

点をベイズ的な方法でエレガントに解決できることを示すものともなっている (Attias, 1999b). 変分ベイズ法を実際に使う際に有用な多くの洞察が得られるので, 読者はこの例を自分で詳しく追ってみることを薦める¹. もっとずっと複雑な分布を持つ他の多くのベイズモデルも, この例の自明な拡張や一般化で解くことができる.

我々の出発点は, 図 9.6 のグラフィカルモデルに示したガウス混合モデルの尤度関数である. 各観測値 $\mathbf{x}_n$ には対応する潜在変数 $\mathbf{z}_n$ があり, これは K 個の要素 z_{nk} ($k = 1, \dots, K$) の中に 1 が一つだけある二値ベクトルである. 前と同様に観測データの集合を $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, 対応する潜在変数を同様に $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ と書く. (9.10) から, 混合比 π が与えられたときの $\mathbf{Z}$ の条件付き分布は次の形式で書ける.

$$p(\mathbf{Z}|\pi) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}}. \quad (10.37)$$

同じように, (9.11) から, 潜在変数と混合要素のパラメータが与えられたときの観測データベクトルの条件付き分布は

$$p(\mathbf{X}|\mathbf{Z}, \mu, \Lambda) = \prod_{n=1}^N \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n | \mu_k, \Lambda_k^{-1})^{z_{nk}} \quad (10.38)$$

となる. ここで, $\mu = \{\mu_k\}$, $\Lambda = \{\Lambda_k\}$ である. 計算がいくらか簡単になるため, 共分散行列ではなく, 精度行列を使っていることに注意されたい.

次に, パラメータ μ , Λ , π の事前分布を導入する. 共役事前分布 (10.4.1 節) を使うことで, 計算がかなり簡単になる. よって, 混合比 π の事前分布にはディリクレ分布を用いる.

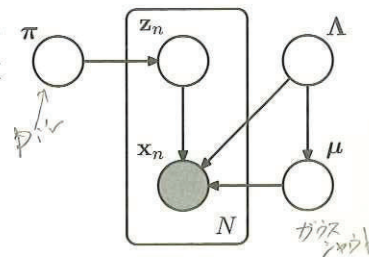
$$p(\pi) = \text{Dir}(\pi | \alpha_0) = C(\alpha_0) \prod_{k=1}^K \pi_k^{\alpha_0 - 1}. \quad (10.39)$$

ここでは要素の対称性を考えて, 各混合要素について同じパラメータ α_0 を用いた. $C(\alpha_0)$ はディリクレ分布の正規化定数であり, (B.23) で定義される. 前に見たように (2.2.1 節), パラメータ α_0 は各混合要素に対する実質的な事前の観測回数に相当すると解釈できる. もし α_0 の値が小さければ, 事後分布は事前分布より主にデータによって決められる.

同様に混合要素のもつガウス分布については, その平均と精度に次で与えられる独立なガウス-ウィシャート事前分布を導入する.

$$\begin{aligned} p(\mu, \Lambda) &= p(\mu | \Lambda) p(\Lambda) \\ &= \prod_{k=1}^K \mathcal{N}(\mu_k | \mathbf{m}_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | \mathbf{W}_0, \nu_0). \end{aligned} \quad (10.40)$$

図 10.5 ベイズ混合ガウス分布モデルの有向非循環グラフ表現. 長方形 (プレート) は N 個の独立同分布の観測値を意味する. μ は $\{\mu_k\}$, Λ は $\{\Lambda_k\}$ を表す.



なぜならば, これは平均も分散も未知の場合の共役事前分布を表しているためである (2.3.6 節). 対称性から, 通常 $\mathbf{m}_0 = \mathbf{0}$ とおく.

結果として, このモデルは図 10.5 に示す有向グラフで表すことができる. Λ から μ へのリンクがあることに注意されたい. これは μ の分布の分散 (10.40) が Λ の関数となっているためである.

この例は, 潜在変数とパラメータの違いをよく示している. プレートの内側にある $\mathbf{z}_n$ のような変数は数がデータ集合のサイズに比例して増えるため, 潜在変数である. これに対して, μ のような変数はプレートの外にあり, データ集合のサイズとは独立に数が決まっているので, パラメータとみなせる. ただし, グラフィカルモデル上はこれらには全く区別はない.

10.2.1 変分事後分布

このモデルを変分ベイズ法で扱うためには, 次にすべての確率変数の同時分布を書き下す必要がある. これは次式で与えられる.

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda) = p(\mathbf{X}|\mathbf{Z}, \mu, \Lambda) p(\mathbf{Z}|\pi) p(\pi) p(\mu|\Lambda) p(\Lambda). \quad (10.41)$$

各因子は上で定義された通りである. 読者は, この分解が実際に図 10.5 の確率的グラフィカルモデルに対応していることを確かめられたい. 観測されているのは変数 $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ のみであることに注意する.

いま, 潜在変数とパラメータに分解した変分近似

$$q(\mathbf{Z}, \pi, \mu, \Lambda) = q(\mathbf{Z}) q(\pi, \mu, \Lambda) \quad (10.42)$$

を考える. これが我々のベイズ混合モデルについて実際に計算可能な解を得る上で, 必要な唯一の仮定であるのは驚くべきことである. 特に, 因子 $q(\mathbf{Z})$ および $q(\pi, \mu, \Lambda)$ の関数形は, 変分近似を最適化することで自動的に導かれる. 分布 q の添え字は (10.41) での p の添え字と同様に省略し, 引数の違いによって異なる分布を表していることに注意されたい.

出できる. 因子 $q(\mathbf{Z})$ についての更新式の導出を考えよう. 最適な因子分布の対数は次で与えられる.

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{\pi, \mu, \Lambda} [\ln p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda)] + \text{const.} \quad (10.43)$$

ここで, 分解 (10.41) を利用する. 右辺で, 変数 $\mathbf{Z}$ への依存関係だけに興味があることに注意しよう. $\mathbf{Z}$ に依存しない項はすべて正規化定数に含まれてしまうため,

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{\pi} [\ln p(\mathbf{Z}|\pi)] + \mathbb{E}_{\mu, \Lambda} [\ln p(\mathbf{X}|\mathbf{Z}, \mu, \Lambda)] + \text{const} \quad (10.44)$$

を得る. 右辺の2つの条件付き分布に実際の関数を代入し, $\mathbf{Z}$ に依存しない項を定数に含めると,

$$\ln q^*(\mathbf{Z}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \ln \rho_{nk} + \text{const} \quad (10.45)$$

を得る. ここで,

$$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2} \mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2} \ln(2\pi) - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k} [(\mathbf{x}_n - \mu_k)^T \Lambda_k (\mathbf{x}_n - \mu_k)] \quad (10.46)$$

と定義した. D はデータ変数 $\mathbf{x}$ の次元である. (10.45) の両辺の指数をとると, 次式を得る.

$$q^*(\mathbf{Z}) \propto \prod_{n=1}^N \prod_{k=1}^K \rho_{nk}^{z_{nk}}. \quad (10.47)$$

この分布は正規化されている必要があることと, 各 n の値について z_{nk} は二値すべての k の値にわたる和が1であることに注意すると, 次式を得る (≡ 演習 10.12).

$$q^*(\mathbf{Z}) = \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{z_{nk}}. \quad (10.48)$$

ここで,

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^K \rho_{nj}} \quad (10.49)$$

とした. 因子 $q(\mathbf{Z})$ の最適解は, 事前分布 $p(\mathbf{Z}|\pi)$ と同じ形をしていることがわかる. ρ_{nk} は実数の指数関数で与えられるため, r_{nk} は必ず非負となり, この和は要請通り1になることに注意しよう.

離散分布 $q^*(\mathbf{Z})$ については, 標準的な公式から

$$\mathbb{E}[z_{nk}] = r_{nk} \quad (10.50)$$

の変数の分布によるモーメントに依存することに注意されたい. ゆえに, 変分ベイズ法の更新式は相互依存関係にあり, 繰り返しで解く必要がある.

いま, 観測データ集合について以下で与えられる, 負担率から計算できる三つの統計量を定義しておくことと便利である.

$$N_k = \sum_{n=1}^N r_{nk} \quad (10.51)$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} \mathbf{x}_n \quad (10.52)$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T. \quad (10.53)$$

これらはガウス混合モデルの最尤推定でのEM アルゴリズムで計算される量と似ていることに注意されたい.

次に, 変分事後分布の因子 $q(\pi, \mu, \Lambda)$ について考えよう. 一般的な結果 (10.9) をまた用いると,

$$\ln q^*(\pi, \mu, \Lambda) = \ln p(\pi) + \sum_{k=1}^K \ln p(\mu_k, \Lambda_k) + \mathbb{E}_{\mathbf{Z}} [\ln p(\mathbf{Z}|\pi)] + \sum_{k=1}^K \sum_{n=1}^N \mathbb{E}[z_{nk}] \ln \mathcal{N}(\mathbf{x}_n | \mu_k, \Lambda_k^{-1}) + \text{const} \quad (10.54)$$

となる. この式の右辺は, π だけを含む項と μ および Λ だけを含む項の和に分解されることがわかる. これは変分事後分布 $q(\pi, \mu, \Lambda)$ が $q(\pi)q(\mu, \Lambda)$ と分解されることを意味する. さらに, μ および Λ を含む項は μ_k と Λ_k を含む項の k についての積からなり, 次のように分解される.

$$q(\pi, \mu, \Lambda) = q(\pi) \prod_{k=1}^K q(\mu_k, \Lambda_k). \quad (10.55)$$

(10.54) の右辺において π に依存する項を探すと, 次を得る.

$$\ln q^*(\pi) = (\alpha_0 - 1) \sum_{k=1}^K \ln \pi_k + \sum_{k=1}^K \sum_{n=1}^N r_{nk} \ln \pi_k + \text{const.} \quad (10.56)$$

ここで (10.50) を用いた. 両辺の指数をとると, $q^*(\pi)$ はディリクレ分布

$$q^*(\pi) = \text{Dir}(\pi|\alpha) \quad (10.57)$$

となることがわかる. ここで, α の要素 α_k は次式である.

$$\alpha_k = \alpha_0 + N_k. \quad (10.58)$$

ないが、乗法定理は必ず適用できるため、 $q^*(\mu_k, \Lambda_k) = q^*(\mu_k | \Lambda_k) q^*(\Lambda_k)$ の形に書くことができる。この二つの因子は、(10.54)の中で μ_k および Λ_k を含む項を調べるとわかる。結果は予想通りガウス-ウィシャート分布となり、次式で与えられる (⇒ 演習 10.13) $\rightarrow$ $W \sim N, \Lambda \sim \text{IW}$

$$q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_k, \nu_k). \quad (10.59)$$

ここで、以下のように定義した。

$$\beta_k = \beta_0 + N_k \quad (10.60)$$

$$m_k = \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k) \quad (10.61)$$

$$W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \quad (10.62)$$

$$\nu_k = \nu_0 + N_k. \quad (10.63)$$

これらの更新式は、混合ガウス分布を最尤推定で解く EM アルゴリズムの M ステップと似ている。モデルのパラメータの変分事後分布を更新するために必要な計算には、データ集合について最尤推定の場合に現れるものと同様の和を求めることになることがわかる。

この変分 M ステップを実行するためには、負担率を表す期待値 $E[z_{nk}] = r_{nk}$ が必要になるが、これは (10.46) で与えられる ρ_{nk} を正規化すれば得られる。この式はパラメータの変分事後分布を使った期待値を含んでいるが、これは容易に計算できて、以下が得られる (⇒ 演習 10.14)。

$$E_{\mu_k, \Lambda_k} [(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] = D\beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k) \quad (10.64)$$

$$\ln \tilde{\Lambda}_k \equiv E[\ln |\Lambda_k|] = \sum_{i=1}^D \psi\left(\frac{\nu_k + 1 - i}{2}\right) + D \ln 2 + \ln |W_k| \quad (10.65)$$

$$\ln \tilde{\pi}_k \equiv E[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}). \quad (10.66)$$

ここで、 $\tilde{\Lambda}_k$ および $\tilde{\pi}_k$ を新しく導入した。 $\psi(\cdot)$ は (B.25) で定義されるディガンマ関数。 $\hat{\alpha} = \sum_k \alpha_k$ である。(10.65) および (10.66) の式はウィシャート分布とディリクレ分布の標準的な性質から得られる (⇒ 付録 B)。

(10.64), (10.65) および (10.66) を (10.46) に代入して (10.49) を使うと、負担率について次の結果が得られる。

$$r_{nk} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{1/2} \exp \left\{ -\frac{D}{2\beta_k} - \frac{\nu_k}{2} (x_n - m_k)^T W_k (x_n - m_k) \right\}. \quad (10.67)$$

この結果は最尤推定での EM アルゴリズムの場合に対応する結果と類似していること

10.13 In order to derive the optimal solution for $q^*(\mu_k, \Lambda_k)$ we start with the result (10.54) and keep only those terms which depend on μ_k or Λ_k to give

$$\begin{aligned} \ln q^*(\mu_k, \Lambda_k) &= \ln \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}) + \ln \mathcal{W}(\Lambda_k | W_0, \nu_0) \\ &\quad + \sum_{n=1}^N E[z_{nk}] \ln \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1}) + \text{const.} \\ &= -\frac{\beta_0}{2} (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) + \frac{1}{2} \ln |\Lambda_k| - \frac{1}{2} \text{Tr}(\Lambda_k W_0^{-1}) \\ &\quad + \frac{(\nu_0 - D - 1)}{2} \ln |\Lambda_k| - \frac{1}{2} \sum_{n=1}^N E[z_{nk}] (\Lambda_k^{-1} (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \\ &\quad + \frac{1}{2} \sum_{n=1}^N E[z_{nk}] \ln |\Lambda_k| + \text{const.} \end{aligned} \quad (140)$$

Using the product rule of probability, we can express $\ln q^*(\mu_k, \Lambda_k)$ as $\ln q^*(\mu_k | \Lambda_k) + \ln q^*(\Lambda_k)$. Let us first of all identify the distribution for μ_k . To do this we need only consider terms on the right hand side of (140) which depend on μ_k , giving

$$\begin{aligned} \ln q^*(\mu_k | \Lambda_k) &= -\frac{1}{2} \mu_k^T \left[\beta_0 + \sum_{n=1}^N E[z_{nk}] \right] \Lambda_k \mu_k + \mu_k^T \Lambda_k \left[\beta_0 m_0 + \sum_{n=1}^N E[z_{nk}] x_n \right] \\ &\quad + \text{const.} \\ &= -\frac{1}{2} \mu_k^T [\beta_0 + N_k] \Lambda_k \mu_k + \mu_k^T \Lambda_k [\beta_0 m_0 + N_k \bar{x}_k] + \text{const.} \end{aligned}$$

where we have made use of (10.51) and (10.52). Thus we see that $\ln q^*(\mu_k | \Lambda_k)$ depends quadratically on μ_k and hence $q^*(\mu_k | \Lambda_k)$ is a Gaussian distribution. Completing the square in the usual way allows us to determine the mean and precision of this Gaussian, giving

$$q^*(\mu_k | \Lambda_k) = \mathcal{N}(\mu_k | m_k, \beta_k \Lambda_k) \quad (141)$$

where

$$\begin{aligned} \beta_k &= \beta_0 + N_k \\ m_k &= \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k). \end{aligned}$$

Next we determine the form of $q^*(\Lambda_k)$ by making use of the relation

$$\ln q^*(\Lambda_k) = \ln q^*(\mu_k, \Lambda_k) - \ln q^*(\mu_k | \Lambda_k).$$

On the right hand side of this relation we substitute for $\ln q^*(\mu_k | \Lambda_k)$ using (140), and we substitute for $\ln q^*(\mu_k, \Lambda_k)$ using the result (141). Keeping only those terms

which depend on Λ_k we obtain

$$\begin{aligned} \ln q^*(\Lambda_k) &= -\frac{\beta_0}{2} (\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) + \frac{1}{2} \ln |\Lambda_k| - \frac{1}{2} \text{Tr}(\Lambda_k W_0^{-1}) \\ &\quad + \frac{(\nu_0 - D - 1)}{2} \ln |\Lambda_k| - \frac{1}{2} \sum_{n=1}^N E[z_{nk}] (\Lambda_k^{-1} (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \\ &\quad + \frac{1}{2} \sum_{n=1}^N E[z_{nk}] \ln |\Lambda_k| + \frac{\beta_k}{2} (\mu_k - m_k)^T \Lambda_k (\mu_k - m_k) \\ &\quad - \frac{1}{2} \ln |\Lambda_k| + \text{const.} \\ &= \frac{(\nu_k - D - 1)}{2} \ln |\Lambda_k| - \frac{1}{2} \text{Tr}(\Lambda_k W_k^{-1}) + \text{const.} \end{aligned}$$

Here we have defined

$$\begin{aligned} W_k^{-1} &= W_0^{-1} + \beta_0 (\mu_k - m_0)(\mu_k - m_0)^T + \sum_{n=1}^N E[z_{nk}] (x_n - \mu_k)(x_n - \mu_k)^T \\ &\quad - \beta_k (\mu_k - m_k)(\mu_k - m_k)^T \\ \nu_k &= \nu_0 + \sum_{n=1}^N E[z_{nk}] \\ &= \nu_0 + N_k, \end{aligned} \quad (142)$$

where we have made use of the result

$$\begin{aligned} \sum_{n=1}^N E[z_{nk}] x_n x_n^T &= \sum_{n=1}^N E[z_{nk}] (x_n - \bar{x}_k + \bar{x}_k)(x_n - \bar{x}_k + \bar{x}_k)^T + N_k \bar{x}_k \bar{x}_k^T \\ &= N_k S_k + N_k \bar{x}_k \bar{x}_k^T \end{aligned} \quad (143)$$

and we have made use of (10.53). Note that the terms involving μ_k have cancelled out in (142) as we expect since $q^*(\Lambda_k)$ is independent of μ_k . Thus we see that $q^*(\Lambda_k)$ is a Wishart distribution of the form

$$q^*(\Lambda_k) = \mathcal{W}(\Lambda_k | W_k, \nu_k).$$

10.16 To derive (10.71) we make use of (10.38) to give

$$\begin{aligned} E[\ln p(D | \mu, \Lambda)] &= \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K E[z_{nk}] \{E[\ln |\Lambda_k|] - E[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] - D \ln(2\pi)\}. \end{aligned}$$