統計数理II

Computer Intensive Methods in Statistics
統計科学における計算機集約技法

ohp20031002 下平英寿 shimo@is.titech.ac.jp

講義情報

- ホームページ
  http://www.is.titech.ac.jp/~shimo/class/
- 担当：下平
- 評価方法： レポート提出
- レポート 課題（1〜2回）
- 質問受け付け： まず質問内容のメールを出すこと．面談が必要な場合はあらかじめメールにてアポイントを取ること．もしくは講義時に直接質問する．

学習内容

- 「統計科学」 データから有用な情報を取り出すための数学的方法論
- “Computer Intensive Methods”（計算機集約技法）
- 数理的側面と現実の応用

- 新たな応用問題に統計科学の手法を適用する能力
- 新たな統計科学の手法を発展させる基礎力

学習計画

1. 確率モデルによる情報処理
   最尤法

2. 情報量規準によるモデル選択
   エントロピー，予測分布，AIC

3. ブート スト ラップ法による信頼性評価
   データのバラツキ，不偏な検定と確率値

4. 回帰分析やバイオインフォマティクスへの応用
   住宅価格データ，哺乳類のミトコンドリアＤＮＡデータ

回帰分析

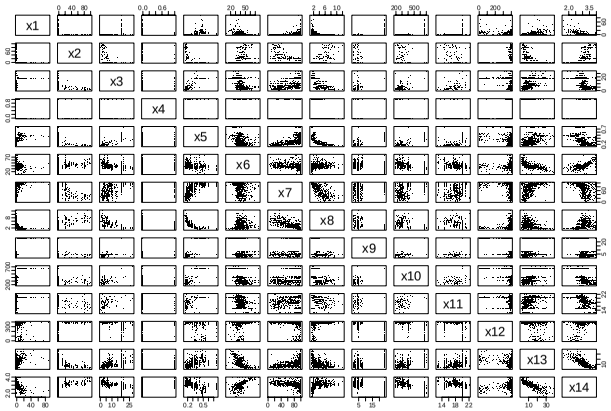
住宅価格データ
ボストンの５０６地域の住宅価格とその共変量

t	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
9	0.21	12.5	7.9	0	0.27	31.7	100.0	6.1	5	311	15.2	387	29.9	2.80
72	0.16	0.0	10.8	0	0.17	35.5	17.5	5.3	4	305	19.2	377	9.9	3.08
89	0.06	0.0	3.4	0	0.24	49.1	86.3	3.4	2	270	17.8	397	5.5	3.16
130	0.88	0.0	21.9	0	0.39	31.8	94.7	2.0	4	437	21.2	397	18.3	2.66
164	1.52	0.0	19.6	1	0.37	70.1	93.9	2.2	5	403	14.7	388	3.3	3.91
269	0.54	20.0	4.0	0	0.33	55.8	52.6	2.9	5	264	13.0	390	3.2	3.77
343	0.02	0.0	1.9	0	0.27	42.8	59.7	6.3	1	422	15.9	390	8.7	2.80
373	8.27	0.0	18.1	1	0.45	34.5	89.6	1.1	24	666	20.2	348	8.9	3.91
451	6.72	0.0	18.1	0	0.51	45.5	92.6	2.3	24	666	20.2	0	17.4	2.60
477	4.87	0.0	18.1	0	0.38	42.0	93.6	2.3	24	666	20.2	396	18.7	2.82

変数の説明：x_1 ＝犯罪率，x_2 ＝宅地割合，x_3 ＝非商用地割合，x_4 ＝チャールス川沿いか，x_5 ＝窒素酸化物濃度の二乗，x_6 ＝平均部屋数の二乗，x_7 ＝1940年より古い住宅の割合，x_8 ＝ビジネス街への距離，x_9 ＝ハイウェイへのアクセス，x_{10} ＝固定資産税，x_{11} ＝生徒と教師の比率，x_{12} ＝黒人の比率を Bk とした$1000(\mathrm{Bk}-0.63)^2$，x_{13} ＝社会的地位の低い者の割合，x_{14} ＝持ち家価格の中央値の対数．

- 出典: D. Harrison and D. L. Rubinfeld (1978) “Hedonic Housing Prices and the Demand for Clean Air,” *Journal of Environmental Economics and Management*, 5, 81-102.
- 入力ミスの修正済みデータ “boston_corrected”
  StatLib Datasets Archive
  http://lib.stat.cmu.edu/datasets/ より入手可能
- ここではデータセットの全サンプルサイズ $n = 506$のうちランダムに選んだ10地点を表示した．
- いくつかの変数に二乗や対数変換を施してある．

散布図 (scatter plot)



正規線形回帰モデル

$$x_{14} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{13} x_{13} + \epsilon$$

- 目的変数： x_{14}
- 説明変数： $x_1, \ldots, x_{13}$
- 回帰係数： $\beta_0, \beta_1, \ldots, \beta_{13}$
- 誤差項： $\epsilon \sim N(0, \sigma^2)$

$$p(\epsilon; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\epsilon-\mu)^2}{2\sigma^2}\right)$$

回帰係数の推定

i	$\hat{\beta}_i$	標準誤差	t統計量	確率値
0	9.937	0.352	28.2	0.000
1	-0.214	0.027	-8.0	0.000
2	0.070	0.031	2.3	0.023
3	0.048	0.040	1.2	0.232
4	0.063	0.021	3.0	0.003
5	-0.194	0.038	-5.1	0.000
6	0.182	0.028	6.5	0.000
7	-0.003	0.035	-0.1	0.930
8	-0.237	0.040	-6.0	0.000
9	0.282	0.055	5.1	0.000
10	-0.260	0.060	-4.3	0.000
11	-0.191	0.027	-7.0	0.000
12	0.092	0.023	4.0	0.000
13	-0.496	0.034	-14.4	0.000

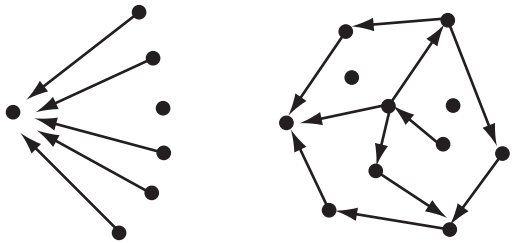
比較を容易にするため，あらかじめ各$x_1, \ldots, x_{14}$をその標準偏差で割ってから回帰分析を行なった．残差の分散は$\hat{\sigma}^2 = 0.20$．

変数選択

順位	k	$\lvert S_k \rvert$	AIC_k	$\Delta\mathrm{AIC}_k$	$\tilde{\alpha}_k$	$\hat{\alpha}_k$	$\tilde{d}_k$	$\tilde{c}_k$	S_kのパタン
1	8124	11	646.7	0.0	0.35	0.72	0.15	0.73	**-***-******
2	8128	12	647.2	0.5	0.22	0.87	−0.37	0.78	******-******
3	8188	12	648.7	2.0	0.11	0.31	0.82	0.33	**-**********
4	8192	13	649.2	2.5	0.08	0.33	0.61	0.19	**********
5	8122	10	649.6	2.9	0.06	0.19	1.57	0.71	*--***-******
6	8126	11	650.7	4.0	0.02	0.14	1.67	0.59	*-****-******
7	8186	11	651.4	4.7	0.03	0.10	1.75	0.46	*--**********
8	8190	12	652.5	5.8	0.01	0.06	1.85	0.32	*-**********
9	8120	11	654.6	7.9	0.02	0.14	1.67	0.60	***-**-******
10	8116	10	654.9	8.2	0.02	0.10	1.96	0.66	**--**-******
17	6076	10	660.2	13.6	0.02	0.15	1.79	0.75	**-***-*****-*
18	6080	11	661.1	14.4	0.01	0.14	1.76	0.69	*******-*****-*
19	6140	11	662.2	15.5	0.01	0.07	2.07	0.56	**-********-*

住宅価格データの変数選択．順位はAICの昇順．$\tilde{\alpha}_k$はブート ストラップ確率($n' = n$)．$\hat{\alpha}_k$はマルチスケール・ブート ストラップ法で計算した近似的に不偏な確率値．この表には8192個のモデルのうち$\hat{\alpha}_k \geq 0.05$となる13個だけが示されている．$\tilde{d}_k$と$\tilde{c}_k$はそれぞれ仮説の符号付距離と曲率を表し，$\hat{\alpha}_k$の計算に用いた．

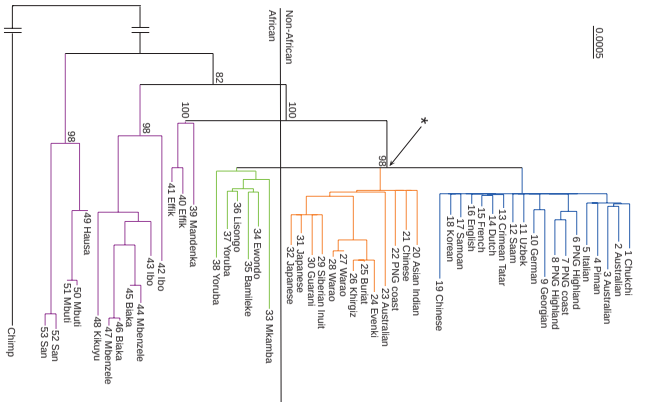
グラフィカルモデル



変数間の関係をグラフで表現

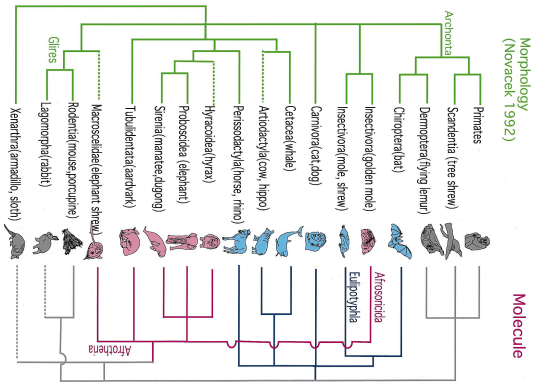
進化系統樹の推定

人類の起源はアフリカ?



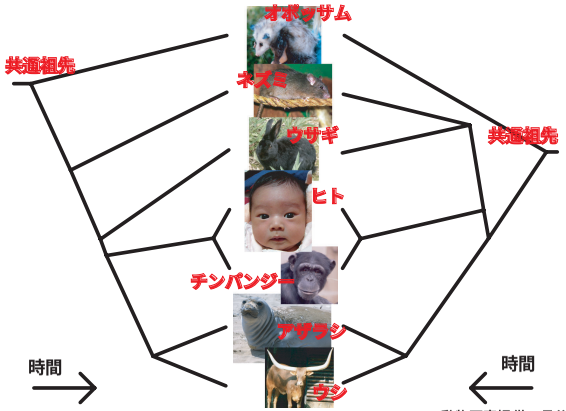
Ingman et al. (2000) Nature, 408.

形態学 vs 分子系統樹



Cao et al. (2000) Gene 259, 149-158.

哺乳類の進化系統樹?



動物写真提供：長谷川政美氏

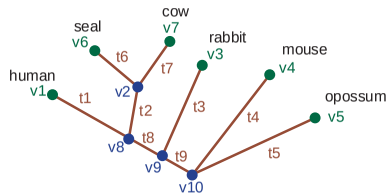
mtDNA配列データ

アミノ酸配列

```
                 2         3         4         5         6         7         8         9
        0123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
human   ERKILGYMQLRKGPNVVGPYGLLQPFADAMKLPTKEPLKPATSTITLYITAPTLALTIALLLWAPLPMPNPLVNLNLGLL
seal    ERKVLGYMQLRKGPNIVGPYGLLQPIADAVKLPTKEPLRPLTSSTTMPIMAPILALALALTMWVPLPMPYPLINMNLGVL
cow     ERKVLGYMQLRKGPNVVGPYGLLQPIADAIKLPIKEPLRPATSSASMPILAPIMALGLALTMWIPLPMPYPLINMNLGVL
rabbit  ERKILGYMQLRKGPNIVGPYGLLQPIADAIKLPTKEPLRPLTSSPLLFIIAPTLALTLALSMWLPIPMPYPLVNLNMGIL
mouse   ERKILGYMQLRKGPNIVGPYGILQPFADAMKLPMKEPMRPLTTSMSLFIIAPTLSLTLALSLWVPLPMPHPLINLNLGIL
opossum ERKVLGYMQLPRKGPNVIGPYGILQPFADALKLPIKEPLRPMTSSISMPTIAPTLALTLAFTIWTPLPMPNALLDLNLGLL
```

シーケンス長 $n = 3414$

系統樹の確率モデル



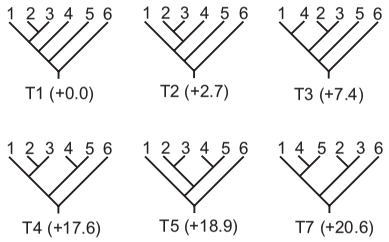
$\theta = (t1, t2, t3, t4, t5, t6, t7, t8, t9, Q, \pi)$

サイト hにおける配列パタンx_hの確率

$$P(x_h; T, \theta) = \sum_{v2,v8,v9,v10} P(v1|v8;t1)P(v2|v8;t2)P(v3|v9;t3) \times P(v4|v10;t4)P(v5|v10;t5)P(v6|v2;t6) \times P(v7|v2;t7)P(v8|v9;t8)P(v9|v10;t9)P(v10)$$

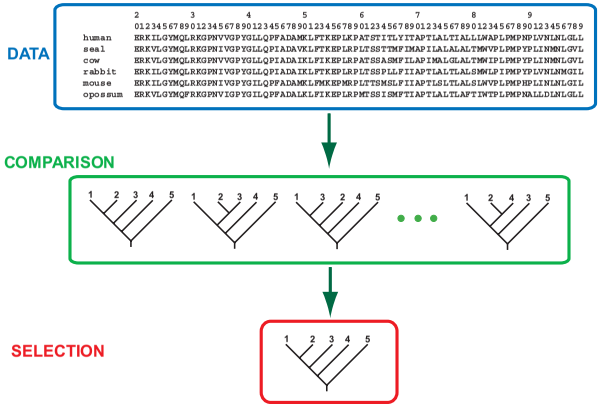
各枝でのマルコフ過程(つまり進化)は独立という仮定

系統樹の対数尤度差



ヒト＝1，アザラシ＝2，ウシ＝3，
ウサギ＝4，マウス＝5，オポッサム＝6

系統樹推定



モデル選択

確率モデルによる情報処理

- **確率モデル（確率密度関数または確率関数）**

$$p(X;\theta)$$

$X =$ **データ**，　$\theta =$ **パラメタ**

- **モデルの候補**

$$p_1(X;\theta_1),\ p_2(X;\theta_2),\ \ldots\ ,p_K(X;\theta_K)$$

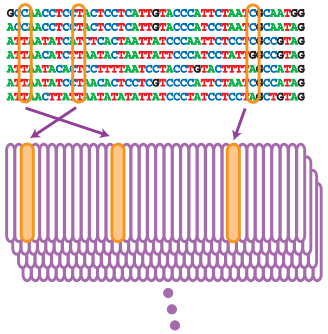
- **モデル選択**

$$p_{\hat{k}}(X;\theta_{\hat{k}})$$

- **信頼性評価**

$\hat{k}$**で本当に良いのか？　データのバラツキは？**

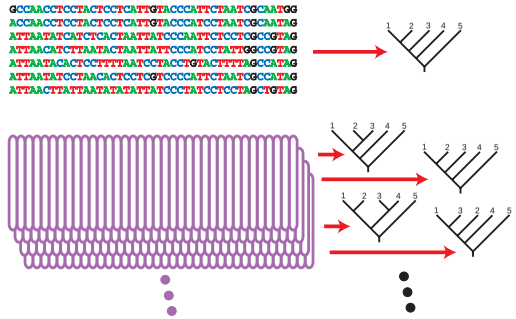
ブートストラップ法



有限長の配列データからのリサンプリング

Efron (1979)

ブートストラップ確率



膨大な計算量 ── 並列計算による高速化