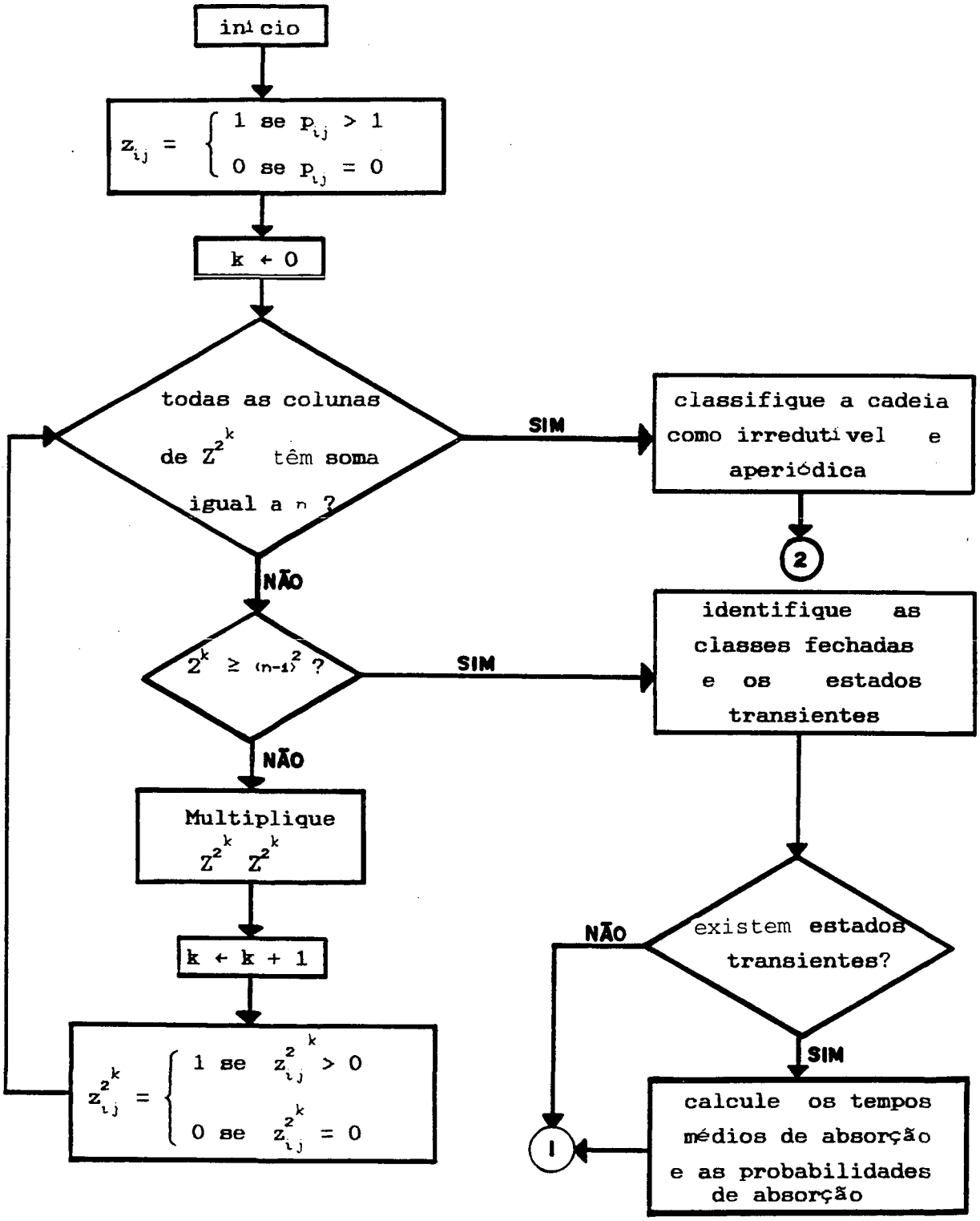
continuação ...



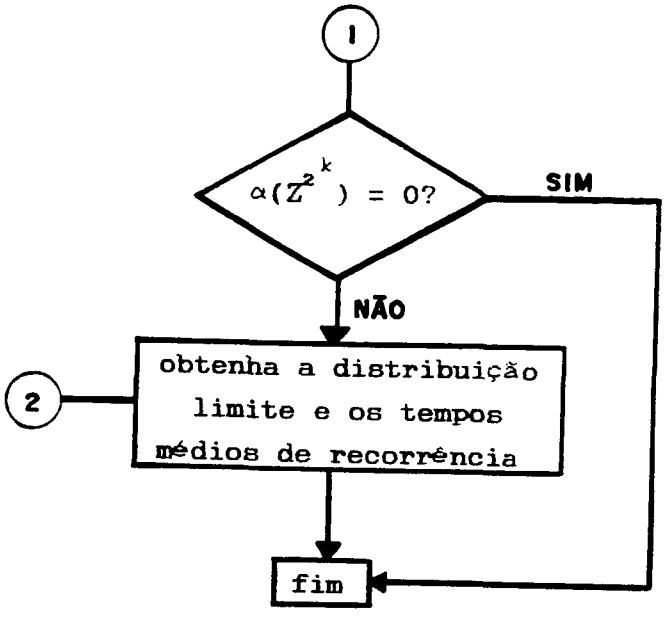
conclusão


4
3
NÃO
$K_c = K_d$?
SIM
$F(a) = \bigcup_{v=0}^{d-1} K_v$
o período de F(a) é d
$D = \bigcup_{v=0}^{d-1} D_v$
NÃO
a = 1 ?
SIM
$z^*_{ij} = 0$ se $i,j \in D$
$z^*_{ij} = \begin{cases} z^M_{ij} & \text{se } j \in D^c \\ 0 & \text{caso contrário} \end{cases}$
NÃO
$\max_{i \in S} \sum_{j \in S} z^*_{ji} = 0$?
SIM
a ← a + 1
1
fim
os estados restantes são transientes

Fluxograma 4: análise completa de uma Cadeia de Markov



conclusão


1
$\alpha(Z^{2^k}) = 0$?
SIM
NÃO
2
obtenha a distribuição limite e os tempos médios de recorrência
fim

BIBLIOGRAFIA

ISAACSON, D. L., MADSEN R. W. *Markov chains*: theory and applications. New York: J. Wiley, 1976.

KARLIN, S., TAYLOR, H. *A first course in stochastic processes*. 2. ed. [S.l.]: Academic Press, 1975.

KEMNY, J. G., SNELL J. L. *Finite markov chains*. [S.l.]: D. Van Nostrand, 1960.

PARZEN, E. *Stochastic processes*. [S.l.]: Halden Day, 1967.

RESUMO

A finalidade deste artigo é apresentar uma discussão sobre uma das técnicas computacionais existentes na análise de Cadeias de Markov. A discussão é restrita ao contexto de Cadeias estacionárias e finitas, tomando–se como base o método proposto por Isaacson e Madsen (1976), o qual consiste de dois algoritmos básicos. Com fundamentação nestes dois algoritmos, propõe–se um único algoritmo para analisar a existência de ergodicidade, a classificação de estados,os tempos médios de absorção, as probabilidades de absorção e os tempos médios de recorrência.

ABSTRACT

The objective of this paper is to present a discussion about one of available computational techniques in the Markov Chains analysis. The discussion restricts to the finite and stationary chains context, and it is based on the method proposed by Isaacson and Madsen (1976), which consists of two basic algorithms. Based on the two algorithms, one single algorithm is proposed to analyse the existence of ergodicity, the classification of the states, the mean absorption times, the absorption probabilities and the mean recurrence times.

POLÍTICA EDITORIAL

A RBEs objetiva promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais através da apresentação, descrição e discussão desses métodos e de suas aplicações, num formato de fácil assimilação pelos membros da comunidade científica. Destina-se também a servir de veículo para troca de idéias entre os especialistas e todos os interessados em análise e desenvolvimento de metodologia estatística. A RBEs tem periodicidade semestral e publica artigos teóricos e/ou aplicados de métodos estatísticos, com ênfase na análise de fenômenos econômicos e sociais. São também aceitos artigos abordando os diversos aspectos do desenvolvimento metodológico relevantes para órgãos produtores de estatísticas, assim como artigos de revisão do estado da arte em temas específicos.

a) delineamento de pesquisas;

b) avaliação de pesquisas e mensuração de erros;

c) uso e combinação de fontes alternativas de informações; novos desenvolvimentos em metodologia de pesquisa;

e) análise de séries de tempo;

f) estudos demográficos;

g) integração de dados;

h) amostragem e estimação;

i) análise de dados;

j) crítica e imputação de dados;

l) disseminação e confiabilidade de dados; e

m) modelos econométricos.

Todos os artigos submetidos serão avaliados pelo Comitê Editorial da RBEs quanto a sua qualidade e relevância, devendo os mesmos serem inéditos. Além disto, não deverão ter sido simultaneamente submetidos a qualquer outro periódico nacional.

A RBEs publicará também resenhas de livros, artigos escritos a convites e ensaios sobre o ensino de Estatística.

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

Os artigos remetidos para publicação deverão ser submetidos em 3 vias (que não serão devolvidas) para:

Djalma G. C. Pessoa
Editor Responsável - RBEs
ENCE/IBGE
Rua André Cavalcanti, 106
Bairro de Fátima
20231-050 - Rio de Janeiro - RJ

- Os artigos submetidos à RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

- Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

Instruções para preparo de originais:

1. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se para cada um a filiação e endereço permanente para correspondência. Agradecimentos a colaboradores, outras instituições e auxílios recebidos devem figurar também nesta página.

2. A segunda página do original deve conter resumos em português e em inglês (Abstract), destacando os pontos relevantes do artigo. Cada resumo deve ser datilografado seguindo o mesmo padrão do restante do texto, em um *único parágrafo*, sem fórmulas, com no máximo 150 palavras.

3. O artigo deve ser dividido em seções numeradas progressivamente, com títulos concisos e apropriados. Subseções, todas, inclusive a primeira, devem ser numeradas e receber título apropriado.

4. A citação de referências no texto e a listagem final das referências devem ser feitas de acordo com as normas da ABNT.

5. As tabelas e gráficos devem ser apresentados em folhas separadas, precedidas de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração tais como as demais seções do trabalho.

6. Gráficos e diagramas para publicação devem ser traçados em papel branco, como nitidez e boa qualidade, para permitir que a redução seja feita mantendo qualidade. Fotocópias não serão aceitas. É fundamental que não existam erros quer no desenho quer nas legendas ou títulos.

7. Serão aceitos originais processados por editores de texto, tais como: CW, Word, Carta Certa, WP e WS.