ガンマダイバージェンスに基づく EM/em アルゴリズム

EM/em algorithm based on Gamma-divergence

数学専攻　古田 旭宏

FURUTA, Akihiro

1　はじめに

統計学における最尤推定は，KL ダイバージェンスと呼ばれる確率分布間の乖離度を表す尺度を用いて幾何学的に捉えることができる．具体的には，リーマン接続ではなく双対接続により KL ダイバージェンスを導くことに着目し，一般化ピタゴラスの定理 [1, 2] を示すことにより，推定を射影と捉えることができる．このことから，KL ダイバージェンスに由来する統計手法は数学的に性質がよい．一方で，現実のデータに対して最尤推定の振る舞いが常に優れているとは限らない．最尤推定の弱点は外れ値を含むデータに対するロバスト性で，理想的なデータに近くないと推定が不安定になってしまうことである．

最尤推定の弱点を補う手法としてガンマダイバージェンス [3] が研究されている．ガンマダイバージェンスは情報幾何学的な性質を多くもち，幾何学的に考察しつつロバスト性を保証するパラメータ推定を行うことができる．近年では，ガンマダイバージェンスが回帰モデル，確率的主成分分析，クラスタリングなどの統計手法に応用されている [4, 5, 6].

従来より統計学において研究されている EM アルゴリズムは，潜在変数を含む統計モデルのパラメータ推定を行うものである．これを情報幾何学によって捉えると，e-射影と m-射影を交互に行なった em アルゴリズム [7] と同等であることが知られている．em アルゴリズムそのものは KL ダイバージェンスを用いて定義されるものである．本研究では，ガンマダイバージェンスのもとでの em アルゴリズムを拡張し，em アルゴリズムに代わるロバストなアルゴリズムを提案する．

2　情報幾何学と一般化ピタゴラスの定理

双対接続に着目してダイバージェンスにおける一般化ピタゴラスの定理 [1] が発見されたことを発端に，様々な研究者によって幾何学的な概念を用いた統計学の理論展開が成されてきた．したがって，情報幾何学は狭い意味では双対接続の微分幾何学のことであると考えることができる．

定義 2.1 (接続) C^∞ 級多様体 M 上の関数を $f \in C^\infty(M)$，ベクトル場 $X, Y \in \mathcal{X}(M)$ の局所座標表示をそれぞれ $X = X^i\partial_i, Y = Y^j\partial_j$ と表す．このとき，

$$\begin{aligned}&\nabla_{X+Y}Z = \nabla_X Z + \nabla_Y Z, \quad \nabla_X(Y+Z) = \nabla_X Y + \nabla_X Z,\\ &\nabla_{fX}Y = f\nabla_X Y, \quad \nabla_X(fY) = (Xf)Y + f\nabla_X Y\end{aligned}$$

を満たす写像 $\nabla : \mathcal{X}(M) \times \mathcal{X}(M) \to \mathcal{X}(M)$ を共変微分という．また，共変微分から接続係数 Γ_{ij}^k が $\nabla_{\partial_i}\partial_j = \Gamma_{ij}^k\partial_k$ により定まる．

M 上のリーマン計量を g と書くとき，内積を保つ接続はリーマン接続に限っており，その接続係数は

$$\Gamma_{ij}^k = g^{kl}\Gamma_{ij,l}, \quad \Gamma_{ij,k} = \frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}\right) \tag{2.1}$$

で与えられる．一方で，情報幾何学ではリーマン接続の代わりに，次で定義される双対接続 $\nabla, \overset{*}{\nabla}$ を用いる．

定義 2.2 (双対接続と双対平坦多様体) 計量 (g_{ij}) に対し，2 つの接続 $\nabla, \overset{*}{\nabla}$ から定まる接続係数を $\Gamma_{ij,k}, \overset{*}{\Gamma}_{ij,k}$ と

する．関係式

$$\partial_i g_{jk} = \Gamma_{ij,k} + \overset{*}{\Gamma}_{ij,k} \tag{2.2}$$

が成り立つとき，$\nabla, \overset{*}{\nabla}$ を双対接続と呼ぶ．また，双対接続 $\nabla, \overset{*}{\nabla}$ に関する曲率と捩率がともに 0 となるとき，M を双対平坦多様体と呼ぶ．

統計学のフィッシャー情報行列はリーマン計量となることが知られている．情報幾何学ではこれをフィッシャー計量と呼び，(g_{ij}) と書く．そして，情報幾何学で双対接続を用いる理由は，KL ダイバージェンスに由来する双対接続である e-接続と m-接続が一般の統計モデル $\mathcal{M}$ に対してそれぞれ

$$\overset{\mathrm{e}}{\Gamma}_{ab,c}(\theta) = \int \partial_a \partial_b \log f(x;\theta) \partial_c f(x;\theta) \mathrm{d}x, \quad \overset{\mathrm{m}}{\Gamma}_{ab,c}(\theta) = \int \partial_a \partial_b f(x;\theta) \partial_c \log f(x;\theta) \mathrm{d}x \tag{2.3}$$

で定義され，これを自然に導出できるからである．(g_{ij}) をフィッシャー計量としたときに，微積分の順序交換を認めることで双対性 $\partial_a g_{bc} = \overset{\mathrm{e}}{\Gamma}_{ab,c} + \overset{\mathrm{m}}{\Gamma}_{ab,c}$ が容易にわかる．また，統計学で最も重要な確率分布族である指数型分布族は双対平坦多様体の典型例である [1, 2]．

M を双対平坦多様体とし，大域的な双対アファイン座標系 $(\theta^i), (\eta_i)$ とルジャンドル変換 ψ, φ により定義される

$$D_\nabla(p,q) = \psi(\theta(p)) + \varphi(\eta(q)) - \theta^i(p)\eta_i(q), \ p, q \in M \tag{2.4}$$

を ∇-ダイバージェンスと呼ぶ．これは M が指数型分布族のときに KL ダイバージェンスに一致する．双対平坦多様体 $\left(M, g, \nabla, \overset{*}{\nabla}\right)$ と D_∇ に対して，次の一般化ピタゴラスの定理 [1, 2] が成り立つ．

定理 2.1 (双対平坦多様体における一般化ピタゴラスの定理) 3 点 $p, q, r \in \left(M, g, \nabla, \overset{*}{\nabla}\right)$ に対して，p, q を結ぶ $\overset{*}{\nabla}$-測地線と q, r を結ぶ ∇-測地線が点 q において，計量 g の意味で直交するならば

$$D_\nabla(p,r) = D_\nabla(p,q) + D_\nabla(q,r) \tag{2.5}$$

が成り立つ．

定理 2.1 により分布族を多様体と捉えたとき，統計学におけるパラメータ推定は射影と考えることができる．なぜなら，経験分布と分布族の間のダイバージェンスが最小になるようなパラメータ (分布族の元) を求めることが推定だからである．

3 ガンマダイバージェンスの定義とその性質

統計学の多くの理論は KL ダイバージェンスに基づいており，その他のダイバージェンスに基づいた理論展開はあまり行われてこなかった．その理由は KL ダイバージェンスから導かれる最尤推定量のもつ漸近的性質の良さにある．一方，最尤推定が確率密度の積で与えられることに起因し，そのロバスト性への脆弱さが常に問題視されてきた．そこで近年，KL ダイバージェンスではなくガンマダイバージェンス [3] を用いてロバストな推定量を与える研究が行われつつある．ガンマダイバージェンスを用いることで既存のロバスト推定と比べて情報幾何学を用いた理論展開が行えるという利点がある．

3.1 ガンマダイバージェンスとピタゴラス関係

KL ダイバージェンスのロバストな拡張としてガンマダイバージェンス [3] が提案されている．

定義 3.1 (ガンマダイバージェンス) $\gamma > 0$ に対し，

$$d_\gamma(g,f) := -\frac{1}{\gamma(\gamma+1)} \int \left(\frac{f(x)}{\|f\|_\gamma}\right)^\gamma g(x) \mathrm{d}x \tag{3.1}$$

$$D_\gamma(g,f) := d_\gamma(g,f) - d_\gamma(g,g) = \frac{1}{\gamma(\gamma+1)} \left(\|g\|_\gamma - \int \left(\frac{f(x)}{\|f\|_\gamma}\right)^\gamma g(x) \mathrm{d}x\right) \tag{3.2}$$

をそれぞれガンマ交差エントロピーおよびガンマダイバージェンスという．ただし，$\|h\|_\gamma := \left\{\int h(x)^{\gamma+1}\,\mathrm{d}x\right\}^{\frac{1}{\gamma+1}}$ である．

g と f を分離するようにヘルダーの不等式を用いて上界を考えれば直ちに $D_\gamma(g,f) \geq 0$ かつ $D_\gamma(g,f) = 0 \Leftrightarrow g = f$ a.e. が示され，ダイバージェンスの定義を満たす [3]．また，ガンマダイバージェンスは $\gamma \to 0$ のとき KL ダイバージェンスに収束する．

ガンマダイバージェンスに関しては，汚染分布 $\delta(x)$ を用いた以下のピタゴラス関係が成り立つ．

定理 3.1 (汚染分布とピタゴラス関係) $\nu_{f_\theta} := \int \delta(x) f_\theta(x)^\gamma\,\mathrm{d}x$，$\nu_{g_0} := \int \delta(x) g_0(x)^\gamma\,\mathrm{d}x$ とおく．このとき $\nu := \max(\nu_{f_\theta}, \nu_{g_0}) \to 0$ のもとで以下が成り立つ:

$$D_\gamma(g_\varepsilon, f_\theta) = D_\gamma(g_\varepsilon, g_0) + D_\gamma(g_0, f_\theta) + O(\nu). \tag{3.3}$$

これは g_0, f_θ を固定して $\nu \to 0$ となるように δ を動かしても式 (3.3) の関係が崩れないことを意味する．ν が小さいことは δ が f_θ や g_0 の裾，すなわち密度の小さいところに値をもつということある．$\varepsilon \to 0$ に限らずとも式 (3.3) が成り立つので，θ の推定量にロバスト性が保証される．もちろん，どの観測値が外れ値であるか予め知っている必要はない．$\gamma \to 0$ のとき，すなわち KL ダイバージェンスのときは $O(\nu)$ を無視できなくなる．

4 EM/em アルゴリズムとガンマダイバージェンス

EM アルゴリズムは, 期待値をとる E ステップ (Expectation), 尤度の最大化を行う M ステップ (Maximization) を繰り返す方法であり，潜在変数を含む統計モデルのパラメータ推定に用いられる．情報幾何学を用いると，EM アルゴリズムはモデル多様体 M からデータ多様体 D への e-射影と D から M への m-射影を繰り返す方法 (em アルゴリズム) と同等であることを示すことができる [7]．本研究では，em アルゴリズムを KL ダイバージェンスからガンマダイバージェンスを用いたものに拡張するとともに，潜在変数を含むモデルに対するロバストなパラメータ推定法を提案する．

確率変数 $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$ に対して，$\boldsymbol{y}$ のみ観測可能とする．興味のあるパラメータを $\boldsymbol{\xi}$ とし，これで規定される確率分布族を $M = \{p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\xi})\}$ と書き，これをモデル多様体と呼ぶ．経験分布を $\hat{q}(\boldsymbol{y})$ とし，確率分布の空間を $\mathcal{P} := \{p(\boldsymbol{y}, \boldsymbol{z})\}$ と書くとき，

$$D := \left\{ q \in \mathcal{P} \;\middle|\; q = \hat{q}(\boldsymbol{y})\, q(\boldsymbol{z}|\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\boldsymbol{y} - \boldsymbol{y}_i)\, q(\boldsymbol{z}|\boldsymbol{y}_i) \right\} \subset \mathcal{P} \tag{4.1}$$

をデータ多様体と呼ぶ．モデル多様体 M とデータ多様体 D，および KL ダイバージェンスを用いて em アルゴリズムを以下のように定義する．

定義 4.1 (em アルゴリズム) 統計モデル M の一点 f_0 をとる．$\{f_t\}_{t\in\mathbb{N}}$ に対して

$$g_t := \operatorname*{argmin}_{g\in D} D_{\mathrm{KL}}(g, f_t) \tag{4.2}$$

を e ステップ，$\{g_t\}_{t\in\mathbb{N}}$ に対して

$$f_{t+1} := \operatorname*{argmin}_{f\in M} D_{\mathrm{KL}}(g_t, f) \tag{4.3}$$

を m ステップと呼び，e ステップと m ステップを交互に繰り返すアルゴリズムを em アルゴリズムと呼ぶ．

e ステップは f_t を D に e-射影することで，m ステップは g_t を M に m-射影することである．そして，e ステップについて以下の定理が成り立つ．

定理 4.1 (e ステップの不変性) モデル多様体 M からデータ多様体 D への e-射影を行うと，$q(\boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\xi})$ が成り立つ．

	KL ダイバージェンス	ガンマダイバージェンス	解
M への射影	$\displaystyle\int \frac{q(\boldsymbol{z}\mid\boldsymbol{y})}{p(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{\xi})}\frac{\partial}{\partial\boldsymbol{\xi}}p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{z}=0$	$\frac{\int q(\boldsymbol{z}\mid\boldsymbol{y})(p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi}))^{\gamma-1}\cdot\frac{\partial}{\partial\boldsymbol{\xi}}p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi})\mathrm{d}\boldsymbol{z}}{\int q(\boldsymbol{z}\mid\boldsymbol{y})(p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi}))^{\gamma}\mathrm{d}\boldsymbol{z}} - \frac{\int (p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi}))^{\gamma}\cdot\frac{\partial}{\partial\boldsymbol{\xi}}p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi})\mathrm{d}\boldsymbol{z}}{\int (p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi}))^{1+\gamma}\mathrm{d}\boldsymbol{z}}=0$	$\hat{\boldsymbol{\xi}}$
D への射影	$q(\boldsymbol{z}\mid\boldsymbol{y})=p(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{\xi})$	$\frac{(q(\boldsymbol{z}\mid\boldsymbol{y}))^{\gamma}}{(\lVert q(\boldsymbol{z}\mid\boldsymbol{y})\rVert_{\gamma})^{1+\gamma}}=\frac{(p(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{\xi}))^{\gamma}}{\int (p(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{\xi}))^{\gamma}\cdot q(\boldsymbol{z}\mid\boldsymbol{y})\mathrm{d}\boldsymbol{z}}$	$\hat{q}$

表 1　潜在変数のあるモデルの推定における方程式の比較.

この定理によって，em アルゴリズムが EM アルゴリズムと同等であることが示される [7]．また，M への射影は方程式

$$\int \frac{q(\boldsymbol{z}\mid\boldsymbol{y})}{p(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{\xi})}\frac{\partial}{\partial\boldsymbol{\xi}}p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi})\,\mathrm{d}z=0 \tag{4.4}$$

を $\boldsymbol{\xi}$ に関して解くことである．em アルゴリズムをガンマダイバージェンスによって拡張することで，表 1 に示したように 2 つの方程式をそれぞれ得ることができる．このとき，以下の定理が成り立つ．

定理 4.2 $\gamma\approx 0$ のとき，

$$\frac{\int q(\boldsymbol{z}\mid\boldsymbol{y})\,(p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi}))^{\gamma-1}\cdot\frac{\partial}{\partial\boldsymbol{\xi}}p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi})\mathrm{d}\boldsymbol{z}}{\int q(\boldsymbol{z}\mid\boldsymbol{y})\,(p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi}))^{\gamma}\,\mathrm{d}\boldsymbol{z}}-\frac{\int (p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi}))^{\gamma}\cdot\frac{\partial}{\partial\boldsymbol{\xi}}p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi})\mathrm{d}\boldsymbol{z}}{\int (p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\xi}))^{1+\gamma}\,\mathrm{d}\boldsymbol{z}}=0 \tag{4.5}$$

$$\frac{(q(\boldsymbol{z}\mid\boldsymbol{y}))^{\gamma}}{(\lVert q(\boldsymbol{z}\mid\boldsymbol{y})\rVert_{\gamma})^{1+\gamma}}=\frac{(p(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{\xi}))^{\gamma}}{\int (p(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{\xi}))^{\gamma}\cdot q(\boldsymbol{z}\mid\boldsymbol{y})\,\mathrm{d}\boldsymbol{z}} \tag{4.6}$$

はそれぞれ式 (4.4) と定理 4.1 の結果に一致する．

定理 4.2 より，提案したアルゴリズムは em アルゴリズムの拡張となっている．また，ガンマダイバージェンスを用いているため $\boldsymbol{\xi}$ の推定は外れ値に対してロバストであることもわかる．

5　まとめと今後の展望

前半では情報幾何学と一般化ピタゴラスの定理を通して，推定を射影として捉えることの正当性について述べた．後半ではガンマダイバージェンスの定義と諸性質について説明した．最後に，提案したアルゴリズムが em アルゴリズムの拡張となっていることを述べた．今後，EM アルゴリズムを用いた統計手法に提案したアルゴリズムを適用し，シミュレーションを行うことで提案手法の妥当性を考察していきたい．また，提案手法のパラメータ推定のロバスト性を示す理論的な不等式などを導くことで，ロバスト性の保証を与えたいと考えている．

参考文献

[1] Amari, S. and Nagaoka, H. (2000), Methods of Information Geometry, American Mathematical Society.

[2] 藤原彰夫 (2021), 情報幾何学の基礎, 共立出版.

[3] Fujisawa, H. and Eguchi, S. (2008), Robust parameter estimation with a small bias against heavy contamination, *Journal of Multivariate Analysis*, **99**, 2053-2081.

[4] Kawashima, T. and Fujisawa, H. (2017), Robust and Sparse Regression via γ-Divergence, *Entropy*.

[5] 森岡 優輝, 谷岡 健資, 宿久 洋 (2020), ガンマダイバージェンスを用いた確率的主成分分析, 日本計算機統計学会第 34 回大会講演論文集.

[6] Notsu, A., Eguchi, S., and Komori, O. (2017), Spontaneous learning for data distributions via minimum divergence.

[7] 甘利俊一 (2019), 情報幾何学の新展開, サイエンス社.