Consistência e identificabilidade no modelo *Grade of Membership*: uma nota metodológica

André Junqueira Caetano*
Carla Jorge Machado**

Introdução

O método *Grade of Membership* (GoM) vem sendo amplamente utilizado no Brasil na área de Demografia e na Saúde Pública. Há, por exemplo, um número expressivo de estudos recentes na área de saúde reprodutiva (MIRANDA-RIBEIRO et al., 2007, entre outros), na vulnerabilidade de pacientes à adesão ao tratamento antiretroviral (BONOLO et al., 2008, entre outros), na área de saúde mental (GUIMARÃES et al., 2008), sendo o *software* GoM3, executável em ambiente DOS, um dos programas mais comumente usados. É gratuito e se encontra disponível na Internet (http://www.stat.unipg.it/stat/statlib/DOS/general/). O algoritmo utilizado é o de Woodbury e Clive (1974) e, sob este algoritmo, os parâmetros g_{ik} e λ_{kjl} são iterativamente estimados. Há aspectos estatísticos que merecem maior detalhamento e atenção e são abordados neste trabalho, com o objetivo de esclarecer as propriedades principais dos estimadores de GoM e das estimativas geradas. O foco, neste texto, é a consistência dos parâmetros obtidos em algumas condições e mediante algumas restrições, bem como a questão de identificabilidade dos modelos. Finalmente, é explicada a necessidade de se procurarem máximos globais, que garantam uma modelagem que possa gerar uma parametrização única e identificável. Antes, contudo, é feita uma breve descrição do método.

Descrição do método

O método se apresenta atrativo, por gerar dois conjuntos de parâmetros: os λ_{kjl}, referentes à probabilidade de que um indivíduo tipo puro de um perfil tenha a probabilidade (λ) de resposta (l) à variável (j), no perfil (k). Estas probabilidades valem apenas para indivíduos considerados "tipos puros" de um determinado perfil, quais sejam: aqueles que possuem pertencimento ao perfil igual a 1, ou seja, $g_{ik}=1{,}00$. Assim, g_{ik} indica o pertencimento (g) do indivíduo (i) ao perfil (k). Este parâmetro não possui interpretação probabilística e pode ser interpretado como um atributo individual, e sua amplitude é de 0,00 a 1,00 (MANTON; WOODBURY; TOLLEY, 1994).

O produto $g_{ik}\ \lambda_{kjl}$ indica a probabilidade de um indivíduo, com g_{ik} variando de 0,00 a 1,00, possuir a resposta l-ésima para a questão j-ésima. A soma dos produtos, para todos os perfis de GoM gerados (K perfis), para cada indivíduo, é dada pelo seguinte somatório:

$$P(x_{ijl}=1)=\sum_{k=1}^{K} g_{ik}\,\lambda_{kjl} \qquad (1)$$

onde $P(x_{ijl}=1)$ é a probabilidade de que o i-ésimo indivíduo possua a l-ésima resposta, como predito pelo produto interno dos k pares de g_{ik} e λ_{kjl} estimados.

Assumindo independência dos indivíduos condicionalmente a λ_{kjl} e g_{ik}, a função de verossimilhança é definida:

$$L=\prod_{i}\prod_{j}\prod_{l}\left(\sum_{k=1}^{K} g_{ik}\,\lambda_{kjl}\right)^{x_{ijl}} \qquad (2)$$

* Professor-adjunto III da PUC-Minas, Programa de Pós-graduação em Ciências Sociais.
** Professora-adjunta III do Departamento de Demografia do Cedeplar, Universidade Federal de Minas Gerias.

Os parâmetros se encontram sob as seguintes restrições:

$0 \leq g_{ik} \leq 1;\ \sum_K g_{ik} = 1;\ 0 \leq \lambda_{kjl} \leq 1;\ \sum_l \lambda_{kjl} = 1$
(MANTON; WOODBURY; TOLLEY, 1994).

Consistência de estimativas de máxima verossimilhança mediante restrições sobre I (tamanho da amostra), J (número de variáveis) e K (número de perfis)

Um estimador consistente, de forma geral, pode ser definido como aquele que mais se aproxima do verdadeiro valor do parâmetro populacional, à medida que se acresce o tamanho da amostra. A consistência do estimador é, geralmente, considerada uma propriedade essencial de um estimador razoável, pois significa que o estimador é não viciado (não contém erro) e que sua variabilidade converge para zero (LARSEN; MARX, 1986).

Uma pergunta que surge seria, então: se esta é uma propriedade tão importante, os parâmetros de GoM (gik e lkjl) são estimados de forma consistente? A resposta imediata é que não se podem esperar estimativas consistentes de valores de g_{ik} (uma vez fixados K e J) com o aumento de I, pois trata-se de parâmetros individuais e todo novo elemento da amostra carrega consigo a sua "história", ou seja, sua heterogeneidade, medida pelo conjunto de respostas às variáveis J.[1] Contudo, segundo Manton, Woodbury e Tolley (2004), H(x), a distribuição de valores de ζ_{ik}, da qual g_{ik} são realizações, pode ser consistentemente estimada, bastando que sejam estimados os momentos de ordem J desta distribuição.

A explicação mais detalhada a este problema foge do escopo deste trabalho, mas os autores demonstram que se, por exemplo, K=2, basta estimar os dois primeiros momentos de H(x) – que refletem a localização na sua distribuição e a dispersão respectivamente.[2] Desta forma, a distribuição H(x) estaria definida e o conjunto de g_{ik} poderia ser estimado. Já no caso dos valores de λ_{kjl}, a resposta seria mais direta: estes parâmetros seriam tão melhores estimados quanto maior for o aumento em I (uma vez fixados K e J). Ou seja, para efeito de adequadamente estimar os parâmetros de probabilidade para os tipos puros, o aumento de I é bastante desejável.

Cabe ainda observar que, como já visto, o número de g_{ik} aumenta diretamente com o tamanho da amostra I, o que adiciona maior complexidade ao modelo. A estabilidade dos g_{ik} só aumenta com o acréscimo de J para I e K fixos (em outras palavras, novas variáveis, em geral, trazem novas informações sobre os indivíduos e quanto maior o número de informações, maior a probabilidade de estimar com maior grau de certeza os parâmetros individuais).

Em resumo, para melhor estimar λ_{kjl} são necessários maiores números de observações (aumento de I) e para melhor estimar g_{ik} precisa-se de maior número de variáveis (aumento de J). Ou seja, a princípio, maior número de variáveis e de observações seria desejável na estimação de um modelo de GoM, adicionando maior complexidade na estimativa.

Identificabilidade dos parâmetros gerados pelo modelo de GoM

Alguns modelos, especialmente os complexos e que dependem de geração de estimativas de forma iterativa, podem ter um conjunto de soluções e não apenas uma (ou seja, mais de um conjunto de valores de parâmetros com igual ajustamento a um conjunto de dados). Nestes casos, as conclusões sobre estes dados podem estar equivocadas se for utilizado um conjunto de parâmetros (e não o(s) outro(s)) e este modelo não deveria ser utilizado para geração de hipóteses a respeito do fenômeno em estudo. Assim, deste ponto de vista, o importante é obter uma parametrização única dos dados disponíveis (BRADY, 1985).

[1] Neste texto, sempre que a referência for feita a uma variável J, está-se pensando em J binária, com dois níveis de resposta.

[2] Ao leitor interessado em se aprofundar um pouco mais na questão de geração de momentos de ordem ou superior, indicamos o capítulo 3 de Larsen e Marx (1986) e o capítulo 2 de Mood, Graybill e Boes (1974).

Questões práticas de estimativas de um modelo identificável

Entendida a importância de um modelo estatístico ser identificável, há que se perguntar se existem garantias de o resultado obtido no modelo de GoM ser identificável, quaisquer que sejam os parâmetros iniciais utilizados na geração do resultado. Normalmente o modelo iterativo tem início com um conjunto de λ_{kjl}, que são aleatoriamente gerados pelo programa. O algoritmo de geração de estimativas de máxima verossimilhança é o EM (*expectation-maximization*), que depende de parâmetros iniciais para que o processo de iteração seja iniciado. Este processo de iteração é necessário para que possam ser aproximadas soluções de equações não lineares, a partir de um valor inicial dado, até que haja convergência a um valor final (MANTON; WOODBURY; TOLLEY, 1994). O objetivo é haver convergência até um valor máximo global. Contudo, dependendo dos parâmetros iniciais, pode-se obter um máximo local iterativamente, o que não é o desejado.

Dado que o algoritmo EM procura encontrar o melhor ajuste de um modelo para um certo conjunto de dados, deve haver alguma forma de se avaliar a adequação deste modelo ajustado aos dados. A premissa básica do algoritmo, assim, é que deve existir um ponto ótimo a partir do qual não é preciso mais qualquer tentativa de um novo modelo, com novos parâmetros e, neste momento, as iterações terminam. Dessa forma, o algoritmo EM continua o processo de convergência rumo a uma solução até que a mudança de probabilidade entre dois conjuntos de estimativas seja desprezível, a partir de critérios de convergência previamente definidos. É comum que os usuários destes programas computacionalmente intensivos não tenham claros os problemas decorrentes de se encontrar um máximo local em vez de um máximo global.

Shephard (2003) fornece um exemplo interessante sobre a diferença entre máximo local e máximo global, que é bastante intuitivo. A melhor forma de descrever este problema, segundo o autor, é se pensar em duas montanhas: o monte Fuji, no Japão, e um dos picos da cadeia de montanhas do Himalaya (por exemplo, o Monte Everest). Considere que a altitude do pico a qualquer ponto da terra é a adequação do ajuste do modelo estimado (ou seja, a verossimilhança) e os valores máximos globais são os topos das montanhas. Se é preciso atingir o ponto mais alto das montanhas a partir de um ponto inicial aleatório, mas com a restrição de se mover unicamente para cima, parece bastante lógico (óbvio!) que não se pode descer. No caso do monte Fuji, seria fácil encontrar o caminho até o topo obedecendo essa regra, pois não existem pequenas depressões entre o ponto de partida e o topo. Contudo, se um indivíduo estiver em um ponto aleatório no Himalaya, seria bem mais difícil alcançar o Monte Everest dada a restrição. Não apenas a posição em que um indivíduo está pode não permitir enxergar o pico o qual está tentando atingir (o do Monte Everest), como também podem existir outros pequenos picos (máximos locais) em seu caminho, e seria necessário escalá-los e, em seguida, descer novamente (o que não é permitido pela regra) antes de se atingir o máximo global, o Monte Everest. Assim, seria muito difícil alcançar o ápice do Monte Everest em decorrência da existência de máximos locais. Infelizmente, como o máximo local não é uma solução única para o problema e o máximo global representa esta solução, a tarefa é tentar, então, evitar que qualquer modelo final selecionado seja aquele para o qual não há garantias de que houve convergência para um máximo global.

No caso do GoM, o melhor procedimento com vistas a atingir esta solução única é analisar o comportamento de convergência do algoritmo por meio de uma série de análises prévias nas quais os pontos de partida (valores iniciais e aleatórios de λ_{kjl}) sejam continuamente alterados (ou seja, vários modelos necessitam ser gerados – sugere-se que em torno de 20 a 30). Com um conjunto de λ_{kjl} fixo, a função de verossimilhança é maximizada e os g_{ik} são gerados. Em seguida, com g_{ik} fixo, a verossimilhança é maximizada novamente, a fim de estimar λ_{kjl}. Esta iteração é repetida (em cada modelo, ou seja, internamente ao programa)

até que os parâmetros não variem mais. O pesquisador repete este processo entre 20 e 30 vezes e compara os vários valores de λ_{kjl}. Com base em múltiplos modelos com diferentes λ_{kjl} iniciais para um valor fixo de K, é possível identificar um conjunto de λ_{kjl} iniciais mais consistentes, utilizá-los como valores iniciais de λ_{kjl} e gerar novos parâmetros. Por exemplo, se entre 30 λ_{kjl} quaisquer, 20 são iguais a 1,00, 5 iguais a 0,80 e 5 iguais a 0,40, parece razoável que um valor inicial para este parâmetro seja mais próximo de 1,00 e este λ_{kjl} igual a 1,00 pode ser assumido como um valor informativo. Outro exemplo seria se, em um conjunto de 30 λ_{kjl}, 28 fossem iguais a 0,20, 1 igual a 0,70 e 1 igual a 0,75. O mais provável é que estes dois últimos valores tenham sido devidos à aleatoriedade e os primeiros 28 (mais comumente obtidos) sejam os informativos, sendo razoável, então, assumir 0,2 como um valor para a iteração final. Assim, obtém-se um conjunto de λ_{kjl} finais a partir deles e, então, define-se o conjunto dos valores iniciais que serão utilizados em um novo processo de modelamento (uma nova "rodada" do Modelo de GoM). Espera-se que estes valores, informativos, possam ser utilizados para a geração de um novo modelo que convirja para um máximo global.

Conclusões

Este trabalho procurou descrever as propriedades dos estimadores de GoM, enfatizando a consistência e a identificabilidade, na busca de um modelo que convirja para uma solução única. Este é um processo intuitivamente fácil e um pouco trabalhoso, pela necessidade de geração de múltiplos modelos para definição de λ_{kjl} informativos, que sirvam para gerar o modelo final (com máximo global).

Esta solução única e identificável é bastante valorizada em estatística, sendo o ponto de partida para formulação de hipóteses sobre determinado fenômeno. O objetivo deste texto foi facilitar o entendimento de que, dado um conjunto de λ_{kjl}, vários valores finais dos parâmetros podem ser obtidos e isso é decorrência da impossibilidade de obtenção de um máximo global logo no primeiro conjunto de iterações. Neste trabalho, procurou-se não apenas mostrar que o modelo dispõe de propriedades estatísticas importantes, mas também indicar a melhor forma de se lidar com a situação de obtenção de máximos locais em vez de máximos globais.

Por fim, objetivou-se elaborar um pouco mais algumas das propriedades de forma teórica. Contudo, o critério de suficiência estatística não foi abordado. Com este objetivo, em um estudo futuro, os autores deste texto pretendem apresentar um exemplo prático e detalhado das estimativas para o método *Grade of Membership,* de tal forma que os pesquisadores possam utilizar o *software* Gom3 sem maiores dificuldades.

Referências

BONOLO, P. F.; MACHADO, C. J.; CÉSAR, C. C.; CECCATO, M. G. B.; GUIMARÃES, M. D. C. Vulnerability and non-adherence to antiretroviral therapy among HIV patients, Minas Gerais State, Brazil. **Cadernos de Saúde Pública**, Rio de Janeiro, 24(11): 2.603-2.613, 2008.

BRADY, H. E. Statistical consistency and hypothesis testing for nonmetric multidimensional scaling. **Psychometrika**, New York, 50(4): 509-537, 1985.

GUIMARÃES, M. D. C.; OLIVEIRA, H. N.; CAMPOS, L. N.; SANTOS, C. A.; GOMES, C. E. R.; OLIVEIRA, S. B.; FREITAS, M. I.; ACURCIO, F. A.; MACHADO, C. J. Reliability and validity of a questionnaire on vulnerability to sexually transmitted infections among adults with chronic mental illness: PESSOAS Project. **Revista Brasileira de Psiquiatria**, São Paulo, 30(1): 55-59, mar. 2008.

LARSEN, R. J.; MARX, M. L. **An introduction to mathematical statistics and its applications**. Second edition. New Jersey: Prentice Hall, 1986.

MANTON, K. G.; WOODBURY, M. A.; TOLLEY, H. D. **Statistical applications**

using fuzzy sets. First edition. New York: Willey Interscience, 1994.

MIRANDA-RIBEIRO, P; SIMÃO, A. B.; CAETANO, A. J.; PERPÉTUO, I. H. O.; LACERDA, M. A.; TORRES, M. E. A. Acesso à contracepção e ao diagnóstico do câncer de colo uterino em Belo Horizonte: uma contribuição metodológica aos estudos quanti-quali. **Revista Brasileira de Estudos de População**, São Paulo, 24(2): 341, jul./dez. 2007.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. Third edition. United States: McGraw Hill, 1974.

SHEPHARD, N. Local & Global Maxima. Disponível em: <http://slack.ser.man.ac.uk/theory/em_maxima.html>. Acesso em: 7 dez. 2008.

WOODBURY, M. A.; CLIVE, J. Clinical pure types as a fuzzy partition. **Journal of Cybernetics**, 4, p. 111-121, 1974.

Recebido para publicação em 11/12/2008.
Aceito para publicação em 20/12/2008.