

# 最大エントロピーと最小ダイバージェンスの双対性

江口 真透 数理推論研究系

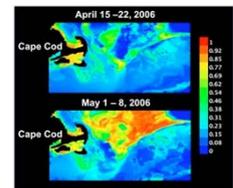
- 概要**
- 生態学、計算言語学において**最大エントロピー法**が広く使われている。特徴変数から統計モデルと同時に統計推定が自動的に得られるのでデータ解析が容易に行うことができる。
  - パッケージMaxEnt (Phillip *et al.* 2006) を使えば、ある生物種の在データと、その調査エリアのGIS データ、気象データなどのリンクからの特徴変数データが構成され、Lassoタイプの変数選択による**生息マップ**の描画が簡単に得られる。
  - 統計学の見地からは、この最大エントロピー法は統計モデルは**指数型分布族**、推定は**最尤法**に限られているのでモデルの誤特定とデータの混濁に対してロバストでない点が挙げられる。
  - この問題を解決する方法は幾つか考えられるが、ここでは**エントロピーの一般化**による最大化を提案したい。このアプローチは生態学では生物多様性を測る別のエントロピーが考えられていることから支持される。
  - 一般化最大エントロピーモデルに対して**最小ダイバージェンス法**を適用すると自然な振る舞いが保持されることを示し、0和ゲームのミニマックス性によって最大エントロピーと最小ダイバージェンスの**双対性**について明らかにする。

## 最大エントロピー法とは

ある生物種の調査エリア  $A$  の生息分布を推定したい。そのために在データ  $\{x_1, \dots, x_n\}$  と同時にGISや気象から特徴変数ベクトル  $\{t(x_1), \dots, t(x_n)\}$  が得られたとする。このとき、統計量  $\bar{t} = n^{-1} \sum t(x_i)$  に対して  $t(X)$  の平均がちょうど  $\bar{t}$  となる分布の集合を考える。この平均の制約の下でエントロピー最大分布は次のようになる。

$$\hat{f} := \arg \max_{f \in \mathcal{F}(\bar{t})} \left\{ - \sum_{x \in A} f(x) \log f(x) \right\} \Rightarrow \hat{f}(x) = \exp \{ \hat{\theta}^T t(x) - \kappa(\hat{\theta}) \},$$

ここで  $\mathcal{F}(\bar{t}) = \{ f : \mathbb{E}_f \{ t(X) \} = \bar{t} \}$ ,  $\kappa(\theta) = \log \sum \exp \{ \theta^T t(x) \}$ ,  $\hat{\theta}$  は制約  $\mathbb{E}_{\hat{f}} \{ t(X) \} = \bar{t}$  から決まるこれより生息分布  $\hat{f}(x)$  による生息マップが得られる。



大西洋セミクジラの生息分布  
<http://www.seascapemodelling.org/>

## 一般化エントロピー

上記の(1)で扱われたエントロピーはBoltzmann-Shannonのものであるが、生態学においては生物多様性の観点からSimpson や Hillのエントロピーが広く用いられている。このようにより広いクラスにエントロピーを拡張しよう。

単調増加で凸な関数  $U(s)$  に対して共役な凸関数  $U^*(t) := \max_{s \in \mathbb{R}} \{ ts - U(s) \}$  を使って、 $U$ -エントロピーを

$$H_U(f) = - \sum U^*(f) = - \sum \{ f \xi(f) - U(\xi(f)) \}, \quad \text{ここで } \xi(t) = (U')^{-1}(t).$$

と定義する。  $M_U = \{ U'(\theta^T t(x)) - \kappa(\theta) : \theta \in \Theta \}$ ,  $\Theta = \{ \theta : \kappa(\theta) < \infty \}$  として、 $M_U$  を  $U$ -モデルと呼ぶ。このとき、

$$\hat{f}_U := \arg \max_{f \in \mathcal{F}(\bar{t})} H_U(f) \Rightarrow \hat{f}_U(x) = U'(\hat{\theta}_U^T t(x) - \kappa(\hat{\theta}_U)), \quad (\kappa(\theta) \text{ は正規化項 } \hat{\theta}_U \text{ は平均制約 } \mathbb{E}_{\hat{f}} \{ t(X) \} = \bar{t} \text{ から決まる})$$

実際、  $f \in \mathcal{F}(\bar{t}) \Rightarrow H_U(\hat{f}_U) = C_U(f, \hat{f}_U) \iff H_U(\hat{f}_U) - H_U(f) = D_U(f, \hat{f}_U) \geq 0$  である。

## 最小ダイバージェンス

$U$ -クロスエントロピーと  $U$ -ダイバージェンスを次のように定める。

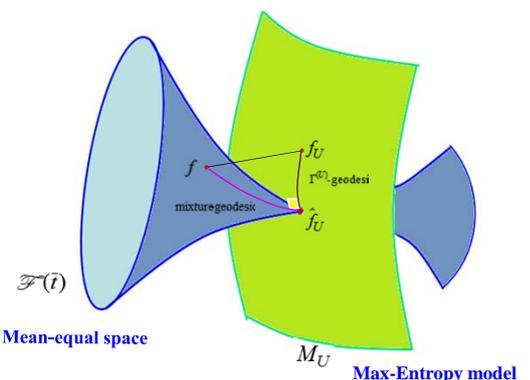
$$C_U(f, g) = \sum \{ f \xi(g) - U(\xi(g)) \}, \quad D_U(f, g) = C_U(f, g) - H_U(f)$$

このとき、最小  $U$ -ダイバージェンスと0和ゲームのミニマックス性:

$$\hat{\theta}_U = \operatorname{argmin}_{\theta \in \Theta} \{ -\theta^T \bar{t} + \kappa(\theta) + \sum U(\theta^T t + \kappa(\theta)) \}$$

$$\max_{f \in \mathcal{F}(\bar{t})} \min_{g \in \mathcal{F}^*} C_U(f, g) = H_U(\hat{f}_U) = \min_{g \in \mathcal{F}^*} \max_{f \in \mathcal{F}(\bar{t})} C_U(f, g)$$

が成立する。



## これからやりたいこと

- 標準的な最大エントロピー法の欠点を一般化最大エントロピー法によって**改善**したい。このためには生成関数  $U$  をデータから有効に選択する方法の開発が必要となる。一つのアイデアとしては幾つかの  $U$  によって得られた生息分布の性能をテストデータによる検証で選択する方法が考えられる。
- 拡張されたエントロピー最大分布は標準の場合の**ギブス分布**とは異なる。この分布の統計的、生態学的な適切な解釈を得たい。同時に在のみデータの**ポアソン過程**としての理解も拡大したい。