Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Utilização dos critérios de informação na seleção de modelos de regressão linear

Thales Esteves Lima Sobral[1]

Laboratório de Controle e Sistemas Inteligentes- LCSI, Departamento de Semicondutores, Instrumentos e Fotônica – DSIF, Faculdade de Engenharia Elétrica e de Computação – FEEC, Universidade Estadual de Campinas – UNICAMP. Av. Albert Einstein, 400, Campinas, SP, Brasil

Gilmar Barreto[2]

Laboratório de Controle e Sistemas Inteligentes- LCSI, Departamento de Semicondutores, Instrumentos e Fotônica – DSIF, Faculdade de Engenharia Elétrica e de Computação – FEEC, Universidade Estadual de Campinas – UNICAMP. Av. Albert Einstein, 400, Campinas, SP, Brasil

Resumo. Este trabalho visa demonstrar o comportamento de seis critérios de informação quando utilizados para a seleção de ordem de modelos do tipo regressão linear.

Palavras-chave. Identificação de sistemas, Modelagem matemática, Seleção de modelos, Regressão Linear, Critério de Informação de Akaike (AIC)

1 Introdução

A modelagem matemática é de grande importância no estudo de fenômenos da natureza, assim como nas áreas de psicologia, botânica, entre outros. Ao definir um modelo matemático para explicar um conjunto de dados em estudo, é possível fazer inferências sobre o fenômeno de interesse. Um modelo é dado como uma estrutura de parâmetros, que são definidos com base nos dados coletados, através de processos de estimação como os Mínimos Quadrados ou o Estimador de Máxima Verossimilhança. O objetivo destes métodos é estimar os parâmetros do modelo de forma a acolher melhor os dados utilizados para a modelagem. Estas técnicas são reconhecidas e largamente utilizadas, de forma que não há muita discussão sobre sua eficácia.

Entretanto, tal paradigma não existe na hora da definição do tipo do modelo, nem da sua ordem: Não existe uma metodologia única, tampouco uma metodologia que

[1] thales.sobral@gmail.com
[2] gbarreto@dsif.fee.unicamp.br

funcione bem em todas as situações. Para resolver esta questão, alguns métodos são mais populares, como a validação cruzada, em que se divide o conjunto de dados a ser estudado em duas ou mais partes, em que uma é utilizada para treinamento do modelo (estimação dos parâmetros), enquanto as outras são utilizadas para validação. Este método é bastante intuitivo, porém ao reduzir o tamanho da amostra a ser estudado, pode-se estar descartando informação útil à modelagem. Utilizar todo o conjunto de dados para efetuar a estimação dos parâmetros é uma forma de trazer mais informação ao modelo, porém desta forma não há como confirmar que o modelo estimado se comportará bem na predição de dados novos. Olhar somente os parâmetros como erro médio quadrádico, ou o valor da verossimilhança tenderá a favorecer os modelos mais complexos, que irão conformar melhor aos dados apresentados, sem que isso signifique um modelo que contenha mais informação (este fenômeno é denominado "*overfit*"). Um modelo que sofra de "overfit" é um modelo que irá descrever muito bem a informação que o gerou, porém não irá se comportar bem ao descrever novos eventos do mesmo sistema o qual ele se propõe a modelar.

Outra alternativa, sugerida por Akaike, é utilizar o Critério de Informação de Akaike (Akaike Information Criterion – AIC) [1]. Este critério é baseado na Divergência de Kullback-Leibler [4], que é uma medida da "distância" entre o modelo identificado e um teórico "modelo real". Como o modelo real não é conhecido, Akaike desenvolveu uma forma de estimar esta distância através dos dados utilizados na modelagem, usando a função de verossimilhança e a ordem do modelo, tendo a seguinte forma:

$$AIC = -2\ln f\left(x\middle|\hat{\theta}\right) + 2k \tag{1}$$

À medida que a verossimilhança aumenta, o termo $-2\ln f\left(x\middle|\hat{\theta}\right)$ decresce, enquanto o termo $2k$ cresce sempre que a ordem do modelo for maior. Dessa forma, o critério de Akaike pondera entre a adequação aos dados e a complexidade do modelo.

Após a apresentação do AIC, vários outros critérios de informação foram propostos, com variações no termo da penalização, com outros contextos teóricos por trás, como por exemplo o BIC [5], KIC [3], e o DIC [7]. Também foram propostos modelos para lidar com modelagens de pequenas amostras, em que os critérios de informação tradicionais, modelados com base em propriedades assintóticas, costumam falhar, levando ao "*overfit*". Tais critérios, como o AICc [8], AICF, KICc, AKICc [6], têm termos de compensação para pequenas amostras, como demonstrado abaixo:

$$AICc = -2\ln f\left(x\middle|\hat{\theta}\right) + \left(\frac{2n}{n-k-1}\right)k \tag{2}$$

O termo $\left(\frac{2n}{n-k-1}\right)$ tende para 2, quando n tende ao infinito, assim o AICc retorna valores muito próximos ao AIC, quando em grandes amostras, e em pequenas amostras tem uma penalização maior, assim alguns autores defendem que o AICc deve ser preferido em relação ao AIC para todos os casos de seleção de modelos [2].

Os outros critérios a serem utilizados neste estudo são:

DOI: 10.5540/03.2016.004.01.0220

$$BIC = -2\ln f\left(x\middle|\hat{\theta}\right) + k\ln n \quad (3)$$

$$KIC = -2\ln f\left(x\middle|\hat{\theta}\right) + 3k \quad (4)$$

$$AKICc = -2\ln f\left(x\middle|\hat{\theta}\right) + \frac{(k+1)(3n-k-2)}{n-k-2} + \frac{k}{n-k} \quad (5)$$

$$AICF = -2\ln f\left(x\middle|\hat{\theta}\right) + 2n\frac{k}{n-2k} \quad (6)$$

2 Modelo de Regressão Linear

O modelo utilizado neste trabalho será o de Regressão Linear. Este modelo caracteriza-se pela saída ser uma combinação linear de valores de entrada deslocados no tempo, multiplicados por constantes, sua forma genérica é:

$$y[n] = a_0x[n] + a_1x[n-1] + a_2x[n-2] + \cdots + a_nx[0] + be[n] \quad (3)$$

em que $e[n]$ é um ruído branco de média nula e variância unitária.

A referência será o modelo

$$y[n] = 0{,}4x[n] - 0{,}25x[n-1] + 0{,}1x[n-2] + e[n] \quad (4)$$

A partir deste modelo, foi gerado o sinal de saída na plataforma MATLAB, e em seguida foram gerados modelos de ordem 1 a 6, com a estimativa dos parâmetos feita através do método dos mínimos quadrados. Em seguida, foram utilizados os critérios AIC, AICc, BIC, KIC, AKICc e AIC para indicar qual dos modelos propostos era o mais adequado, ou seja, que iria prever dados futuros com maior fidelidade. Esta rotina foi repetida 1000 vezes, e anotados quantas vezes cada critério escolheu determinada ordem de modelo. Os tamanhos de amostras utilizados foram de 30, 300 e 3000, para avaliar o desempenho em pequenas amostras, e como os critérios se comportam à medida que o tamanho da amostra aumenta.

3 Resultados dos experimentos

3.1 Experimento com 30 amostras

Nos experimentos com 30 amostras, pode-se ver que as versões corrigidas para pequenas amostras se comportaram melhor que os critérios originais.

DOI: 10.5540/03.2016.004.01.0220

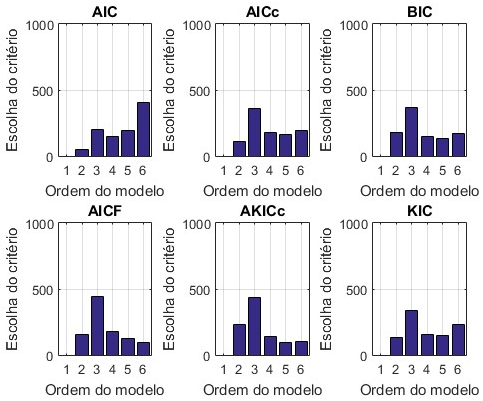

Figura 1: Escolhas dos critérios de informação – 30 amostras

O critério AIC sofreu fortemente de "overfit", escolhendo o modelo de ordem 6 na maioria das vezes.

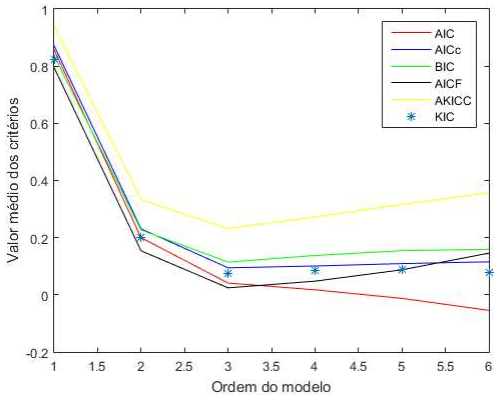

Figura 2: Valores médios dos critérios para as ordens dos modelos – 30 amostras

Como pode ser visto na Figura 2, o valor do AIC continua caindo à medida que a ordem do modelo é maior. Desta forma, ele escolheria modelos de ordem ainda maiores, caso houvesse algum. Portanto, o AIC não é adequado para a escolha de ordem de modelos que foram gerados a partir de dados com poucas amostras, sendo recomendável o uso da sua variante AICc. O BIC tem um fator de penalização maior que os critérios AIC (a partir de 8 amostras, $\ln 8 \cong 2{,}07$) e do KIC (a partir de 21 amostras, $\ln 21 \cong 3{,}04$), então nessas pequenas amostras seu termo de penalização é comparável ao dos critérios corrigidos (que crescem à medida que a amostra diminui). No que então parece ser uma grande vantagem, já que o BIC foi tão bem quanto os critérios corrigidos em pequenas amostras, no caso de se tentar modelar um conjunto de dados com uma arquitetura de modelo diferente do modelo gerador (diferente do exemplo do artigo, em que usou-se o modelo de regressão linear para modelar um conjunto de dados gerado por regressão linear), o decaimento do termo $-2ln\, f(x|\hat{\theta})$ pode não se dar de forma tão abrupta, e o BIC irá sofrer gravemente de "*underfit*" (escolher um modelo com ordem menor que a correta). Os critérios AICF e AKICc tiveram os melhores resultados, escolhendo a ordem correta em quase 50% dos casos. De qualquer forma, nenhum dos critérios conseguiu identificar a ordem correta na maior parte das vezes, isso se dá ao número muito baixo de amostras, o que pode denotar que os modelos são

de baixa qualidade. Daí, outra constatação muito importante é deduzida: os critérios de informação podem servir para comparação entre os modelos, porém não dizem se algum deles é realmente bom. Se todos forem ruins, ele só vai identificar o "menos pior". Daí, uma inferência importante pode ser feita: Se os dados contiverem muito ruído, haverá uma degradação dos modelos gerados, com consequências inesperadas para a indicação dos critérios: Se houverem modelos muito complexos como candidatos, eles podem ser escolhidos por "acolherem" o ruído como parte dos seus parâmetros.

3.2 Experimento com 300 amostras

Com um conjunto de dados maior, a identificação deve ser de melhor qualidade, e como o tamanho da amostra passa a ser consideravelmente maior que o tamanho do modelo, a diferença de desempenho entre os critérios corrigidos para pequenas amostras e suas versões originais deve ser menor.

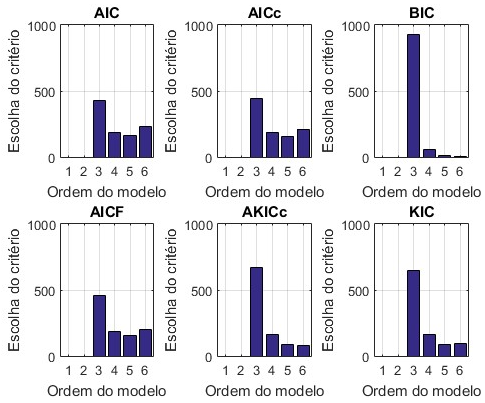

Figura 3: Escolhas dos critérios de informação – 300 amostras

A Figura 3 demonstra claramente uma melhora no desempenho dos critérios, com particular melhor do BIC, KIC e AKICc, que escolheram a ordem correta em mais de 50% das vezes. O BIC acertou a ordem em mais de 90% dos experimentos. Nota-se que há uma consistência no crescimento do valor dos critérios à medida que a ordem do modelo cresce, confirmando que a ordem correta do modelo é 3, conforme Figura 4.

Também é possível ver que o desempenho dos critérios AICc e AICF ficou muito próximo do AIC, da mesma forma que o AKICc selecionou modelos similares ao KIC.

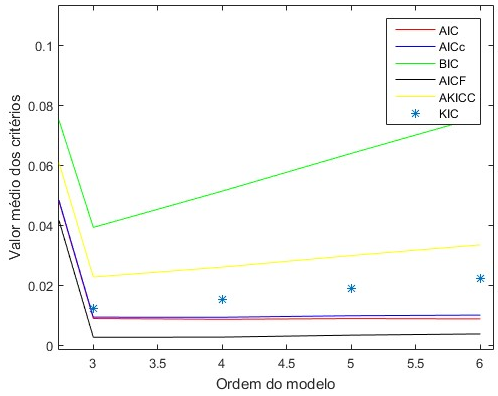

Figura 4: Valores médios dos critérios para as ordens dos modelos – 300 amostras

DOI: 10.5540/03.2016.004.01.0220

3.3 Experimento com 3000 amostras

Ao repetir o experimento com um número de amostras maior, deve favorecer o critério BIC, que é assintoticamente consistente (escolhe o modelo correto com probabilidade 1, à medida que o tamanho da amostra cresce).

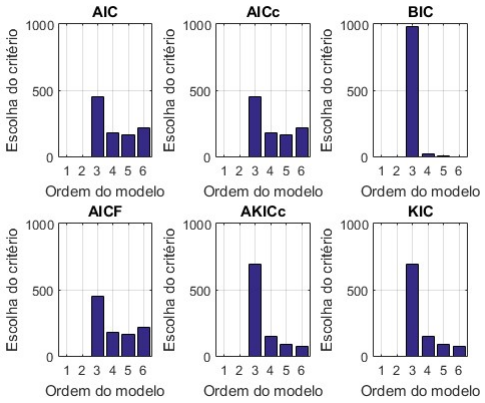

Figura 5: Escolhas dos critérios de informação – 3000 amostras

A Figura 5 confirma a afirmação, com o BIC escolhendo a ordem correta em mais de 96% das vezes, mostrando que a tendência do critério é acertar cada vez mais. Embora esta tendência possa induzir a afirmar que o BIC é superior aos outros critérios, devemos lembrar que na modelagem de fenômenos reais nem sempre a afirmação de que o modelo "correto" está dentre o conjunto de candidados, na verdade, não se pode afirmar nem que tal modelo existe. Com isso, nessas situações, o BIC pode não fazer as melhores escolhas, tendo o pesquisador que ter o cuidado de não confiar cegamente no critério de informação.

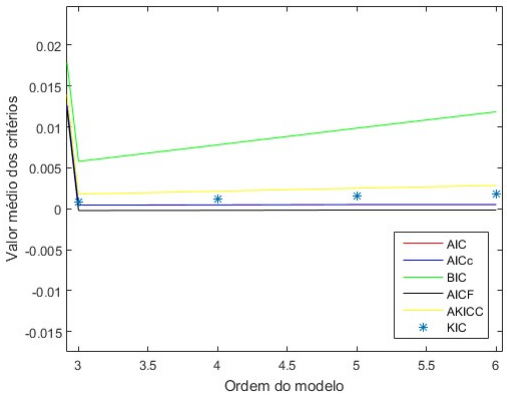

Figura 6: Valores médios dos critérios para as ordens dos modelos – 3000 amostras

A Figura 6 confirma a maior penalização do BIC em relação aos outros critérios.

DOI: 10.5540/03.2016.004.01.0220

4 Conclusões

Neste trabalho, foram analisados seis critérios de informação no tocante a desempenho em pequenas amostras e evolução das características asssintóticas de modelos de regressão linear de ordem de 1 a 6.

Foi demonstrado que as modificações para pequenas amostras contribuem para uma melhora na seleção dos modelos quando o conjunto de dados é pequeno. À medida que o conjunto de dados cresce, o desempenho das variantes de pequenas amostras convergem para o de suas versões originais, o que indica que elas podem ser utilizadas em substituição aos critérios originais (AIC, KIC), sem nenhuma desvantagem aparente.

O critério BIC, por sua vez, teve excelente desempenho com amostras maiores, sendo recomendável sua aplicação na seleção de ordem de modelos de regressão linear.

Agradecimentos

Agradecimentos à UNICAMP, pela estrutura para desenvolver este trabalho.

Referências

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, Proc. 2nd International Symposiom on Information Theory (eds. B.N. Petrov and F. Csaki), 267-281, (1973). DOI:10.1007/978-1-4612-1694-0_15

[2] K. Burnham and D. Anderson, Model Selection and Multimodel Inference: A Practical Information-Teoretic Approach, 2nd ed., New York: Springer, (2002).

[3] J. Cavanaugh, A large-sample model selection criterion based on Kullback's symmetric divergence, Statistics & Probability Letters, vol. 42, nº 4, 333-343, (1999). DOI:10.1016/S0167-7152(98)00200-4

[4] S. Kullback and R. Leibler, On information and sufficiency, Annals of Mathematical Statistics, vol. 22, nº 1, 79-86, (1951). DOI:10.1214/aoms/1177729694

[5] G. Schwartz, Estimating the dimension of a model, Annals of Statystics, vol. 6, 461-464, (1978). DOI:10.1214/aos/1176344136

[6] A. -K. Seghouane, A Small Sample Model Selection Criterion Based on Kullback's Symmetric Divergence, IEEE transactions on signal processing, vol. 52, nº 12, 3314-3323, December (2004). DOI: 10.1109/TSP.2004.837416

[7] B. N. C. B. V. d. L. A. Spiegelhalter D.J., Bayesian Measures of Model Complexity and Fit, Journal of the Royal Statistical Society, vol. 64, nº 4, 583-616, (2002). DOI:10.1111/1467-9868.00353

[8] N. Sugiura, Further analysis of the data by Akaike's information criterion and the finite corrections, Communications in Statistics, 13-26, (1978). DOI:10.1080/03610927808827599

