数理解析研究所講究録
第1560巻 2007年 109-120

Gradient modeling and information geometry

東京大学・情報理工学系研究科　清 智也 (Tomonari SEI)
School of Information Scienece and Technology,
The University of Tokyo

概 要

量的データのための統計的モデリングとして, 勾配モデリングを提案する. 勾配モデリングでは, 確率密度を勾配写像で表現する. 指数型モデルとの大きな違いは, 正規化定数が不要であることである. 勾配写像が一つ定まると, 対応する確率密度に従うサンプルの発生は容易に行なえる. この点も従来のモデルにはない大きな特徴である. また, g-平坦モデルを定義し, g-平坦モデルに対する最尤法が凸最適化問題であることを述べる. 情報幾何学的な側面として, g-平坦性と相対エントロピーから導かれる g-計量を導入し, g-計量に基づくスコア法を提案する.

キーワード 逆関数法, 高次相互作用, 勾配表現, g-平坦モデル, スコア法, 凸関数.

1 はじめに：3次の相互作用を持つ多変量分布

多変量の連続分布で最も扱いやすい分布は (多変量) 正規分布であろう. ところが, 正規分布は3次以上の相互作用を表現することができない. 例えば, 図1の (b) と (c) のように, 第3変数の正負によって第1, 第2変数の相関が変わるような分布を, 正規分布で表現することは不可能である. このような制約は, 量的データの解析において必ずしも現実的ではない (例えば宮川 1997, 7.3節).

ここでは3次の相互作用を測る基準として3次のキュムラントを用いる. 正規分布では3次以上のキュムラントは0になる. すなわち, m 変量正規分布の密度関数を

$$\phi(x|\mu,\Sigma) \;=\; (2\pi)^{-m/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right), \quad \mu\in\mathbb{R}^m, \quad \Sigma\in\mathbb{R}^{m\times m}$$

とおくと,

$$\log\left[\int \mathrm{e}^{\mathrm{i}t^\top x}\phi(x|\mu,\Sigma)\mathrm{d}x\right] \;=\; \mathrm{i}\mu^\top t-\frac{1}{2}t^\top\Sigma t$$

となり, t に関する3次以上の項は現れない.

3次の相互作用を持たせたければ, 例えば

$$p(x_1,x_2,x_3) \;=\; \exp(f(x_1,x_2,x_3)-C), \quad f(x_1,x_2,x_3) \;:=\; \frac{1}{2}\|x\|^2+(x_1x_2x_3\wedge 1)\vee(-1)$$

のような分布を考えればよいかもしれないが, 正規化定数 C が陽に求まらないという欠点がある.

そこで, 3次の相互作用を持つ多変量分布で, しかも密度関数が陽に書けるようなものを構成しよう. 天下り的だが, 次のような $\mathbb{R}^3$ 上の関数を考える.

$$\psi_\epsilon(x) \;=\; \frac{1}{2}\|x\|^2+\epsilon\sum_{\lambda=1}^{4}\arctan(e_\lambda^\top x), \quad (e_1,e_2,e_3,e_4)=\begin{pmatrix}1&1&-1&-1\\1&-1&1&-1\\1&-1&-1&1\end{pmatrix}.$$

パラメータ ϵ が十分 0 に近いとき ψ_ϵ は凸だから, その勾配写像 $\nabla\psi_\epsilon = (\partial\psi_\epsilon/\partial x_i)_{i=1}^3$ は全単射である. そこで, 方程式

$$\nabla\psi_\epsilon(X) \sim N(0, I_3)$$

によって確率変数 X の分布を定める. 図 1 は, X の散布図の例である. X の密度関数の形は, 変数変換の公式から

$$p_\epsilon(x) := \phi(\nabla\psi_\epsilon(x)) \det[\nabla\nabla^\top\psi_\epsilon(x)]$$

と陽に書ける (ϕ は標準正規分布の密度). また, $\epsilon \to 0$ の下で X の 3 次キュムラント κ_{ijk} を漸近展開すると,

$$\kappa_{123} = 0.1237\epsilon + O(\epsilon^2), \quad \kappa_{111} = \kappa_{112} = \cdots = O(\epsilon^2)$$

となり (Sei 2006), X は ϵ のオーダーの 3 次キュムラントを持っていることが分かる.

このように, 勾配写像を用いて新しい確率密度関数を作ると, これまであまり解析できなかった高次の相互作用を表現することができる. 2 節では, 勾配写像に基づいた統計的モデルの性質を簡単に述べる. 3 節では情報幾何との関連を述べ, また 4 節で最尤推定アルゴリズムについて考察する.

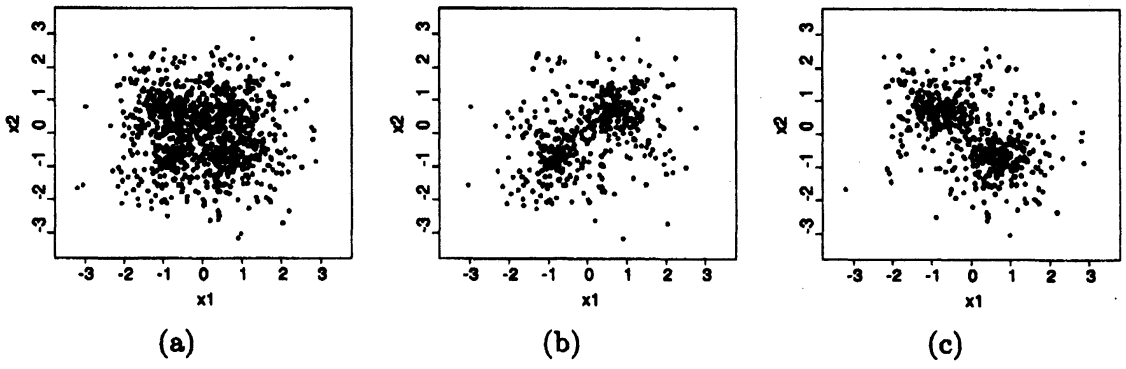

図 1: 散布図の例 ($\epsilon = 0.35$). (a) $p_\epsilon(x_1, x_2)$, (b) $p_\epsilon(x_1, x_2|x_3 > 0)$, (c) $p_\epsilon(x_1, x_2|x_3 < 0)$.

2 勾配表現と g-平坦モデル

この節では, 確率密度の勾配表現と g-平坦モデルを定義し, その性質を見る. 詳しくは Sei (2006) を参照のこと.

2.1 確率密度の勾配表現

まず, 任意の密度関数が勾配写像を用いて表現できることを述べた次の定理を紹介する. ここで $\nabla = (\partial/\partial x_1, \ldots, \partial/\partial x_m)^\top$ である.

定理 1 (Brenier 1991, McCann 1995). $\mathbb{R}^m$ 上のルベーグ測度に対する任意の密度関数 $p(x)$, $q(x)$ が与えられたとき, ある凸関数 $\psi(x)$ が存在して

$$p(x) \;=\; q(\nabla\psi(x))\det[\nabla\nabla^\top\psi(x)] \tag{1}$$

が成り立つ. 言い換えれば, 任意の連続分布を持つ確率変数 X と Y に対し, ある凸関数 $\psi(x)$ が存在して $\nabla\psi(X)$ と Y の分布は等しくなる. また, ψ は定数を除いて一意的に定まる. □

(1) 式は Monge-Ampère 方程式と呼ばれ, 偏微分方程式や輸送問題の観点からよく調べられている (詳しくは Villani 2003 を参照). 定理 1 を精密化した, 次の定理もある.

定理 2 (Villani 2003, Theorem 4.14). もし $p(x)$, $q(x)$ がいたるところ正の密度を持ち, 有界かつ $C^{0,\alpha}$ 級 (α はヘルダー指数を表す) ならば, 定理 1 の ψ は $C^{2,\alpha}$ 級となる. □

定理 1 において,

$$q(x) \;=\; \phi(x) \;:=\; (2\pi)^{-m/2}\exp(-\|x\|^2/2) \quad \text{(標準正規分布)}$$

と固定すると, p と $\nabla\psi$ は一対一に対応する. そこで, $\nabla\psi$ を p の**勾配表現**と呼ぶことにしよう. 統計的モデルとは, 確率密度関数の集合であった. 密度関数の集合を考える代わりに, その勾配表現の集合を考えてもよい. 本研究の目的は, 勾配写像の集合としての統計的モデルの性質を調べることである.

勾配表現のクラスとして, 本稿では

$$\mathcal{G}_{\mathrm{ALL}} \;:=\; \left\{\nabla\psi \,\middle|\, \psi : \text{convex},\ \psi\in C^2,\ \nabla\psi(\mathbb{R}^m)=\mathbb{R}^m,\ \nabla\nabla^\top\psi \succ 0\right\} \tag{2}$$

を用いる. 定理 2 より, $\mathcal{G}_{\mathrm{ALL}}$ は十分広いクラスである. 任意の $g\in\mathcal{G}_{\mathrm{ALL}}$ に対して, 勾配表現が g となるような密度関数は

$$p[g](x) \;:=\; \phi(g(x))\det[G(x)], \quad G(x) \;:=\; \nabla g(x)^\top$$

と陽に表される.

次の定理は, 勾配表現が与えられたとき, 対応する密度関数からの標本は容易に得られることを示したものである. これは乱数発生における**逆関数法**の一般化と見なすことができる.

定理 3 (逆関数法). Y を m 次元標準正規分布に従う確率変数とする. このとき, 次の最適化問題の解 X は $p[\nabla\psi](x)$ に従う:

$$X \;=\; \operatorname*{argmin}_{x\in\mathbb{R}^m}\left[\psi(x)-Y^\top x\right]$$

Proof. $Y=\nabla\psi(X)$ が成り立つので明らか. □

2.2 g-平坦モデル

まず, (2) で定義した $\mathcal{G}_{\mathrm{ALL}}$ が凸集合であることを示そう.

補題 1. $\mathcal{G}_{\mathrm{ALL}}$ は凸集合である.

Proof. 任意の $g_1, g_2 \in \mathcal{G}_{\mathrm{ALL}}$ と $c_1, c_2 > 0$ に対して $c_1 g_1 + c_2 g_2 \in \mathcal{G}_{\mathrm{ALL}}$ であることを示す. $\mathcal{G}_{\mathrm{ALL}}$ の定義より, ある凸関数 $\psi_1, \psi_2 \in C^2$ が存在して

$$g_i = \nabla\psi_i, \quad \nabla\nabla^\top\psi \succ 0, \quad \nabla\psi_i(\mathbb{R}^m) = \mathbb{R}^m \quad (i = 1, 2)$$

を満たす. $\psi = c_1\psi_1 + c_2\psi_2$ とおけば, $\psi \in C^2$ かつ $\nabla\nabla^\top\psi \succ 0$ であることは明らかなので, $\nabla\psi(\mathbb{R}^m) = \mathbb{R}^m$ を示せばよい. 勾配写像 $x \mapsto \nabla\psi(x)$ が全射であるための必要十分条件は ψ が次の条件 (super-linearlity) を満たすことである:

$$\forall x \in \mathbb{R}^m \setminus \{0\}, \quad \lim_{\lambda\to\infty} \frac{\psi(\lambda x)}{\lambda} = \infty$$

ψ_1, ψ_2 が super-linearlity を満たすので ψ も super-linearlity を満たす. よって示された. □

$\mathcal{G}_{\mathrm{ALL}}$ の有限次元部分凸集合を **g-平坦モデル**と呼ぶ. つまり, g-平坦モデルとは次の形を持つ統計的モデルのことである:

$$\mathcal{M} = \left\{ p[g](x) \,\middle|\, g = \sum_{i=1}^{d} \theta^i g_i,\ \theta \in \Theta \right\}, \quad g_i \in \mathcal{G}_{\mathrm{ALL}}.$$

ただし Θ は $\mathbb{R}^d$ の凸部分集合とする.

例 1 (正規分布). 多変量正規分布全体は, 次の形で与えられる g-平坦モデルである.

$$\mathcal{M} = \left\{ p[g](x) \mid g(x) = Ax + b,\ A \in \mathrm{Sym}_+(\mathbb{R}^m),\ b \in \mathbb{R}^m \right\}$$

実際, $p[Ax + b](x)$ は平均 $-A^{-1}b$, 分散 A^{-2} の正規分布に等しい. □

例 2 (3 次の相互作用). 1 節の p_ϵ は, ϵ をパラメータとする g-平坦モデルである. □

g-平坦モデルの特長として, 特に次の性質は著しい.

定理 4. g-平坦モデルの負の対数尤度は凸である.

Proof. g-平坦モデルを $g(x|\theta) = \theta^i g_i(x)$ と書くと, 直接計算により

$$\frac{\partial^2}{\partial\theta^i\partial\theta^j}(-\log p[g](x)) = g_i(x)^\top g_j(x) + \mathrm{tr}[G(x|\theta)^{-1}G_i(x)G(x|\theta)^{-1}G_j(x)]$$

となる. これは非負定値である. □

この結果から, g-平坦モデルの最尤推定量は比較的容易に求まることが分かる. その他, g-平坦モデルは, 独立性の記述や錐型モデルへの拡張においても良い性質を持っていることが知られている (Sei 2006).

3 勾配モデリングと情報幾何

情報幾何学 (Amari 1985) では, 統計的モデル (= 確率密度の集合) を多様体と見なす. すなわち, 1 つ 1 つの確率密度を「点」とみなし, 情報量の観点から距離や平行移動などを定義する. 本節では, 勾配表現と情報幾何学の関連性を考察する.

便宜上, 一般の統計モデルのパラメータを $u = (u^a)$, 何らかの意味で平坦なモデルのパラメータを $\theta = (\theta^i)$ と表すことにする. また, アインシュタインの縮約規則 ($u^a v_a := \sum_{a=1}^{d} u^a v_a$ など) を用いる.

3.1 接空間の勾配表現

統計モデルを $\{p(x|u) \mid u \in U\}$ とする. 多様体の各点 $p = p(\cdot|u)$ における接空間は, スコア関数の張るベクトル空間として定義される:

$$T_u^{(c)} = \mathrm{span}\{\partial_1 \log p(x|u), \ldots, \partial_d \log p(x|u)\}.$$

ここで, $u = (u^a)$ は局所座標を表し, $\partial_a := \partial/\partial u^a$ と定義する. 数学的には, 接空間は微分作用素 $\{\partial_a\}_{a=1}^d$ の張るベクトル空間であり, 上の定義はその具体的な表現の 1 つにすぎない. したがって他にも表現方法はあり, 例えば $F : \mathbb{R}_+ \to \mathbb{R}$ を単射な関数として

$$T_u^{(F)} = \mathrm{span}\{\partial_1 F(p(x|u)), \ldots, \partial_d F(p(x|u))\}$$

を接空間と考えてもよい.

ここでは, 勾配表現に基づいて接空間を表してみよう. すなわち, 確率密度 $p(x|u)$ を勾配表現 $g(x|u)$ で表し, 接空間を

$$T_u^{(g)} = \mathrm{span}\{\partial_1 g(x|u), \ldots, \partial_d g(x|u)\}$$

と表してみる. $T_u^{(g)}$ と $T_u^{(c)}$ が一対一かどうかは自明ではない ($\to$ 5 節) が, とりあえず一対一と仮定して以下は議論する. 接空間の勾配表現は次節で g-計量を定義する際に用いられる.

3.2 フィッシャー計量, preferred point 計量と g-計量

フィッシャー計量は次式で定義されるテンソル量であった.

$$J_{ab}(u) := \int p(x|u) \left[\frac{\partial}{\partial u^a} \log p(x|u)\right] \left[\frac{\partial}{\partial u^b} \log p(x|u)\right] \mathrm{d}x.$$

ここで $\mathrm{d}x$ は基準測度である. フィッシャー計量は, 以下のような性質を持ち, 情報量と呼ぶにふさわしいものである.

- 観測値 x の変数変換に対して不変である.
- n 個の独立同一標本の情報量は 1 個の情報量の n 倍である.
- データ圧縮に対して情報量は単調減少である. つまり, 任意の (単射とは限らない) 関数 $t(x)$ に対して, 確率変数 X の情報量は $t(X)$ の情報量より大きい.
- Cramér-Rao 不等式を満たす : $\mathrm{E}_u[T^a] = u^a$ ならば $\mathrm{Cov}[T^a, T^b] \succeq J^{ab}$ である.

一方, 情報量という観点を一旦忘れれば, 多様体上の計量としては様々なものを考えることができる. 上記の性質をいくつか犠牲にして新しい計量を作ってみよう. 新しい計量を導入する利点は 4 節で明らかになる.

まず議論の出発点として, 指数型モデルと相対エントロピーのなすヘッセ構造を利用して, 任意のモデルのフィッシャー計量を導く. 指数型モデルとは次の形をもつ統計的モデルであった.

$$p(x|\theta) = \exp(\theta^i t_i(x) - C(\theta)), \quad \theta \in \Theta \subset \mathbb{R}^d, \quad C(\theta) := \log \int \exp(\theta^i t_i(x)) \mathrm{d}x.$$

任意の確率密度関数 $p_0(x)$ を固定すると, 相対エントロピー (Kullback-Leibler divergence)

$$D(p_0\|p(\cdot|\theta)) := \int p_0(x)\log\frac{p_0(x)}{p(x|\theta)}\,\mathrm{d}x = \int p_0(x)[\log p_0(x) - \theta^i t_i(x)]\,\mathrm{d}x + C(\theta)$$

は θ について凸である[1]. 実際,

$$\partial_i\partial_j D(p_0\|p(\cdot|\theta)) = \partial_i\partial_j C(\theta) = \int[\partial_i \log p(x|\theta)][\partial_j \log p(x|\theta)]p(x|\theta)\mathrm{d}x \succeq 0$$

となり, これはフィッシャー計量そのものである. 次に, 指数型とは限らない一般のモデル $\{p(x|u) \mid u \in U\}$ を考える. 各点 u における計量は, 指数型モデルで一次近似することにより自然に導かれる. このことを説明しよう. いま任意の点 u_0 を固定する. u_0 におけるスコア (対数尤度の 1 回微分) を $t_a(x) = \partial_a \log p(x|u_0)$ とおく. ここで, u をパラメータとする指数型モデル $q(x|u) = p(x|u_0)\exp((u-u_0)^a t_a(x) - C(u))$ を新たに定義すれば, そのスコアは

$$\partial_a \log q(x|u)|_{u=u_0} = t_a(x) - C(u_0) = t_a(x)$$

となり, もとのモデルのスコアと一致する. つまり二つのモデル $\{p(x|u)\}$, $\{q(x|u)\}$ の u_0 における接空間は同一視できる. このイメージを図 2 に示す. いま指数型モデル $q(x|u)$ には計量が定義されていたから, もとのモデル $p(x|u)$ にも同じ計量を使えばよく, これがフィッシャー計量に一致する:

$$\int \partial_a \log q(x|u_0)\partial_b \log q(x|u_0) q(x|u_0)\,\mathrm{d}x = \int t_a(x)t_b(x)p(x|u_0)\,\mathrm{d}x = J_{ab}(u_0).$$

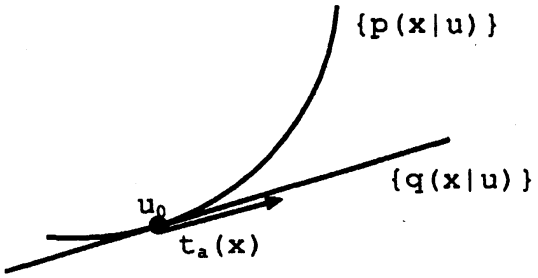

図 2: 共通の接空間を持つ平坦モデルの構成.

ここまでの議論をまとめると,

- 指数型モデルについては, 相対エントロピーが凸関数となり, そのヘッセ形式からフィッシャー計量が導かれる.
- 一般のモデル $p(x|u)$ については, 各点 u_0 において接空間が共通となるような指数型モデル $q(x|u)$ を用意する. その指数型モデルのフィッシャー計量を計算すると, 元のモデルのフィッシャー計量に一致する.

[1] p_0 は任意なので, 「負の対数尤度が凸である」と言い換えても同じことである.

上の手順において, 指数型モデルの代わりに別のモデルを考えれば, 別の計量が導かれる. 例として混合型モデル

$$p(x|\theta) \;=\; p(x|0) + \theta^i t_i(x), \quad \int t_i(x)\mathrm{d}x = 0$$

を考えると, その相対エントロピーは θ に関して凸である：

$$\partial_i\partial_j D(p_0\|p(\cdot|\theta)) \;=\; \int p_0(x)\frac{t_i(x)t_j(x)}{\{p(x|0)+\theta^i t_i(x)\}^2}\,\mathrm{d}x \;\succeq\; 0.$$

このヘッセ形式は混合型モデルの計量として使える. 次に, 混合モデルとは限らない一般のモデル $\{p(x|u)\}$ を考える. 新たに混合モデル $q(x|u) = p(x|u_0) + (u-u_0)^a t_a(x)$ $(t_a(x) := \partial_a p(x|u_0))$ を考えれば, 2つのモデル $\{p(x|u)\}$, $\{q(x|u)\}$ の u_0 における接空間は一致する. よって, 後者の計量から前者の計量が導かれる：

$$J^{(\mathrm{m})}_{ab}(u_0) \;:=\; \int p_0(x)\frac{t_a(x)t_b(x)}{p(x|u_0)^2}\,\mathrm{d}x \;=\; \int p_0(x)[\partial_a \log p(x|u_0)\partial_b \log p(x|u_0)]\,\mathrm{d}x$$

この計量は preferred-point 計量 (Critchley et al. 1993) と呼ばれる.

最後に g-平坦モデルから計量を導こう. 前節までに述べたとおり, g-平坦モデルとは

$$p(x|\theta) \;=\; \phi(g(x|\theta))\det(G(x|\theta)), \quad g(x|\theta) \;=\; g(x|0) + \theta^i t_i(x)$$

の形で書けるモデルであった. 相対エントロピーは θ に関して凸になる：

$$\partial_i\partial_j D(p_0\|p(\cdot|\theta)) \;=\; \int p_0(x)\left[t_i(x)^{\mathsf{T}} t_j(x) + \mathrm{tr}[G(x|\theta)^{-1}T_i(x)G(x|\theta)^{-1}T_j(x)]\right]\mathrm{d}x \;\succeq\; 0.$$

ただし $G := \nabla g^{\mathsf{T}}$, $T_i := \nabla t_i^{\mathsf{T}}$ とおいた. このヘッセ形式は g-平坦モデルの計量として使える. g-平坦でないモデルについては, その勾配表現を $g(x|u)$ とおき, 接ベクトル $t_a(x) := \partial_a g(x|u_0) \in T^{(\mathrm{g})}_{u_0}$ が共通となるような g-平坦モデルを新たに用意する：

$$q(x|u) \;=\; \phi(g(x|u_0) + (u-u_0)^a t_a(x))\det(G(x|u_0) + (u-u_0)^a T_a(x)).$$

対応する計量は

$$J^{(\mathrm{g})}_{ab}(u) \;:=\; \int p_0(x)\left[\partial_a g(x|u)^{\mathsf{T}}\partial_b g(x|u) + \mathrm{tr}[G(x|u)^{-1}(\partial_a G(x|u))G(x|u)^{-1}(\partial_b G(x|u))]\right]\mathrm{d}x$$

となる. これを **g-計量**と呼ぶことにする.

簡単のため, フィッシャー計量, preferred point 計量をそれぞれ **e-計量**, **m-計量**と呼び, $J^{(\mathrm{c})}_{ab}$, $J^{(\mathrm{m})}_{ab}$ と表す.

補題 2. *$p_0(x) = p(x|u)$ のとき, $J^{(\mathrm{c})}_{ab}(u) = J^{(\mathrm{m})}_{ab}(u) = J^{(\mathrm{g})}_{ab}(u) =$ (フィッシャー計量).*

Proof. $s \in \{\mathrm{e},\mathrm{m},\mathrm{g}\}$ とする. 点 u_0 で接空間が共通な s-平坦モデルを $q(x|u)$ とおけば, s-計量の定義より

$$\begin{aligned} J^{(s)}_{ab}(u_0) \;&=\; \int q(x|u_0)\left[-\frac{\partial^2}{\partial u^a\partial u^b}\log q(x|u_0)\right]\mathrm{d}x \\ &=\; \int q(x|u_0)\left[\frac{\partial}{\partial u^a}\log q(x|u_0)\frac{\partial}{\partial u^b}\log q(x|u_0)\right]\mathrm{d}x \\ &=\; \int p(x|u_0)\left[\frac{\partial}{\partial u^a}\log p(x|u_0)\frac{\partial}{\partial u^b}\log p(x|u_0)\right]\mathrm{d}x \end{aligned}$$

と書ける. □

3.3 g-計量に関する双対接続

微分幾何において大切な概念としてベクトル場の平行移動, あるいは接続係数がある. ここでは, g-計量にともなう接続係数として, 以下で g-接続を導入し, またその双対接続を求めてみる. g-接続の統計学的な利用価値は今のところ定かではないが, 数学的には興味のある対象であると思われる.

g-平坦モデルに対しては, ベクトル場の平行移動は自然に定義される. つまり, g-平坦モデルのアファイン座標に対する接続係数を $\Gamma_{ij,k}=0$ と定義すればよい. この Γ を**g-接続**と呼ぶことにする. また一般に, 平坦な多様体にヘッセ計量 $\partial_i\partial_j\varphi(\theta)$ が与えられているとき, 双対接続が $\tilde{\Gamma}_{ij,k}=\partial_i\partial_j\partial_k\varphi(\theta)$ によって定義される (例えば志賀 2001). g-平坦モデル $g(x|\theta)=g(x|0)+\theta^i t_i(x)$ の場合, $G=\nabla g^\top$, $T_i:=\nabla t_i^\top$ として

$$\tilde{\Gamma}_{ij,k} \;=\; -2\int p_0 \operatorname{tr}\left[G^{-1}T_iG^{-1}T_jG^{-1}T_k\right]\,\mathrm{d}x$$

と表される. 一般の座標系に対する接続係数は, $g_{ab}:=\partial_a\partial_b g(x|u)$ などと略記すると

$$\Gamma_{ab,c} \;=\; \int p_0\left[g_{ab}^\top g_c+\operatorname{tr}\left[G^{-1}G_{ab}G^{-1}G_c\right]\right]\,\mathrm{d}x$$

となり, また双対接続は,

$$\tilde{\Gamma}_{ab,c} \;=\; \Gamma_{ab,c}-2\int p_0\operatorname{tr}\left[G^{-1}G_aG^{-1}G_bG^{-1}G_c\right]\,\mathrm{d}x$$

となる.

勾配表現 $g(x|u)$ に対して, その**双対勾配表現**を

$$\tilde{g}(x|u) \;=\; p_0(x)g(x|u)+\sum_i\frac{\partial}{\partial x^i}(p_0(x)G(x|u)^{-1})_{\cdot i}$$

と定義する. このとき, $g_a:=\partial_a g(x|u)$, $\tilde{g}_a:=\partial_a\tilde{g}(x|u)$ などと略記すると, 部分積分を用いて

$$\begin{aligned}\int g_a^\top\tilde{g}_b\,\mathrm{d}x \;&=\; \int g_a^\top\left[p_0g_b-\sum_i\frac{\partial}{\partial x^i}(p_0G^{-1}G_bG^{-1})_{\cdot i}\right]\,\mathrm{d}x\\ &=\; \int p_0\left[g_a^\top g_b+\operatorname{tr}[G_aG^{-1}G_bG^{-1}]\right]\,\mathrm{d}x\\ &=\; J_{ab}^{(\mathrm{g})}\end{aligned}$$

同様に

$$\int(\partial_a\partial_b g)^\top(\partial_c\tilde{g})\,\mathrm{d}x \;=\; \Gamma_{ab,c},\quad \int(\partial_c g)^\top(\partial_a\partial_b\tilde{g})\,\mathrm{d}x \;=\; \tilde{\Gamma}_{ab,c}$$

が成り立つ. 特に, 双対勾配表現がアファイン関係 $\tilde{g}(x|\theta)=\tilde{g}(x|0)+\theta^i t_i(x)$ となるようなモデルでは, 双対接続が $\tilde{\Gamma}_{ij,k}=0$ となる.

4 g-計量を用いた最尤推定アルゴリズムの大域的性質

フィッシャー計量を利用した最尤推定アルゴリズムとしてフィッシャーのスコア法がある. この節では, g-計量 $J_{ab}^{(\mathrm{g})}$ を利用した最尤推定アルゴリズムを考えてみる.

4.1 g-スコア法

まず, 任意の密度関数を $p_0(x)$ とおき, 相対エントロピー

$$D(p_0\|p(\cdot|u)) = \int p_0(x)\log\frac{p_0(x)}{p(x|u)}\,\mathrm{d}x$$

の最小化問題を考える. 最尤法は, n 個の標本 $X_1,\ldots,X_n$ の経験密度 $\hat{p}_n(x)=n^{-1}\sum_{k=1}^n\delta_{X_i}(x)$ を p_0 とした場合に相当するので, 以下では一般に p_0 のまま話を進めることにする. スコア関数を

$$S_a=\int p_0(x)\partial_a\log p(x|u)\,\mathrm{d}x$$

とおく. e-, m-, g-計量はそれぞれ

$$\begin{aligned}
J^{(\mathrm{e})}_{ab}(u) &= \int p(x|u)\partial_a\log p(x|u)\partial_b\log p(x|u)\,\mathrm{d}x,\\
J^{(\mathrm{m})}_{ab}(u) &= \int p_0(x)\partial_a\log p(x|u)\partial_b\log p(x|u)\,\mathrm{d}x,\\
J^{(\mathrm{g})}_{ab}(u) &= \int p_0(x)[(\partial_a g)(\partial_b g)+\mathrm{tr}[G^{-1}\partial_a G G^{-1}\partial_b G]]\,\mathrm{d}x
\end{aligned}$$

と定義されていた.

e-, m-, g-スコア法

次式によるパラメータ u の更新方法を, それぞれ e-スコア法, m-スコア法, g-スコア法と呼ぶ.

$$\begin{aligned}
u &\leftarrow u+\Delta u^{(\mathrm{e})},\quad \Delta u^{(\mathrm{e})}:=(J^{(\mathrm{e})})^{-1}S,\\
u &\leftarrow u+\Delta u^{(\mathrm{m})},\quad \Delta u^{(\mathrm{m})}:=(J^{(\mathrm{m})})^{-1}S,\\
u &\leftarrow u+\Delta u^{(\mathrm{g})},\quad \Delta u^{(\mathrm{g})}:=(J^{(\mathrm{g})})^{-1}S.
\end{aligned}$$

e-スコア法はフィッシャーのスコア法と呼ばれるアルゴリズムと同じである.

補題 3. $s\in\{\mathrm{e},\mathrm{m},\mathrm{g}\}$ とする. s-平坦モデルに対する s-スコア法はニュートン法に一致する.

Proof. s-平坦モデル $\{p(x|\theta)\}$ の s-計量は $\mathrm{E}_{p_0}[-\partial_i\partial_j\log p(x|\theta)]$ で定義されていた. また $S_i=\mathrm{E}_{p_0}[\partial_i\log p(x|\theta)]$ であるから, 更新則 $\Delta u^{(s)}=(J^{(s)})^{-1}S$ は目的関数 $\mathrm{E}_{p_0}[-\log p(x|\theta)]$ に対するニュートン法に一致する. □

仮に真のパラメータを u_0 とすれば, 経験密度 $\hat{p}_n(x)$ は n が大きいとき $p(x|u_0)$ に収束するから, 以後は極限 $p_0(x)=p(x|u_0)$ を考えることにする. 補題 2 から, u が真のパラメータ u_0 に近いときはどのアルゴリズムの挙動も大差はない. 問題は, u と u_0 が大きく異なる場合である. 普通, アルゴリズムの初期値は真値 u_0 から大きく異なるはずだから, そのような場合での挙動を調べておくことには意味があるだろう.

4.2 位置母数モデルの場合

簡単のため, 1 次元の位置母数モデルを考える. つまり, 密度関数が

$$p(x|u) = f(x-u), \quad f(x) = \phi(g(x))g'(x).$$

のように書けるものとする. $g(x)$ は $f(x)$ の勾配表現である. 1 次元の場合は f から g が陽に求まる:

$$g(x) = \Phi^{-1}(F(x)), \quad \Phi(y) = \int_{-\infty}^{y} \phi(\eta)\mathrm{d}\eta, \quad F(x) = \int_{-\infty}^{x} f(\xi)\mathrm{d}\xi$$

また, 位置母数モデルの場合, フィッシャー計量 $J^{(\mathrm{c})}$ は u によらず定数になることに注意しておく.

まず, 分布の裾が軽い場合 (short tail) について考えると次の定理を得る.

定理 5 (short tail). *f は偶関数で, 次の性質を満たすと仮定する*[2].

$$\exists C, \gamma > 0, \quad f(x) \sim \mathrm{e}^{-C|x|^{\gamma}} \quad (|x| \to \infty). \tag{3}$$

$u_0 = 0$ のとき, $u \to \infty$ の下で

$$\begin{aligned} \Delta u^{(\mathrm{c})} &\sim -C\gamma (J^{(\mathrm{c})})^{-1} u^{\gamma-1}, \\ \Delta u^{(\mathrm{m})} &\sim -(C\gamma)^{-1} u^{1-\gamma}, \\ \Delta u^{(\mathrm{g})} &\sim -2\gamma^{-1} u \end{aligned}$$

が成り立つ. それぞれについて, スコア法が発散しない十分条件 ($\exists U > 0, \forall u \geq U,\ -2 < \Delta u/u < 0$), 大域的に一次収束する必要条件 ($-2 < \lim_{u\to\infty} \Delta u/u < 0$) は以下の通りである.

種類	Δu	発散しない十分条件	大域的に一次収束する必要条件
e	$-C\gamma (J^{(\mathrm{c})})^{-1} u^{\gamma-1}$	$0 < \gamma < 2$ or $\gamma = 2,\ C(J^{(\mathrm{c})})^{-1} < 1$	$\gamma = 2,\ C(J^{(\mathrm{c})})^{-1} < 1$
m	$-(C\gamma)^{-1} u^{1-\gamma}$	$\gamma > 0$	$\emptyset$
g	$-2\gamma^{-1} u$	$\gamma > 1$	$\gamma > 1$

Proof. まず, $f'(x)/f(x) \sim -\operatorname{sgn}(x) C\gamma |x|^{\gamma-1}$ より, スコア関数は $u \to \infty$ のとき

$$S = -\int f(x) \frac{f'(x-u)}{f(x-u)} \mathrm{d}x \sim -\frac{f'(-u)}{f(-u)} \sim -C\gamma u^{\gamma-1}$$

となる. また, m-計量は

$$J^{(\mathrm{m})} = \int f(x) \left(\frac{f'(x-u)}{f(x-u)} \right)^2 \mathrm{d}x \sim \left(\frac{f'(-u)}{f(-u)} \right)^2 \sim (C\gamma)^2 u^{2\gamma-2}$$

となる. よって $(J^{(\mathrm{m})})^{-1} S \sim -(C\gamma)^{-1} u^{-\gamma+1}$ を得る. 次に, $g(x) \sim (2C)^{1/2} |x|^{\gamma/2}$ を示すことができる. よって, g-計量は

$$J^{(\mathrm{g})} = \int f(x) \left((g'(x-u))^2 + ((\log g')'(x-u))^2 \right) \mathrm{d}x \sim g'(-u)^2 \sim 2^{-1} C\gamma^2 u^{\gamma-2}$$

となり, $(J^{(\mathrm{g})})^{-1} S \sim -2\gamma^{-1} u$ を得る. □

[2] 厳密には, 漸近的同値性 $\sim$ と微分の順序が交換可能, などの条件も仮定しなければならないが, 省略する.

裾が重い場合 (long tail; ただし期待値の存在は仮定) については, 次の定理が成り立つ.

定理 6 (long tail). f は偶関数かつ次の性質を満たすと仮定する.

$$\exists A > 0,\ \exists \alpha > 2, \quad f(x) \sim A|x|^{-\alpha} \quad (|x| \to \infty).$$

$u_0 = 0$ のとき, $u \to \infty$ の下で

$$\begin{aligned} \Delta u^{(\mathrm{c})} &\sim -A\alpha (J^{(\mathrm{c})})^{-1} u^{-1}, \\ \Delta u^{(\mathrm{m})} &\sim -(A\alpha)^{-1} u, \\ \Delta u^{(\mathrm{g})} &\sim -\alpha u \end{aligned}$$

が成り立つ. それぞれの適用可能範囲は以下の通りである.

種類	Δu	大域的に収束する必要条件	大域的に一次収束する必要条件
e	$-\alpha (J^{(\mathrm{c})})^{-1} u^{-1}$	any $\alpha > 2$	$\emptyset$
m	$-\alpha^{-1} u$	any $\alpha > 2$	any $\alpha > 2$
g	$-\alpha u$	$\emptyset$	$\emptyset$

Proof. まず $f'(x)/f(x) \sim -\operatorname{sgn}(x)\alpha|x|^{-1}$ より, スコア関数は $u \to \infty$ のとき

$$S = -\int f(x) \frac{f'(x-u)}{f(x-u)} \mathrm{d}x \sim -\frac{f'(-u)}{f(-u)} \sim -\alpha u^{-1}$$

となる. また, m-計量は

$$J^{(\mathrm{m})} = \int f(x) \left(\frac{f'(x-u)}{f(x-u)} \right)^2 \mathrm{d}x \sim \left(\frac{f'(-u)}{f(-u)} \right)^2 \sim \alpha^2 u^{-2}$$

となる. よって $(J^{(\mathrm{m})})^{-1} S \sim -\alpha^{-1} u$ を得る. また, $g(x) \sim \operatorname{sgn}(x)\sqrt{2(\alpha-1)\log|x|}$ を示すことができる. よって g-計量は

$$J^{(\mathrm{g})} = \int f(x) \left((g'(x-u))^2 + ((\log g')'(x-u))^2 \right) \mathrm{d}x \sim ((\log g')'(-u))^2 \sim u^{-2}$$

となるが示される. よって $(J^{(\mathrm{g})})^{-1} S \sim -\alpha u$ を得る. □

以上, 2 つの定理から, 裾が軽い分布には g-スコア法が, 裾が重い分布には m-スコア法がよく機能することが示された.

5 今後の課題

今後解決すべき数学的な問題を列挙してみる. なお, 筆者の不勉強により, 簡単な問題も含まれているかと思われるので注意されたい.

問題 (勾配写像の正則性)

関数 $(x,u) \mapsto p(x|u)$ がどの程度の微分可能性を持てば $(x,u) \mapsto \psi(x|u)$ の微分可能性が保証されるか?

問題 (接空間の勾配表現)

勾配写像をパラメータ微分してできる関数は, 接空間と一対一であるか?つまり, $g(x|u)$ を統計モデルの勾配表現とするとき, $\partial_a g \in T_u^{(g)}$ $(a = 1, \ldots, d)$ が一次独立ならば

$$\partial_a \log p[g] \;=\; -g^\top \partial_a g + \mathrm{tr}[G^{-1}\partial_a G] \;\in\; T_u^{(c)} \quad (a = 1, \ldots, d)$$

は一次独立か?

問題 (他の平坦モデル)

相対エントロピー $D(p_0 \| p(\cdot|\theta))$ が θ に関して凸関数となるようなモデルは, 指数型モデル, 混合型モデル, g-平坦モデルの他にどのようなものがあるだろうか?

問題 (ポテンシャルの基本形)

次の集合は, 凸関数全体の中で (適当な位相に関して) 稠密か?

$$\left\{ \sum_{\lambda \in \Lambda} f_\lambda(e_\lambda^\top x) \;\middle|\; \Lambda \text{ は有限集合}, \; f_\lambda : \mathbb{R} \to \mathbb{R} \text{ は凸関数}, \; e_\lambda \in \mathbb{R}^m \right\}.$$

参考文献

[1] Amari, S. (1985). *Differential-Geometrical Methods in Statistics*, Springer-Verlag, Tokyo.

[2] Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions, *Comm. Pure Appl. Math.*, **44**, 375–417.

[3] Critchley, F., Marriott, P. and Salmon, M. (1993). Preferred point geometry and statistical manifolds, *Ann. Statist.*, **21**, 1197–1224.

[4] McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps, *Duke Math. J.*, **80**, 309-323.

[5] 宮川 雅巳 (1997). グラフィカルモデリング, 朝倉書店.

[6] Sei, T. (2006). Parametric modeling based on the gradient maps of convex functions, Technical Report METR2006-51, Department of Mathematical Engineering and Information Physics, The University of Tokyo.

[7] 志賀 裕彦 (2001). ヘッセ幾何学, 裳華房.

[8] Villani, C. (2003). *Topics in Optimal Transportation*, AMS, Providence.