混合正規分布とギブスサンプラーの例

渡辺澄夫　東京工業大学

1　確率モデルと事前分布

(1) データを $x \in \mathbb{R}^N$ とする。
(2) パラメータは $a = \{a_k\}, b = \{b_k\}, s = \{s_k\}$ で、$s_k > 0$, $a_k > 0$, $\sum_{k=1}^K a_k = 1$, $b_k \in \mathbb{R}^N$ を満たすものとする。
(3) 確率モデルは

$$p(x|a,b,s) = \sum_{k=1}^K a_k \Big(\frac{s_k}{2\pi}\Big)^{N/2} \exp\Big(-\frac{s_k}{2}\|x-b_k\|^2\Big).$$

(4) 事前分布は $\{\alpha_k > 0, \mu_k > 0, \rho_k > 0, r \geq N/2\}$ をハイパーパラメータとして

$$\begin{aligned} \varphi(a) &= \frac{1}{z_1}\prod_{k=1}^K (a_k)^{\alpha_k - 1} \\ \varphi(b,s) &= \frac{1}{z_2}\prod_{k=1}^K (s_k)^r \exp\Big(-\frac{s_k}{2}(\rho_k + \mu_k\|b_k\|^2)\Big) \end{aligned}$$

ここで z_1, z_2 は正規化定数 (ハイパーパラメータの関数）.

2　競合的変数

K を 1 以上の整数とする。競合変数の集合 $\mathcal{C}$ を次式で定義する。

$$\mathcal{C} = \{y = (y_1, y_2, ..., y_K)\ ;\ y_k \text{ はどれかひとつが } 1 \text{ で他は } 0\}$$

$(x,y) \in \mathbb{R}^N \times \mathcal{C}$ の確率モデルを

$$p(x,y|a,b,s) = \prod_{k=1}^K \Big[a_k \Big(\frac{s_k}{2\pi}\Big)^{N/2} \exp\Big(-\frac{s_k}{2}\|x-b_k\|^2\Big)\Big]^{y_k}$$

と定義すると、

$$p(x|a,b,s) = \sum_{y\in\mathcal{C}} p(x,y|a,b,s)$$

が成り立つ．つまり $p(x|a,b,s)$ は $p(x,y|a,b,s)$ で y についての情報が得られない場合の確率モデルに相当することがわかった。

3　事後分布

真の分布 $q(x)$ から独立な n 個のデータ $x^n = \{x_i \in \mathbb{R}^N; i = 1, 2, ..., n\}$ が得られたとする。事後分布

$$p(a, b, s|x^n) \propto \varphi(a)\varphi(b, s) \prod_{i=1}^{n} p(x_i|a, s, b)$$

に従う (a, b, s) をサンプリングすることが目標であるが、直接にサンプリングするのは容易ではない。そこで $y^n = \{y_i \in \mathcal{C}; i = 1, 2, ..., n\}$ を用いて (a, b, s, y^n) の分布を

$$p(a, b, s, y^n|x^n) \propto \varphi(a)\varphi(b, s) \prod_{i=1}^{n} p(x_i, y_i|a, b)$$

と定義すると

$$p(a, b, s|x^n) = \sum_{y_1 \in \mathcal{C}} \sum_{y_2 \in \mathcal{C}} \cdots \sum_{y_n \in \mathcal{C}} p(a, b, s, y^n|x^n)$$

が成り立つので、$p(a, b, s, y^n|x^n)$ から (a, b, s, y^n) をサンプリングして y^n を捨てて (a, b, s) を残せば目的が達成できる。

目標：$P \equiv p(a, b, s, y^n|x^n)$ からサンプリングを行い、得られた $\{(a_k.b_k, s_k)\}$ をベイズ推測に利用する。

このために次のギブスサンプラーを作る。

ギブスサンプラー：(I) $p(y^n|a, b, s, x^n)$ から y^n をサンプルすることと (II) $p(a, b, s|x^n, y^n)$ から (a, b, s) をサンプルすることとをサンプルすることを繰り返す。

(I) まず (a, b, s) が与えられたときの y_i のサンプリング法.

$$\begin{aligned} P &\propto \varphi(a)\varphi(b, s) \prod_{i=1}^{n} p(x_i, y_i|a, b) \\ &\propto \prod_{k=1}^{K} (a_k)^{\alpha_k - 1} (s_k)^r \exp\Big(-\frac{s_k}{2}(\rho_k + \mu_k \|b_k\|^2)\Big) \\ &\quad \times \prod_{i=1}^{n} \Big[a_k (s_k/2\pi)^{N/2} \exp\Big(-\frac{s_k}{2}\|x_i - b_k\|^2\Big)\Big]^{y_{ik}} \end{aligned}$$

各 i ごとに $\{y_{ik}\ ;\ k = 1, 2, ..., K\} \in \mathcal{C}$ を独立に多項分布

$$P(y_{ik}|a, b, s, x_i) \propto \Big[a_k (s_k/2\pi)^{N/2} \exp\Big(-\frac{s_k}{2}\|x_i - b_k\|^2\Big)\Big]^{y_{ik}}$$

からサンプリングできる．(出る目の種類が K 個のサイコロで [ ] の中身が確率であるようなもの)。

(II) 次に y^n が与えられた時の (a, b, s) のサンプリング法を導出する。

$$\begin{aligned} P &\propto \prod_{k=1}^{K}(a_k)^{\alpha_k-1}(s_k)^r \exp\Big(-\frac{s_k}{2}(\rho_k+\mu_k\|b_k\|^2)\Big) \\ &\quad \times\prod_{i=1}^{n}\Big[a_k(s_k/2\pi)^{N/2}\exp\Big(-\frac{s_k}{2}\|x_i-b_k\|^2\Big)\Big]^{y_{ik}} \\ &\propto \Big[\prod_{k=1}^{K}(a_k)^{\alpha_k-1+\sum_{i=1}^n y_{ik}}\Big] \\ &\quad \times\Big[\prod_{k=1}^{K}(s_k)^{r+(N/2)\sum_i y_{ik}}\exp\Big(-\frac{s_k}{2}(\rho_k+\mu_k\|b_k\|^2+\sum_{i=1}^{n}y_{ik}\|x_i-b_k\|^2)\Big)\Big] \end{aligned}$$

上式の exp の中身を b についての式とみると次の式の k についての和

$$\begin{aligned} H_k(b,s) &= \frac{s_k}{2}(\rho_k+\mu_k\|b_k\|^2+\sum_i y_{ik}\|x_i-b_k\|^2) \\ &= \frac{s_k}{2}\Big(\rho_k+(\mu_k+n_k)\|b_k\|^2-2\Big(\sum_{i=1}^{n}y_{ik}x_i\Big)b_k+\sum_{i=1}^{n}y_{ik}\|x_i\|^2\Big) \\ &= \frac{s_k}{2}\Big(\rho_k+(\mu_k+n_k)\Big\|b_k-\Big(\sum_i y_{ik}x_i\Big)/(\mu_k+n_k)\Big\|^2 \\ &\quad +\sum_{i=1}^{n}y_{ik}\|x_i\|^2-\frac{1}{(\mu_k+n_k)}\|\sum_{i=1}^{n}y_{ik}x_i\|^2)\Big) \end{aligned}$$

になる．ここで $n_k=\sum_{i=1}^n y_{ik}$ とおいた。さらに

$$\begin{aligned} B_k &= \Big(\sum_i y_{ik}x_i\Big)/(\mu_k+n_k) \\ C_k &= \rho_k+\sum_{i=1}^{n}y_{ik}\|x_i\|^2-\frac{1}{(\mu_k+n_k)}\|\sum_{i=1}^{n}y_{ik}x_i\|^2 \\ D_k &= r+(N/2)n_k-N/2 \end{aligned}$$

とおくと

$$\begin{aligned} P &\propto \Big[\prod_{k=1}^{K}(a_k)^{\alpha_k-1+n_k}\Big]\Big[\prod_{k=1}^{K}(s_k)^{r+(N/2)n_k}\exp(-\frac{s_k}{2}((\mu_k+n_k)\|b_k-B_k\|^2+C_k))\Big] \\ &= \mathrm{Dir}(a|\alpha+n)\prod_{k=1}^{K}\Big[\ \mathrm{Nor}(b_k|B_k, I/(s_k(\mu_k+n_k)))\ \mathrm{Gam}(s_k|D_k+1, 2/C_k)\Big] \end{aligned}$$

ここで Dir, Nor, Gam は次のディリクレ分布、正規分布、ガンマ分布を表している。

$$\begin{aligned} \mathrm{Dir}(a|\alpha+n) &\propto \prod_{k=1}^{K}(a_k)^{\alpha_k+n_k-1} \\ \mathrm{Nor}(b_k|B_k, I/(s_k(\mu_k+n_k))) &= \text{平均 } B_k, \text{ 分散共分散 } I/(s_k(\mu_k+n_k)) \text{ の正規分布} \\ \mathrm{Gam}(s_k|D_k+1, 2/C_k) &\propto (s_k)^{D_k}\exp(-s_kC_k/2) \end{aligned}$$

これより $a=\{a_k\}$ はディリクレ分布からサンプリングし、s_k はガンマ分布からサンプリングし、b_k は得られた s_k を用いて正規分布からサンプリングすればよいことがわかった。