1A1-4

高次元確率空間における高精度期待値ベイズ推定の検討

A Bayesian estimation of statistical expectation in a high-dimensional probability space

松田 衆治 Shuji Matsuda　Nguyen Ha Hon　鷲尾 隆 Takashi Washio　河原 吉伸 Yoshinobu Kawahara　清水 昌平 Shohei Shimizu　猪口 明博 Akihiro Inokuchi

大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka Univ.

Along development of computer and sensing technology, the requirements to analyze super-high-dimensional data and to estimate its associated expectation value are increased. However, in a super-high-dimensional probability space, the curse of dimensionality makes hard to estimate statistical expectation by conventional methods. In this paper, we analyze the mechanism of the curse of dimensionality, and investigate a Bayesian estimation method of statistical expectation in a high-dimensional probability space based on the analysis.

1. 序論

データの要約や解釈をおこなう統計的手段として，データ集合の標本分布と，そのデータの母集団確率密度に関する背景知識に基づくデータの尤度から，データの標本分布のみに基づくよりも妥当な母集団確率密度を推定する事が考えられる．このように標本分布が与える事前確率密度と背景知識に基づく尤度から得られる確率密度を事後確率密度と呼ぶ．本稿では，特に事後確率密度そのものよりも，その平均期待値を推定する問題を取り上げる．

一方，近年の計算機技術や計測技術の発展に伴い，超高次元データを扱う機会が増えている．また，それと同時に超高次元データに関する期待値推定が必要となる機会も増えている．例えば，コンビニエンスストアなどの小売店においては，顧客が商品を購入した日時，商品名，その時の天候，曜日，原価など多くの項目を含む高次元購買履歴データが蓄積されている．そこで，背景知識に基づき個々の商品の需要予測を行い，その予測と実際の購買履歴データを用いて，さらに精度の高い需要予測を行う事は，逐次適切な商品の仕入総量を決める上で非常に有益である．

しかしながら、後述するように次元の呪いと呼ばれる現象により，有限な標本分布は母集団分布をうまく反映せず，また，各事例間の確率密度比は低次元の場合に比べて非常に大きい．これらの効果により，従来手法では期待値を高精度に推定することができない場合が多い．そこで，本研究ではそのような精度低下が起こる機構を考察し，次にそれの問題軽減手法を提案し，更にその手法の性能に関して評価実験と考察を行う．

2. 研究の背景と目的

2.1 期待値ベイズ推定問題

ある p 次元確率空間 $\Omega \in R^p$ において，ある未知の母集団確率密度関数 $P_D(\mathbf{X})(\mathbf{X} \in \Omega)$ に従って分布するデータ集合 $D = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n\}$ が与えられた時，背景知識に基づく尤度関数 $P_L(\mathbf{X})$ と，データ集合 D を用いて事後確率密度に関する平均期待値 $\mathbf{F}$ を推定する問題を期待値ベイズ推定問題と呼ぶ．ここで，$\mathbf{F}$ は以下のように定義される．

$$\mathbf{F} = \frac{1}{\int_\Omega P_L(\mathbf{X}) * P_D(\mathbf{X}) d\mathbf{X}} \int_\Omega \mathbf{X} P_L(\mathbf{X}) * P_D(\mathbf{X}) d\mathbf{X} \quad (1)$$

2.2 次元の呪い

データの次元が非常に大きい場合に**「次元の呪い」**という現象が発生する事が指摘されている [Silverman 86] が，その詳細な内容にはまだ十分な研究が進んでいない．これに対して我々は超高次元で以下の現象が起こる事を発見した．p 次元確率空間 Ω の一点 $\mathbf{X} = [x_1, x_2, ...x_p]^T$ の原点からの 2 乗距離は以下で与えられる．

$$|\mathbf{X}|^2 = \sum_{k=1}^{p} x_k^2$$

また，$\mathbf{X}$ が Ω 内に確率密度関数 $P(\mathbf{X})$ に従って分布し，かつ $P(\mathbf{X})$ が各確率変数 x_k について平均 0，標準偏差 σ の独立同一分布をする以下の p 次元正規分布とする．

$$P(\mathbf{X}) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_i^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^p e^{-\frac{|\mathbf{X}|^p}{2\sigma^2}}$$

これより，以下は自由度 p のカイ二乗分布に従い，平均 p，標準偏差 $\sqrt{2p}$ である．

$$\frac{|X|^2}{\sigma^2} = \sum_{k=1}^{p} \left(\frac{x_k}{\sigma^2}\right)^2$$

従って，点 $\mathbf{X}$ の原点からの距離 $|\mathbf{X}|^2$ の平均は $p\sigma^2$，標準偏差は $\sqrt{2p}\sigma^2$ であり，平均と標準偏差の比率は

$$\frac{\text{標準偏差}}{\text{平均}} = \frac{\sqrt{2p}\sigma^2}{p\sigma^2} = \sqrt{\frac{2}{p}} \to 0 \quad (p \to \infty)$$

となり $p \to \infty$ において相対的バラツキは零に収束する．よって，超高次元確率空間で各次元の分散が σ の多次元ガウス分布からサンプリングした事例は半径 $\sqrt{p}\sigma$ の超球面上にその半径に比して相対的に集中する事が分かる．このように超高次元確率空間では標本分布が母集団分布をうまく反映しない場合が多い．この現象を本稿では**次元の呪い 1** と呼ぶ．

さらに，我々は超高次元確率空間にて相違なる 2 点の確率密度比について起こる以下の現象を発見した．ただし，この現

連絡先: 松田 衆治，大阪大学 産業科学研究所，567-0047 大阪府茨木市美穂ヶ丘 8-1，smatsuda@ar.sanken.osaka-u.ac.jp

象は [Leeuwen 03],[Snyder 08] によって部分的に指摘されている．p 次元確率空間 Ω の確率変数 $\mathbf{X} = [x_1, x_2, ...x_n]^T$ の確率密度関数を $P(\mathbf{X})$ とし，今までの説明と同様に各次元平均 0, 標準偏差 σ のガウス分布の独立同一分布とする．$P(\mathbf{X})$ に従う相違なる 2 点 $\mathbf{X_i}, \mathbf{X_j}(i, j = 1, 2, ..., n)$ の確率密度比

$$\frac{P(\mathrm{X}_i)}{P(\mathrm{X}_j)} = \exp\left\{\frac{1}{2}\left(\frac{|\mathbf{X}_i|^2}{\sigma^2} - \frac{|\mathbf{X}_j|^2}{\sigma^2}\right)\right\}$$

の両辺の対数をとり，二乗すると次式になる．

$$(\log P(\mathbf{X}_i) - \log P(\mathbf{X}_j))^2 = \left\{\frac{1}{2}\left(\frac{|\mathbf{X}_i|^2}{\sigma^2} - \frac{|\mathbf{X}_j|^2}{\sigma^2}\right)\right\}^2$$

前述の次元の呪い 1 の説明より $\dfrac{|\mathbf{X}_i|^2}{\sigma^2} = \displaystyle\sum_{k=1}^{p}\left(\frac{x_k^i}{\sigma}\right)^2$ は自由度 p のカイ二乗分布に従い，その平均は p，標準偏差は $\sqrt{2p}$ である．よって，$\dfrac{1}{2}\left(\dfrac{|\mathbf{X}_i|^2}{\sigma^2} - \dfrac{|\mathbf{X}_j|^2}{\sigma^2}\right)$ は平均が 0, 標準偏差が $\sqrt{p}$ のカイ二乗分布に従う．$p \to \infty$ の時，このカイ二乗分布は正規分布 $\mathrm{N}(0, p)$ に近づくので，

$$\left(\frac{\log P(\mathbf{X}_i) - \log P(\mathbf{X}_j)}{\sqrt{p}}\right)^2 = \left\{\frac{1}{2\sqrt{p}}\left(\frac{|\mathbf{X}_i|^2}{\sigma^2} - \frac{|\mathbf{X}_j|^2}{\sigma^2}\right)\right\}^2$$

は自由度 1 のカイ二乗分布し，平均 1, 標準偏差 $\sqrt{2}$ である．これより $(\log P(\mathbf{X}_i) - \log P(\mathbf{X}_j))^2$ は平均 p，標準偏差 $\sqrt{2}p$ を有する．従って，

$$E[(\log P(\mathbf{X}_i) - \log P(\mathbf{X}_j))^2] = p \to \infty$$

となり，$p \to \infty$ の時，相違なる 2 点 X_i と X_j うちどちらか一方が少しでも確率密度が高い位置にある時，2 点の密度比は無限に大きくなる．このように確率空間の次元が大規模であるほど，事例の尤度や重みの違いは大きくなり，少数または 1 個の事例データの重みが支配的になる．本稿ではこの現象を次元の呪い 2 と呼ぶ．

2.3 従来法とその問題点

従来の標準的方法である標本分布法では，まずデータ集合 D に基づく $P_D(\mathbf{X})$ の標本分布近似である

$$\tilde{P}_D(X) = \frac{1}{n}\sum_{\mathbf{X}_i \in D}\delta(\mathbf{X} - \mathbf{X}_i)$$

を用い，F のデータ分布による推定値

$$\begin{aligned}\tilde{\mathbf{F}} &= \frac{1}{\int_\Omega P_L(\mathbf{X}) * \tilde{P}_D(\mathbf{X})d\mathbf{X}}\int_\Omega \mathbf{X}P_L(\mathbf{X}) * \tilde{P}_D(\mathbf{X})d\mathbf{X} \\ &= \frac{1}{\int_\Omega P_L(\mathbf{X})\frac{1}{n}\sum_{\mathbf{X}_i \in D}\delta(\mathbf{X} - \mathbf{X}_i)d\mathbf{X}} \\ &\times \int_\Omega \mathbf{X}P_L(\mathbf{X})\frac{1}{n}\sum_{\mathbf{X}_i \in D}\delta(\mathbf{X} - \mathbf{X}_i)d\mathbf{X} \\ &= \frac{1}{\sum_{\mathbf{X}_i \in D} P_L(\mathbf{X}_i)}\sum_{\mathbf{X}_i \in D}\mathbf{X}_i P_L(\mathbf{X}_i)\end{aligned}$$

を計算する．

超高次元確率空間中で $P_D(\mathbf{X})$ が分散 σ_D の多次元ガウス分布である場合，次元の呪い 1 により D に含まれるデータのほとんどは半径 $\sqrt{p}\sigma_D$ の超球表面近傍にしか存在せず，さらに次元の呪い 2 により他のデータより，たまたま尤度 $P_L(\mathbf{X}_i)$ の大きなデータ $\mathbf{X}_i$ が推定値 $\tilde{\mathbf{F}}$ を支配してしまう．そのため，$P_D(\mathbf{X})$ の平均から $\sqrt{p}\sigma_D$ だけ離れ，かつたまたま尤度 $P_L(\mathbf{X})$ の大きなデータ $\mathbf{X}_i$ が推定値として得られてしまい，正しい推定を行うことができない． このように超高次元確率空間では従来の標準的方法である標本分布法によっては高精度な平均期待値推定を行うことが難しい．

2.4 従来の問題軽減手法

本稿では，このような次元の呪いの問題を軽減する従来手法として以下で述べる周辺分布法 [Everitt 02] と呼ばれる手法を説明する．$P_L(\mathbf{X})$ と $P_D(\mathbf{X})$ の積 $P_L(\mathbf{X})P_D(\mathbf{X})$ が変数分離形，すなわち各次元 x_j(j=1,2.....,p) に関する確率密度の積

$$P_L(\mathbf{X})P_D(\mathbf{X}) = \prod_{j=1}^{p} P_L(x_j)P_D(x_j)$$

で表すことができるとする．この仮定と式 (1) より平均期待値ベクトル $\mathbf{F}$ の j 次元目の要素 F_j は

$$\begin{aligned}F_j &= \frac{1}{\int_\Omega P_L(\mathbf{X})P_D(\mathbf{X})d\mathbf{X}}\int_\Omega x_j P_L(\mathbf{X})P_D(\mathbf{X})d\mathbf{X} \quad (2) \\ &= \frac{1}{\int_\Omega \prod_{i=1}^{p} P_L(x_i)P_D(x_i)dx_1dx_2.....dx_p} \\ &\times \int_\Omega x_j \prod_{i=1}^{p} P_L(x_i)P_D(x_i)dx_1dx_2.....dx_p \\ &= \frac{1}{\int_\Omega P_L(x_j)P_D(x_j)dx_j}\int_\Omega x_j P_L(x_j)P_D(x_j)dx_j \quad (3)\end{aligned}$$

で与えられる．この式 (3) に $P_D(x_j)$ の標本分布近似

$$\tilde{P}_D(x_j) = \frac{1}{n}\sum_{x_{ij} \in D_j}\delta(x_j - x_{ij})$$

を代入し，以下の近似値 $\tilde{F}_j$ を得る．ただし，データ集合 D 中の各データの j 次元の要素のみから成る集合を $D_j = \{x_{1j}, ..., x_{nj}\}$ とする．

$$\begin{aligned}\tilde{F}_j &= \frac{1}{\int_\Omega P_L(x_j)\frac{1}{n}\sum_{x_{ij} \in D_j}\delta(x_j - x_{ij})dx_j} \\ &\times \int_\Omega x_j P_L(x_j)\frac{1}{n}\sum_{X_i \in D}\delta(x_j - x_{ij})dx_j \\ &= \frac{1}{\sum_{x_{ij} \in D_j} P_L(x_{ij})}\sum_{x_{ij} \in D_j} x_{ij}P_L(x_{ij}) \\ &= \sum_{x_{ij} \in D} x_{ij}w_{ij}(x_{ij})\end{aligned}$$

ただし，

$$w_{ij}(x_{ij}) = \frac{P_L(x_{ij})}{\sum_{x_{ij} \in D_j} P_L(x_{ij})}$$

とする．D_j は D と違って次元の呪いの影響を受けず，母集団分布をより反映した分布を持つと期待される．また，$P_L(x_j)$ の比は $P_L(\mathbf{X})$ の比ほど大きくはならず，少数の点が推定値の中で支配的になる事はない．したがって，D_j のベイズ推定平均期待

値は $\sum_{x_{ij}\in D_j} x_{ij}w_{ij}(x_{ij})$ を計算することで高精度に推定することができる．他の各次元についても同様に計算することで，$\mathbf{F}$ 全体の平均期待値を高精度に推定することができる．

以上で説明したように周辺分布法では，$P_L(\mathbf{X})P_D(\mathbf{X})$ が変数分離形か，またそれに近い関数形である場合，高精度に平均期待値を推定することができる．一方，それ以外の場合には式 (2) と式 (3) の等号関係が成立しないので，各次元の平均期待値を正確に推定することができず，ベクトル $\mathbf{F}$ の推定精度は悪化すると考えられる．

3. 提案手法

本研究では次元の呪いの問題を軽減する手法として，以下で述べるプロポーザル分布法を提案する．人為的に $P(X) = P_L(\mathbf{X})P_D(\mathbf{X})$ の高確率密度領域に一定数以上のデータを集中させたデータ集合 $D' = \{\mathbf{X}'_1, \mathbf{X}'_2, ...\mathbf{X}'_n\}$ をデータ集合 $D = \{\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_n\}$ と既知確率密度関数 $P_L(\mathbf{X})$ を基に作成する．D' に含まれるデータの分布を母集団分布 $Q_{D'}(\mathbf{X})$ の標本分布とする．$Q_{D'}(\mathbf{X})$ をプロポーザル分布と呼ぶ．期待値 $\mathbf{F}$ はこの $Q_{D'}(\mathbf{X})$ を用いて以下のように書き換えることができる．

$$\begin{aligned}\mathbf{F} &= \frac{1}{\int_\Omega P_L(\mathbf{X})P_D(\mathbf{X})d\mathbf{X}}\int_\Omega \mathbf{X}P_L(\mathbf{X})P_D(\mathbf{X})d\mathbf{X}\\ &= \frac{1}{\int_\Omega \frac{P_L(\mathbf{X})P_D(\mathbf{X})}{Q_{D'}(\mathbf{X})}Q_{D'}(\mathbf{X})d\mathbf{X}}\\ &\times \int_\Omega \mathbf{X}\frac{P_L(\mathbf{X})P_D(\mathbf{X})}{Q_{D'}(\mathbf{X})}Q_{D'}(\mathbf{X})d\mathbf{X}\\ &= \frac{1}{\int_\Omega R(\mathbf{X})Q_{D'}(\mathbf{X})d\mathbf{X}}\int_\Omega \mathbf{X}R(\mathbf{X})Q_{D'}(\mathbf{X})d\mathbf{X} \quad (4)\end{aligned}$$

ただし，$R(\mathbf{X})$ を

$$R(\mathbf{X}) = \frac{P_L(\mathbf{X})P_D(\mathbf{X})}{Q_{D'}(\mathbf{X})} \quad (5)$$

とする．データ集合 D は既知であり，それから生成した D' も既知であるが，$P_D(\mathbf{X})$ および $Q_{D'}(\mathbf{X})$ は未知である．そこで，期待値 $\mathbf{F}$ の推定値を得るため $P_D(\mathbf{X})$，$Q_{D'}(\mathbf{X})$ を何らかの方法で近似する必要がある．人為的に作成したデータ $\mathbf{X}'_i \in D'$ を推定に用いるため，式（4）中の $Q_{D'}(\mathbf{X})$ には標本分布近似

$$\tilde{Q}_{D'}(\mathbf{X}) = \frac{1}{n}\sum_{\mathbf{X}_i\in D'}\delta(\mathbf{X}-\mathbf{X}_i) \quad (6)$$

を用いる．ただし，標本分布近似は確率密度の実値を与えるものではないので，式 (5) 中の $P_D(\mathbf{X})$ や $Q_{D'}(\mathbf{X})$ に標本分布近似を用いると，確率密度比 $R(\mathbf{X})$ の誤差が過大となってしまう．故に，式 (5) 中の $P_D(\mathbf{X})$ や $Q_{D'}(\mathbf{X})$ には以下のノンパラメトリックカーネル近似 $\hat{P}_D(\mathbf{X})$, $\hat{Q}_{D'}(\mathbf{X})$ を用いる [Silverman 86].

$$\hat{P}_D(\mathbf{X}) = \frac{1}{n}\sum_{X_j\in D}K(\mathbf{X}-\mathbf{X}_j, h)$$

$$\hat{Q}_{D'}(\mathbf{X}) = \frac{1}{n}\sum_{X'_j\in D'}K(\mathbf{X}-\mathbf{X'}_j, h')$$

ただし，h と h' はカーネル関数幅であり，それぞれデータ集合 D 及び D' から決められる．この $\hat{P}_D(\mathbf{X})$, $\hat{Q}_{D'}(\mathbf{X})$ を用いて $R(\mathbf{X})$ を以下のように近似する．

$$\hat{R}(\mathbf{X}) = \frac{P_L(\mathbf{X})\hat{P}_D(\mathbf{X})}{\hat{Q}_{D'}(\mathbf{X})}$$

この $\hat{R}(\mathbf{X})$ と式 (6) を用いて期待値 $\mathbf{F}$ のプロポーザル分布法による以下の推定値を得る．

$$\begin{aligned}\hat{\mathbf{F}} &= \frac{1}{\int_\Omega \hat{R}(\mathbf{X})\frac{1}{n}\sum_{\mathbf{X}_i\in D'}\delta(\mathbf{X}-\mathbf{X}_i)d\mathbf{X}}\\ &\times \int_\Omega \mathbf{X}\hat{R}(\mathbf{X})\frac{1}{n}\sum_{\mathbf{X}_i\in D'}\delta(\mathbf{X}-\mathbf{X}_i)d\mathbf{X}\\ &= \frac{1}{\sum_{\mathbf{X}'_i\in D'}\hat{R}(\mathbf{X}'_i)}\sum_{\mathbf{X}'_i\in D'}\mathbf{X}'_i\hat{R}(\mathbf{X}'_i)\\ &= \sum_{\mathbf{X}'_i\in D'}w_p(\mathbf{X}'_i)\mathbf{X}'_i \quad (7)\end{aligned}$$

ただし，

$$w_p(\mathbf{X}'_i) = \frac{\hat{R}(\mathbf{X}'_i)}{\sum_{\mathbf{X}'_i\in D'}\hat{R}(\mathbf{X}'_i)}$$

とする．この計算により，期待値 $\mathbf{F}$ の推定値 $\hat{\mathbf{F}}$ を得る．

次にデータ集合 D' の作成方法について説明する．データ集合 $D = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n\}$ から人為的に高確率密度領域に一定数のデータ点を集中させたデータ集合を作成するため，データ集合 D から全体の $\frac{1}{3}$ の割合でランダムにデータ集合 D_α を選ぶ．そのデータ集合の各点 $\mathbf{X}_k^{(\alpha)}$ を初期値にして，最大勾配法により $P_L(\mathbf{X})\hat{P}_D(\mathbf{X})$ の極大点 $\mathbf{X'}_k^{(\alpha)}$ を探す．ここで $\hat{P}_D(\mathbf{X})$ は $P_D(\mathbf{X})$ のノンパラメトリックカーネル近似である．この極大点 $\mathbf{X'}_k^{(\alpha)}$ から成るデータを D'_α とし，D の残り $\frac{2}{3}$ のデータ D_β とから，新たなデータ集合 $D' = D'_\alpha + D_\beta$ を得る．

式 (7) による推定値 $\hat{\mathbf{F}}$ には以上で説明した方法により，高確率密度領域に一定数のデータを集中させたデータ集合 D' を用いるので，従来の標本分布法のように僅かなデータ点のみが支配的な推定計算が行われる事はない．また，データ集合 D' の分布は事後分布である $P_L(\mathbf{X})P_D(\mathbf{X})$ をより反映した分布になっている．したがって，D' を用いたプロポーザル分布法による推定値 $\hat{\mathbf{F}}$ は D のみを用いた標本分布法による推定値 $\tilde{\mathbf{F}}$ よりも高精度であると期待される．

4. 評価実験

4.1 実験手順

本節では，以下で示す二つの条件下でベイズ推定平均期待値を標本分布法，周辺分布法，プロポーザル分布法を用いて推定し，それぞれの手法の精度分析を行う．ここで，2.1 節の (1) の理論値を $\mathbf{F}$ として，推定結果を $\breve{\mathbf{F}}$ とすると，推定値の誤差の評価は以下の指標で行う．

$$\text{推定誤差} = ||\breve{\mathbf{F}} - \mathbf{F}_{true}||$$

$||\mathbf{A}||$ はベクトル $\mathbf{A}$ のノルムである．

比較実験 1. 本実験では，$P_L(\mathbf{X})$ を多次元ガウス分布，$P_D(\mathbf{X})$ を $P_D(\mathbf{X}) = \sum_{k=1}^{3}\alpha_k P_{Dk}(\mathbf{X})$ とする三つの多次元ガウス分布の重ね合わせとする．データ集合 D のデータ数 n を 80，多次元混合ガウス分布 $P_D(\mathbf{X})$ の混合比を $\alpha_1 = 4/7, \alpha_2 =$

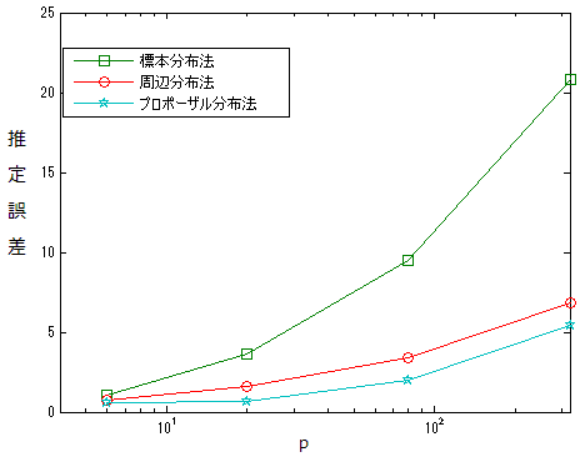

図 1　比較実験 1 の結果

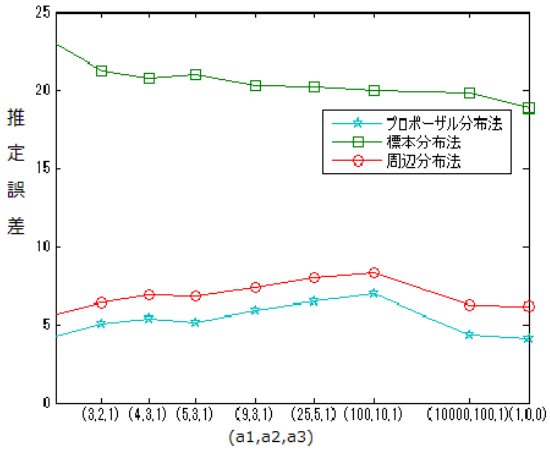

図 2　比較実験 2 の結果

$2/7, \alpha_3 = 1/7$ とし，各 $P_{Dk}(\mathbf{X})$ の平均 $\mathbf{E}_k$ を

$$\mathbf{E}_1 = (\sqrt{p}, 0, 0, ...0, \sqrt{2}, ..., \sqrt{2})$$
$$\mathbf{E}_2 = (0, \sqrt{p}, 0, ...0, \sqrt{2}, ..., \sqrt{2})$$
$$\mathbf{E}_3 = (0, 0, \sqrt{p}, ...0, \sqrt{2}, ..., \sqrt{2})$$

とする．つまり $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$ の後半半分の要素全てを $\sqrt{2}$，最初の 3 要素を順に $\sqrt{p}$ にし，他の要素を全て 0 とする．各 $P_{Dk}(\mathbf{X})$ の分散 σ_{Dk} を互いに大きく異なる値とし，確率密度関数全体の対称性を損なわせる事によって $P_L(\mathbf{X})P_D(\mathbf{X})$ が変数分離形で表される形状から大きく異なるようにするために，以下のように分散を設定する．σ^2_{D2} を 1 とし，$\sigma^2_{D1}, \sigma^2_{D3}$ を $r_1^2\sigma^2_{D2}, r_2^2\sigma^2_{D2}$ とする．ただし

$$r_1 = \sqrt{\frac{2}{p}} + 1, r_2 = \sqrt{\frac{4}{p}} + 1$$

とする．$P_L(\mathbf{X})$ の平均 $\mathbf{E}_L$ を $\mathbf{0}$ とし，各次元の分散を 1 とする．そして，次元数 p がそれぞれの手法の推定精度にどの程度影響を及ぼすかを分析するため，確率変数 $\mathbf{X}$ の次元数 p を

$$p = 6, 20, 80, 320$$

と変化させて，精度分析を行う．

比較実験 2. 確率変数 $\mathbf{X}$ の次元数 p を 320, データ集合 D の総数 n を 80,$\Delta E = 2p$ とし，$P_L(\mathbf{X})$ と各 $P_{Dk}(\mathbf{X})$ を比較実験 *1* と同様に設定する．そして，

$$\alpha_k = \frac{a_k}{a_1 + a_2 + a_3} \quad (k = 1, 2, 3)$$

とし，(a_1, a_2, a_3) を $(1,1,1); (3,2,1); (4,2,1); (5,3,1); (9,3,1); (25,5,1); (100,10,1); (10000,100,1); (1,0,0)$ と変化させ，精度分析を行った．この比較実験では，α_1 を徐々に 1 に近づけ，$P_L(\mathbf{X})P_D(\mathbf{X})$ を変数分離系に近づける事で周辺分布法の推定精度がどの程度改善されるかを分析する．

4.2　比較実験の結果と考察

実験によって得られた結果を図 1 ,2 に示す．図 1 を見ると，標本分布法は次元数 p が大きくなると推定精度が他の手法より悪化している事が分かる．これより．次元数 p が増加すると次元の呪いの影響を強く受け，標本分布法では高精度な推定が難しい事がわかる．また，図 1 ではプロポーザル分布法が一番推定誤差が小さい．これより $P_L(\mathbf{X})P_D(\mathbf{X})$ が変数分離系でない場合，プロポーザル分布法で最も高精度な推定ができる事が分かる．図 2 を見ると，$(\alpha_1, \alpha_2, \alpha_3)$ の比率を変えても，標本分布の推定誤差は変わらず他の手法と比べて大きいままである．これは $(\alpha_1, \alpha_2, \alpha_3)$ によらずデータ集合 D は $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$ を中心とする，それぞれ半径が $\sqrt{p}\sigma_{D1}, \sqrt{p}\sigma_{D2}, \sqrt{p}\sigma_{D3}$ の超球面上にほぼ全てののデータが集まり，また，その中でたまたま尤度の高い数点のみが推定値を支配してしまうため，高精度な推定が行えない事が原因である．また，周辺分布法よりプロポーザル分布法の方が推定精度が良い．これより，$P_L(\mathbf{X})P_D(\mathbf{X})$ が変数分離形でない場合，周辺分布法よりプロポーザル分布法の方が高精度に推定できる事が分かる．

5.　まとめ

従来法である標本分布法では，次元の呪いにより超高次元確率空間における期待値のベイズ推定を高精度に行う事が難しい．本研究では，問題軽減手法であるプロポーザル分布法がこの問題に対してどの程度有効かを分析するために精度分析を行った．精度分析の結果，従来の問題軽減手法である周辺分布法に比して，提案手法であるプロポーザル分布法が，特に $P_L(\mathbf{X})P_D(\mathbf{X})$ が変数分離形でない場合に高精度な推定を可能にする事が分かった．

参考文献

[Silverman 86] Bernard W. Silverman: Density Estimation for Statistics and Data Analysis, Chap. 3 and 4, Chapman and Hall (1986).

[Leeuwen 03] Peter J. van Leeuwen: A variance-minimizing Filter for large-scale applications, Mon. Wea. Rev., Vol.131, pp. 2071-2084 (2003).

[Snyder 08] Chris Snyder, Thomas Bengtsson, Peter Bickel and JeR Anderson: Obstacles to High-Dimensional Particle Filtering, Monthly Weather Review, Vol.136, No.12, pp.4629-4640 (2008).

[Everitt 02] Brian S. Everitt: The Cambridge Dictionary of Statistics. Cambridge University Press (2002).