海洋生態学のための関数データ解析

江口　真透　　数理推論研究系

紹介： 生態学の主な目的の一つはある地域の生態システムの健全さの評価を行うことにある．その地域と対象となる生物たちに着目して定期的に観測が継続された結果，関数 $\{\boldsymbol{X}(t): t\in T\}$ が得られる．その生態系が健全であるかないかで 2値の確率変数 Y が 0 または 1 を取るとしよう．このようにして．n 組のデータ $(\{\boldsymbol{x}_i(t): t\in T_i\}, y_i)\quad i=1,\cdots, n$ が得られる．このデータ解析のために関数一般化線形モデルなどについて考察したい．特に線形判別解析とロジスティック回帰モデルと混合正規モデルの関係を関数データの文脈で拡張を試みたい．水産資源の内容でおおよそ200系群を超える約55年間のデータが蓄えられている公開データベース‘**RAM Legacy Stock Assessment**’の全域解析を試みる．

関数ロジスティックモデル： 一般化線形モデルの内容で線形予測量が

$$\eta = \alpha + \int_T \beta(t)X(t)\,dt$$

と表されるとし，リンク関数 g と分散関数 σ^2 によって次のように表されると仮定する：

$$\mathbb{E}(Y\,|\,X(t),t\in T) = g(\eta),\quad \mathrm{Var}(\mathrm{Y}\,|\,X(t),t\in T) = \sigma^2(\eta)$$

さらに基底関数 $\{\phi_j(t): j=1,...,p\}$ に対してパラメトリック形：$\beta(t)=\beta_1\phi_1(t)+\cdots+\beta_J\phi_J(t)=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\varphi}(t)$ を仮定する．
ロジステック回帰モデルは

$$\mathbb{E}(Y=y\,|\,X(t),t\in T) \;=\; \frac{\exp\{y(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\int_T\boldsymbol{\varphi}(t)X(t)dt)\}}{1+\exp(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\int_T\boldsymbol{\varphi}(t)X(t)dt)}$$

となる．ここでは共変量が単一の関数の場合を扱っているが，J 本の関数 $\{\boldsymbol{X}_j(t): j=1,...,J,\ t\in T\}$ の場合や通常の共変量ベクトル $\boldsymbol{x}$ が追加されても線形予測量は同様に記述される：

$$\eta = \alpha + \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{\beta}_1^{\mathrm{T}}\int_T\boldsymbol{\varphi}_1(t)X_1(t)\,dt+\cdots+\boldsymbol{\beta}_J^{\mathrm{T}}\int_T\boldsymbol{\varphi}_J(t)X_J(t)\,dt$$

関数線形判別分析： 線形判別分析の内容でが $Y=y$ が与えられた $\{X(t)\}_{t\in T}$ の条件付分布を正規分布

$$\{X(t)\}_{t\in T}\,|\,Y=y \;\sim\; N(\{\boldsymbol{\mu}_y^{\mathrm{T}}\boldsymbol{\varphi}(t)\}_{t\in T},\Omega+\sigma^2\mathrm{I})$$

を仮定すると線形判別関数は

$$\eta = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\int_T\boldsymbol{\varphi}(t)X(t)\,dt$$

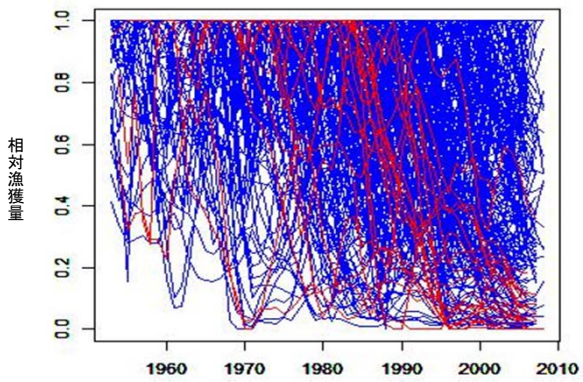
関数クラスタリング： モデルによるクラスタリングは正規混合モデル

$$\{X(t)\}_{t\in T} \sim \pi_1 N(\{\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\varphi}(t)\}_{t\in T},\Omega+\sigma^2\mathrm{I})+\pi_0 N(\{\boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\varphi}(t)\}_{t\in T},\Omega+\sigma^2\mathrm{I})$$

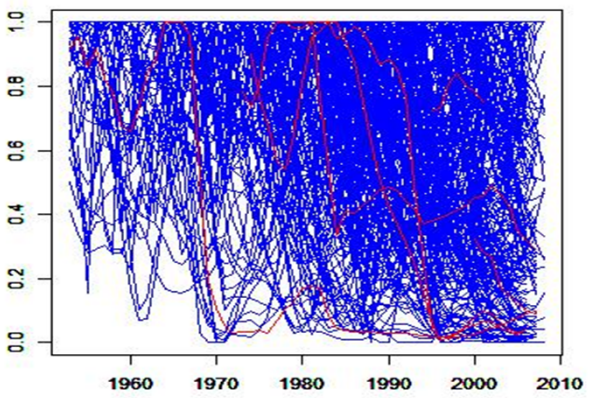
を仮定し，事後分布の推定によって実行される．

3つの方法の関連： 上記の3つのモデルの尤度解析は対数尤度関数の意味で線形判別解析はロジスティック回帰と正規混合モデルに分解される．正確には　$L_{\mathrm{J}}(\boldsymbol{\theta}) = L_{\mathrm{C}}(\boldsymbol{\beta})+L_{\mathrm{M}}(\boldsymbol{\theta})$ の関係で特徴付けられる．

RAM Legacy Stock Assessment： **RAMの232ストックの1953-2008間の相対漁獲データ**　（t **年の相対漁獲量**　$R(t)=\dfrac{X(t)}{\max\{X(s): s\le t\}}$　）



真のラベル
（赤：枯渇　青：非枯渇）



FGLMの予測ラベル, AUC = 0.81