公開講座 2000年7月6日

統計数理研究所 下平英寿

(I) モデル信頼集合

(II) 情報量規準の拡張

参考：「モデル選択理論の新展開」

(統計数理 47，1999)

モデル選択，情報量規準の研究の紹介

1．モデル信頼集合

変数選択問題

モデル選択のバラツキと一致性

モデル選択の信頼性

2．情報量規準の拡張

さまざまな推測方式に応じて規準を構成する

罰金付き最尤法，ベイズ予測分布，不完全データ，実験計画

1　モデル信頼集合

- モデル選択 $\widehat{k}$ ── 点推定

  AICの差が小さければ，どちらのモデルもおなじくらい良いハズ

- モデル信頼集合 $\mathcal{T}$ ── 区間推定

  モデル選択のバラツキを考慮したうえで，良いモデルを複数選ぶ

とくにノンネストなモデルの比較に有効

1.1　回帰モデルの変数選択問題 (例: 説明変数が３個)

$$Y \sim N\big(\mu(X_{\mathsf{a}}, X_{\mathsf{b}}, X_{\mathsf{c}}), \sigma^2\big)$$

$$\mu(x_{\mathsf{a}}, x_{\mathsf{b}}, x_{\mathsf{c}}) = \beta_0 + \beta_{\mathsf{a}} x_{\mathsf{a}} + \beta_{\mathsf{b}} x_{\mathsf{b}} + \beta_{\mathsf{c}} x_{\mathsf{c}}$$

変数選択問題

$$x_{\mathsf{a}}, x_{\mathsf{b}}, x_{\mathsf{c}}$$

の中から適切な変数だけを選ぶ．

モデル α は選択する変数の集合を表す

$$\alpha \subset \{\mathsf{a}, \mathsf{b}, \mathsf{c}\}$$

パラメタのとり得る値とパラメタ数は

$$\Theta_\alpha = \{(\sigma, \boldsymbol{\beta}) | \sigma > 0, \beta_i = 0, i \in \alpha^c\}, \quad |\alpha| + 2$$

- 候補となるモデルαの集合を$\mathcal{M}$で表す．

$$\alpha \in \mathcal{M}$$

[例] フルモデル $\{\mathrm{a},\mathrm{b},\mathrm{c}\}$
すべての組み合わせ $\mathcal{M} = \{\phi, \{\mathrm{a}\}, \{\mathrm{b}\}, \{\mathrm{c}\}, \{\mathrm{a},\mathrm{b}\}, \{\mathrm{a},\mathrm{c}\}, \{\mathrm{b},\mathrm{c}\}, \{\mathrm{a},\mathrm{b},\mathrm{c}\}\}$
２個の変数を使う $\mathcal{M} = \{\{\mathrm{a},\mathrm{b}\}, \{\mathrm{a},\mathrm{c}\}, \{\mathrm{b},\mathrm{c}\}\}$

- 候補モデルに番号をつける

$$\alpha(k), \quad k \in \{1, 2, \ldots, K\}$$

各モデルは集合αまたは番号kで示される

- パラメトリック・モデル

$$M_\alpha = f(\cdot|\Theta_\alpha), \quad \alpha \in \mathcal{M}$$

$$M_1, M_2, \ldots, M_K$$

モデルの包含関係

二つのモデルM_αとM_βを考える

- $\alpha \subset \beta$ **のとき，**M_α**は**M_β**の部分モデル**

  このときM_α**は**M_β**に「ネスト」していると言う**

  [例] $\phi \subset \{\mathrm{a}\} \subset \{\mathrm{a},\mathrm{b}\} \subset \{\mathrm{a},\mathrm{b},\mathrm{c}\}$

- $\alpha \subset \beta$**でも**$\beta \subset \alpha$**でも無いとき，**M_α**と**M_β**は「ノンネスト」と言う**

  [例] $\{\mathrm{a},\mathrm{b}\}$**と**$\{\mathrm{b},\mathrm{c}\}$**は互いにネストの関係に無い（ノンネスト）**

モデル選択とモデル信頼集合

- **期待平均対数尤度**

$$l_n^*(\alpha) \quad \text{または} \quad l_n^*(k)$$

を最小にするα^***または**k^***が「良いモデル」と考える．**

- **バイアス補正した対数尤度**

$$l(\hat{\boldsymbol{\theta}}_\alpha) - \dim M_\alpha \quad \text{または} \quad l(\hat{\boldsymbol{\theta}}_k) - \dim M_k$$

をおおきくするモデル$\hat{\alpha}$**または**$\hat{k}$**を選択する．**

モデル信頼集合 $\mathcal{T}$**は，**$\mathcal{M}$**の部分集合（**$\mathcal{T} \subset \mathcal{M}$**）で，**

$$P\{\alpha^* \in \mathcal{T}\} \text{ または } P\{k^* \in \mathcal{T}\} \geq 1 - P^*$$

を満たすもの．ただしP^***は有意水準**

数値例

HALDのセメントデータ

目的変数：セメント発熱量
説明変数：４種の主成分の混合率
データ数： $n = 13$
モデル数： $K = 16$ **（すべての説明変数の組み合わせ** $2^4 = 16$ **個）**

BOSTONの住宅価格データ

目的変数：住宅価格（各地区の中央値の対数）
説明変数：各地区の１３個の指標
a＝犯罪率，f＝平均部屋数，k＝教師数:生徒数の比，m＝社会階層の構成比
データ数： $n = 506$
モデル数： $K = 286$ **（説明変数を３個だけ使う）**

HALDデータのAICとp-値

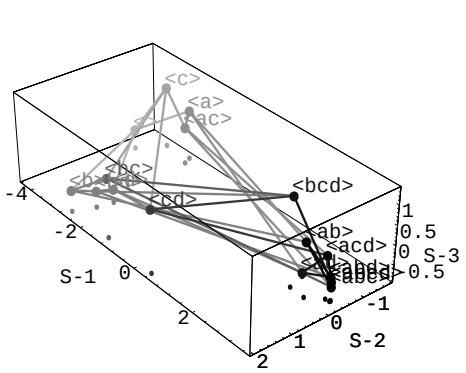
順位	α	△AIC	正規検定	多重比較	選択確率	LR検定
1	<abd>	0	−	0.972	0.180	0.863
2	<abc>	0.04	0.493	0.938	0.256	0.796
3	<ab>	0.45	0.438	0.911	0.255	0.290
4	<acd>	0.75	0.382	0.924	0.158	0.376
5	<abcd>	1.97	0.088	0.451	0.002	−
6	<ad>	3.77	0.169	0.540	0.094	0.055
7	<bcd>	5.60	0.136	0.353	0.052	0.018
8	<cd>	14.88	0.019	0.074	0.004	0.000
9	<bc>	26.06	0.001	0.000	0.000	0.000
10	<d>	33.88	0.000	0.000	0.000	0.000
11	<b>	34.20	0.000	0.000	0.000	0.000
12	<bd>	35.66	0.000	0.000	0.000	0.000
13	<a>	38.55	0.000	0.000	0.000	0.000
14	<ac>	40.14	0.000	0.000	0.000	0.000
15	<c>	44.09	0.000	0.000	0.000	0.000
16	<>	46.47	0.000	0.000	0.000	0.000

BOSTONデータにおける上位20個のモデルのAICとp-値

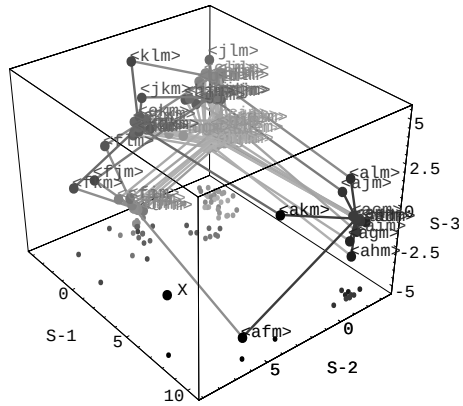
			p-value × 100					
順位	α	LK	BA	BP	KH	MC	MS	LR
1	<afm>	0	70.3	48.8	−	99.3	99.4	0.0
2	<akm>	0.9	29.7	39.5	46.3	98.8	98.7	0.0
3	<ahm>	9.3	0.0	7.3	18.0	86.2	77.1	0.0
4	<agm>	12.4	0.0	0.5	9.0	79.8	53.9	0.0
5	<adm>	12.7	0.0	1.5	11.1	80.7	55.1	0.0
6	<alm>	16.1	0.0	0.0	5.0	74.0	22.1	0.0
7	<ajm>	17.7	0.0	0.0	2.4	71.2	4.6	0.0
8	<abm>	19.2	0.0	0.0	1.9	66.6	10.0	0.0
9	<aim>	19.8	0.0	0.0	1.4	65.8	8.3	0.0
10	<acm>	20.4	0.0	0.0	1.0	64.3	3.0	0.0
11	<aem>	20.5	0.0	0.0	1.1	63.8	3.5	0.0
12	<klm>	23.4	0.0	1.4	8.3	55.9	41.5	0.0
13	<fjm>	24.4	0.0	0.2	2.9	55.7	26.0	0.0
14	<jkm>	27.0	0.0	0.0	3.2	47.9	12.5	0.0
15	<fkm>	27.1	0.0	0.1	3.4	47.9	31.4	0.0
16	<flm>	29.9	0.0	0.4	3.7	40.6	27.7	0.0
17	<hjm>	30.3	0.0	0.1	2.6	40.1	20.5	0.0
18	<dkm>	31.0	0.0	0.0	2.8	38.1	15.7	0.0
19	<bkm>	31.3	0.0	0.0	2.3	37.2	13.0	0.0
20	<ekm>	32.4	0.0	0.0	1.7	34.7	6.6	0.0

LK:△log L
BA:ベイズ事後確率
BP:選択確率
KH:正規検定
MC:多重比較(おもみ無し)
MS:多重比較
LR:尤度比検定

モデル地図



HALDデータ



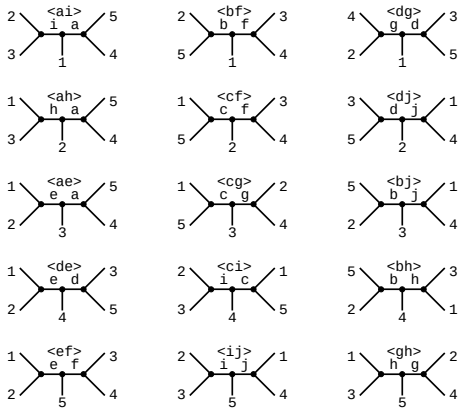
BOSTONデータ

分子系統樹

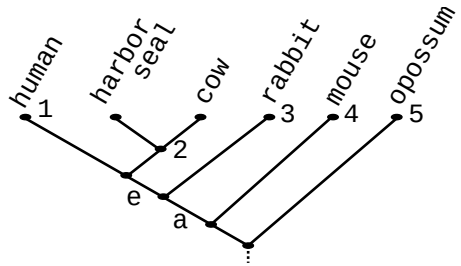
ミトコンドリアのDNAシーケンス

```
          1                   h=20                                  50
  human   ANLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEPLKP
   seal   INIISLIIPILLAVAFLTLVERKVLGYMQLRKGPNIVGPYGLLQPIADAVKLFTKEPLRP
    cow   INILMLIIPILLAVAFLTLVERKVLGYMQLRKGPNVVGPYGLLQPIADAIKLFIKEPLRP
 rabbit   INTLLLILPVLLAMAFLTLVERKILGYMQLRKGPNIVGPYGLLQPIADAIKLFTKEPLRP
  mouse   INILTLLVPILIAMAFLTLVERKILGYMQLRKGPNIVGPYGILQPFADAMKLFMKEPMRP
opossum   INLLMYIIPILLAVAFLTLVERKVLGYMQFRKGPNVIGPYGILQPFADALKLFIKEPLRP
                              Xh
```

シーケンス数: 6種の哺乳類
human, (harbor seal, cow), rabbit, mouse, opossum
説明変数: 10個の "split"
データ数: $n = 3416$のアミノ酸
モデル数: $K = 15$の系統樹



可能な系統樹



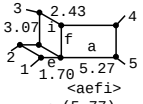
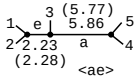
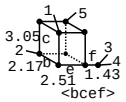
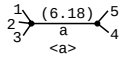
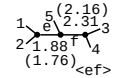
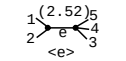
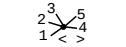
最尤系統樹

a = {123|45}, b = {134|25}, c = {234|15}, d = {124|35}, e = {12|345},
f = {34|125}, g = {24|135}, h = {13|245}, i = {23|145}, j = {14|235}.

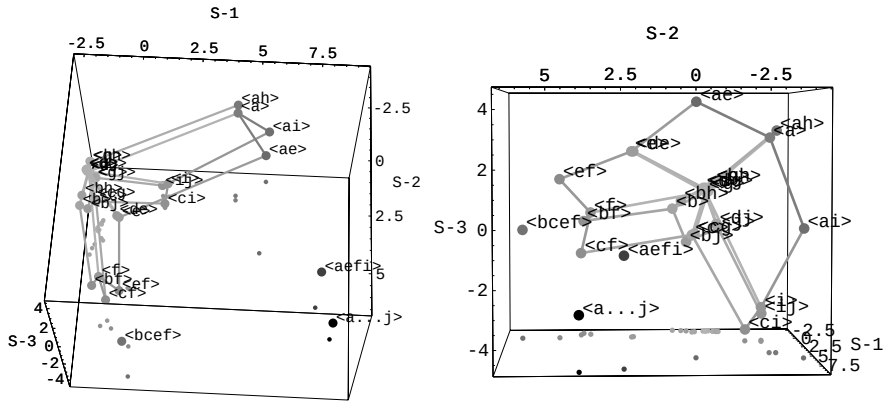
Mammal Phylogeny: p-values

	α	$\Delta \log L$	p-values LR	BA	BP	KH	MC	MS	MB	tree
1	{a, e}	0.0	.000	.934	.583	–	.941	.944	.85	(((12)3)45)
2	{a, i}	2.7	.000	.065	.317	.360	.811	.805	.57	((1(23))45)
3	{a, h}	7.4	.000	.001	.038	.121	.577	.422	.27	(((13)2)45)
4	{e, f}	17.6	.000	.000	.012	.040	.169	.203	.15	((12)(34)5)
5	{c, f}	18.9	.000	.000	.030	.066	.139	.296	.18	(1(2(34))5)
6	{c, i}	20.1	.000	.000	.006	.050	.109	.100	.09	(1((23)4)5)
7	{b, f}	20.6	.000	.000	.011	.048	.107	.248	.19	((1(34))25)
8	{i, j}	22.2	.000	.000	.001	.032	.070	.048	.08	((14)(23)5)
9	{d, e}	25.4	.000	.000	.000	.001	.029	.013	.01	(((12)4)35)
10	{b, j}	26.3	.000	.000	.002	.018	.032	.124	.09	(((14)3)25)
11	{b, h}	28.9	.000	.000	.000	.008	.017	.069	.08	(((13)4)25)
12	{d, j}	31.6	.000	.000	.000	.003	.006	.032	.04	(((14)2)35)
13	{c, g}	31.7	.000	.000	.000	.003	.006	.035	.04	(1((24)3)5)
14	{g, h}	34.7	.000	.000	.000	.001	.002	.012	.03	((13)(24)5)
15	{d, g}	36.2	.000	.000	.000	.000	.001	.007	.02	((1(24))35)

LR:尤度比検定，BA:ベイズ事後確率，BP:選択確率，KH:正規検定,
MC:多重比較(おもみ無し)，MS:多重比較，MB:New Test

α		$\Delta \log L$	$\hat{\theta}_\alpha$
	full	**95.2**	(6.14,2.32,2.78,2.28,2.24,2.94,1.21,1.90,2.13,2.23)
{aefi}		**61.3**	(5.27, , , ,1.70,3.07, , ,2.43,)
{ae}		**38.5**	(5.86, , , ,2.23, , , , ,)
{bcef}		**34.6**	(,2.17,3.05, ,2.51,1.43, , , ,)
{a}		**28.4**	(6.18, , , , , , , , ,)
{ef}		**21.0**	(, , , ,1.88,2.31, , , ,)
{e}		**12.7**	(, , , ,2.52, , , , ,)
{}		**0**	(, , , , , , , , ,)

モデル地図



1.2　モデル選択のバラツキと一致性

- 選択されたモデル $\hat{k}$ はデータ x の関数

$$\hat{k} = \hat{k}(\boldsymbol{x})$$

したがって $\hat{k}(X)$ は確率変数

- $\hat{k}(X)$ は k^* の周辺に分布している

- モデル選択の一致性

$$P\left(\hat{k}(\boldsymbol{X}) = k^*\right) \to 1 \quad (n \to \infty)$$

AICによるモデル選択は，一致性がない場合があることが知られている

これは必ずしもAICの欠点ではない

モデル選択のバラツキ（教科書のフーリエ級数の例）

表 4.6　AIC最小となるk ($=\hat{k}$) の分布 ($n = 500$)

k_A	2	4	6	8	10	12	14	16	18	20	22	計
度数	0	0	0	326	342	173	91	27	17	14	10	1000

表 4.7　$n = 2000$のときの$\hat{k}$の分布

k_A	2	4	6	8	10	12	14	16	18	20	22	計
度数	0	0	0	6	210	359	264	68	58	21	14	1000

なお，$l^*_{500}(k)$最小は$k^* = 10$，$l^*_{2000}(k)$最小は$k^* = 12$のとき

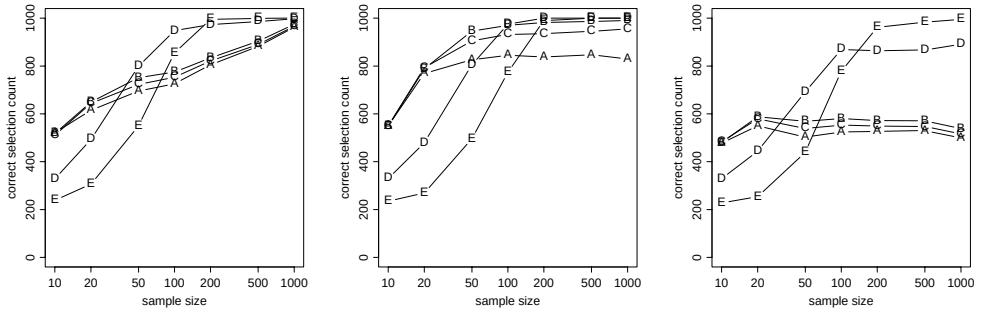
モデル選択の一致性（変数選択の例）

シミュレーション：一番良いモデルが選ばれた回数（1000回中）

データサイズ $n = 10, 20, 50, 100, 200, 500, 1000$

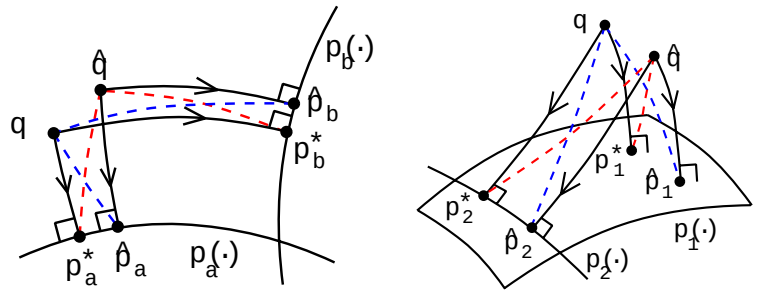
$$\mathrm{IC}_\alpha = -2 \times l(\hat{\boldsymbol{\theta}}_\alpha) + c_n \dim M_\alpha$$

係数c_n：$A = 2$ (AIC)，$B = \log n$ (BIC)，$C = 2n^{0.1}$，$D = 2n^{0.5}$，$E = 2n^{0.6}$



ケース１　ケース２　ケース３

モデルと真の分布の関係（一般の場合）

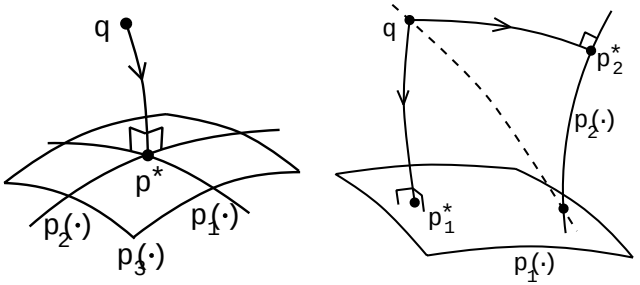


ケース１（ノンネスト）　ケース１’（ネスト）

数値例のパラメタ値 $\boldsymbol{\beta} = (\beta_\mathrm{a}, \beta_\mathrm{b}, \beta_\mathrm{c})$

ケース１：$\boldsymbol{\beta} = (1.0, 0.9, 0.1)$，$\mathcal{M} = \{\phi, \{\mathrm{a}\}, \{\mathrm{b}, \mathrm{c}\}\}$

モデルと真の分布の関係 (特殊な場合)



ケース2 **ケース3**

ケース2: $\beta = (1,1,0)$, $\mathcal{M} = \{\phi, \{\mathtt{a}\}, \{\mathtt{a},\mathtt{c}\}\}$

ケース3: $\beta = (1,1,0)$, $\mathcal{M} = \{\phi, \{\mathtt{a}\}, \{\mathtt{b},\mathtt{c}\}\}$

ICの差

二つのモデル M_α **と** M_β

$$\Delta\mathrm{IC} = 2 \times \left(l(\hat{\boldsymbol{\theta}}_\alpha) - l(\hat{\boldsymbol{\theta}}_\beta)\right) - c_n \times \left(\dim M_\alpha - \dim M_\beta\right)$$

$\Delta\mathrm{IC} > 0$**なら** M_α**を選択，**$\Delta\mathrm{IC} < 0$**なら** M_β**を選択**

$$\begin{aligned}\Delta\mathrm{IC} &\approx 2 \times \left(l(\boldsymbol{\theta}^*_\alpha) - l(\boldsymbol{\theta}^*_\beta)\right) \\ &\quad + \left(\Delta(\boldsymbol{\theta}^*_\alpha, \hat{\boldsymbol{\theta}}_\alpha) - \Delta(\boldsymbol{\theta}^*_\beta, \hat{\boldsymbol{\theta}}_\beta)\right) - c_n \times \left(\dim M_\alpha - \dim M_\beta\right)\end{aligned}$$

ただし対数尤度の $\hat{\theta}_\alpha$**の周りでのテーラ展開は (**θ**を** M_α**の自由パラメタに制限して)**

$$\begin{aligned}l(\boldsymbol{\theta}) &\approx l(\hat{\boldsymbol{\theta}}_\alpha) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_\alpha)\left[\frac{\partial l}{\partial \boldsymbol{\theta}'}\right]_{\hat{\theta}_\alpha} \\ &\quad + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_\alpha)\left[\frac{\partial^2 l}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\right]_{\hat{\theta}_\alpha} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_\alpha)'\end{aligned}$$

これより

$$l(\hat{\boldsymbol{\theta}}_\alpha) \approx l(\boldsymbol{\theta}_\alpha) + \frac{1}{2}\Delta(\boldsymbol{\theta}^*_\alpha, \hat{\boldsymbol{\theta}}_\alpha)$$

ICの差の第1項

$$l(\boldsymbol{\theta}^*_\alpha) - l(\boldsymbol{\theta}^*_\beta) = \sum_{i=1}^{n} \left(\log f(x_i|\boldsymbol{\theta}^*_\alpha) - \log f(x_i|\boldsymbol{\theta}^*_\beta)\right)$$

中心極限定理より，十分大きな n**で**

$$l(\boldsymbol{\theta}^*_\alpha) - l(\boldsymbol{\theta}^*_\beta) \sim N\left(l^*(\boldsymbol{\theta}^*_\alpha) - l^*(\boldsymbol{\theta}^*_\beta), nJ_{\alpha,\beta}\right)$$

$$l^*(\boldsymbol{\theta}^*_\alpha) - l^*(\boldsymbol{\theta}^*_\beta) = n \times \left(I(g; f(\boldsymbol{\theta}^*_\beta)) - I(g; f(\boldsymbol{\theta}^*_\alpha))\right)$$

$$\begin{aligned}J_{\alpha,\beta} &= V\left[\log f(X|\boldsymbol{\theta}^*_\alpha) - \log f(X|\boldsymbol{\theta}^*_\beta)\right] \\ &\approx I\left(f(\boldsymbol{\theta}^*_\alpha), f(\boldsymbol{\theta}^*_\beta)\right) + I\left(f(\boldsymbol{\theta}^*_\beta), f(\boldsymbol{\theta}^*_\alpha)\right)\end{aligned}$$

分散の推定は

$$\hat{J}_{\alpha,\beta} = \frac{1}{n}\sum_{i=1}^{n}\left(\log f(x_i|\hat{\boldsymbol{\theta}}_\alpha) - \log f(x_i|\hat{\boldsymbol{\theta}}_\beta) - \left(l(\hat{\boldsymbol{\theta}}_\alpha) - l(\hat{\boldsymbol{\theta}}_\beta)\right)/n\right)^2$$

ICの差 ケース1

十分大きな n**では第1項が支配的**

$$\frac{1}{2}\Delta\mathrm{IC} \approx \left(l(\boldsymbol{\theta}^*_\alpha) - l(\boldsymbol{\theta}^*_\beta)\right) \sim N\left(l^*(\boldsymbol{\theta}^*_\alpha) - l^*(\boldsymbol{\theta}^*_\beta), nJ_{\alpha,\beta}\right)$$

$I(g; f(\boldsymbol{\theta}^*_\beta)) - I(g; f(\boldsymbol{\theta}^*_\alpha)) > 0$**なら**

$$P\left(\Delta\mathrm{IC} > 0\right) \to 1$$

$I(g; f(\boldsymbol{\theta}^*_\beta)) - I(g; f(\boldsymbol{\theta}^*_\alpha)) < 0$**なら**

$$P\left(\Delta\mathrm{IC} < 0\right) \to 1$$

もし $I(g; f(\boldsymbol{\theta}^*_\beta)) - I(g; f(\boldsymbol{\theta}^*_\alpha)) = 0$**ならば，ケース2かケース3に相当**

ICの差 ケース2

$f(\cdot|\boldsymbol{\theta}^*_\alpha) = f(\cdot|\boldsymbol{\theta}^*_\beta)$**のとき，**

$$J_{\alpha,\beta} = V\left[\log f(X|\boldsymbol{\theta}^*_\alpha) - \log f(X|\boldsymbol{\theta}^*_\beta)\right] = 0$$

$$I(g; f(\boldsymbol{\theta}^*_\beta)) - I(g; f(\boldsymbol{\theta}^*_\alpha)) = 0$$

xなので $\triangle$IC**の第1項は0になる**

$$\triangle\mathrm{IC} \approx \left(\triangle(\boldsymbol{\theta}^*_\alpha, \hat{\boldsymbol{\theta}}_\alpha) - \triangle(\boldsymbol{\theta}^*_\beta, \hat{\boldsymbol{\theta}}_\beta)\right) - c_n \times \left(\dim M_\alpha - \dim M_\beta\right)$$

$$\triangle(\boldsymbol{\theta}^*_\alpha, \hat{\boldsymbol{\theta}}_\alpha) \sim \chi^2_{\dim M_\alpha}, \quad \triangle(\boldsymbol{\theta}^*_\beta, \hat{\boldsymbol{\theta}}_\beta) \sim \chi^2_{\dim M_\beta}$$

もし

$$d = \dim M_\alpha - \dim M_\beta > 0$$

ならば M_β**の方が良いモデル．もし** $c_n \to \infty$**ならば** M_β**を選択する確率は**

$$P(\triangle\mathrm{IC} < 0) \to 1$$

ICの差 ケース2（ネスト）

もしネスト（$M_\beta \subset M_\alpha$**）ならば**

$$\triangle\mathrm{IC} \sim \chi^2_d - c_n d, \quad d = \dim M_\alpha - \dim M_\beta$$

M_β**を選ぶ確率は**

$$P(\triangle\mathrm{IC} < 0) = P(\chi^2_d < c_n d)$$

$c_n = 2, d = 1$**なら** $P(\triangle\mathrm{IC} < 0) \approx 0.84$**．**

$c_n = 2, d = 10$**なら** $P(\triangle\mathrm{IC} < 0) \approx 0.97$**．**

固定した d**について** $c_n \to \infty$**のとき** $P(\triangle\mathrm{IC} < 0) \to 1$**．**

ICの差 ケース3

$I(g; f(\boldsymbol{\theta}^*_\beta)) - I(g; f(\boldsymbol{\theta}^*_\alpha)) = 0$**かつ** $f(\cdot|\boldsymbol{\theta}^*_\alpha) \neq f(\cdot|\boldsymbol{\theta}^*_\beta)$**のとき**

$$\triangle\mathrm{IC} \approx 2 \times \left(l(\boldsymbol{\theta}^*_\alpha) - l(\boldsymbol{\theta}^*_\beta)\right) - c_n \times \left(\dim M_\alpha - \dim M_\beta\right)$$

でケース1とほぼ同じなのだが，

$$l^*(\boldsymbol{\theta}^*_\alpha) - l^*(\boldsymbol{\theta}^*_\beta) = 0$$

なので

$$l(\boldsymbol{\theta}^*_\alpha) - l(\boldsymbol{\theta}^*_\beta) \sim N\left(0, nJ_{\alpha,\beta}\right)$$

もし

$$d = \dim M_\alpha - \dim M_\beta > 0$$

ならば M_β**の方が良いモデル．もし** $c_n/n^{\frac{1}{2}} \to \infty$**ならば** M_β**を選択する確率は**

$$P(\triangle\mathrm{IC} < 0) \to 1$$

モデル選択が一致性を持つための条件

モデル選択規準として

$$\mathrm{IC}_\alpha = -2 \times l(\hat{\boldsymbol{\theta}}_\alpha) + c_n \dim M_\alpha$$

を使うことにする．

ケース1

$$c_n/n \to 0$$

ケース2

$$c_n \to \infty$$

ケース3

$$c_n/n^{\frac{1}{2}} \to \infty$$

AIC（$c_n = 2$**）はケース1で一致性がある**

BIC（$c_n = \log n$**）はケース1と2で一致性がある**

1.3 モデル選択の信頼性

モデル選択の一致性の議論は実用上あまり意味が無い

- **一般的な状況(ケース1)ではどのモデル選択規準も一致性をもつ**
- **有限のnではど規準を使ってもモデル選択のバラツキがある**

モデル選択のバラツキを定量的に評価して，選択の信頼性を表す数値で示す

$P_1, P_2, \ldots, P_K$; **各モデルのp-value，p-値**

モデルをひとつだけ選ぶのではなく，同程度に良いモデルをすべて選び出す

モデル信頼集合

確率値 (p-value)

各モデルM_k，$k = 1, \ldots, K$に$[0, 1]$の範囲の数値

$$P_k$$

を与え，それが1に近いほど良いモデルである可能性を表す

いろいろな考え方がある

- **モデルが選択される確率の推定 (ブート スト ラップ選択確率)**
- **モデル信頼集合 (それに対応する検定の確率値 p-value)**

リサンプリング

データ

$$\boldsymbol{x} = \{x_1, \ldots, x_n\}$$

選択されたモデル

$$\widehat{k}(\boldsymbol{x})$$

データからのリサンプリング (各 $\tilde{x}_i$ は x からランダムに選ぶ)

$$\tilde{\boldsymbol{x}} = \{\tilde{x}_1, \ldots, \tilde{x}_n\}$$

選択されるモデル

$$\widehat{k}(\tilde{\boldsymbol{x}})$$

ブート スト ラップ選択確率

リサンプリングを多数繰り返す (例えば$B = 1000$)

$$\tilde{\boldsymbol{x}}[1], \tilde{\boldsymbol{x}}[2], \ldots, \tilde{\boldsymbol{x}}[B]$$

選択されるモデル

$$\widehat{k}(\tilde{\boldsymbol{x}}[1]), \widehat{k}(\tilde{\boldsymbol{x}}[2]), \ldots, \widehat{k}(\tilde{\boldsymbol{x}}[B])$$

モデルM_kのブート スト ラップ選択確率

$$P_k = \frac{\#\left\{\widehat{k}(\tilde{\boldsymbol{x}}[b]) = k,\, b = 1, \ldots, B\right\}}{B}, \quad k = 1, \ldots, K$$

$$\sum_{k=1}^{K} P_k = 1,\ 0 \leq P_k \leq 1$$

すなわちM_kが選択された頻度．P_kが大きいモデルほど良いモデルと考えられる．

モデル選択と仮説検定

- パラメトリック・モデル

$$M_\alpha = f(\cdot|\Theta_\alpha), \quad \alpha \in \mathcal{M}$$

$$M_1, M_2, \ldots, M_K$$

- モデル選択 $\hat{k}$では，どのM_kが真の分布に近いかを推定する．

  相対的にどのモデルが良いか？

- 仮説検定では，どのM_kが真の分布を含むかを調べる．

  どのモデルが正しいか？

モデル選択の検定

- モデルM_kが$M_1, \ldots, M_K$の中で一番良いという仮説を

$$H_k : k^* = k$$

  であらわす．($l_n^*(k) = l_n^*(k^*)$となるタイのkすべてでH_kがいえるとする．)

- モデル選択の検定

$$H_1, H_2, \ldots, H_K$$

  のうち，どの仮説が正しいか？

- モデル信頼集合

$$\mathcal{T} = \left\{ k \in \{1, \ldots, K\} \,|\, H_k\text{が有意水準}\, P^*\text{で棄却されない} \right\}$$

モデル選択，仮説検定，モデル選択の検定

これらは互いに関連しているが，同じものではない．

- モデル選択 ── どのモデルが（相対的に）良いかの推定
- 仮説検定 ── どのモデルが正しいかの検定
- モデル選択の検定 ── どのモデルが（相対的に）良いかの検定

モデルを比較する目的はなにか

良いモデルを選択して予測精度をあげる？ ── モデル選択よりモデル混合

モデル選択を通して対象の理解を深める？ ── 解釈可能なモデルを利用

AICの差の有意性

二つのモデルM_αとM_β

$$\Delta\mathrm{AIC} = 2 \times \left(l(\hat{\boldsymbol{\theta}}_\alpha) - l(\hat{\boldsymbol{\theta}}_\beta) \right) - 2 \times \left(\dim M_\alpha - \dim M_\beta \right)$$

$\Delta\mathrm{AIC} > 0$ならM_αを選択，$\Delta\mathrm{AIC} < 0$ならM_βを選択

AICの差の有意性を検定する ── 統計 (Linhart 1988)，計量経済 (Vuong 1989)，分子生物 (Kishino-Hasegawa 1989) などによって提案

$l_n^*(\alpha) = l_n^*(\beta)$かつ$I(f(\boldsymbol{\theta}_\alpha^*); f(\boldsymbol{\theta}_\beta^*)) > 0$のとき，十分大きな$n$に対して

$$\frac{\Delta\mathrm{AIC}}{\sqrt{\hat{V}\{\Delta\mathrm{AIC}\}}} \sim N(0,1)$$

ネストの場合はあまり良い近似ではないが，ノンネストの場合に有効．

分散の推定は

$$\hat{V}\{\Delta\mathrm{AIC}\} = 4n\hat{J}_{\alpha,\beta} + v_{\alpha,\beta}$$

モデル選択の検定（二つのモデル）

$$\mathrm{AIC}_\beta - \mathrm{AIC}_\alpha > 0$$

のとき M_α が選択されるが，

$$\frac{\mathrm{AIC}_\beta - \mathrm{AIC}_\alpha}{\sqrt{\widehat{V}\{\mathrm{AIC}_\beta - \mathrm{AIC}_\alpha\}}} < c$$

ならば M_α と M_β の良さの差は有意ではないと判断される．ただし c は有意水準 P^* より

$$\Phi(c) = 1 - P^*$$

によって定める．

モデル選択の検定（$K \geq 3$）── 選択バイアスを無視した場合

K 個のモデルを AIC の小さい順に並べて，

$$M_1, M_2, \ldots, M_K$$

$$\mathrm{AIC}_1 \leq \mathrm{AIC}_2 \leq \cdots \leq \mathrm{AIC}_K$$

このうち，

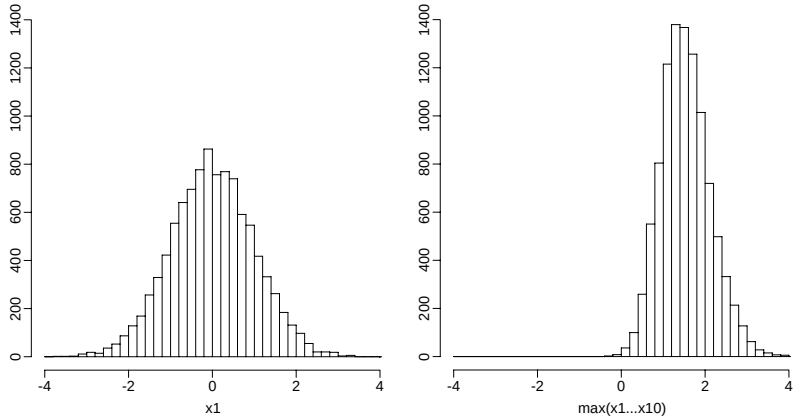
$$\frac{\mathrm{AIC}_k - \mathrm{AIC}_1}{\sqrt{\widehat{V}\{\mathrm{AIC}_k - \mathrm{AIC}_1\}}} < c$$

を満たす M_k は，M_1 とのモデルの良さの差が有意でないと考えられる．

実際には，$\mathrm{AIC}_k - \mathrm{AIC}_1$ は選択バイアスによって大きくなる傾向がある．

$$\mathrm{AIC}_k - \mathrm{AIC}_1 = \max_{k'=1,\ldots,K} (\mathrm{AIC}_k - \mathrm{AIC}_{k'})$$

選択バイアス



x_1 $\qquad$ $\max(x_1, \ldots, x_{10})$

$(x_1, \ldots, x_{10}) \sim N(0, I)$ $\qquad$ 10000 **samples**

多重比較による p-値の計算

$$\mathrm{aic} = (\mathrm{AIC}_1, \mathrm{AIC}_2, \ldots, \mathrm{AIC}_K) \sim N(E\{\mathrm{aic}\}, V\{\mathrm{aic}\})$$

と近似的にみなし，$E\{\mathrm{AIC}_k\}$ を最小にする k^* の信頼集合を「多重比較」で構成する．

$$\delta_k(\mathrm{aic}) = \max_{k'=1,\ldots,k-1,k+1,\ldots,K} \frac{\mathrm{AIC}_k - \mathrm{AIC}_{k'}}{\sqrt{\widehat{V}\{\mathrm{AIC}_k - \mathrm{AIC}_{k'}\}}}$$

(1) リサンプリングを B 回行い，各モデルの AIC の複製を計算する．

$$\mathrm{aic}(\tilde{\boldsymbol{x}}[b]), \quad b = 1, \ldots, B$$

(2) AIC の平均 $\widehat{E}\{\mathrm{aic}\}$ を求め，それを引いたうえで δ_k の複製を計算する．

$$\tilde{\delta}_k[b] = \delta_k\left(\mathrm{aic}(\tilde{\boldsymbol{x}}[b]) - \widehat{E}\{\mathrm{aic}\}\right), b = 1, \ldots, B$$

(3) 各モデル M_k 毎に，H_k の検定に関する p-値を計算する．

$$P_k = \frac{\#\left\{\tilde{\delta}_k[b] > \delta_k(\mathrm{aic}), b = 1, \ldots, B\right\}}{B}, \quad k = 1, 2, \ldots, K$$

H_kの検定

あらかじめ与えた有意水準P^*に対して，

$P_k < P^*$ならH_kを棄却する

$P_k \geq P^*$ならH_kを棄却しない

ただし，　$H_k : \delta_k(E\{\mathrm{aic}\}) \leq 0$，もしくは

$$H_k : E(\mathrm{AIC}_1) \geq E(\mathrm{AIC}_k), \ldots, E(\mathrm{AIC}_K) \geq E(\mathrm{AIC}_k)$$

である．もしH_kが真ならば，H_kが誤って棄却される確率は

$$P\left(P_k < P^*\right) \leq P^*$$

等号は

$$E(\mathrm{AIC}_1) = E(\mathrm{AIC}_2) = \cdots = E(\mathrm{AIC}_K)$$

のとき．

モデル信頼集合の構成

モデル信頼集合は

$$\mathcal{T} = \{k \in \{1, 2, \ldots, K\} \mid P_k \geq P^*\}$$

これがk^*を含む確率（被覆確率）は

$$P\left(k^* \in \mathcal{T}\right) \geq 1 - P^*$$

(証明)

$$k^* \in \mathcal{T} \Leftrightarrow P_k \geq P^* \Leftrightarrow H_k\text{が棄却されない}$$

$\mathcal{T}$に含まれるモデルは，$\hat{k}$より有意に悪いとは言えない $\mathcal{T}$に含まれないモデルは，$\hat{k}$より有意に悪いとは言える

下平（1993)，Shimodaira（1998)など

確率分布のベクトル表現

モデル$f(\cdot\|\boldsymbol{\theta})$を$n$次元ベクトル $$\boldsymbol{\xi}(\boldsymbol{\theta}) = (\log f(x_1\|\boldsymbol{\theta}), \ldots, \log f(x_n\|\boldsymbol{\theta}))$$ で表す．

各モデルM_kは$\dim M_k$次元の集合

$$\{\boldsymbol{\xi}(\boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta_k\}$$

最尤モデルは

$$\boldsymbol{\xi}(\hat{\boldsymbol{\theta}}_k)$$

で表される．

実際の描画では主成分分析(PCA)等により，低次元に射影する

2　情報量規準の拡張

真の分布: $X_1, X_2, \ldots, X_n \sim g(\cdot)$ モデル: $X_1, X_2, \ldots, X_n \sim f(\cdot\|\boldsymbol{\theta})$ データ: $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ 予測分布: $f(\cdot\|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))$

$f(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))$が平均的にどれだけ$g(\cdot)$に近いか？

$$E\left\{I(g(\cdot); f(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{X})))\right\}$$

の推定量としての情報量規準

データから予測分布を与える方式(手続き)には，さまざまな形式が考えられる

$$f(\cdot|\boldsymbol{x})$$

の平均的な良さ

$$E\left\{I(g(\cdot); f(\cdot|\boldsymbol{X}))\right\}$$

の推定量は，情報量規準の拡張になる．

色々な情報量規準

- **パラメタ推定による予測分布の場合**
  AIC，TIC，GIC，EIC，RIC (罰金付き最尤法) など
- **ベイズ予測分布の場合**
  曲率との関係
- **不完全データの場合**
  EMアルゴリズムとの関係，PDIO
- **説明変数の分布が変化する場合**
  実験計画，重み付き最尤法
- **ベイズ的方法**

2.1　パラメタ推定による予測分布の場合

分布 $g(\cdot)$ **の空間から** Θ **への汎関数**

$$T(g(\cdot))$$

を考える．特に，モデルが真の分布を含む場合は，

$$T(f(\cdot|\boldsymbol{\theta}^*)) = \boldsymbol{\theta}^*$$

とする．パラメタの推定量として

$$T(\hat{g}(\cdot))$$

を用いる．

$$\mathrm{GIC} = -l(T(\hat{g})) + E\left\{\frac{\partial \log f(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big|_{\theta^*} T^{(1)}(X|g)\right\}$$

解析的に解く代わりに，ブート ストラップを使って数値的に計算する方法がある (EIC)．

M-推定量

$$\sum_{i=1}^{n} \psi(x_i|\hat{\boldsymbol{\theta}}) = 0$$

$$T^{(1)}(x|g) = M(\psi|g)^{-1}\psi(x|\boldsymbol{\theta}^*); \quad M(\psi|g) = -E\left\{\frac{\partial \psi(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big|_{\theta^*}\right\}$$

$$GIC = -l(\hat{\boldsymbol{\theta}}) + \mathrm{tr}\left(E\left\{\psi(X|\boldsymbol{\theta}^*)\frac{\partial \log p(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big|_{\theta^*}\right\} \cdot M(\psi|g)^{-1}\right)$$

罰金付き最尤法

$$l(\boldsymbol{\theta}) - \lambda k(\boldsymbol{\theta})$$

を最大にする推定量で，λ**をどうやって選ぶか?**

$$\psi(x|\boldsymbol{\theta}) = \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \lambda\frac{\partial k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

2.2　ベイズ予測分布の場合

パラメタの事前分布: $f(\boldsymbol{\theta})$
パラメタの事後分布: $f(\boldsymbol{\theta}|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})$

$$f(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\boldsymbol{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

ベイズ予測分布 $f(\cdot|x)$ **は**

$$f(z|\boldsymbol{x}) = \int f(z|\boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{x})\,d\boldsymbol{\theta}$$

$$\mathrm{IC} = -\sum_{i=1}^{n} \log f(x_i|\boldsymbol{x}) + \mathrm{tr}(GH^{-1})$$

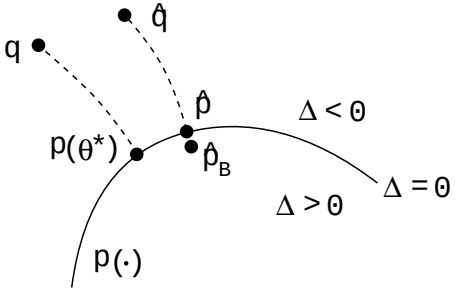
[参考] ベイズ予測分布と最尤推定の関係

ベイズ予測分布のICは書き直せて，

$$\mathrm{IC} \approx -\sum_{i=1}^{n} \log f(x_i|\hat{\boldsymbol{\theta}}) + \frac{1}{2}\left(\mathrm{tr}(GH^{-1}) + \dim \boldsymbol{\theta}\right)$$

一方，最尤推定のICは

$$\mathrm{IC} = -\sum_{i=1}^{n} \log f(x_i|\hat{\boldsymbol{\theta}}) + \mathrm{tr}(GH^{-1})$$



したがってこれらの差は$\Delta/2$，ただし

$$\Delta = \mathrm{tr}(GH^{-1}) - \dim \boldsymbol{\theta}$$

2.3 不完全データの場合

完全データ$x = (y, z)$のうち yしか観測出来ず，zが観測できない確率変数

完全データのモデル: $f(x|\boldsymbol{\theta})$
観測データのモデル: $f(y|\boldsymbol{\theta})$

$$f(y|\boldsymbol{\theta}) = \int f(y, z|\boldsymbol{\theta})\, dz$$

最尤推定$\hat{\theta}$は次の対数尤度から求める（例えばEMアルゴリズム等を使って）

$$l_Y(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta})$$

AICは$f(y|\hat{\theta})$の$g(y)$からの平均的な隔たりを測る

$$\mathrm{AIC} = -l_Y(\hat{\boldsymbol{\theta}}) + \dim \boldsymbol{\theta}$$

ただし

$$g(y) = \int g(y, z)\, dz$$

完全データに関するモデルの良さを測る規準

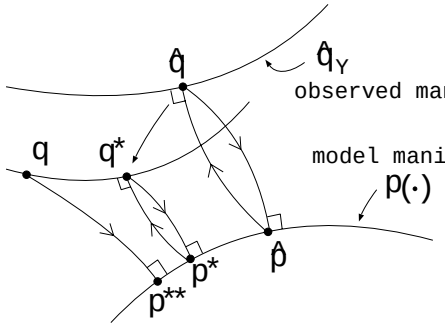
モデル$f(x|\hat{\theta})$の真の分布$g(x)$からの平均的な隔たりを測るには

$$\mathrm{IC} = -l_Y(\hat{\boldsymbol{\theta}}) + \mathrm{tr}\left(I_X I_Y^{-1}\right)$$

が使える．

$$I_X(\boldsymbol{\theta}) = -\int f(x|\boldsymbol{\theta}) \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\, dx, \quad I_Y(\boldsymbol{\theta}) = -\int f(y|\boldsymbol{\theta}) \frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\, dy$$

はそれぞれxとyに関するθのFisher情報行列．



情報量の差

$$I_{Z|Y} = I_X - I_Y$$

とおくと，

$$\mathrm{tr}(I_X I_Y^{-1}) = \dim \theta + \mathrm{tr}(I_{Z|Y} I_Y^{-1})$$

2.4 説明変数の分布が変化する場合

説明変数zの分布がデータを観測した時と予測分布の良さを評価する時で異なる場合

回帰モデル: $f(y|z, \boldsymbol{\theta})$
目的変数: y
説明変数: z

$z \sim g_0(z)$ **データ**, $\quad z \sim g_1(z)$ **評価**

重み付き対数尤度

$$l_w(\boldsymbol{\theta}) = \sum_{i=1}^{n} w(z_i) \log f(y_t|z_t, \boldsymbol{\theta})$$

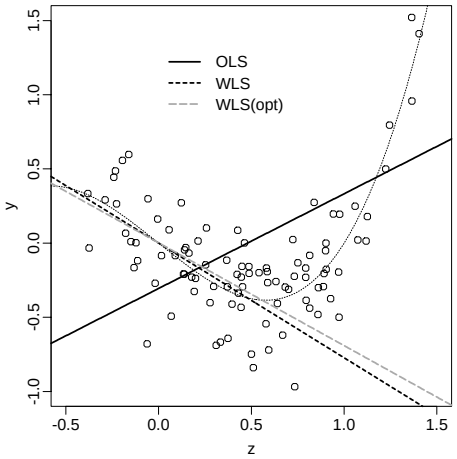
これを最大にする重み付き最尤推定を$\hat{\theta}_w$とかく．予測分布

$$f(y|z, \hat{\boldsymbol{\theta}}_w) g_1(z)$$

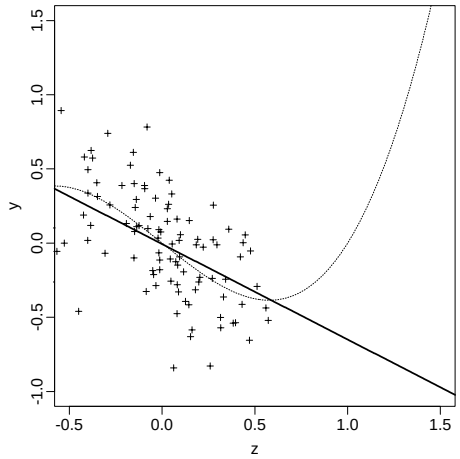
が$g(y|z)g_1(z)$に近いかを評価する．

数値例: 多項式回帰モデル

$$y = -z + z^3 + \epsilon, \quad \epsilon \sim N(0, 0.3^2)$$



$z \sim N(0.5, 0.5^2)$
$\lambda = 0, 0.77, 1$



$z \sim N(0, 0.3^2)$
$\lambda = 0$

重み付き最尤推定の良さを測る規準

データが $(y, z) \sim g(y|z)g_0(z)$ **のとき，**
$f(y|z, \hat{\theta}_w)g_1(z)$ **が** $g(y|z)g_1(z)$ **にどれだけ近いかを測るには**

$$\mathrm{IC}_w = -\sum_{i=1}^{n} \frac{g_1(z_i)}{g_0(z_i)} \log f(y_i|z_i, \hat{\boldsymbol{\theta}}_w) + \mathrm{tr}(J_w H_w^{-1})$$

が使える．ただし

$$\begin{aligned} \hat{J}_w &= \frac{1}{n}\sum_{i=1}^{n} \frac{g_1(z_i)}{g_0(z_i)} w(z_i) \frac{\partial \log f(y_i|z_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\hat{\theta}_w} \frac{\partial \log f(y_i|z_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big|_{\hat{\theta}_w} \\ \hat{H}_w &= -\frac{1}{n}\sum_{i=1}^{n} w(z_i) \frac{\partial^2 \log f(y_i|z_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big|_{\hat{\theta}_w} \end{aligned}$$

実際には重み関数として

$$w(z) = \left(\frac{g_1(z)}{g_0(z)}\right)^{\lambda}, \quad \lambda \in [0, 1]$$

2.5 ベイズ的方法

$$x_1, x_2, \ldots, x_n, \ldots \sim f(x|\boldsymbol{\theta})$$

事前分布 $f(\theta)$ **を考えると**

$$(x_1, \ldots, x_n) \sim f(x_1, \ldots, x_n) = \int f(x_1|\boldsymbol{\theta}) \cdots f(x_n|\boldsymbol{\theta}) f(\boldsymbol{\theta})\, d\theta$$

十分大きな n **に対して**

$$\mathrm{BIC} = -\sum_{i=1}^{n} \log f(x_i|\hat{\boldsymbol{\theta}}) + \frac{\log n}{2} \dim \boldsymbol{\theta}$$

は $-\log f(x_1, \ldots, x_n)$ **の近似になる．(あまり良い近似ではないが．)**

BICを小さくするモデルは，近似的に事後確率を大きくする

AICとBICの関係

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_n|x_1, \ldots, x_{n-1}) f(x_{n-1}|x_1, \ldots, x_{n-2}) \cdots \\ &\quad \cdots f(x_3|x_1, x_2) f(x_2|x_1) f(x_1) \end{aligned}$$

より

$$-\log f(x_1, \ldots, x_n) = -\sum_{i=1}^{n} \log f(x_i|x_1, \ldots, x_{i-1})$$

各時刻でのベイズ予測分布の良さはICで測れて，各項に

$$\frac{1}{2i} \dim \boldsymbol{\theta}$$

の補正項が出て来る．これを $i = 1, \ldots, n$ **まで足し合わせて**

$$\sum_{i=1}^{n} \frac{1}{2i} \dim \boldsymbol{\theta} \approx \frac{\log n}{2} \dim \boldsymbol{\theta}$$

というBICの第2項が出て来る．これに対して，AICは

$$x_1, \ldots, x_n$$

から与える予測分布 $f(x_{n+1}|\hat{\theta}(x_1, \ldots, x_n))$ **の良さを測っている．**