第7章 最尤推定法におけるモデル選択

7.3 赤池の情報量基準
7.4 竹内の情報量基準

06T4029T 斉藤優

復習

- カルバック・ライブラー情報量

$$KL(p\|\hat{p}) = \underbrace{E_x[\log p(x)]}_{\text{負のエントロピー}} - \underbrace{E_x[\log \hat{p}(x)]}_{\text{期待対数尤度}}$$

  - カルバック・ライブラー情報量が小さい
    → $\hat{p}(x)$は$p(x)$の良い近似
  - 期待対数尤度を直接求めることができない
    → 推定法を利用して求める
  - 期待対数尤度を平均対数尤度で近似
    → 平均対数尤度を最大にするモデルを選択
      → 常に最も複雑なモデルが選んでしまう

赤池の情報量基準(1)

- 赤池の情報量基準

$$AIC := \underbrace{-\sum_{i=1}^{n} \log q(x_i; \hat{\theta}_{ML})}_{\text{負の対数尤度}} + \underbrace{t}_{\text{モデルのパラメータの数}}$$

  - 精度の良い期待対数尤度の推定量
  - 相反する単調関数を足し合わせたもの(図1)

    パラメータ数の増加→負の対数尤度の減少
    負の対数尤度の増加→パラメータ数の減少

  - AICが小さくなるモデルを選択(図1)

    → 程よい複雑さのモデルが選択

赤池の情報量基準(2)

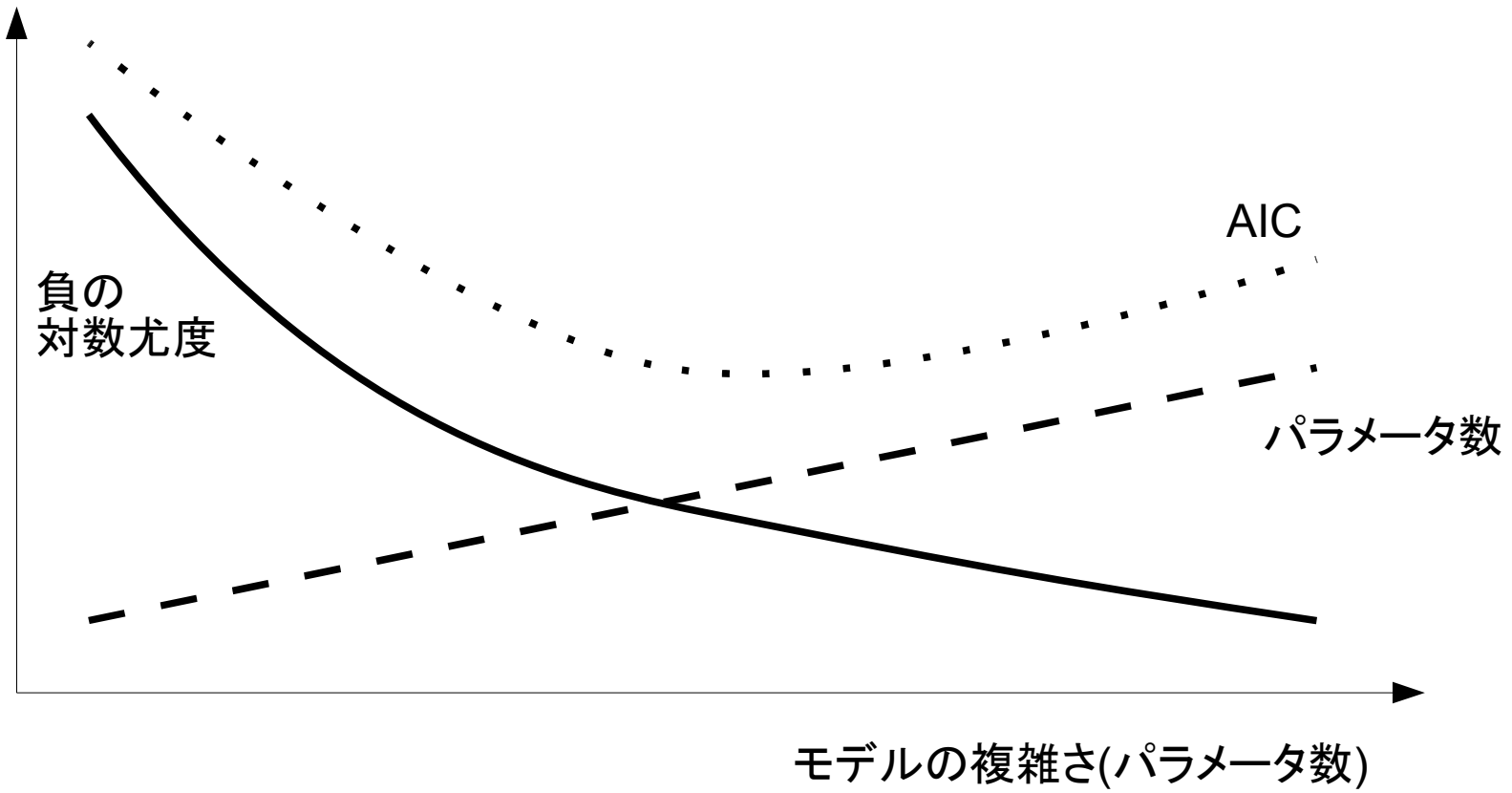

図1：赤池の情報量基準(AIC)

AICの理論的な妥当性(1)

- 準備
  - 負の期待対数尤度

$$J := -E_x[\log q(x;\hat{\theta}_{ML})]$$

  - Jを最小にするパラメータ値

$$\theta^* := \underset{\theta}{argmin}\, J$$

  - Jを 訓練標本$\{x_i\}_{i=1}^n$のとり方に関して期待値をとった量

$$E[J] = -E\left[\frac{1}{n}\sum_{i=1}^{n}\log q(x_i;\hat{\theta}_{ML})\right] + \frac{1}{n}tr(Q(\theta^*)G(\theta^*)^{-1}) + o(n^{-1})$$

AICの理論的妥当性(2)

- 準備(続き)
  - フィッシャー情報行列

$$Q_{\ell,\ell'}(\theta) := E_x\left[\frac{\partial}{\partial\theta^{(\ell)}}\log q(x;\theta)\frac{\partial}{\partial\theta^{(\ell')}}\log q(x;\theta)\right]$$

  - 負の期待ヘッセ行列

$$G_{\ell,\ell'}(\theta) := -E_x\left[\frac{\partial^2}{\partial\theta^{(\ell)}\partial\theta^{(\ell')}}\log q(x;\theta)\right]$$

  - 負の期待ヘッセ行列の変形

$$G_{\ell,\ell'}(\theta) = -\int \frac{\frac{\partial^2}{\partial\theta^{(\ell)}\partial\theta^{(\ell')}}q(x;\theta)}{q(x;\theta)}p(x)dx + Q_{\ell,\ell'}(\theta)$$

AICの理論的妥当性(3)

- AICの導出

(1) パラメトリックモデルが真の確率密度関数 $p(x)$を含む場合を仮定

$$q(x;\theta^*)=p(x)$$

(2) このときの負の期待ヘッセ行列

$$G_{\ell,\ell'}(\theta^*)=-\int \frac{\partial^2}{\partial\theta^{(\ell)}\partial\theta^{(\ell')}} q(x;\theta)\Big|_{\theta=\theta^*} dx+Q_{\ell,\ell'}(\theta^*)$$

(3) パラメトリックモデル $q(x;\theta)$は確率密度関数

$$\int q(x;\theta)\,dx=1$$

AICの理論的妥当性(4)

- AICの導出(続き)

(4) 両辺を$\theta^{(\ell)}$と$\theta^{(\ell')}$で偏微分

$$\frac{\partial^2}{\partial\theta^{(\ell)}\partial\theta^{(\ell')}}\int q(x;\theta)\,dx=0$$

(5) 偏微分と積分の順序が入れ替えられることを仮定

$$\int \frac{\partial^2}{\partial\theta^{(\ell)}\partial\theta^{(\ell')}} q(x;\theta)\,dx=0$$

(6) (2)の負の期待ヘッセ行列より

$$G_{\ell,\ell'}(\theta^*)=Q_{\ell,\ell'}(\theta^*)$$

ＡICの理論的妥当性(5)

- AICの導出(続き)

  (7) 次式が成立

$$tr\left(G_{\ell,\ell'}(\theta^*)Q_{\ell,\ell'}(\theta^*)^{-1}\right)=tr\left(I_t\right)=t$$

- 結論
  - 真の確率密度関数のパラメータθ^*でAICの第2項目が得られる

竹内の情報量基準(1)

- 竹内の情報量基準

$$TIC := -\sum_{i=1}^{n} \log q(x_i ; \hat{\theta}_{ML}) + tr(\hat{Q}(\hat{\theta}_{ML}) G(\hat{\theta}_{ML})^{-1})$$

  - 一般化した規準

  - $\hat{Q}(\hat{\theta}_{ML})$と$\hat{G}(\hat{\theta}_{ML})$の定義

$$\hat{Q}_{\ell,\ell'}(\hat{\theta}_{ML}) := \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta^{(\ell)}} \log q(x_i;\theta) \frac{\partial}{\partial \theta^{(\ell')}} \log q(x_i;\theta) \Big|_{\theta=\hat{\theta}_{ML}}$$

$$\hat{G}_{\ell,\ell'}(\hat{\theta}_{ML}) := -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^{(\ell)} \partial \theta^{(\ell')}} \log q(x_i;\theta) \Big|_{\theta=\hat{\theta}_{ML}}$$

竹内の情報量規準(2)

- 竹内の情報量規準(続き)
  - $\hat{Q}(\hat{\theta}_{ML})$と$\hat{G}(\hat{\theta}_{ML})$は$Q(\theta^*)$と$G(\theta^*)$の一致推定量

$$\hat{Q}(\hat{\theta}_{ML}) \xrightarrow{p} Q(\theta^*),\ \ \hat{G}(\hat{\theta}_{ML}) \xrightarrow{p} G(\theta^*)$$

  - TICの$1/n$倍の期待値はJの期待値に収束

$$\frac{1}{n}E[TIC] = E[J] + \underbrace{o(n^{-1})}_{\text{誤差}}$$

負の平均対数尤度をそのままJの推定量として用いた場合

$$E\left[-\frac{1}{n}\sum_{i=1}^{n}\log q(x_i;\hat{\theta}_{ML})\right] = E[J] + O(n^{-1})$$

竹内の情報量規準(3)

- 結論
  - *TIC* は負の平均対数尤度よりも J の推定精度が良い
  - 理論的にはAICよりもTICの方が汎用的
    → 簡便性から現実問題ではAICの方が利用頻度が高い