数理解析研究所講究録
第1910巻 2014年 29-42

Bayes予測における尤度とエントロピーの双対性

九州大学・経済学研究院　大西俊郎
Toshio Ohnishi
Faculty of Economics, Kyushu University

§1. Introduction.

本稿の目的は，尤度最大化とエントロピー最大化の間にある非常に興味深い関係を明らかにすることである．尤度最大化は統計学において基本的な原理である．標準的な教科書，例えばHogg *et al.* (2012, Ch.6)などに記述されているとおり，最尤推定量は漸近有効性などのよい性質をもつ．一方，Shannonエントロピーは情報学においては最も基本的な量であり (Cover & Thomas, 2006, Ch.8)，統計学では指数型分布族の特徴づけに用いられる (Kagan *et al.*, 1973, Ch.13)．エントロピー増大則（いわゆる第2法則）は熱力学において基本的な原理の1つである (Callen, 1985, Ch.1)．

本稿で対象とするモデルは次のようなものである．

1. Bayesモデル $p(x;\theta)\pi(\theta;c)$ において超事前分布 $\lambda(c)$ を仮定する場合．
2. モデル $p(x;\theta,\tau)$ において，まず θ に事前分布 $\pi(\theta|\tau)$ を仮定し，次に τ に事前分布 $\lambda(\tau)$ を仮定する場合．
3. 上の2つの両方を含む場合．

これらは実際のBayes分析でしばしば遭遇する状況である．

上記の3つの場合を統一的に取り扱うために，**Bayesian model averaging** (Hoeting

et al., 1999) の枠組みで考える．具体的には，インデックス ξ をもつ Bayes モデル

$$p_\xi(x;\theta)\pi_\xi(\theta), \quad \xi \in \Xi$$

を $\lambda(\xi)$ で平均化すると考える．モデル ξ に対する事前信念 $\lambda(\xi)$ を **prior averaging density** と呼ぶことにする．

Bayesian model averaging では，モデル ξ に対する事後信念が重要な役割を果たす．Bayes の定理により，事後信念は次のように求めることができる．

$$\lambda(\xi|x) = \frac{\lambda(\xi)m_\xi(x)}{m(x)}. \tag{1.1}$$

ただし，$m_\xi(x)$ および $\pi_\xi(\theta|x)$ はそれぞれモデル ξ における周辺密度および事後密度であり，$m(x) = \mathrm{E}\big[m_\xi(x) \,\big|\, \lambda(\xi)\big]$ は「全体」の周辺密度である．ここで $\mathrm{E}[f \,|\, p]$ は確率密度 p に関する f の期待値を表す．(1.1) の $\lambda(\xi|x)$ を **posterior averaging density** と呼ぶことにする．

本稿では **Bayes 予測問題**として問題を定式化する．Bayes 予測問題とは一言でいえば，推定問題の一般化である．Bayes モデル $p(x;\theta)\pi(\theta)$ において，将来の確率変数 y に対する確率密度 $p(y;\theta)$ を予測分布 $q(y|x)$ によって推定することである．$\hat{\theta}(x)$ を推定量とすると，推定問題では予測分布が $p\big(y;\hat{\theta}(x)\big)$ の形にに限定されることに注意されたい．

推定の良さを測るための損失関数として **α-divergence** を採用する．これは Kullback-Leibler divergence の一般化であり，次のように定義される．

$$\mathrm{D}_\alpha(p,q) := \mathrm{E}\left[u_\alpha\left(\frac{q}{p}\right) \,\middle|\, p\right].$$

ただし，

$$u_\alpha(r) := \begin{cases} -\log r & (\alpha = -1) \\ \frac{4}{1-\alpha^2}\left(1 - r^{\frac{1+\alpha}{2}}\right) & (-1 < \alpha < 1) \\ r\log r & (\alpha = 1) \end{cases} \tag{1.2}$$

である．粗く言えば，α-divergence は確率密度の比 q/p をべき乗したものの期待値ということができる．

Kullback-Leibler divergence $\mathrm{KL}(p,q) = \mathrm{E}[\log\{p/q\}\,|\,p]$ は非対称であること，すなわち，一般に $\mathrm{KL}(p,q) \neq \mathrm{KL}(q,p)$ であることが知られている．Amari & Nagaoka (2000) は「$\mathrm{KL}(q,p)$ と $\mathrm{KL}(p,q)$ は双対である」と表現している．本稿のタイトルにある「双対性」も Kullback-Leibler divergence がもつこの非対称性に起因するものである．Amari & Nagaoka (2000) に従い，記号の約束として $\alpha = 1$ を e，$\alpha = -1$ を m と書くことにすれば，

$$\mathrm{D}_e(p,q) = \mathrm{KL}(q,p) \ \text{ and } \ \mathrm{D}_m(p,q) = \mathrm{KL}(p,q)$$

となる．これらはそれぞれ **e-divergence** および **m-divergence** と呼ばれる．本稿では α を $+1$ から -1 まで変化させる．損失関数を変化させることによって尤度最大化と Shannon エントロピー最大化の関係が明らかになるからである．

§2. Formulating a Bayes risk minimization problem.

α-divergence 損失の下での Bayes 予測問題を Bayes リスク最小問題として定式化すると次のようになる．

$$\min \mathrm{E}\Big[\mathrm{D}_\alpha\big(p_\xi(y;\theta),\, q(y|x)\big) \;\Big|\; p_\xi(x;\theta)\pi_\xi(\theta)\lambda(\xi)\Big]. \tag{2.1}$$

Bayes の定理に関する等式 $p_\xi(x;\theta)\pi_\xi(\theta)\lambda(\xi) = \pi_\xi(\theta|x)\lambda(\xi|x)m(x)$ から，Bayes 予測問題 (2.1) は次のように等価変形できる．

$$\min \mathrm{E}\Big[\mathrm{D}_\alpha\big(p_\xi(y;\theta),\, q(y|x)\big) \;\Big|\; \pi_\xi(\theta|x)\lambda(\xi|x)\Big]. \tag{2.2}$$

参考としてモデル ξ における Bayes 予測問題を記しておく．

$$\min \mathrm{E}\Big[\mathrm{D}_\alpha\big(p_\xi(y;\theta),\, q(y|x)\big) \;\Big|\; \pi_\xi(\theta|x)\Big]. \tag{2.3}$$

モデル ξ における Bayes 予測問題 (2.3) の最適解は

- $-1 \leq \alpha < 1$ のとき

$$q_\xi^\alpha(y|x) \propto \Big\{\mathrm{E}\big[\{p_\xi(y;\theta)\}^{\frac{1-\alpha}{2}} \;\big|\; \pi_\xi(\theta|x)\big]\Big\}^{\frac{2}{1-\alpha}}$$

- $\alpha = 1$ のとき

$$q^e_\xi(y|x) \propto \exp\Big\{\mathrm{E}\big[\log p_\xi(y;\theta) \;\big|\; \pi_\xi(\theta|x)\big]\Big\}$$

によって与えられることが知られている (Aitchison, 1975; Corcuera & Giummole, 1999). これらは確率密度のさまざまな平均と言える. 例えば, $\alpha = -1$ のケース $q^m_\xi(y|x)$ は「算術平均」であり, $\alpha = 1$ のケース $q^e_\xi(y|x)$ は「幾何平均」である. このような平均については Hardy *et al.* (1988) が詳しいので参考文献として挙げておく. 最適解を記述するには α に関する場合分け「$-1 \le \alpha < 1$ および $\alpha = 1$」が必要であった. 以下の節において, この場合分けまたは別の場合分け「$\alpha = -1$ および $-1 < \alpha \le 1$」が頻出する.

事後リスク最小問題 (2.2) は次のように等価な最小問題に書き換えられる.

$$\min \mathrm{E}\Big[\mathrm{D}\big(q^e_\xi(y|x),\, q(y|x)\big) \;\Big|\; \lambda(\xi|x)\Big]. \tag{2.4}$$

この書き換えの根拠は次の等式である.

$$\mathrm{E}\Big[\mathrm{D}_\alpha\big(p_\xi(y;\theta),\, q(y|x)\big) \;\Big|\; \pi_\xi(\theta|x)\Big] = \mathrm{D}_\alpha\big(q^\alpha_\xi(y|x),\, q(y|x)\big) + (q(y|x) \text{ に依存しない項}).$$

詳細については Yanagimoto & Ohnishi (2009) を参照されたい. この書き換えが意味するところは, モデル ξ のすべての確率密度を考える必要はなく, モデル ξ の最適解のみ考えれば十分ということである.

事後リスク最小問題 (2.4) を少し一般化する. すなわち, posterior averaging density $\lambda(\xi|x)$ を一般の確率密度 $h(\xi)$ に置き換える.

$$\min \mathrm{E}\Big[\mathrm{D}_\alpha\big(q^\alpha_\xi(y|x), q(y|x)\big) \;\Big|\; h(\xi)\Big]. \tag{2.5}$$

確率密度 $h(\xi)$ を **canonical weight** と呼ぶことにする. 事後リスク最小問題 (2.4) の解は, (2.5) の解において $h(\xi)$ を $\lambda(\xi|x)$ に置き換えると得られる. 一般化しておくことのメリットは, 事前信念または事後信念をいろいろ変えたときにも対応できることである. 「正しい」$\lambda(\xi)$ は誰も知らないので, 「正しい」$\lambda(\xi|x)$ も分からないことに注意されたい.

次節以降において重要な役割を果たす予測分布を定義しておく.

Definition 1 (α-mixture). 次によって定義される予測分布 $f^\alpha(y|x;h)$ を **α-mixture** と呼ぶ. ただし, $K^\alpha_x(h)$ および $K^e_x(h)$ は規格化定数である.

- $-1 \leq \alpha < 1$ のとき

$$f^{\alpha}(y|x;h) := \frac{1}{K_x^{\alpha}(h)} \Big[\mathrm{E}\big[\{q_{\xi}^{\alpha}(y|x)\}^{\frac{1-\alpha}{2}} \mid h(\xi) \big] \Big]^{\frac{2}{1-\alpha}}.$$

- $\alpha = 1$ のとき

$$f^{e}(y|x;h) := \frac{1}{K_x^{e}(h)} \exp\Big\{ \mathrm{E}\big[\log q_{\xi}^{e}(y|x) \mid h(\xi) \big] \Big\}.$$

$f^{\alpha}(y|x;h)$ は h の汎関数になっていることに注意されたい．規格化定数 $K_x^{\alpha}(h)$ および $K_x^{e}(h)$ も汎関数である．

次節での議論において確率密度 $h(\xi)$ を変化させたときに α-mixture などの汎関数がどのように変化するかを計算することになる．そのために次の定義を与えておく．

Definition 2 (Gateaux differential). h_1, h_2 を確率密度とする．h_1 における増分 $h_2 - h_1$ に対する汎関数 $F(h)$ の **Gateaux 微分**を

$$\delta_G F(h_1; h_2 - h_1) = \lim_{\beta \to 0} \frac{F\big(h_1 + \beta(h_2 - h_1)\big) - F(h_1)}{\beta}$$

によって定義する．

h_1 に増分 $h_2 - h_1$ を足した後も確率密度になっていなければいけないことに注意されたい．その点で普通の Gateaux 微分の定義 (Luenberger, 1997, Ch.7) と異なっている．$h(\xi)$ が離散型のとき，Gateaux 微分は普通の偏微分になり，本稿の結果は Ohnishi & Yanagimoto (2013) に帰着する．

§3. 6 つの定理

本節では本稿の主張を 6 つの定理の形で述べる．まず初めに，前節で定義した α-mixture の最適性について述べる．

Theorem 1 (Optimal predictor). Definition 1 の α-mixture $f^{\alpha}(y|x;h)$ は，リス

ク最小問題 (2.5) の最適解である．特に，(1.1) の posterior averaging density $\lambda(\xi|x)$ を $h_x^*(\xi)$ と定義すると，$f^\alpha(y|x;h_x^*)$ は事後リスク最小問題 (2.4) の最適解である．

最適解は面白い等式を満たす．等式は 2 つの量のバランスを意味し，その一方は divergence の期待値である．

Theorem 2 (Average saddlepoint equality). Theorem 1 の最適解は次の等式を満たす．

- $\alpha = -1$ のとき

$$\mathrm{E}\Big[\mathrm{H}\big[f^m(y|x;h)\big] - \mathrm{H}\big[q_\xi^m(y|x)\big] - \mathrm{D}_m\big(q_\xi^m(y|x),\, f^m(y|x;h)\big) \;\Big|\; h(\xi)\Big] = 0.$$

  ただし，$\mathrm{H}[p] := \mathrm{E}[-\log p\,|p]$ は確率密度 p の Shannon エントロピーを表す．

- $-1 < \alpha \leq 1$ のとき

$$\mathrm{E}\left[u_{-\alpha}\left(\frac{q_\xi^\alpha(x|x)}{f^\alpha(x|x;h)}\right) - \mathrm{D}_\alpha\big(q_\xi^\alpha(y|x), f^\alpha(y|x;h)\big) \;\middle|\; h(\xi)\right] = 0.$$

Yanagimoto & Ohnishi (2009), Ohnishi & Yanagimoto (2013) に倣って Theorem 2 の等式を平均鞍点等式と呼ぶことにする．平均鞍点等式において divergence 損失と平均的にバランスしている量を **divergence 共役量**と呼ぶことにする．$\alpha = -1$ のとき，divergence 共役量は Shannon entropy 差

$$\mathrm{H}\big[f^m(y|x;h)\big] - \mathrm{H}\big[q_\xi^m(y|x)\big]$$

である．$-1 < \alpha \leq 1$ のとき，divergence 共役量は

$$u_{-\alpha}\left(\frac{q_\xi^\alpha(x|x)}{f^\alpha(x|x;h)}\right)$$

である．u_α ではなく，$u_{-\alpha}$ であることに注意されたい．$q_\xi^\alpha(x|x)$，$f^\alpha(x|x;h)$ は確率密度のデータ x における値である．そこでこれらを尤度と呼ぶことにする．関数 $u_{-\alpha}(r)$ の定義 (1.2) により，この関数は粗く言えば「べき乗」関数である．したがって，$-1 < \alpha \leq 1$ のとき，divergence 共役量は尤度比の「べき乗」と言ってよい．

リスク最小問題 (2.5) の最小値は重要な役割を果たす．この最小値を $-\psi_x^\alpha(h)$ とおく．これは canonical weight $h(\xi)$ の汎関数である．Theorem 2 から，canonical weight に関する divergence 共役量の期待値がリスク最小問題の最小値と一致することが分かる．この事実を利用してリスクの最小値を計算すると次のようになる．

- $\alpha = -1$ のとき，$-\psi_x^m(h) = \mathrm{E}\Big[\mathrm{H}\big[q_\xi^m(x|x)\big] \;\Big|\; h(\xi)\Big] - \mathrm{H}\big[f^m(y|x;h)\big]$.
- $-1 < \alpha \leq 1$ のとき，$-\psi_x^\alpha(h) = u_{-\alpha}\big(K_x^\alpha(h)\big)$.

次に canonical weight と対になる概念を導入する．

Definition 3 (Mean weight). 次によって定義される量 $t_x^\alpha(\xi;h)$ を **mean weight** と呼ぶ．

- $-1 \leq \alpha < 1$ のとき

$$t_x^\alpha(\xi;h) := -\mathrm{D}_\alpha\big(q_\xi^\alpha(y|x), f^\alpha(y|x;h)\big) + u_\alpha\big(f^\alpha(x|x;h)\big). \tag{3.1}$$

- $\alpha = 1$ のとき

$$t_x^e(\xi;h) := -\mathrm{D}_e\big(q_\xi^e(y|x),\, f^e(y|x;h)\big) - \mathrm{H}\big[f^e(y|x;h)\big]. \tag{3.2}$$

Mean weight はリスクの最小値 $-\psi_x^\alpha(h)$ の Gateaux 微分に現れる．

$$\delta_G\psi_x^\alpha(h_1; h_2 - h_1) = \mathrm{E}\big[t_x^\alpha(\xi;h_1) \;\big|\; h_2(\xi) - h_1(\xi)\big].$$

これは指数型分布族の正準パラメータと平均パラメータの関係と同じである．Amari & Nagaoka (2000) は，正準パラメータと平均パラメータの関係を双対と呼んでいる．これに倣えば，「canonical weight と mean weight は双対である」と表現することができる．

リスク最小問題 (2.5) は制約条件なしの最小問題である．これを等価な制約条件つき最大問題に書き換える．ここでの等価性は 2 つの問題が同一の最適解をもつことを意味する．このような等価性の定義は Courant & Hilbert (1989, Ch. 4) に見られる．

Theorem 3 (Equivalent maximization problem with constraints). リスク最

小問題 (2.5) と次の問題は，$s(\xi) = t_x^\alpha(\xi; h)$ のときに限り，同一の最適解 $f^\alpha(y|x; h)$ をもつ．

- $-1 \leq \alpha < 1$ のとき

$$\begin{aligned} &\max \ -u_\alpha\big(q(x|x)\big) \\ &\text{s.t.} \ -\mathrm{D}_\alpha\big(q_\xi^\alpha(y|x), q(y|x)\big) + u_\alpha\big(q(x|x)\big) = s(\xi) \end{aligned}$$

- $\alpha = 1$ のとき

$$\begin{aligned} &\max \mathrm{H}\big[q(y|x)\big] \\ &\text{s.t.} \ -\mathrm{D}_e\big(q_\xi^\alpha(y|x), q(y|x)\big) - \mathrm{H}\big[q(y|x)\big] = s(\xi) \end{aligned}$$

証明の本質だけの述べることにする．次のような一般的な問題を考えよう．$d(A, B)$ を 2 つの点 A, B の乖離度とし，X をいろいろ動かして $d(A, X)$ と $d(B, X)$ を同時に小さくしたいとする．この問題のアプローチとして次の 2 つ

1. 適当に h を決め，$(1-h)d(A, X) + hd(B, X)$ を最小化する．
2. $d(B, X) - d(A, X) = t$ を固定し，$d(A, X)$ を最小化する．

が考えられ，両者は Lagrange の未定乗数法で結ばれている．つまり，Theorem 3 は Lagrange の未定乗数法の逆プロセスを行ったということである．

Theorem 3 は，状況に応じて原理を等価変形する熱力学の原理に似ている．熱力学において平衡状態は次のように特徴づけられることが知られている (Callen, 1985, Ch.5).

- Energy minimum principle:
  エントロピーが一定のとき，平衡状態では内部エネルギーが最小化される．
- Helmholtz potential minimum principle:
  温度が一定のとき，平衡状態では Helmholtz potential が最小化される．

Theorem 1 において canonical weight の 1 つとして posterior averaging density $h_x^*(\xi) = \lambda(\xi|x)$ を考えた．以下，別の 2 つの canonical weight を考え，それらに関する定理を述べる．まず 1 つ目として，divergence 共役量を最大化する（停留させる）canonical weight を $h_x^{\alpha\dagger}(\xi)$ と定義する．関数 $u_{-\alpha}(r)$ の単調性などから divergence 共役量の最大化は，

- $\alpha = -1$ のとき，Shannon エントロピー $\mathrm{H}\big[f^m(y|x;h)\big]$ の最大化と一致し，
- $-1 < \alpha \leq 1$ のとき，尤度 $f^\alpha(x|x;h)$ の最大化と一致する．

もう 1 つとして，リスクの最小値 $-\psi_x^\alpha(h)$ を最大化する（停留させる）canonical weight を $h_x^{\alpha c}(\xi)$ と定義する．最小値の最大化というアイデアは Courant & Hilbert (1989, Ch. 4) に見られる．

Canonical weight $h_x^*(\xi) = \lambda(\xi|x)$ は divergence 共役量と divergence を平均的にバランスさせた．一方，$h_x^{\alpha\dagger}(\xi)$ は，divergence 共役量と divergence を exact に一致させる．

Theorem 4 (Exact saddlepoint equality). Canonical weight $h_x^{\alpha\dagger}(\xi)$ に対応する α-mixture は次の等式を満たす（厳密鞍点等式と呼ぶことにする）．

- $\alpha = -1$ のとき

$$\mathrm{H}\big[f^m(y|x;h_x^{m\dagger})\big] - \mathrm{H}\big[q_\xi^m(y|x)\big] = \mathrm{D}_m\big(q_\xi^m(y|x),\, f^m(y|x;h_x^{m\dagger})\big).$$

- $-1 < \alpha < 1$ のとき

$$u_{-\alpha}\left(\frac{q_\xi^\alpha(x|x)}{f^\alpha(x|x;h_x^{\alpha\dagger})}\right) = \mathrm{D}_\alpha\big(q_\xi^\alpha(y|x),\, f^\alpha(y|x;h_x^{\alpha\dagger})\big).$$

Divergence 共役量と divergence を次の意味でバランスさせる予測分布 $f^\alpha(y|x;h)$ の集合 $\mathcal{Q}^\alpha$ を考える．

- $\alpha = -1$ のとき

$$\mathrm{E}\Big[\mathrm{H}\big[f^m(y|x;h)\big] - \mathrm{H}\big[q_\xi^m(y|x)\big] - \mathrm{D}_m\big(q_\xi^m(y|x),\, f^m(y|x;h)\big) \;\Big|\; \lambda(\xi|x)m(x)\Big] = 0. \tag{3.3}$$

- $-1 < \alpha \leq 1$ のとき

$$\mathrm{E}\left[u_{-\alpha}\left(\frac{q_\xi^\alpha(x|x)}{f^\alpha(x|x;h)}\right) - \mathrm{D}_\alpha\big(q_\xi^\alpha(y|x),\, f^\alpha(y|x;h)\big) \;\middle|\; \lambda(\xi|x)m(x)\right] = 0. \tag{3.4}$$

確率密度 $\lambda(\xi|x)m(x)$ で期待値を計算しているので，$\mathcal{Q}^\alpha$ を定義する等式 (3.3) および (3.4) は，Theorem 1 の平均鞍点等式および Theorem 4 の厳密鞍点等式より弱い．強い等式から弱い等式へ並べると，厳密鞍点等式, 平均鞍点等式, (3.3) および (3.4) の順番に

なる．したがって，2 つの canonical weight $h_x^*(\xi)$ および $h_x^{\alpha\dagger}(\xi)$ に対応する α-mixture は $\mathcal{Q}^\alpha$ に属する．次の定理はこの両者が $\mathcal{Q}^\alpha$ の中で両極端な予測分布になっていることを主張している．

Theorem 5 (Best & worst). 最適解 $f^\alpha\big(y|x;h_x^*\big)$ は $\mathcal{Q}^\alpha$ に属し，$f^\alpha\big(y|x;h_x^{\alpha\dagger}\big)$ は $\mathcal{Q}^\alpha$ の中で最悪である．

$\mathcal{Q}^\alpha$ が最適解 $f^\alpha(y|x;h_x^*)$ を含むこと，および，$\mathcal{Q}^\alpha$ の定義式がある種の不偏性を意味するので，$\mathcal{Q}^\alpha$ は「優秀な」予測分布の集合と考えてもよいように思われる．Yanagimoto & Ohnishi (2011, 2013) はこの不偏性に着目して情報量基準を考察している．予測分布 $f^\alpha(y|x;h_x^*)$ は最適であるが，その最適性は「正しい」posterior averaging density の選択に依存している．「間違った」posterior averaging density を選択すると最適でないどころか，「優秀クラス」にさえ入らない可能性がある．それに対して予測分布 $f^\alpha\big(y|x;h_x^{\alpha\dagger}\big)$ は posterior averaging density の選択によらず「優秀クラス」に入る．ただし，常に「優秀クラス」の「ビリ」である．

次に $h_x^{\alpha\mathrm{c}}(\xi)$ に対応する α-mixture の性質を述べる．予測分布 $f^\alpha(y|x;h_x^{\alpha\mathrm{c}})$ を用いると，ある種の頑健性が保証される．

Theorem 6 (Constant risk). 任意の canonical weight $h(\xi)$ に対して

$$\mathrm{E}\Big[\mathrm{D}_\alpha\big(q_\xi^e(y|x),\, f^\alpha(y|x;h_x^{\alpha\mathrm{c}})\big)\;\Big|\;h(\xi)\Big]$$

は一定である．

以上の 6 つの定理をまとめておく．一言で表現するならば，尤度最大化と Shannon entropy 最大化は双対である．

	最良	最悪	リスク一定
$\alpha=-1=m$	maxL'	maxE	maxminR
$-1<\alpha<1$	maxL'	maxL	maxminR
$\alpha=1=e$	maxE'	maxL	maxminR

ただし，記号の意味は次のとおりである．

- maxL: $f^{\alpha}(x|x;h)$ の最大化
- maxL': $q(x|x)$ の制約つき最大化
- maxE: $\mathrm{H}\big[f^{\alpha}(y|x;h)\big]$ の最大化
- maxE': $\mathrm{H}\big[q(y|x)\big]$ の制約つき最大化
- maxminR: リスク最小値 $-\psi_x^{\alpha}(h)$ の最大化

§4. Implications

4.1 Information criteria.

情報量基準との関連を論じる．第 1 節で述べたように canonical weight はモデルに対する信念である．BIC (Schwarz, 1978) は $h_x^*(\xi) = \lambda(\xi|x)$ の近似値として得られる．Mean weight の定義式 (3.1) および (3.2) から，mean weight は粗く言えば予測分布とモデル ξ との距離である．$p_\xi(x;\theta)$ が指数型分布族の場合，$h_x^{e\dagger}(\xi)$ に対する mean weight は

$$t_x^e(\xi; h_x^{e\dagger}) = (\text{最大対数尤度}) - (\text{罰則項})$$

となり，AIC (Akaike, 1973) に似ている．$p_\xi(y;\theta)$ がいわゆる混合分布の場合，$h_x^{m\dagger}(\xi)$ に対する mean weight は上と双対な形

$$t_x^m(\xi; h_x^{m\dagger}) = (\text{最大 Shannon entropy}) - (\text{罰則項})$$

になる．

4.2 An empirical Bayes method.

経験 Bayes 法の 1 つは，周辺尤度 $m_\xi(x)$ の最大化する $\hat{\xi}_M$ を求め，$q_\xi^e(y|x)$ または $q_\xi^m(y|x)$ に plug-in することである．これは canonical weight として Dirac の δ 関数

$$h(\xi) = \delta(\xi - \hat{\xi}_M)$$

を使うことに相当する．経験 Bayes 法は，Theorem 3 で考察した最大化，すなわち，mean weight を指定し，$\mathrm{H}\big[q(y|x)\big]$ または $\log q(x|x)$ を最大化することとは一般に異なることに注意されたい．

4.3 Semi-group structure.

Bayes リスク最小問題は半群の構造を持っている．§2 において，「全体」の最適化問題 (2.2) が「部分」の最適解を用いた最適化問題 (2.4) に等価変形できることを示した．最適解 $f^{\alpha}\big(y|x;h_x^*\big)$ は「全体」の平均になっていることを示すことができる．一方，$f^{\alpha}\big(y|x;h_x^*\big)$ は Theorem 1 で見たように「部分の平均」の平均になっていた．これらの状況は，相転移を論じる統計物理学における**くりこみ群**と似ている．

4.4 Generalization of likelihood and Shannon entropy.

§3 の議論は，通常の Bayes モデル $p(x;\theta)\pi(\theta)$ にも適用できる．そのためにはデルタ関数を用い，Bayes モデル $p(x;c)\delta(c-\theta)$ を $\pi(\theta)$ で平均化すると考えればよい．この考え方は，尤度の拡張につながる．事前分布 $\pi(\theta)$ の汎関数という見方である．具体的には，

$$f^{e}(x;\pi)=\frac{1}{K^{e}(\pi)}\exp\big\{\mathrm{E}[p(x;\theta)|\pi(\theta)]\big\}$$

を拡張尤度とみなすのである．同様に Shannon エントロピーの概念も拡張できる．標本分布として $p(y;\theta)$ を，事前分布として $\pi(\theta)$ を仮定したときの Shannon エントロピーを

$$\mathrm{H}[f^{m}(y;\pi)]=\mathrm{H}\big[\mathrm{E}[p(y;\theta)|\pi(\theta)]\big]$$

によって定義するのである．

REFERENCES

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267-281 in B.N. Petrov and F. Csaki (editors) *Second International Symposium on Information Theory.* Akademiai Kiado, Budapest.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547-554.

Amari, S-I. and Nagaoka, H. (2000). *Methods of Information Geometry.* American Mathematical Society, Load Island.

Callen, H. B. (1985). *Thermodynamics and an Introduction to Thermostatistics, 2nd ed.* John Wiley & Sons.

Corcuera, J.M. and Giummole, F. (1999). A generalized Bayes rule for prediction. *Scandinavian Journal of Statistics*, **26**, 265-279

Courant, R. and Hilbert, D. (1989). *Methods of Mathematical Physics.* Wiley-VCH.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory, 2nd ed.* Wiley-Interscience.

Hardy, G. H., Littlewood, J. E. and Polya, G. (1988). *Inequalities.* Cambridge University Press.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382-417.

Hogg, R. V., McKean, J., and Craig, A. T. (2012). *Introduction to Mathematical Statistics, 7th ed.*. Prentice Hall.

Kagan, A. M., Linnik, Y. V. and Rao, C. R. (1973). *Characterization Problems in Mathematical Statistics.* John Wiley & Sons, New York.

Luenberger, D. G. (1997). *Optimization by Vector Space Methods.* Wiley-Interscience.

Ohnishi, T. and Yanagimoto, T. (2013). Twofold structure of duality in Bayesian model averaging. *Journal of the Japan Statistical Society*, **43**, 29-55.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Yanagimoto, T. and Ohnishi, T., (2009). Bayesian prediction of a density function

in terms of e-mixture. *Journal of Statistical Planning and Inference*, **139**, 3064-3075.

Yanagimoto, T. and Ohnishi, T., (2011). Saddlepoint condition on a predictor to reconfirm the need for the assumption of a prior distribution. *Journal of Statistical Planning and Inference*, **141**, 1990-2000.