水産資源データ解析における NUOPTおよびVMStudioの利用

庄野　宏
(独)水産総合研究センター
遠洋水産研究所)

1.はじめに 1).発表の構成

1. はじめに (資源評価手法・問題の背景など)
2. 水産資源データ解析におけるNUOPTの利用 (体長組成の年齢分解(成分数推定)・構造モデルとフルモデルの融合・生物情報の利用)
3. 水産資源評価におけるVMStudioの利用例 (まぐろ類の漁獲効率(CPUE)に関する解析)
4. まとめ (NUOPTやVMStudio等に関する感想)

A. 付録 (有限正規混合分布の理論的考察(続))
B. 補遺 (サポートベクターマシーンのコンセプト)

1.はじめに 2).まぐろ・かつお類の水産資源解析

a. 遠洋水産研究所 (独)水研センター) の役割
- 日本の領海外の漁業資源の水産資源解析
  →国際漁業委員会などで一括して管理する

b. 水産資源解析の2つの目的
- 資源評価 (海の中にいる魚の数を推定する)
  – プロダクションモデルやコホート解析などを使用
- 資源管理 (評価結果に基づき漁獲量を決定)
  – 再生産関係 (親子関係) や海洋環境などを考慮

1.はじめに 3).資源評価手法および取り扱う問題

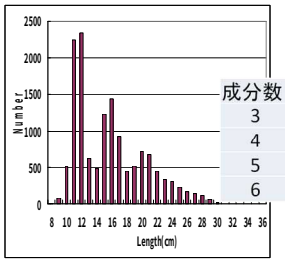
- 魚の絶対量を推定するための資源評価手法
  – 微分方程式をベースにした水産資源動態モデル
  – コホート解析等に基づく年別年齢別漁獲量モデル

1. Catch-at-Size(CAS)->Catch-at-age(CAA) →第2章
2. CPUE (catch per unit effort:漁獲効率) の観測値とモデルからの予測値のフィッティング →第3章

$$LL = -\frac{1}{2}\sum_{a,y}\{log\left(\frac{1}{2}\pi\sigma^2\right) + \frac{\left[\log\left(CPUE_{a,y}\right) - \log(q_a N_{a,y})\right]}{2\sigma^2}\}$$

2.水産資源解析におけるNUOPT利用 1).体長組成データの年齢分解問題

a.キダイ体長組成(田中1956)



b. 年齢分解問題の条件設定

・有限混合正規分布を仮定して成分数およびパラメーター(平均・分散・混合比率)を推定

成分数	MLL	BIC	Bayes	AIC
3	75479.6	75556	75486.6	75495.6
4	75367.6	75472.6	75379.6	75389.6
5	75351.6	75485.2	75371.6	75379.6
6	75342	75504.4	75372.2	75376

・Catch-at-size→Catch-at-age 変換にて成分数推定が重要

2.水産資源解析におけるNUOPT利用 2).過去の知見および生物情報の利用

a. 過去の知見 (～1990年代)
- 成分数は目で見て判断！
- EMアルゴリズムの利用??? (成分数を固定後、各々の正規分布の平均・分散等の推定に際して効果を発揮!)
- 成分数推定に際しては、カイ二乗検定が利用可能???
- 成分数推定に際しては、情報量規準は利用不可?????

水産資源分野の奇妙な知見

b. 生物的な補助情報の利用
- 耳石の年輪・日輪等の活用
- 背鰭や尻鰭等の年齢形質

正規分布の平均に対して上記VB成長曲線を仮定して母数 (K, t0, L(∞)) を推定→構

$$\mu_t = L(\infty)[1-\exp\{-K(t-t_0)\}]$$

$$L(\theta) = \prod_{i=1}^{n}\left[\sum_{j=1}^{m}\pi_j N(X_i|\mu_j,\sigma_j^2)\right]$$

2.水産資源解析におけるNUOPT利用
3).成分数推定にχ2検定は使用不可

例：$X_1,\cdots,X_n$ (i.i.d.)～

$\alpha\ \mathrm{N}(x|\mu_1,\sigma_1^2) + (1\text{-}\alpha)\mathrm{N}(x|\mu_2,\sigma_2^2)\ (0\le\alpha\le1)$

$H_0:\ \alpha=1$ v.s. $H_1:\ 0\le\alpha<1 \Rightarrow -2\log\lambda_n \to \chi^2_{3\text{-}2}(\text{as } n\to\infty)$?

対数尤度の過剰な振る舞いゆえに不成立！

- 認定可能性の欠如(非識別性)
- 帰無仮説の下でFisher情報行列が特異
- 対数尤度が∞になる場合も存在(非有界性)

2.水産資源解析におけるNUOPT利用
4).罰金付き最尤推定法(Leroux,1992)

$l(\theta|X)$ (対数尤度関数) $=\sum_{i=1}^{n}\log\left[\sum_{j=1}^{m}\pi_j N(x_i|\mu_j,\sigma_j^2)\right]$

$L:=-2\,l(\theta|X)+2a_{m,n}$ とおき、成分数mの推定量$\hat{m}$を

最小化問題 $\hat{m}=\min_m\{\min_\theta L\}$ の解として求める

但し $a_{m,n}$ は $a_{m,n}>0,\ a_{m+1,n}>a_{m,n},\ \frac{a_{m,n}}{n}\to0\ (n\to0)$

を満たす実数列 とする (注)θはmに依存している)

2.水産資源解析におけるNUOPT利用
5).情報量規準による成分数の推定

罰金付き最尤法(Leroux,1992)により求められた
成分数(m) は
漸近的に過小推定でない → 一致性とは異なる

$a_{m,n}=3m-1$ → AIC

$a_{m,n}=\frac{(3m-1)\log n}{2}$ → BIC

正規分布以外 $\sum_{j=1}^{m}\pi_j h(x,\omega_j)(p.d.f.)$ でも適用可能!

2.水産資源解析におけるNUOPT利用
6).フルモデル(制約無しモデル)の特徴

- 長所
  - 柔軟なパラメーター推定が可能
- 短所
  - モデルの正則性(有界性,識別性含む)の問題
  - 情報量規準による成分数推定の根拠の問題
    - Leroux (*Annals of Statistics*, 1992) の[学位]論文
      "罰則付き尤度関数(AICやBICを含む)による成分数の推定は、漸近的に過小推定にはならない"
      ⇒ 過大推定の可能性は言及なし→片手落ちなのでは?

2.水産資源解析におけるNUOPT利用
7).構造モデル(制約付モデル)の特徴

- 長所
  - モデルの安定性
    - パラメーターの次元が縮約されているため、識別性が保たれ、(対数) 尤度関数がパラメーター空間で有界
- 短所
  - 成分数の推定が難しい
  - モデルの柔軟性 (特定化に失敗する恐れも)
    e.g. VB曲線よりもゴンペルツ曲線が適当な場合 etc.
    ⇒成分数を固定すれば情報量規準により選択可能

2.水産資源解析におけるNUOPT利用
8).新モデル(Eguchi and Yoshioka,2001)

コンセプト：フルモデルと構造モデルの融合

- 長所
  - "対数尤度の過剰な振る舞い"の問題解決
    →罰則付き尤度関数(フルモデルの尤度と構造モデルの尤度の結合形)の有界性を数学的に証明！
  - 成分数(山の数)が推定可能
  - 構造(VB型の成長曲線など)が推定可能
- 短所
  - 計算手順や背景となる理論が非常に複雑

2.水産資源解析におけるNUOPT利用 9).新しいモデルの計算手順概略-(1)

Step1: 構造モデルH: $\theta=\theta(\xi)$ の下で最尤法により
パラメーターの推定値 $\theta(\hat{\xi})$ を求める

Step2: 罰則付き尤度(構造及びフルモデルの融合)

$$l_\lambda(\theta)=(1-\lambda)l(\theta)-\lambda D^{(J)}(\theta(\hat{\xi}),\theta)$$

を最大にするパラメーターθの値を求める

Step3: クロスバリデーション(or近似手法)により
チューニングパラメーターλの値を求める

$l(\cdot)$: フルモデルの対数尤度関数
$D^{(J)}(\cdot,\cdot)$: 構造モデルとフルモデルのK-L情報量

2.水産資源解析におけるNUOPT利用 10).新しいモデルの計算手順概略-(2)

Step4: 最後に下式により成分数mの推定を行なう

$$\hat{m}=\min_m CV(\lambda,m) \quad \text{or} \quad \min_m ACV(\lambda,m)$$

・考えうる成分数を持つ複数のモデルを最初に仮定

- クロス・バリデーション（全データから第k番目の観測値をx(k)(k=1,...,n)を抜いての計算実行）はデータ数nや成分数mが大きくなると実行不可能

↓

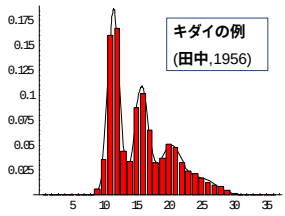
近似的手法(ACV)やグリッドサーチの利用

2.水産資源解析におけるNUOPT利用 11).モデルの仮定とNUOPTの適用事例

- 構造モデルの仮定
  - 平均:VB成長曲線
  - 分散:平均のべき乗
  - 比率:減少率一定

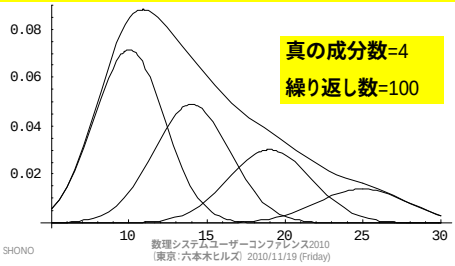
$$\mu(t\,|\,\inf,K,\tau)=e^{\inf}(1-\tau e^{-Kt})$$
$$(=L(\infty)[1-\exp\{-K(t-t_0)\}])$$
$$\sigma^2(t\,|\,c,p)=c\{\mu(t)\}^p$$
$$Ratio(t\,|\,a)=e^{-at}/\sum_{k=i}^{i+l}e^{-it}$$

推定値 L(∞)=48.7, K=0.13, t0= -1.08, c=0.62, p=0.24, a=0.55, **成分数は4** (r1=0.41, r2=0.29, r3=0.19, r4=0.11)



2.水産資源解析におけるNUOPT利用 12).新しいモデルを含む計算機実験

$$X_1,\cdots,X_n(i.i.d)\sim p.d.f.\mathrm{f(x)}=0.4\mathrm{N(x|10,5)}+0.3\mathrm{N(x|14,6)}+0.2\mathrm{N(x|19,7)}+0.1\mathrm{N(x|25,8)}$$



真の成分数=4
繰り返し数=100

2.水産資源解析におけるNUOPT利用 13).計算機実験による結果の比較

- 新しいモデル（サンプルサイズ=250）

成分数	2	3	4	5	6
ACV		2	97	1	

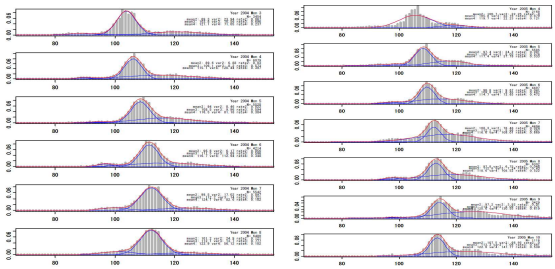
- フルモデル（サンプルサイズ=3000）

成分数	2	3	4	5	6
AIC			**90**	9	1
BIC		11	**89**		
Bayes*		1	**94**	5	

*-Bayes型規準では混合比率に対するDilichlet事前分布を仮定

2.水産資源解析におけるNUOPT利用 14).豪州畜養ミナミマグロの体長組成

a.2004年3-8月のage2-4の組成 **b.2005年4-10月のage2-4の組成**



2.水産資源解析におけるNUOPT利用
15).文献（有限正規混合分布モデル）

- Chernoff, H. 1954: *Ann. Math. Statist.*, **25**, 573-578.
- Eguchi, S. and Yoshioka, K. 2001: *The Institute of Mathematical Statistics.*, **809**, 30pp.
- Henna, J. 1985: *Ann. Inst. Math. Statist.*, **37**, 235-240.
- Leroux, B. 1992: *Ann. Statist.*, **20**(3), 1350-1360.
- Shapiro, A. 1985: *Biometrika*, **72**, 133-144.
- Shapiro, A. 1988; *Inter. Statist. Rev.*, **56**, 49-62.
- **庄野宏**. 2006: **計量生物学**, **27**(1), 55-67.
- **田中昌一**. 1956: **東海区水産研究所報告**, **14**, 1-4.

3.資源評価におけるVMStudioの利用
1).CPUE(catch per unit effort)の定義

- CPUEとは?→漁獲効率を表す一般的な指標!

$$CPUE=\frac{\text{Catch(漁獲量：漁獲尾数や漁獲重量で表わす)}}{\text{Effort(努力量：延縄船 針数，巻網船 操業日数等)}}$$

- 人間の側から見ると→漁獲効率・漁獲能率
- 生物の側から見ると→資源密度（ミクロ的）・相対資源量（マクロ的）（年トレンドの把握）
- CPUE標準化とは?→資源の年変動に対応する部分を取り出す作業（加工していないCPUEから資源密度以外の様々な要因を取り除く）
- 標準化CPUEは資源評価モデルに利用される

3.資源評価におけるVMStudioの利用
2).統計モデルを利用したCPUE解析

- LogCPUE-Normalモデル(共分散分析)

$\log(CPUE_{ijk})=Intercept+Year_i+Area_j+Month_k+(Year*Area)_{ij}+error_{ijk}$, $error_{ijk}\sim N(0,\sigma^2)$

Year, Area, Month：順序のないカテゴリカル変数

- Catch-Poisson(or NegativeBinomial)モデル(GLM)

$E[Catch_{ijk}]=Effort*\exp(Intercpet+Year_i+Area_j+Month_k+Year*Area_{ij})$, $Catch_{ijk}\sim Po(\lambda)$ or $NB(a,b)$

- CPUE年トレンド抽出法（要因分析法）

$$CPUE_y=\exp\left\{IC+YEAR_y+\overline{AREA}+(\overline{YEAR*AREA})_{y\cdot}\cdots\right\}$$

3.資源評価におけるVMStudioの利用
3).マイニング手法による解析の動機

- 統計的学習理論(データマイニング)への関心
- 線形モデル(統計モデル)から非線形モデルへ
- 教師付き学習(応答変数が存在)でのイメージ
  - 回帰問題(neural network>サポートベクター回帰)
  - 判別問題(サポートベクターマシン>neural network)
- 実データにおけるvalidationを通じた性能比較
- 予測値に基づく簡便な要因分析法(CPUE年傾向抽出法)の提案->ブラックボックスからの脱却

3.資源評価におけるVMStudioの利用
4).Case Study:計算手順およびデータ

1. 出力変数と入力変数を以下のように設定
   - 出力変数：SBT-CPUE(catch at age4+/1000hooks)
   - 入力変数：年(1969-2007),月(4-9), 統計海区(4-9:経度の情報含む), 緯度(30°S-50°S: 5°刻み)
2. データを学習用(80%)と検証用(20%)に分割
3. 学習用データで3手法によるルールを作成
   - サポートベクター回帰
   - ニューラルネットワーク（誤差逆伝搬法:教師付）
   - 樹形モデル（CART algorithm, If...thenで2分割）
4. 上記ルールにより検証用データで性能評価

3.資源評価におけるVMStudioの利用
5).Case Study: 計算に関する条件設定

サポートベクター回帰	ニューラルネットワーク	樹形モデル
カーネルの種類 (Gaussian Kernel)	Active Function (Sigmoid function)	Algorithm (CART)
上記パラメータ (Sigma=1.0)	Objective Function (SS: sum of square)	分岐方法 (Info Gain)
Slack変数 (C=1.0)	# of hidden layer (5)	複雑度係数 (1)
誤差項の大きさ (Error=1.0)	# of repeat=1000 Decay=0	節点の不純度 (0.01)

3.資源評価におけるVMStudioの利用
6).Case Study:モデルの性能評価-(1)

- 観測値と(それに対応する)予測値との差異
  - MAE= Σ |CPUEpre(i)-CPUEobs(i)| /N (絶対誤差)
  - MSE= Σ (CPUEpre(i)-CPUEobs(i))^2/N (平方誤差)

SBT	Suppot vector machine		Neural network		Tree regression model	
data	MSE	MAE	MSE	MAE	MSE	MAE
Training	103	8.07	67.5	5.81	373	12.7
Testing	175*	9.48	188	9.27*	339	12.1
Total	118	8.35	91.7	6.50	366	12.6

注) *印(黄色塗りつぶし)のマイニング手法の性能が一番良い(値が小さいほど性能が良い)

3.資源評価におけるVMStudioの利用
7).Case Study:モデルの性能評価-(2)

- 観測値と予測値の差異をPearson相関係数と順位相関係数(Kendoll-τ, Spearman-ρ)で判断

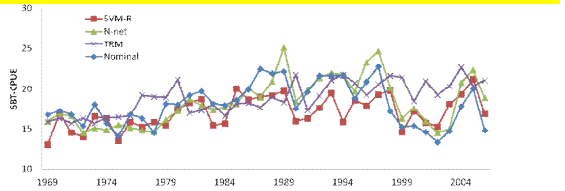
SBT	Support vector machine			Neural network			Tree regression model		
data	Pearson	Kendoll τ	Spearman ρ	Pearson	Kendoll τ	Spearman ρ	Pearson	Kendoll τ	Spearman ρ
Training	0.910	0.498	0.676	0.937	0.564	0.724	0.565	0.362	0.490
Testing	0.829*	0.474*	0.653*	0.821	0.470	0.637	0.61	0.374	0.516
Total	0.894	0.494	0.672	0.913	0.544	0.706	0.574	0.364	0.494

注) *印(黄色塗りつぶし)のマイニング手法の性能が一番良い(値が大きいほど性能が良い)

3.資源評価におけるVMStudioの利用
8).予測値ベースの年トレンド抽出法

- 要因分析が非常に難しい→ブラックボックス
- 簡便な手法を提案(→質的説明変数に着目)

$$CPUE_{year}^{pre} = \left(\frac{1}{N_m}\right)\sum_{month=1}^{N_m}\left(\frac{1}{N_a}\right)\sum_{area=1}^{N_a}\left(\frac{1}{N_l}\right)\sum_{latitude=1}^{N_l} CPUE_{year,month,area,latitude}^{pre}$$



4.まとめ
1).NUOPTに関する個人的な印象

- 大規模問題に有利な内点法を使用しているためか、初期値依存度が低く、性能が良い！
  - c.f. 準ニュートン法(収束は早いが初期依存性大)
  - c.f. シンプレックス法(収束が遅くて時間がかかる)
- 数式に似た記述法(文法)が非常に分かり易く、制約条件や解の制限等の追加記述も容易！
- パラメータの区間推定機能を装備して欲しい
  (Hessian, Profile-Likelihood, Bootstrap etc.)

4.まとめ
2).VMStudioに関する個人的な印象

- 装備しているデータマイニング手法が幅広く、各々の手法に関する機能やオプションも多彩
  - e.g. SVMやSVM-Rに関する多彩なカーネル関数
  - e.g. Naïve-Bayes・Boosting等最先端機能を装備
  - e.g. 複数手法の結果をまとめるモデル統合機能
- GUIがユーザーフレンドリーで操作性が高い
- NUOPTの最適化アルゴリズムおよびTrial and Errorによる停止規則などを取り入れてほしい

4.まとめ
3).統計ソフトに関する個人的な印象

Software	Advantage	Disadvantage
S-Plus/R	カスタマイズが得意・最先端の機能を装備	インタプリタゆえに、実行速度が遅い*
SAS	大規模な定型的処理	ライセンス料が高額
SPSS	ユーザーフレンドリーなGUI機能を有する	CUIが使いにくくモデル選択などが難しい

S-Plus	・Validation/Rに無い機能も(多変量GARCH等)
R	・最先端の機能が多く存在するがバグも多い

END

Thank you for your kind attention!

【謝辞:数理システム株式会社の皆様方】
特に田辺様、佐藤様、橋本様、徐様、中園様、田澤様、小木様、福田様、友廣様

A.付録 (有限混合分布の理論的背景)
1).カイ二乗バー分布(Shapiro,1985)

特別な場合に $-2\log\lambda_n$ (λ_n:尤度比検定統計量)の漸近分布がカイ二乗分布の重み付き平均の分布(カイ二乗バー分布)になる

例：$X_1,\cdots,X_n$ (i.i.d.)$\sim p.d.f.\sum_{j=1}^{m}\pi_j g_j(x)$

$0\le\pi_j\le 1,\sum_{j=1}^{m}\pi_j=1$ 分布が既知の場合

成分数3つ $\Rightarrow -2\log\lambda\to\frac{1}{6}\chi_0^2+\frac{1}{2}\chi_1^2+\frac{1}{3}\chi_2^2$

A.付録 (有限混合分布の理論的背景)
2). 罰金付き最小距離法(Chen,1993)

mの推定量 $\hat{m}=\min_m[\min_\theta\{F_n(x),F(x|\theta))-c_n\sum_{j=1}^{m}\log\pi_j\}]$

$d(F_n(x),F(x|\theta^*))=o_p(c_n),\quad c_n=o(1),c_n>0,c_n$:実数列

$F(x|\theta)$:分布関数., $F_n(x)$:経験分布関数 (例:$c_n=1/n$)

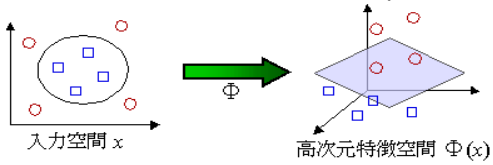
c.f. Minimum Distance (Henna,1985)

$$\hat{m}=\min_m\left[m\in\mathbb{N}\mid\min_\theta\left\{\int\left(F(x\mid\theta)-F_n(x)\right)^2dF_n(x)\right\}<\frac{\lambda^2(n)}{n}\right]$$

$$\lambda(n)\to\infty,\frac{\lambda^2(n)}{n}\to 0(n\to\infty),\sum_{n=1}^{\infty}\frac{\lambda^2(n)}{n}\exp\{-2\lambda^2(n)\}<\infty$$

B.補遺
1).サポートベクターマシンのイメージ

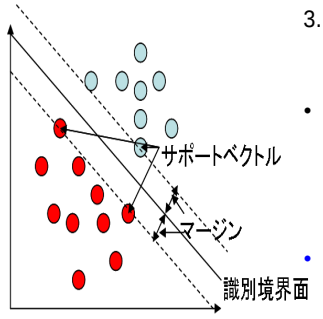
サポートベクターマシンのコンセプト(計算手順)



- 超平面 (1本の直線) で分離したい→How?

1. カーネル関数 e.g. $K(x,y)=\exp(-|x-y|^2/\sigma^2)$ etc.
2. 高次元空間へ移す i.e. 地球から月に移動

B.補遺
2).サポートベクターマシンのイメージ(続)



3. マージン最大化を用いて一意に分離可能 (道路の幅を広くする)
- 高次元空間での操作を元の空間からカーネル関数を通じて行う (月での操作を地球上からリモートで行う)
- カーネルトリック (検証データに対する予測性能が非常に良い)