

Seleção de modelos

Critério de Informação de Akaike

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Novembro, 2020

Introdução

- ▶ AIC, o critério de informação Akaike, é geralmente considerado como o primeiro critério de seleção do modelo.
- ▶ Hoje, o AIC continua a ser a mais amplamente conhecida e usada ferramenta de seleção de modelos entre os profissionais.
- ▶ AIC foi introduzido por Hirotugu Akaike em 1973 no artigo "Information Theory and an Extension of the Maximum Likelihood Principle" (in: B. N. Petrov and F. Csaki, eds., 2nd International Symposium on Information Theory, Akademia Kiado, Budapest, pp. 267-281).

Introdução

- ▶ O paradigma de máxima verossimilhança tradicional, aplicada a modelagem estatística, fornece um mecanismo para estimar os parâmetros desconhecidos de um modelo com uma estrutura e dimensão específica.
- ▶ Akaike estendeu esse paradigma ao considerar uma estrutura em que a dimensão do modelo também é desconhecida e deve portanto, ser determinada a partir dos dados.
- ▶ Assim, Akaike propôs uma estrutura em que ambos os modelos: o modelo estimado e o modelo selecionado podem ser simultaneamente realizados.

Introdução

- ▶ Para um modelo de candidato paramétrico de interesse, a função de verossimilhança reflete a conformidade do modelo com os dados observados.
- ▶ À medida que a complexidade do modelo aumenta, o modelo torna-se mais capaz de se adaptar às características dos dados.
- ▶ Assim, selecionando o modelo ajustado que maximiza a verossimilhança empírica invariavelmente levará alguém a escolher o mais complexo modelo na coleção de candidatos.
- ▶ A seleção de modelos com base no princípio da verossimilhança, portanto, requer uma extensão do paradigma de verossimilhança tradicional.

Introdução

Estrutura

- ▶ Estrutura da seleção de modelos
- ▶ Informação de Kullback-Leibler
- ▶ Derivação do AIC
- ▶ Propriedades e limitações do AIC
- ▶ Uso do AIC
- ▶ Aplicações

Estrutura da seleção de modelos

- ▶ Suponha que uma coleção de dados y tenha sido gerada de acordo a um modelo desconhecido ou densidade $g(y)$.
- ▶ Nos esforçamos para encontrar um modelo paramétrico que forneça uma aproximação adequada para $g(y)$.
- ▶ Seja $\mathcal{F}_k = \{f(y, \theta_k) \mid \theta_k \in \Theta_k\}$ a classe paramétrica k -dimensional, ou seja, uma classe de densidades em que o espaço paramétrico Θ_k consiste em vetores k -dimensionais cujos componentes são funcionalmente independentes.
- ▶ Seja $\hat{\theta}_k$ o vetor de estimativas obtidas pela maximização da função de verossimilhança $f(y; \theta_k)$ sobre Θ_k .
- ▶ Seja $f(y; \hat{\theta}_k)$ o modelo ajustado correspondente.

Estrutura da seleção de modelos

- ▶ Suponha que nosso objetivo seja pesquisar entre uma coleção de classes $\mathcal{F} = \{\mathcal{F}_{k_1}, \mathcal{F}_{k_2}, \dots, \mathcal{F}_{k_L}\}$ para o modelo ajustado $f(y; \hat{\theta}_k)$ com $k \in \{k_1, k_2, \dots, k_L\}$, que serve como a melhor aproximação de $g(y)$.
- ▶ Nota: o AIC pode ser usado para delinear entre diferentes modelos ajustados com a mesma dimensão. Por exemplo, modelos de regressão ajustados com base em matrizes de planejamento com o mesmo tamanho, mas com espaços de coluna diferentes.
- ▶ Para simplificar, nossa notação e estrutura pressupõem que cada classe de modelos candidatos $\mathcal{F}_k$ e o correspondente modelo ajustado $f(y; \hat{\theta}_k)$ são distinguidos pela dimensão k .
- ▶ Nosso problema de seleção de modelos pode, portanto, ser visto como um problema de determinação de dimensão.

Estrutura da seleção de modelos

- ▶ Modelo verdadeiro ou gerador: $g(y)$.
- ▶ Candidatos ou modelos aproximados: $f(y; \theta_k)$.
- ▶ Modelos ajustados: $f(y; \hat{\theta}_k)$.
- ▶ Família dos candidatos: $\mathcal{F}_k$.
- ▶ Para determinar qual dos modelos ajustados

$$\left\{ f(y; \hat{\theta}_{k_1}), f(y; \hat{\theta}_{k_2}), \dots, f(y; \hat{\theta}_{k_L}) \right\}$$

melhor se aproxima de $g(y)$, exigimos uma medida que forneça um reflexo adequado da disparidade entre o modelo verdadeiro $g(y)$ e um modelo aproximado $f(y; \theta_k)$.

- ▶ A informação de Kullback-Leibler é uma dessas medidas.

Informação de Kullback-Leibler

- ▶ Para duas densidades paramétricas arbitrárias $g(y)$ e $f(y)$, a informação de Kullback-Leibler ou a divergência dirigida de Kullback entre $g(y)$ e $f(y)$ em relação a $g(y)$ é definida como

$$I(g, f) = E \left(\ln \frac{g(Y)}{f(Y)} \right),$$

onde E denota a esperança sob $g(y)$.

- ▶ $I(g, f) \geq 0$ com igualdade se, e somente se, g e f forem densidades iguais.
- ▶ $I(g, f)$ reflete a separação entre g e f .
- ▶ $I(g, f)$ não é uma medida de distância formal.

Informação de Kullback-Leibler

- ▶ Para nossos propósitos, consideraremos a informação de Kullback-Leibler entre o modelo verdadeiro $g(y)$ e o modelo aproximador $f(y; \theta_k)$ em relação a $g(y)$, que denotaremos como $I(\theta_k)$:

$$I(\theta_k) = E \left(\ln \frac{g(Y)}{f(Y; \theta_k)} \right).$$

- ▶ Na expressão anterior e daí em diante, E denota a esperança sob a densidade ou modelo verdadeiro.

Informação de Kullback-Leibler

- ▶ Seja

$$d(\theta_k) = E(-2 \ln(f(Y; \theta_k))) .$$

- ▶ $d(\theta_k)$ é chamada de discrepância de Kullback.
- ▶ Observe que podemos escrever

$$2l(\theta_k) = d(\theta_k) - E(-2 \ln(g(y))) .$$

- ▶ Como $E(-2 \ln(g(y)))$ não depende de θ_k , qualquer classificação de um conjunto de modelos candidatos correspondendo a valores de $l(\theta_k)$ seria idêntica a uma classificação correspondente a valores de $d(\theta_k)$.
- ▶ Portanto, com o propósito de discriminar entre vários modelos candidatos, $d(\theta_k)$ serve como um substituto válido para $l(\theta_k)$.

Derivação do AIC

► A medida

$$d(\hat{\theta}_k) = E \left(-2 \ln (f(Y; \theta_k)) \right) \Big|_{\theta_k = \hat{\theta}_k}$$

reflete a separação entre o modelo gerador $g(y)$ e um modelo ajustado $f(y; \hat{\theta}_k)$.

- Avaliar $d(\hat{\theta}_k)$ não é possível, requer conhecimento de $g(y)$.
- O trabalho de Akaike (1973), entretanto, sugere que

$$-2 \ln (f(Y; \theta_k)),$$

serve como um estimador enviesado de $d(\hat{\theta}_k)$ e o viés

$$E(d(\hat{\theta}_k)) - E(-2 \ln (f(y; \hat{\theta}_k))),$$

é estimado assintoticamente como duas vezes a dimensão de θ_k .

Derivação do AIC

- Uma vez que k denota a dimensão de θ_k , sob condições apropriadas, o valor esperado de

$$AIC = -2 \ln (f(y; \hat{\theta}_k)) + 2k$$

deve se aproximar assintoticamente do valor esperado de $d(\hat{\theta}_k)$, que chamaremos

$$\Delta_k = E(d(\hat{\theta}_k)).$$

- Especificamente, pode-se estabelecer que

$$E(AIC) + o(1) = \Delta_k.$$

- AIC, portanto, fornece um estimador não enviesado assintoticamente de Δ_k .

Derivação do AIC

- ▶ Δ_k é frequentemente chamada de discrepância esperada de Kullback.
- ▶ Δ_k reflete a separação média entre o modelo gerador $g(y)$ e os modelos ajustados com a mesma estrutura que $f(y; \hat{\theta}_k)$.
- ▶ A propriedade do AIC ser não enviesado assintoticamente requer a suposição de que $g(y) \in \mathcal{F}_k$.
- ▶ Essa suposição implica que o verdadeiro modelo é um membro da classe $\mathcal{F}_k$ e pode ser escrito como $f(y; \theta_0)$, $\theta_0 \in \Theta_k$.
- ▶ De uma perspectiva prática, a suposição de $f(y; \theta_0) \in \mathcal{F}_k$ implica que o modelo ajustado $f(y; \hat{\theta}_k)$ está corretamente especificado ou superdimensionado.
- ▶ Esta é uma suposição problemática?

Derivação do AIC

- ▶ Para justificar a imparcialidade assintótica de AIC, considere escrever Δ_k da seguinte forma:

$$\begin{aligned}\Delta_k &= E(d(\hat{\theta}_k)) \\ &= E(-2 \ln(f(y; \hat{\theta}_k))) \\ &\quad + \left(E(-2 \ln(f(y; \theta_0))) - E(-2 \ln(f(y; \hat{\theta}_k))) \right) \quad (1) \\ &\quad + \left(E(d(\hat{\theta}_k)) - E(-2 \ln(f(y; \theta_0))) \right) \cdot \quad (2)\end{aligned}$$

- ▶ O seguinte lema afirma que tanto (1) quanto (2) satisfazendo serem de ordem $o(1)$ conforme k .

Derivação do AIC

- Assumimos as condições de regularidade necessárias para garantir a consistência e normalidade assintótica do vetor de estimadores de máxima verossimilhança $\hat{\theta}_k$.

Lema

$$E(-2 \ln(f(y; \theta_0))) - E(-2 \ln(f(y; \hat{\theta}_k))) = k + o(1) \quad (1)$$

$$E(d(\hat{\theta}_k)) - E(-2 \ln(f(y; \theta_0))) = k + o(1) \quad (2)$$

Derivação do AIC

Demonstração



$$I(\theta_k) = E \left(- \frac{\partial^2 \ln (f(y; \theta_k))}{\partial \theta_k \partial \theta_k^\top} \right)$$

e

$$\mathcal{I}(\theta_k; y) = \left(- \frac{\partial^2 \ln (f(y; \theta_k))}{\partial \theta_k \partial \theta_k^\top} \right)$$

- ▶ $I(\theta_k)$ é a matriz de informação de Fisher esperada.
- ▶ $\mathcal{I}(\theta_k; y)$ é a matriz de informação Fisher observada.

Derivação do AIC

Demonstração

- ▶ Fazer a expansão até segunda ordem de $-2 \ln(f(y; \theta_0))$ a respeito de $\hat{\theta}_k$ e avaliar a esperança do resultado.
- ▶ Obtemos

$$\begin{aligned} E(-2 \ln(f(y; \theta_0))) &= E(-2 \ln(f(y; \hat{\theta}_k))) \\ &+ E\left((\hat{\theta}_k - \theta_0)^\top \mathcal{I}(\hat{\theta}_k; y)(\hat{\theta}_k - \theta_0)\right) + o(1). \end{aligned}$$

- ▶ Então

$$\begin{aligned} E(-2 \ln(f(y; \theta_0))) - E(-2 \ln(f(y; \hat{\theta}_k))) &= \\ &= E\left((\hat{\theta}_k - \theta_0)^\top \mathcal{I}(\hat{\theta}_k; y)(\hat{\theta}_k - \theta_0)\right) + o(1). \quad (3) \end{aligned}$$

Derivação do AIC

Demonstração

- ▶ Agora, considere fazer uma expansão de segunda ordem de $d(\hat{\theta}_k)$ a respeito de θ_0 e avaliar a esperança do resultado.
- ▶ Obtemos

$$\begin{aligned} E(d(\hat{\theta}_k)) &= E(-2 \ln(f(y; \theta_0))) \\ &+ E\left((\hat{\theta}_k - \theta_0)^\top I(\theta_0)(\hat{\theta}_k - \theta_0)\right) + o(1). \end{aligned}$$

- ▶ Então

$$\begin{aligned} E(d(\hat{\theta}_k)) - E(-2 \ln(f(y; \theta_0))) &= \\ &= E\left((\hat{\theta}_k - \theta_0)^\top I(\theta_0)(\hat{\theta}_k - \theta_0)\right) + o(1). \quad (4) \end{aligned}$$

Derivação do AIC

Demonstração

- ▶ As formas quadráticas

$$(\hat{\theta}_k - \theta_0)^\top \mathcal{I}(\hat{\theta}_k; y) (\hat{\theta}_k - \theta_0)$$

e

$$(\hat{\theta}_k - \theta_0)^\top \mathcal{I}(\theta_0) (\hat{\theta}_k - \theta_0),$$

convergem para variáveis aleatórias com distribuição qui-quadrado centrais com k graus de liberdade.

- ▶ Lembremos, estamos assumindo que $\theta_0 \in \Theta_k$.
- ▶ Assim, as esperanças de ambas as formas quadráticas são de ordem $o(1)$ conforme k .
- ▶ Este fato junto com (3) e (4) estabelece (1) e (2).

Correção do vício

- ▶ O AIC nos fornece um estimador aproximadamente não enviesado de Δ_k em ambientes onde n é grande e k é comparativamente pequeno.
- ▶ Em situações onde n é pequeno e k é comparativamente grande, por exemplo, $k \approx n/2$, $2k$ é frequentemente muito menor do que o ajuste do vício, tornando o AIC substancialmente negativo como um estimador de Δ_k .
- ▶ Em aplicações de pequenas amostras, melhores estimadores do que $2k$, do vício, estão disponíveis.
- ▶ AIC é assintoticamente eficiente no sentido descrito nos artigos de Shibata (1980, 1981), mas não é consistente.

Correção do vício

- ▶ Se o AIC subestimar severamente Δ_k para modelos estimados de dimensões superiores no conjunto de candidatos, o critério pode favorecer os modelos de dimensão mais elevadas, mesmo quando a discrepância esperada entre esses modelos e o modelo gerador for bastante grande.
- ▶ Exemplos que ilustram esse fenômeno aparecem em Linhart and Zucchini (1986), que comentam "em alguns casos o critério simplesmente continua a diminuir à medida que o número de parâmetros no modelo de aproximação aumenta".

Eficiência assintótica

- ▶ Suponha que o modelo gerador seja de dimensão finita e que esse modelo esteja representado na coleção de candidatos em consideração. Um critério consistente selecionará assintoticamente o modelo estimado com a estrutura correta com probabilidade um.
- ▶ Por outro lado, suponha que o modelo gerador seja de dimensão infinita e, portanto, esteja fora da coleção de candidatos em consideração. Um critério assintoticamente eficiente selecionará assintoticamente o modelo ajustado candidato que minimiza o erro quadrático médio de predição.

Uso do AIC

- ▶ AIC é aplicável em uma ampla gama de estruturas de modelagem, uma vez que sua justificação requer apenas propriedades convencionais dos estimadores de máxima verossimilhança em grandes amostras
- ▶ A aplicação do critério não requer a suposição de que um dos modelos candidatos seja o verdadeiro modelo ou o modelo correto, embora a derivação implique o contrário.
- ▶ Em uma aplicação de seleção de modelo, o modelo ajustado ótimo é identificado pelo valor mínimo de AIC.
- ▶ No entanto, os valores do critério são importantes; modelos com valores semelhantes devem receber a mesma "classificação" na avaliação de preferências de critérios.

Uso do AIC

- ▶ "Uma vantagem substancial em usar critérios teóricos da informação é que eles são válidos para modelos não aninhados. Obviamente, os testes tradicionais de razão de verossimilhança são definidos apenas para modelos aninhados, e isso representa outra limitação substancial no uso de testes de hipótese na seleção de modelos." (Burnham and Anderson, 2002)
- ▶ Contudo, não constantes devem ser descartadas do termo de bondade $-2 \ln(f(y; \hat{\theta}_k))$, como $n \ln(2\pi)$.
- ▶ Lembre-se de que certos softwares estatísticos descartam rotineiramente constantes na avaliação de critérios de seleção baseados na verossimilhança: por exemplo, na regressão linear normal, $n \ln(\hat{\sigma}^2)$ é muitas vezes usado como o termo de bondade de ajuste, como oposto a $n \ln(\hat{\sigma}^2) + n(\ln(2\pi) + 1)$.

Referências

- ▶ Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov and F. Csáki, eds., 2nd International Symposium on Information Theory, Akadémia Kiadó, Budapest, pp. 267-281.
- ▶ Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control AC-19, 716-723.
- ▶ Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Annals of Statistics 80, 147-164.
- ▶ Shibata, R. (1981). An optimal selection of regression variables. Biometrika 68, 45-54.

Referências

- ▶ Zucchini, W., Claeskens, G. and Nguefack-Tsague, G. (2011). Model Selection, Springer Berlin Heidelberg.
- ▶ Linhart, H. and Zucchini, W. (1986). Model Selection, John Wiley & Sons.
- ▶ Burnham, K.P. and Anderson, D.R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Springer-Verlag.
- ▶ Hurvich, C.M. and Tsai, C.L. (1989). Regression and Time Series Model Selection in Small Samples, Biometrika, 76.
- ▶ Burnham, K.P. and Anderson, D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection, Sociological methods and research, 33, pp. 261–304.

Hirotsugu Akaike (1927 - 2009)

