EM-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений.

Теоретический обзор

В. Ю. Королев

11 октября 2007 г.

Содержание

Введение **5**

1. Предварительные сведения **8**

2. Общее описание EM-алгоритма **12**

3. Монотонность EM-алгоритма **14**

3.1. Проксимальные алгоритмы (PP-алгоритмы) и их свойства 15

3.2. EM-алгоритм как проксимальный алгоритм 16

3.3. Неподвижные точки EM-алгоритма и локальные максимумы функции правдоподобия 18

4. Некоторые свойства смесей распределений вероятностей **20**

4.1. Основные определения 20

4.2. Идентифицируемость смесей вероятностных распределений 24

5. Решение задачи разделения смесей вероятностных распределений с помощью EM-алгоритма **30**

6. Разделение конечных смесей нормальных распределений с помощью EM-алгоритма **36**

7. Модификации EM-алгоритма **40**

7.1. Медианные модификации EM-алгоритма 40

7.2. SEM-алгоритм 44

7.3. CEM-алгоритм 50

7.4. MCEM и SAEM-алгоритмы 50

7.4.1. Классический MCEM-алгоритм 50

7.4.2. MC-модификация SEM-алгоритма 52

7.4.3. SAEM-алгоритм 52

8. Приближенное разделение конечных смесей с помощью метода фиксированных компонент для выбора начального приближения EM-алгоритма 55

8.1. Основная идея метода фиксированных компонент 55

8.2. Разделение конечных смесей вероятностных распределений с фиксированными компонентами при помощи метода наименьших квадратов 59

8.3. Разделение конечных смесей вероятностных распределений с фиксированными компонентами при помощи метода наименьших модулей 63

8.3.1. Решение в смысле минимума sup-нормы 63

8.3.2. Решение в смысле минимума L_1-нормы 64

8.4. Разделение конечных смесей вероятностных распределений с фиксированными компонентами при помощи "усеченного" EM-алгоритма 65

8.5. Разделение конечных смесей вероятностных распределений с фиксированными компонентами при помощи байесовской классификации 67

9. Выбор модели (определение типа и числа компонент смеси) 68

9.1. Некоторые сведения из теории проверки сложных статистических гипотез 69

9.2. Проверка значимости динамической составляющей волатильности с помощью критерия отношения правдоподобия 72

9.3. "Последовательный" критерий отношения правдоподобия для определения числа компонент смеси 74

9.4. Определение числа компонент смеси с помощью SEM-алгоритма . 77

9.5. Информационные критерии выбора модели (числа компонент и типа смеси) 78

9.5.1. Информационный критерий Акаике (AIC) 78

9.5.2. Байесовский информационный критерий (BIC) 83

Список литературы 86

Введение

EM-алгоритмом принято называть довольно работоспособную схему построения процедур итерационного типа для численного решения задачи поиска экстремума целевой функции в разнообразных задачах оптимизации. В частности, в прикладной статистике эта схема довольно эффективна для поиска оценок максимального правдоподобия и родственных им в ситуациях, когда функция правдоподобия имеет сложную структуру, из-за которой другие методы оказываются неэффективными или вообще не применимыми.

По-видимому, впервые итерационная процедура типа EM-алгоритма, позволяющая находить численное решение задачи максимизации функции правдоподобия при разделении смесей распределений вероятностей, была предложена в работе (McKendrick, 1926). Затем после довольно большого перерыва эта идея вновь возникла в работах (Healy and Westmacott, 1956), (Шлезингер, 1965), (Шлезингер, 1968), (Day, 1969a), (Day, 1969b), (Wolfe, 1970) , а затем развита и систематически исследована в работах (Dempster, Laird and Rubin, 1977), (Kazakos, 1977). Само название *EM-алгоритм* было предложено в работе (Dempster, Laird and Rubin, 1977), посвященной применению метода максимального правдоподобия к статистическому оцениванию по неполным статистическим данным. Возможно, поэтому зарубежные источники традиционно ссылаются на эту статью (Dempster, Laird and Rubin, 1977) как на первую работу по EM-алгоритму.

Основные свойства EM-алгоритма описаны еще в работе (Шлезингер, 1965). Позднее в работах (Dempster, Laird and Rubin, 1977), (Everitt and Hand, 1981), (Wu, 1983), (Boyles, 1983), (Redner and Walker, 1984) эти свойства были передоказаны и развиты.

Литература по EM-алгоритму и его применениям к решению задач из конкретных областей обширна. Перечисление всех работ практически невозможно. Мы ограничимся упоминанием лишь книг, посвященных соб-

ственно EM-алгоритму (Литтл и Рубин, 1991), (McLachlan and Krishnan, 1997), книг, в которых EM-алгоритму уделено значительное место (Айвазян и др., 1989), (Tanner, 1993), а также двух обстоятельных работ (Bilmes, 1998) и (Figueiredo, 2004).

В работе (Dempster, Laird and Rubin, 1977) была предложена концепция EM-алгоритма как метода работы с неполными данными (или данными с пропусками). Эта концепция удобна с методической точки зрения и хорошо проясняет смысл метода. Именно этой концепции мы и будем придерживаться при описании EM-алгоритма.

Как правило, EM-алгоритм применяется при решении задач двух типов. К первому типу можно отнести статистические задачи, связанные с анализом *действительно* неполных данных, когда некоторые статистические данные отсутствуют в силу каких-либо причин. Методы решения таких задач описаны в книге (Литтл и Рубин, 1991).

К другому типу задач можно отнести статистические задачи, в которых функция правдоподобия имеет вид, не допускающий удобных аналитических методов исследования, но допускающий серьезные упрощения, если в задачу ввести дополнительные "ненаблюдаемые" ("отсутствующие", скрытые, латентные) величины. Примерами прикладных задач второго типа являются задачи распознавания образов, реконструкции изображений. Математическую суть этих прикладных задач составляют задачи кластерного анализа, классификации и разделения смесей вероятностных распределений.

Метод скользящего разделения смесей лежит в основе предложенного недавно подхода к исследованию стохастической структуры хаотических информационных потоков в сложных телекоммуникационных сетях (Батракова, Королев, 2006), (Батракова, Королев и Шоргин, 2007). Этот подход основан на стохастической модели телекоммуникационной сети, в рамках которой она представляется в виде суперпозиции некоторых простых последовательно-параллельных структур. Принцип максимума энтропии в комбинации с предельными теоремами теории вероятностей естественно приводят к тому, что такая модель порождает смеси гамма-распределений для времени выполнения (обработки) запроса сетью. Параметры получаемой смеси гамма-распределений характеризуют стохастическую структуру информационных потоков в сети. Для решения задачи статистического оценивания параметров смесей экспоненциальных и гамма-распределений (задачи разделения смесей) используется EM-алгоритм. Чтобы проследить изменение стохастической структуры информационных потоков во времени, EM-алгоритм применяется в режиме скользящего окна. В рамках

этого подхода чрезвычайно важно правильно подобрать нужную версию EM-алгоритма, обеспечивающую высокое быстродействие и удобную интерпретацию получаемых результатов. В данной работе довольно подробно разобраны свойства EM-алгоритма и его наиболее часто используемых модификаций, а также предложены новые приемы, направленные на повышение точности, устойчивости EM-алгоритма и удобство интерпретации результатов его работы при решении задач разделения смесей. При этом основное внимание уделяется применению EM-алгоритма для разделения смесей нормальных законов. Задачи разделения таких смесей составляют ядро метода декомпозиции волатильности финансовых индексов (Королев, 2007), (Королев, Ломской, Пресняков и Рэй, 2005) и турбулентной плазмы (Korolev and Rey, 2005), (Королев и Скворцова, 205).

Автор считает своей приятной обязанностью выразить благодарность В. Е. Бенингу, В. А. Ломскому, Е. В. Непомнящему и С. Я. Шоргину за полезное обсуждение отдельных разделов данной работы.

Работа над данной книгой проходила при поддержке РФФИ, проекты 05-01-00396, 05-01-00535 и 05-07-90103.

1.

Предварительные сведения

Пусть $\mathbf{X}$ и $\mathbf{Y}$ – случайные величины, заданные на одном и том же измеримом пространстве $(\Omega, \mathcal{A})$. Для определенности будем считать, что $\mathbf{X}$ и $\mathbf{Y}$ принимают свои значения в $\mathbb{R}^n$ и $\mathbb{R}^m$, соответственно, где $n \geq 1$, $m \geq 1$. Другими словами, будем считать, что $(\mathbf{X}, \mathbf{Y})$ – это $(n+m)$-мерный случайный вектор.

Предположим, что на σ-алгебре $\mathcal{A}$ задано семейство вероятностных мер $\{\mathsf{P}_\theta : \theta \in \Theta\}$, где Θ – множество, вообще говоря, произвольной природы.

Пусть $\mathcal{B}_n$ и $\mathcal{B}_m$ – борелевские σ-алгебры на $\mathbb{R}^n$ и $\mathbb{R}^m$, соответственно и пусть $\mu_{\mathbf{X}}$ и $\mu_{\mathbf{Y}}$ – некоторые σ-конечные меры, определенные, соответственно, на $\mathcal{B}_n$ и $\mathcal{B}_m$ (напомним, что мера ν, заданная на борелевской σ-алгебре $\mathcal{B}_k$ подмножеств множества $\mathbb{R}^k$, $k \geq 1$, называется *σ-конечной*, если существует последовательность множеств $A_1, A_2, \ldots$ такая, что $A_i \in \mathcal{B}_k$, $\mathbb{R}^k = A_1 \cup A_2 \cup \ldots$ и $\nu(A_i) < \infty$, $i \geq 1$). Пусть μ – мера, заданная на σ-алгебре $\sigma(\mathcal{B})$, порожденной множеством $\mathcal{B}$ измеримых прямоугольников,

$$\mathcal{B} = \{A \times B : \, A \in \mathcal{B}_n, \, B \in \mathcal{B}_m\},$$

и удовлетворяющая условию

$$\mu(A \times B) = \mu_{\mathbf{X}}(A) \cdot \mu_{\mathbf{Y}}(B), \quad A \in \mathcal{B}_n, \, B \in \mathcal{B}_m.$$

Если меры $\mu_{\mathbf{X}}$ и $\mu_{\mathbf{Y}}$ σ-конечны, то, во-первых, указанное условие определяет меру μ однозначно и, во-вторых, мера μ является σ-конечной (см., например, теорему 2 в (Халмош, 1953), с. 143). Так определенная мера μ называется *произведением* мер $\mu_{\mathbf{X}}$ и $\mu_{\mathbf{Y}}$.

Предположим, что при каждом $\theta \in \Theta$ совместное распределение $(n+m)$-мерного случайного вектора $(\mathbf{X}, \mathbf{Y})$ абсолютно непрерывно относительно введенной выше меры μ.

Плотность распределения $(n+m)$-мерного случайного вектора $(\mathbf{X},\,\mathbf{Y})$ относительно меры μ обозначим $f_\theta(\boldsymbol{x},\boldsymbol{y})$, $\boldsymbol{x}\in\mathbb{R}^n$, $\boldsymbol{y}\in\mathbb{R}^m$, $\theta\in\Theta$. Это означает, что при каждом $\theta\in\Theta$ для любых борелевских множеств $A\in\mathcal{B}_n$, $B\in\mathcal{B}_m$ справедливо представление

$$\mathsf{P}_\theta(\mathbf{X}\in A,\,\mathbf{Y}\in B)=\int\limits_{A\times B} f_\theta(\boldsymbol{x},\boldsymbol{y})\mu(d\boldsymbol{x}\times d\boldsymbol{y}). \tag{1.1}$$

По теореме Фубини (см., например, (Халмош, 1953), § 36) интеграл (1.1) можно записать в виде

$$\mathsf{P}_\theta(\mathbf{X}\in A,\,\mathbf{Y}\in B)=\int\limits_{A\times B} f_\theta(\boldsymbol{x},\boldsymbol{y})\mu(d\boldsymbol{x}\times d\boldsymbol{y})=$$

$$=\int\limits_B\left[\int\limits_A f_\theta(\boldsymbol{x},\boldsymbol{y})\mu_{\mathbf{X}}(d\boldsymbol{x})\right]\mu_{\mathbf{Y}}(dy)=\int\limits_A\left[\int\limits_B f_\theta(\boldsymbol{x},\boldsymbol{y})\mu_{\mathbf{Y}}(d\boldsymbol{y})\right]\mu_{\mathbf{X}}(d\boldsymbol{x}). \tag{1.2}$$

Так как соотношения (1.2) справедливы для любых множеств $A\in\mathcal{B}_n$, $B\in\mathcal{B}_m$, то, полагая в них $A=\mathbb{R}^n$ или $B=\mathbb{R}^m$, мы, соответственно, получаем, что функция

$$f_\theta^{\mathbf{X}}(\boldsymbol{x})=\int\limits_{\mathbf{R}^m} f_\theta(\boldsymbol{x},\boldsymbol{y})\mu_{\mathbf{Y}}(d\boldsymbol{y})$$

является *маргинальной* плотностью случайной величины $\mathbf{X}$ относительно меры $\mu_{\mathbf{X}}$, а функция

$$f_\theta^{\mathbf{Y}}(\boldsymbol{y})=\int\limits_{\mathbf{R}^n} f_\theta(\boldsymbol{x},\boldsymbol{y})\mu_{\mathbf{X}}(d\boldsymbol{x})$$

является *маргинальной* плотностью случайной величины $\mathbf{Y}$ относительно меры $\mu_{\mathbf{Y}}$.

В дальнейшем в качестве мер $\mu_{\mathbf{X}}$ и $\mu_{\mathbf{Y}}$ мы будем рассматривать либо меру Лебега, либо считающую меру.

Условная плотность случайной величины $\mathbf{Y}$ при условии $\mathbf{X}=\boldsymbol{x}$ определяется как

$$f_\theta(\boldsymbol{y}|\boldsymbol{x})=\frac{f_\theta(\boldsymbol{x},\boldsymbol{y})}{f_\theta^{\mathbf{X}}(\boldsymbol{x})},\quad \boldsymbol{y}\in\mathbb{R}^m. \tag{1.3}$$

Выражение (1.3), очевидно, имеет смысл, если $f_\theta^{\mathbf{X}}(\boldsymbol{x}) \neq 0$. Аналогично определяется условная плотность случайной величины $\mathbf{X}$ при условии $\mathbf{Y} = \boldsymbol{y}$:

$$f_\theta(\boldsymbol{x}|\boldsymbol{y}) = \frac{f_\theta(\boldsymbol{x}, \boldsymbol{y})}{f_\theta^{\mathbf{Y}}(\boldsymbol{y})}, \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{1.4}$$

Выражение (1.4) имеет смысл, если $f_\theta^{\mathbf{Y}}(\boldsymbol{y}) \neq 0$.

Соотношение (1.3) по традиции используется в качестве определения условной плотности, если обе случайные величины $\mathbf{X}$ и $\mathbf{Y}$ абсолютно непрерывны относительно меры Лебега. Если же случайные величины $\mathbf{X}$ и $\mathbf{Y}$ дискретны, то есть являются абсолютно непрерывными относительно считающей меры, то соотношение (1.3) превращается в определение условной вероятности:

$$f_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{\mathsf{P}_\theta(\mathbf{X} = \boldsymbol{x},\, \mathbf{Y} = \boldsymbol{y})}{\mathsf{P}_\theta(\mathbf{X} = \boldsymbol{x})},$$

поскольку в таком случае

$$f_\theta^{\mathbf{X}}(\boldsymbol{x}) = \int\limits_{\mathbf{R}^m} f_\theta(\boldsymbol{x}, \boldsymbol{y}) \mu_{\mathbf{Y}}(d\boldsymbol{y}) = \sum_{\boldsymbol{y}:\, \mathsf{P}_\theta(\mathbf{Y}=\boldsymbol{y})>0} \mathsf{P}_\theta(\mathbf{X} = \boldsymbol{x},\, \mathbf{Y} = \boldsymbol{y}) = \mathsf{P}_\theta(\mathbf{X} = \boldsymbol{x}).$$

Из соотношений (1.3) и (1.4) вытекает, что

$$f_\theta(\boldsymbol{x}, \boldsymbol{y}) = f_\theta(\boldsymbol{y}|\boldsymbol{x}) f_\theta^{\mathbf{X}}(\boldsymbol{x}) = f_\theta(\boldsymbol{x}|\boldsymbol{y}) f_\theta^{\mathbf{Y}}(\boldsymbol{y}). \tag{1.5}$$

Будем считать, что случайная величина $\mathbf{X}$ имеет смысл наблюдаемых данных, в то время как ненаблюдаемая (скрытая) случайная величина $\mathbf{Y}$ играет вспомогательную роль.

Зная совместную плотность $f_\theta(\boldsymbol{x}, \boldsymbol{y})$ и значение $\boldsymbol{x}$ наблюдаемой величины $\mathbf{X}$, можно формально определить *полную* функцию правдоподобия

$$L(\theta; \boldsymbol{x}, \boldsymbol{y}) = f_\theta(\boldsymbol{x}, \boldsymbol{y}), \quad \theta \in \Theta. \tag{1.6}$$

При этом функцию

$$L(\theta; \boldsymbol{x}) = f_\theta^{\mathbf{X}}(\boldsymbol{x}), \tag{1.7}$$

являющуюся классической функцией правдоподобия параметра θ при заданных наблюдениях $\boldsymbol{x}$, можно считать *неполной* функцией правдоподобия или функцией правдоподобия параметра θ при неполных данных.

EM-алгоритм предназначен для отыскания значения θ, максимизирующего функции (1.6) или (1.7) при неизвестном значении $\mathbf{Y}$. Другими словами, с помощью EM-алгоритма можно найти оценки максимального правдоподобия параметра θ.

2.

Общее описание EM-алгоритма

В результате действия EM-алгоритма, представляющего собой итерационную процедуру, вычисляется последовательность значений $\{\theta^{(m)}\}_{m\geq 1}$ параметра θ. Если задано некоторое значение $\theta^{(m)}$, то вычисление следующего значения $\theta^{(m+1)}$ можно условно подразделить на два этапа, аббревиатура наименований которых и дала название всей процедуре. Опишем эти этапы.

1°. Этап вычисления математического ожидания (*E-этап*, от английского Expectation – ожидание).

Определим функцию $Q(\theta;\theta^{(m)})$ как условное математическое ожидание логарифма полной функции правдоподобия при известном значении наблюдаемой компоненты $\mathbf{X}$:

$$Q(\theta;\theta^{(m)}) = \mathsf{E}_{\theta^{(m)}}[\log f_\theta(\mathbf{X},\mathbf{Y})|\mathbf{X}]. \tag{2.1}$$

В этом определении θ является аргументом функции $Q(\theta;\theta^{(m)})$, $\mathbf{X}$ и $\theta^{(m)}$ являются параметрами, так что в соотношении (2.1) символ $\mathsf{E}_{\theta^{(m)}}$ означает усреднение по $\mathbf{Y}$ относительно меры $\mathsf{P}_{\theta^{(m)}}$.

При известном значении $\mathbf{X} = \boldsymbol{x}$ функцию $Q(\theta;\theta^{(m)})$ можно вычислить по формуле

$$Q(\theta;\theta^{(m)}) = \int\limits_{\mathbf{R}^m} [\log f_\theta(\boldsymbol{x},\boldsymbol{y})] f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x})\mu_{\mathbf{Y}}(d\boldsymbol{y}). \tag{2.2}$$

2°. Этап максимизации (*M-этап*, от слова Maximization). На этом этапе вычисляется

$$\theta^{(m+1)} = \arg\max_\theta Q(\theta;\theta^{(m)}).$$

Итерационный процесс останавливается в соответствии с заранее согласованным критерием остановки. Например, заранее выбирается какая-нибудь метрика $\rho(\theta_1, \theta_2)$ и фиксируется малое положительное число ϵ. Процесс останавливается на m-ом шаге, если $\rho(\theta^{(m)}, \theta^{(m-1)}) < \epsilon$.

Заметим, что иногда название "EM-алгоритм" объясняют не упомянутой выше аббревиатурой английских слов *Expectation-Maximization*, но возводят к термину *Estimation-Maximization* (см., например, (Айвазян и др., 1989)). По всей вероятности, первый термин все же имеет большее отношение к сути EM-алгоритма.

3.

Монотонность EM-алгоритма

Свойство монотонности EM-алгоритма было впервые установлено в работе (Шлезингер, 1965). Впоследствии это свойство обобщенных и модифицированных версий EM-алгоритма систематически исследовалось в работах (Dempster, Laird and Rubin, 1977), (Everitt and Hand, 1981), (Boyles, 1983), (Wu, 1983), (Redner and Walker, 1984), (Jordan and Xu, 1996), (Xu and Jordan, 1996).

Недавно в работах (Neal and Hinton, 1998) и (Chrétien and Hero, 2000) было замечено, что EM-алгоритм принадлежит к классу так называемых *проксимальных алгоритмов* (или *PP-алгоритмов*, от английского термина Proximal Point algorithms) (в статье (Neal and Hinton, 1998) отмечено соответствующее ключевое свойство EM-алгоритма, но при этом EM-алгоритм формально не идентифицирован как PP-алгоритм). Это замечание существенно упрощает исследование свойства монотонности EM-алгоритма. При этом главную роль играет следующее представление функции $Q(\theta;\theta^{(m)})$.

Из соотношения (1.5) вытекает, что

$$\log f_\theta(\boldsymbol{x},\boldsymbol{y}) = \log f_\theta(\boldsymbol{y}|\boldsymbol{x}) + \log f_\theta^{\mathbf{X}}(\boldsymbol{x}).$$

Поэтому

$$Q(\theta;\theta^{(m)}) = \int\limits_{\mathbf{R}^m} [\log f_\theta(\boldsymbol{x},\boldsymbol{y})] f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) \mu_{\mathbf{Y}}(d\boldsymbol{y}) =$$

$$= \int\limits_{\mathbf{R}^m} [\log f_\theta(\boldsymbol{y}|\boldsymbol{x}) + \log f_\theta^{\mathbf{X}}(\boldsymbol{x})] f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) \mu_{\mathbf{Y}}(d\boldsymbol{y}) =$$

$$= \log f_\theta^{\mathbf{X}}(\boldsymbol{x}) + \int\limits_{\mathbf{R}^m} [\log f_\theta(\boldsymbol{y}|\boldsymbol{x})] f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) \mu_{\mathbf{Y}}(d\boldsymbol{y}). \tag{3.1}$$

Согласно принятой выше терминологии, обычная статистическая процедура поиска оценок максимального правдоподобия направлена на максимизацию по $\theta \in \Theta$ логарифма неполной функции правдоподобия (1.7), который при известном значении $\mathbf{X} = \boldsymbol{x}$ равен первому слагаемому в правой части (3.1).

3.1. Проксимальные алгоритмы (PP-алгоритмы) и их свойства

В общем виде проксимальный алгоритм (PP-алгоритм) можно описать следующим образом. Рассмотрим задачу максимизации функции $I(\theta) : \Theta \to \mathbb{R}$. Обобщенный проксимальный алгоритм задается рекуррентным соотношением

$$\theta^{(m+1)} = \arg\max_\theta \{I(\theta) - b_m \pi(\theta, \theta^{(m)})\}, \tag{3.2}$$

где $\{b_m\}_{m\geq 1}$ – последовательность положительных чисел, а $\pi(\theta, \theta^{(m)})$ – штрафная функция, удовлетворяющая условиям

1° $\pi(\theta, \theta^{(m)}) \geq 0$;

2° $\pi(\theta, \theta^{(m)}) = 0$ тогда и только тогда, когда $\theta = \theta^{(m)}$.

Самая первая версия проксимального алгоритма с $\Theta \subseteq \mathbb{R}^d$ при некотором $d \geq 1$ и $\pi(\theta, \theta^{(m)}) = \|\theta - \theta^{(m)}\|^2$ была предложена и изучена в (Martinet, 1970) и (Rockafellar, 1976). Свойства проксимальных алгоритмов подробно описаны, например, в (Васильев, 2002).

Свойство монотонности проксимального алгоритма устанавливается очень просто. Оно вытекает из определения (3.2) и свойств штрафной функции. Действительно, из (3.2) следует, что

$$I(\theta^{(m+1)}) - b_m \pi(\theta^{(m+1)}, \theta^{(m)}) = \max_\theta \{I(\theta) - b_m \pi(\theta, \theta^{(m)})\} \geq$$

$$\geq I(\theta^{(m)}) - b_m \pi(\theta^{(m)}, \theta^{(m)}) = I(\theta^{(m)}). \tag{3.3}$$

Последнее равенство имеет место, так как $\pi(\theta^{(m)}, \theta^{(m)}) = 0$ в силу свойства 2° штрафной функции. Но из (3.3) в силу свойства 1° штрафной функции вытекает, что

$$I(\theta^{(m+1)}) - I(\theta^{(m)}) \geq b_m \pi(\theta^{(m+1)}, \theta^{(m)}) \geq 0,$$

то есть

$$I(\theta^{(m+1)}) \geq I(\theta^{(m)}),$$

что и означает наличие у любого проксимального алгоритма свойства монотонности.

3.2. EM-алгоритм как проксимальный алгоритм

В этом разделе мы убедимся, что, если положить $I(\theta) = \log f_\theta^{\mathbf{X}}(\boldsymbol{x})$, специальным образом выбрать штрафную функцию $\pi(\theta, \theta^{(m)})$ и положить $b_m \equiv 1$, то описанный выше EM-алгоритм будет алгоритмом проксимального типа.

Рассмотрим *расстояние Кульбака–Лейблера* (Kullback and Leibler, 1951) (см. также, например, (Кульбак, 1967), (Cover and Thomas, 1991)) между условными плотностями $f_\theta(\boldsymbol{y}|\boldsymbol{x})$ и $f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x})$, определяемое как условное математическое ожидание логарифма отношения правдоподобия при известном значении наблюдаемой компоненты $\mathbf{X} = \boldsymbol{x}$ относительно меры $\mathsf{P}_{\theta^{(m)}}$:

$$\mathcal{D}_{KL}[f_\theta(\,\cdot\,|\boldsymbol{x}); f_{\theta^{(m)}}(\,\cdot\,|\boldsymbol{x})] = \mathsf{E}_{\theta^{(m)}}\left[\log \frac{f_{\theta^{(m)}}(\mathbf{Y}|\mathbf{X})}{f_\theta(\mathbf{Y}|\mathbf{X})}\middle|\mathbf{X} = \boldsymbol{x}\right].$$

Расстояние Кульбака–Лейблера можно вычислить по формуле

$$\mathcal{D}_{KL}[f_\theta(\,\cdot\,|\boldsymbol{x}); f_{\theta^{(m)}}(\,\cdot\,|\boldsymbol{x})] = \int\limits_{\mathbf{R}^m} f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) \log \frac{f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x})}{f_\theta(\boldsymbol{y}|\boldsymbol{x})} \mu_{\mathbf{Y}}(d\boldsymbol{y}).$$

Как известно, расстояние Кульбака–Лейблера удовлетворяет условиям

3° $\mathcal{D}_{KL}[f_\theta(\,\cdot\,|\boldsymbol{x}); f_{\theta^{(m)}}(\,\cdot\,|\boldsymbol{x})] \geq 0$;

4° $\mathcal{D}_{KL}[f_\theta(\,\cdot\,|\boldsymbol{x}); f_{\theta^{(m)}}(\,\cdot\,|\boldsymbol{x})] = 0$ тогда и только тогда, когда $\theta = \theta^{(m)}$.

Положив

$$\pi(\theta, \theta^{(m)}) = \mathcal{D}_{KL}[f_\theta(\,\cdot\,|\boldsymbol{x}); f_{\theta^{(m)}}(\,\cdot\,|\boldsymbol{x})],$$

мы таким образом замечаем, что свойства 3° и 4° расстояния Кульбака–Лейблера соответствуют условиям 1° и 2°, которые определяют штрафную функцию в определении проксимального алгоритма.

Осталось убедиться, что с $I(\theta) = \log f_\theta^{\mathbf{X}}(\boldsymbol{x})$, $b_m \equiv 1$ и расстоянием Кульбака–Лейблера в качестве штрафной функции соотношение (3.2),

определяющее проксимальный алгоритм, трансформируется в соотношение, определяющее M-этап EM-алгоритма. Действительно, при таких $I(\theta)$ и $\pi(\theta, \theta^{(m)})$ мы имеем

$$I(\theta) - \pi(\theta, \theta^{(m)}) = \log f_\theta^{\mathbf{X}}(\boldsymbol{x}) - \mathcal{D}_{KL}[f_\theta(\,\cdot\,|\boldsymbol{x}); f_{\theta^{(m)}}(\,\cdot\,|\boldsymbol{x})] =$$

$$= \log f_\theta^{\mathbf{X}}(\boldsymbol{x}) - \int\limits_{\mathbb{R}^m} f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) \log \frac{f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x})}{f_\theta(\boldsymbol{y}|\boldsymbol{x})} \mu_{\mathbf{Y}}(d\boldsymbol{y}) = \log f_\theta^{\mathbf{X}}(\boldsymbol{x}) -$$

$$- \int\limits_{\mathbb{R}^m} f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) \log f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) \mu_{\mathbf{Y}}(d\boldsymbol{y}) + \int\limits_{\mathbb{R}^m} f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) \log f_\theta(\boldsymbol{y}|\boldsymbol{x}) \mu_{\mathbf{Y}}(d\boldsymbol{y}).$$

Заметим, что второе слагаемое в правой части последнего соотношения (равное дифференциальной энтропии условного распределения $f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x})$) не зависит от θ. Поэтому

$$\arg\max_\theta \{I(\theta) - \pi(\theta, \theta^{(m)})\} =$$

$$= \arg\max_\theta \left\{ \log f_\theta^{\mathbf{X}}(\boldsymbol{x}) + \int\limits_{\mathbb{R}^m} f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) \log f_\theta(\boldsymbol{y}|\boldsymbol{x}) \mu_{\mathbf{Y}}(d\boldsymbol{y}) \right\}. \tag{3.4}$$

Но выражение в фигурных скобках в правой части (3.4) оказывается в точности равным функции $Q(\theta; \theta^{(m)})$ (см. (3.1)), откуда вытекает требуемое соотношение

$$\arg\max_\theta \{I(\theta) - \pi(\theta, \theta^{(m)})\} = \arg\max_\theta Q(\theta; \theta^{(m)})$$

Другими словами, рекуррентные соотношения

$$\theta^{(m+1)} = \arg\max_\theta \{I(\theta) - \pi(\theta, \theta^{(m)})\}$$

и

$$\theta^{(m+1)} = \arg\max_\theta Q(\theta; \theta^{(m)})$$

задают одну и ту же последовательность, то есть EM-алгоритм является специальным проксимальным алгоритмом. А так как любой проксимальный алгоритм обладает свойством монотонности (см. раздел 4.1), то свойство монотонности оказывается присущим и EM-алгоритму в том смысле, что, если последовательность $\{\theta^{(m)}\}_{m\geq 1}$ вычисляется в соответствии с правилами, определяющими EM-алгоритм, то

$$\log f_{\theta^{(m+1)}}^{\mathbf{X}}(\boldsymbol{x}) \geq \log f_{\theta^{(m)}}^{\mathbf{X}}(\boldsymbol{x}),$$

то есть

$$L(\theta^{(m+1)}; \boldsymbol{x}) \geq L(\theta^{(m)}; \boldsymbol{x}), \qquad i \geq 1.$$

3.3. Неподвижные точки EM-алгоритма и локальные максимумы функции правдоподобия

В предыдущем разделе мы убедились, что каждая итерация EM-алгоритма гарантированно увеличивает функцию правдоподобия. Однако этого недостаточно, чтобы утверждать, что последовательность оценок параметров, построенная EM-алгоритмом, гарантированно сходится к локальному максимуму функции правдоподобия $L(\theta; \boldsymbol{x})$. Чтобы установить такую сходимость, приходится предполагать, что рассматриваемые распределения удовлетворяют дополнительным условиям регулярности и, в частности, условиям гладкости. Простейшими условиями такого типа являются условия дифференцируемости по θ плотности $f_\theta^{\mathbf{X}}(\boldsymbol{x})$ и расстояния Кульбака–Лейблера $\mathcal{D}_{KL}[f_\theta(\,\cdot\,|\boldsymbol{x}); f_{\theta^{(m)}}(\,\cdot\,|\boldsymbol{x})]$ (см., например, (Tanner, 1993), (Chrétien and Hero, 2000)).

Неподвижная точка EM-алгоритма $\theta^{(\infty)}$ определяется условием

$$\theta^{(\infty)} = \arg\max_\theta \left\{ \log f_\theta^{\mathbf{X}}(\boldsymbol{x}) + \int\limits_{\mathbf{R}^m} f_{\theta^{(\infty)}}(\boldsymbol{y}|\boldsymbol{x}) \log f_\theta(\boldsymbol{y}|\boldsymbol{x}) \mu_{\mathbf{Y}}(d\boldsymbol{y}) \right\}. \tag{3.5}$$

В предположении дифференцируемости указанных функций по θ это означает, что точка $\theta^{(\infty)}$ должна быть стационарной точкой функции, стоящей в фигурных скобках в соотношении (3.5), то есть

$$\left.\frac{\partial \log f_\theta^{\mathbf{X}}(\boldsymbol{x})}{\partial\theta}\right|_{\theta=\theta^{(\infty)}} - \left.\frac{\partial \mathcal{D}_{KL}[f_\theta(\,\cdot\,|\boldsymbol{x}); f_{\theta^{(\infty)}}(\,\cdot\,|\boldsymbol{x})]}{\partial\theta}\right|_{\theta=\theta^{(\infty)}} = 0.$$

Но поскольку расстояние Кульбака–Лейблера $\mathcal{D}_{KL}[f_\theta(\,\cdot\,|\boldsymbol{x}); f_{\theta^{(m)}}(\,\cdot\,|\boldsymbol{x})]$ имеет минимум (равный нулю) при $\theta = \theta^{(\infty)}$, его частная производная по θ в точке $\theta = \theta^{(\infty)}$ равна нулю. Отсюда вытекает, что неподвижные точки EM-алгоритма являются стационарными точками функции правдоподобия, то есть при некоторых условиях регулярности последовательность оценок параметров, построенная EM-алгоритмом, действительно сходится к локальному максимуму функции правдоподобия $L(\theta; \boldsymbol{x})$.

Одновременно свойство монотонности EM-алгоритма свидетельствует о его сильной зависимости от выбора начального (стартового) приближения.

Вопросы скорости сходимости EM-алгоритма обсуждались, например, в работах (Dempster, Laird and Rubin, 1977), (Wu, 1983), (Redner and Walker, 1984), (Jordan and Xu, 1996), (Xu and Jordan, 1996).

Известны многочисленные модификации EM-алгоритма. В частности, такие, у которых на M-этапе функция $Q(\theta;\theta^{(m)})$ не максимизируется, но находится некоторое $\theta^{(m+1)}$, для которого $Q(\theta^{(m+1)};\theta^{(m)}) \geq Q(\theta;\theta^{(m)})$ (см., например, (McLachlan and Krishnan, 1997)). Соответствующие модификации EM-алгоритма принято называть *обобщенными EM-алгоритмами* или GEM-алгоритмами (от английского Generalized EM-algorithm – обобщенный EM-алгоритм). Некоторые другие модификации EM-алгоритма будут рассмотрены ниже.

4.

Некоторые свойства смесей распределений вероятностей

4.1. Основные определения

Ниже мы будем интенсивно использовать некоторые специальные свойства смесей распределений вероятностей, в первую очередь, смесей нормальных распределений. Чтобы систематически исследовать эти свойства, сначала надо напомнить строгое определение смеси вероятностных распределений.

Рассмотрим функцию $F(x, \boldsymbol{y})$, определенную на множестве $\mathbb{R} \times \mathbb{Y}$. Для простоты мы будем предполагать, что $\mathbb{Y}$ – это некоторое подмножество m-мерного евклидова пространства, $\mathbb{Y} \subseteq \mathbb{R}^m$ при некотором $m \geq 1$, причем множество $\mathbb{Y}$ снабжено борелевской σ-алгеброй Σ. Более того, предположим, что при каждом фиксированном $\boldsymbol{y}$ функция $F(x, \boldsymbol{y})$ является функцией распределения по x, а при каждом фиксированном x функция $F(x, \boldsymbol{y})$ измерима по $\boldsymbol{y}$, то есть для любых $x \in \mathbb{R}$ и $c \in \mathbb{R}$ выполнено условие $\{\boldsymbol{y} : \ F(x, \boldsymbol{y}) < c\} \in \Sigma$. Пусть Q – вероятностная мера, определенная на измеримом пространстве $(\mathbb{Y}, \Sigma)$. Функция распределения

$$H(x) = \int_{\mathbb{Y}} F(x, \boldsymbol{y}) \mathsf{Q}(d\boldsymbol{y}), \quad x \in \mathbb{R},$$

называется *смесью* функции распределения $F(x, \boldsymbol{y})$ по $\boldsymbol{y}$ относительно Q. Распределение $F(x, \boldsymbol{y})$ называется *смешиваемым*, в то время как мера Q задает *смешивающее* распределение. Если $\mathbf{Y}$ – m-мерная тождественная случайная величина (то есть $\mathbf{Y}(y) \equiv y$, $y \in \mathbb{Y}$), определенная на вероятностном пространстве $(\mathbb{Y}, \Sigma, \mathsf{Q})$, то функция распределения $H(x)$ может

быть записана в виде

$$H(x) = \mathsf{E}F(x, \mathbf{Y}), \quad x \in \mathbb{R}.$$

Если $f(x, \boldsymbol{y})$ – плотность распределения, соответствующая функции распределения $F(x, \boldsymbol{y})$,

$$f(x, \boldsymbol{y}) = \frac{d}{dx} F(x, \boldsymbol{y}),$$

то смеси $H(x)$ соответствует плотность

$$h(x) = \mathsf{E}f(x, \mathbf{Y}) = \int_{\mathbb{Y}} f(x, \boldsymbol{y}) \mathsf{Q}(d\boldsymbol{y}), \quad x \in \mathbb{R}.$$

Если случайный вектор $\mathbf{Y}$ имеет дискретное распределение и принимает значения $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ с вероятностями, соответственно, $p_1, p_2, \ldots$, то мы получаем смесь вида

$$H(x) = \mathsf{E}F(x, \mathbf{Y}) = \sum_{j \geq 1} p_j F(x, \boldsymbol{y}_j), \quad x \in \mathbb{R},$$

называемую *дискретной*. В таком случае функции распределения $F(x, \boldsymbol{y}_j)$ называются *компонентами* смеси $H(x)$, а числа p_j называются *весами* соответствующих компонент, $j \geq 1$. Если в дискретной смеси число ненулевых весов конечно (то есть случайный вектор $\mathbf{Y}$ принимает конечное число значений), то дискретная смесь называется *конечной*.

Если в таком случае функции распределения $F(x, \boldsymbol{y})$ соответствует плотность $f(x, \boldsymbol{y})$, то дискретной смеси $H(x)$ соответствует плотность

$$h(x) = \mathsf{E}f(x, \mathbf{Y}) = \sum_{j \geq 1} p_j f(x, \boldsymbol{y}_j), \quad x \in \mathbb{R}.$$

Особо подчеркнем, что при разных значениях параметра $\boldsymbol{y}$ функции распределения $F(x, \boldsymbol{y})$ могут относиться к *разным* типам распределения. Например, если смесь дискретна, то есть мера Q приписывает вероятности p_j точкам $\boldsymbol{y}_j$, $j = 1, 2, \ldots$, причем $F(x, \boldsymbol{y}_j) = F_j(x)$, где $F_j(x)$ – функции распределения, возможно, соответствующие разным типам при разных j, то

$$H(x) = \mathsf{E}F(x, \mathbf{Y}) = \sum_{j \geq 1} p_j F_j(x), \quad x \in \mathbb{R}.$$

Более того, если в таком случае компоненты F_j абсолютно непрерывны и имеют плотности f_j, то смесь $H(x)$ также будет абсолютно непрерывной с плотностью

$$h(x) = \mathsf{E} f(x, \mathbf{Y}) = \sum_{j \geq 1} p_j f_j(x), \quad x \in \mathbb{R}.$$

В дальнейшем особую роль будут играть сдвиг/масштабные смеси. Формально они определяются следующим образом. Пусть в определении, сформулированном выше, $m = 2$. Предположим, что вектор $\boldsymbol{y}$ имеет вид

$$\boldsymbol{y} = (u, v),$$

где $u > 0$ и $v \in \mathbb{R}$, так что функция распределения $F(x, \boldsymbol{y})$ допускает представление

$$F(x, \boldsymbol{y}) = F\Big(\frac{x - v}{u}\Big), \quad x \in \mathbb{R}.$$

Тогда $\mathbb{Y}$ – это положительная полуплоскость, то есть $\mathbb{Y} = \mathbb{R}^+ \times \mathbb{R}$, и функция распределения

$$H(x) = \int_{\mathbb{Y}} F\Big(\frac{x - v}{u}\Big) \mathsf{Q}(du, dv), \quad x \in \mathbb{R}, \tag{4.1}$$

называется *сдвиг/масштабной смесью* функции распределения F относительно Q. Здесь u – это параметр масштаба, а v – параметр сдвига (положения). Если функция распределения F имеет плотность f, то функции распределения $F((x - v)/u)$ соответствует плотность

$$f(x, \boldsymbol{y}) = \frac{1}{u} f\Big(\frac{x - v}{u}\Big),$$

так что смеси (4.1) соответствует плотность

$$h(x) = \int_{\mathbb{Y}} \frac{1}{u} f\Big(\frac{x - v}{u}\Big) \mathsf{Q}(du, dv), \quad x \in \mathbb{R}.$$

Если X, U и V – случайные величины, заданные на одном и том же достаточно богатом вероятностном пространстве, так что при каждом фиксированном значении (u, v) пары случайных величин (U, V) случайная величина X имеет функцию распределения $F((x-v)/u)$, то смесь (4.1) может быть записана в виде

$$H(x) = \mathsf{E} F\Big(\frac{x - V}{U}\Big), \quad x \in \mathbb{R}.$$

Более того, в таком случае из теоремы Фубини вытекает, что функция распределения $H(x)$ соответствует случайной величине $X \cdot U + V$, где случайная величина X и случайный вектор $\mathbf{Y} = (U, V)$ стохастически независимы. Кстати, легко убедиться, что в таком контексте чисто сдвиговая смесь $H(x) = \mathsf{E}F(x - V)$ является не чем иным, как функцией распределения суммы двух независимых случайных величин X и V, то есть сверткой их функций распределения. В то же время, чисто масштабная смесь $H(x) = \mathsf{E}F(x/U)$ является не чем иным, как функцией распределения произведения двух независимых случайных величин X и U.

Чтобы определить дискретную сдвиг/масштабную смесь функции распределения $F(x)$, положим $\boldsymbol{y}_j = (\sigma_j, a_j)$, где $a_j \in \mathbb{R}$, $\sigma_j > 0$, $j = 1, \ldots, k$, и в качестве специального случая приведенного выше определения дискретной смеси получим

$$H(x) = \sum_{j=1}^{k} p_j F\Big(\frac{x - a_j}{\sigma_j}\Big), \quad x \in \mathbb{R}. \tag{4.2}$$

Если при этом $a_j = 0$, $j = 1, \ldots, k$, то мы получаем чисто масштабную конечную смесь

$$H(x) = \sum_{j=1}^{k} p_j F\Big(\frac{x}{\sigma_j}\Big), \quad x \in \mathbb{R}.$$

Если функция распределения F абсолютно непрерывна и имеет плотность $f = F'$, то смеси (4.2) функций распределения соответствует смесь плотностей

$$h(x) = \sum_{j=1}^{k} \frac{p_j}{\sigma_j} f\Big(\frac{x - a_j}{\sigma_j}\Big), \quad x \in \mathbb{R}.$$

Физический смысл понятия смеси вероятностных распределений может быть проиллюстрирован на примере дискретной смеси. Рассмотрим некую популяцию, которая не является однородной и, в свою очередь, состоит из некоторого числа, скажем, k суб-популяций. Предположим, что наблюдаемый признак или наблюдаемая характеристика внутри j-й суб-популяции распределен в соответствии с функцией распределения $F_j(x) \equiv F(x, \boldsymbol{y}_j)$, которую можно интерпретировать как условную вероятность того, что значение наблюдаемого признака у случайно выбранного индивидуума будет меньше, чем x, при условии, что случайно выбранный индивидуум является представителем j-й суб-популяции. Пусть вероятность того, что при случайном выборе индивидуума из всей (генеральной) популяции будет выбран представитель именно j-й суб-популяции, равна p_j ($p_j \geq 0$,

$p_1 + \ldots + p_k = 1$). Тогда по формуле полной вероятности безусловная вероятность того, что значение наблюдаемого признака у индивидуума, случайно выбранного из всей генеральной популяции, будет меньше, чем x, окажется равной

$$H(x) = \sum_{j=1}^{k} p_j F_j(x).$$

В определенном смысле операция смешивания вероятностных распределений обеспечивает возможность формально интерпретировать популяции, реально являющиеся неоднородными, как однородные.

Очень часто нельзя непосредственно определить тип суб-популяции, к которой принадлежит очередное наблюдение. Вследствие этого вся (генеральная) популяция вынужденно считается однородной, хотя на самом деле она таковой не является и содержит индивидуумов, принадлежащих к существенно различным типам. Именно такая ситуация типична для анализа финансовых временных рядов или процессов плазменной турбулентности. Поэтому чрезвычайно важно иметь возможность осуществить операцию, в некотором смысле обратную операции смешивания, а именно, операцию разделения (расщепления) смесей. Статистические процедуры, реализующие эту операцию, описываются ниже. Эти процедуры в значительной степени зависят от свойства идентифицируемости смесей вероятностных распределений.

4.2. Идентифицируемость смесей вероятностных распределений

Понятие идентифицируемой смеси интенсивно используется в прикладных задачах, связанных с декомпозицией (разделением, разложением, расщеплением) совокупностей (популяций). В качестве примеров можно упомянуть задачи классификации, распознавания образов или идентификации вероятностных распределений. Библиография по этим вопросам обширна, см., например, обзоры (Исаенко и Урбах, 1976), (Круглов, 1991), или книги (Titterington, Smith and Makov, 1987), (Айвазян, Бухштабер, Енюков и Мешалкин, 1989), (Prakasa Rao, 1992) и списки литературы в указанных источниках.

Напомним определение идентифицируемых семейств смесей распределений вероятностей. Оно было предложено в работе (Teicher, 1961).

Пусть функция $F(x, \boldsymbol{y})$ определена на множестве $\mathbb{R} \times \mathbb{Y}$. Для простоты

предположим, что $\mathbb{Y} \subseteq \mathbb{R}^m$ при некотором $m \geq 1$ и множество $\mathbb{Y}$ снабжено борелевской σ-алгеброй Σ. Как и ранее, мы предполагаем, что функция $F(x, \boldsymbol{y})$ измерима по $\boldsymbol{y}$ при каждом фиксированном x и является функцией распределения как функция аргумента x при каждом фиксированном $\boldsymbol{y}$. Пусть $\mathcal{Q}$ – семейство случайных величин, принимающих значения во множестве $\mathbb{Y}$. Обозначим

$$\mathcal{H} = \{H_Q(x) = \mathsf{E}F(x, Q), x \in \mathbb{R} : Q \in \mathcal{Q}\}. \tag{4.3}$$

Семейство $\mathcal{H}$, определяемое ядром F и множеством $\mathcal{Q}$, называется *идентифицируемым*, если из равенства

$$\mathsf{E}F(x, Q_1) = \mathsf{E}F(x, Q_2), \qquad x \in \mathbb{R},$$

с $Q_1 \in \mathcal{Q},\ Q_2 \in \mathcal{Q}$ вытекает, что $Q_1 \stackrel{d}{=} Q_2$ (здесь и далее символ $\stackrel{d}{=}$ обозначает равенство по распределению, то есть совпадение распределений).

К примеру, идентифицируемыми являются конечные смеси нормальных распределений, показательных распределений, пуассоновских распределений и распределений Коши. Однако легко привести очень простые примеры неидентифицируемых семейств. В частности, в качестве ядра рассмотрим равномерное распределение и связанные с ним смеси

$$\frac{1}{3} \cdot 3 \cdot \mathbb{I}_{[0,\frac{1}{3})}(x) + \frac{2}{3} \cdot \frac{3}{2} \cdot \mathbb{I}_{[\frac{1}{3},1)}(x) = \frac{1}{2} \cdot 2 \cdot \mathbb{I}_{[0,\frac{1}{2})}(x) + \frac{1}{2} \cdot 2 \cdot \mathbb{I}_{[\frac{1}{2},1)}(x).$$

Здесь

$$\mathbb{I}_{[a,b)}(x) = \begin{cases} 1, & \text{если } x \in [a, b), \\ 0, & \text{если } x \notin [a, b). \end{cases}$$

В данной работе мы главным образом рассматриваем сдвиг/масштабные смеси, в которых $\boldsymbol{y} = (u, v)$, $F(x, \boldsymbol{y}) = F((x - v)/u)$. Поэтому все, что на самом деле нам нужно, – это результаты об идентифицируемости семейств сдвиг/масштабных смесей одномерных распределений.

Если X, U и V – случайные величины, определенные на одном и том же достаточно богатом вероятностном пространстве так, что (i) случайная величина X стохастически независима от пары (U, V) и (ii) для любых фиксированных значений u случайной величины U и v случайной величины V случайная величина X имеет функцию распределения $F((x - v)/u)$, то приведенное выше определение идентифицируемости применительно к сдвиг/масштабным смесям сводится к следующему. Сдвиг/масштабная

смесь $H(x) = \mathsf{E}F((x-V)/U)$, порожденная ядром F (сдвиг/масштабная смесь функции распределения F), идентифицируема, если из соотношения

$$X_1 \cdot U_1 + V_1 \stackrel{d}{=} X_2 \cdot U_2 + V_2,$$

где (X_i, U_i, V_i), $i = 1, 2$, – тройки случайных величин, обладающих точно такими же свойствами, что присущи описанным выше случайным величинам X, U, V, вытекает, что

$$(U_1, V_1) \stackrel{d}{=} (U_2, V_2).$$

Сужение определения идентифицируемости на класс конечных сдвиг/масштабных смесей сводится к следующему.

Семейство смесей

$$\mathcal{H} = \Big\{ \sum_{j=1}^{k} p_j F\Big(\frac{x-a_j}{\sigma_j}\Big) : \\ k \geq 1;\, p_j \geq 0,\, p_1 + \ldots + p_k = 1;\, a_j \in \mathbb{R},\, \sigma_j > 0,\, j = \overline{1,k} \Big\},$$

порожденное ядром F, идентифицируемо, если из равенства

$$\sum_{j=1}^{k} p_j F\Big(\frac{x-a_j}{\sigma_j}\Big) = \sum_{i=1}^{m} q_i F\Big(\frac{x-b_i}{\delta_i}\Big)$$

вытекает, что

(i) $k = m$;

(ii) для каждого индекса $j \in \{1, \ldots, k\}$ существует индекс $i \in \{1, \ldots, k\}$ такой, что

$$p_j = q_i, \quad a_j = b_i, \quad \sigma_j = \delta_i.$$

Теперь мы перейдем к рассмотрению условий идентифицируемости. Сначала рассмотрим условия идентифицируемости смесей, в которых смешивание производится либо по параметру сдвига, либо по параметру масштаба. Другими словами, сначала мы рассмотрим семейства однопараметрических смесей. Условия идентифицируемости таких семейств хорошо известны. Напомним некоторые из них.

Семейство функций распределения $\{F(x, y) : y > 0\}$ называется *аддитивно замкнутым*, если для любых $y_1 > 0, y_2 > 0$ справедливо соотношение

$$F(x, y_1) * F(x, y_2) \equiv F(x, y_1 + y_2). \tag{4.4}$$

Здесь символ $*$ обозначает свертку распределений: если F_1 и F_2 – функции распределения, то

$$F_1(x) * F_2(x) = \int_{-\infty}^{\infty} F_1(x-y)dF_2(y) = \int_{-\infty}^{\infty} F_2(x-y)dF_1(y).$$

Иногда свойство (4.4) семейств распределений вероятностей называется *воспроизводимостью* по параметру y.

Семейство нормальных законов с нулевым математическим ожиданием $\mathcal{N}_0 = \{\Phi(x/\sqrt{s}),\ s > 0\}$ является очевидным примером аддитивно замкнутого семейства (относительно дисперсии s).

Следующие результаты принадлежат Г. Тейчеру (Teicher, 1961).

Теорема 4.1. *Семейство смесей* (4.3) *функций распределения* $F(x,\cdot)$ *из аддитивно замкнутого семейства является идентифицируемым.*

Отсюда немедленно вытекает, что семейство $\mathcal{N}_0$ масштабных смесей нормальных законов с нулевым средним идентифицируемо.

Смеси, порождаемые ядрами из аддитивно замкнутых семейств, конечно же, не исчерпывают все примеры идентифицируемых смесей. Рассмотрим масштабные смеси распределений, сосредоточенных на неотрицательной полупрямой.

Теорема 4.2. *Пусть* $F(x,y) = F(xy)$, $y \geq 0$, $F(0) = 0$. *Предположим, что преобразование Фурье функции* $G^*(y) = F(e^y)$, $y \geq 0$, *нигде не обращается в нуль. Тогда семейство смесей*

$$\mathcal{H} = \{H_Q(x) = \mathsf{E}F(xQ), x \geq 0 : \mathsf{P}(Q > 0) = 1\}$$

идентифицируемо.

Аналогичное свойство присуще некоторым семействам сдвиговых смесей распределений вероятностей.

Теорема 4.3. *Пусть* $\mathcal{Q}$ – *множество всех случайных величин. Семейство сдвиговых смесей*

$$\mathcal{H} = \{H_Q(x) = \mathsf{E}F(x-Q),\ x \in \mathbb{R} : Q \in \mathcal{Q}\}$$

идентифицируемо, если характеристическая функция, соответствующая функции распределения $F(x)$, *нигде не обращается в нуль.*

Теорема 4.1 гарантирует, что наряду со смесями нормальных законов с нулевым средним, идентифицируемыми являются семейства масштабных смесей любых строго устойчивых законов.

Теорема 4.3 гарантирует идентифицируемость семейств сдвиговых смесей любых безгранично делимых (в том числе устойчивых) законов.

Здесь мы упомянули только те идентифицируемые семейства, которые так или иначе рассматриваются в данной работе в рамках описания CPC-метода. Многочисленные примеры других идентифицируемых семейств можно найти в работах (Medgyessy, 1961), (Teicher, 1961), (Teicher, 1963), (Yakowitz and Spragins, 1968).

Некоторые критерии (то есть необходимые и достаточные условия идентифицируемости семейств смесей доказаны в работе (Tallis, 1969), также см. обзор (Круглов, 1991). Итоги исследований в области идентифицируемости семейств смесей вероятностных распределений подведены в обстоятельной книге (Prakasa Rao, 1992), где подробно обсуждаются как условия идентифицируемости, так и соответствующие примеры.

К сожалению, примеры ядер, порождающих идентифицируемые семейства сдвиг/масштабных смесей, в общей ситуации (то есть без каких-либо дополнительных условий на смешивающие распределения) не известны. Такие примеры известны лишь для дискретных смешивающих законов. В частности, справедливо следующее утверждение, доказанное Г. Тейчером (Teicher, 1963).

Теорема 4.4. *Семейство конечных сдвиг/масштабных смесей нормальных законов идентифицируемо.*

Покажем, что семейство сдвиг/масштабных смесей нормальных законов при произвольном (двумерном) смешивающем законе не является идентифицируемым. Пусть $X_1,\ X_2,\ U_1$ и U_2 – независимые случайные величины такие, что X_1 и X_2 имеют одно и то же стандартное нормальное распределение, $\mathsf{P}(U_1 = 1) = 1$, а случайная величина U_2 не вырождена, то есть ни при каком u не имеет места соотношение $\mathsf{P}(U_2 = u) = 1$. Более того, предположим, что $\mathsf{P}(U_2 > 0) = 1$. Тогда распределение случайной величины $Z = X_1U_2 + X_2$ может быть записано как в виде

$$\mathsf{P}(Z < x) = \int_0^\infty \int_{-\infty}^\infty \Phi\left(\frac{x-v}{u}\right) d\mathsf{P}(X_2 < v) d\mathsf{P}(U_2 < u), \tag{4.5}$$

так и в виде

$$\mathsf{P}(Z < x) = \int_0^\infty \int_{-\infty}^\infty \Phi\left(\frac{x-v}{u}\right) d\mathsf{P}(X_1U_2 < v) d\mathsf{P}(U_1 < u). \tag{4.6}$$

Легко видеть, что смешивающие двумерные распределения в представлениях (4.5) и (4.6) различны. Таким образом, мы можем заключить, что в общем случае задача разделения сдвиг/масштабных смесей, то есть задача реконструкции двумерного смешивающего распределения по распределению смеси, является некорректной, так как она может иметь несколько различных решений. Это справедливо даже для случая сдвиг/масштабных смесей нормальных законов.

5.

Решение задачи разделения смесей вероятностных распределений с помощью EM-алгоритма

Задача отыскания наиболее правдоподобных оценок параметров смесей вероятностных распределений (задача разделения смесей), по-видимому, является одним из самых популярных приложений EM-алгоритма. Скорее всего, это обусловлено тем, что другие методы решения этой задачи оказываются малоэффективными и неустойчивыми.

Для наглядности, не ограничивая общности, мы будем рассматривать смеси одномерных распределений.

Базовым предположением в рамках данной задачи является то, что плотность наблюдаемой случайной величины X имеет вид

$$f_\theta^X(x) = \sum_{i=1}^{k} p_i \psi_i(x; t_i), \tag{5.1}$$

где $k \geq 1$ – известное натуральное число, $\psi_1, \ldots, \psi_k$ – известные плотности распределения, неизвестный параметр θ имеет вид $\theta = (p_1, \ldots, p_k, t_1, \ldots, t_k)$, причем $p_i \geq 0$, $i = 1, \ldots, k$, $p_1 + \ldots + p_k = 1$, t_i, $i = 1, \ldots, k$, – вообще говоря, многомерные параметры. Плотности $\psi_1, \ldots, \psi_k$ будем называть *компонентами* смеси (5.1), параметры $p_1, \ldots, p_k$ будем называть *весами* соответствующих компонент.

Задачей разделения смеси (5.1) принято называть задачу статистиче-

ского оценивания параметров $\theta = (p_1, \ldots, p_k, t_1, \ldots, t_k)$ по известным реализациям случайной величины X. Необходимым условием существования осмысленного решения задачи разделения смеси вероятностных распределений вида (5.1) является идентифицируемость смеси, то есть неизбежное совпадение неизвестных параметров при тождественном совпадении смесей вида (5.1) (как функций аргумента x) с одинаковыми известными параметрами. Более подробно об этом см. в предыдущем разделе, а также в работах (Teicher, 1961), (Teicher, 1963) или, например, (Айвазян и др., 1989). Многочисленные примеры, связанные с разделением различных смесей, можно найти в книгах (Everitt and Hand, 1981), (Titterington, Smith and Makov, 1987), (McLachlan and Basford, 1988), (Prakasa Rao, 1992), (McLachlan and Peel, 2000).

Предположим, что в нашем распоряжении имеется независимая выборка значений $\boldsymbol{x} = (x_1, \ldots, x_n)$ наблюдаемой случайной величины X, относительно которой, не ограничивая общности, мы будем предполагать, что ее распределение абсолютно непрерывно относительно меры Лебега (откуда необходимо вытекает, что $\psi_1, \ldots, \psi_k$ – это также плотности относительно меры Лебега). В рамках модели (5.1) логарифм классической (неполной) функции правдоподобия параметра θ имеет вид

$$\log L(\theta; \boldsymbol{x}) = \log \prod_{j=1}^{n} f_\theta^X(x_j) = \sum_{j=1}^{n} \log \Bigg(\sum_{i=1}^{k} p_i \psi_i(x_j; t_i) \Bigg).$$

Непосредственный поиск точки максимума этой функции весьма затруднителен. Однако, если мы будем трактовать наблюдения $\boldsymbol{x}$ как *неполные*, то функцию правдоподобия можно записать в намного более удобном виде.

Предположим, что наряду с наблюдаемой случайной величиной X задана *ненаблюдаемая* случайная величина Y, значения которой содержат информацию о номерах компонент, в соответствии с которыми "генерируются" наблюдения $\boldsymbol{x} = (x_1, \ldots, x_n)$. А именно, будем считать, что наблюдения организованы следующим образом. При очередном, скажем, j-ом наблюдении ($j = 1, \ldots, n$) сначала реализуется значение $y_j \in \{1, 2, \ldots, k\}$ ненаблюдаемой случайной величины Y. Это значение y_j имеет смысл номера той компоненты смеси, которая затем выбирается в качестве распределения наблюдаемой случайной величины X при j-ом измерении, результатом которого является значение x_j. Такая схема типична для задач кластерного или дискриминантного анализа, в которых каждое наблюдение может быть порождено одной и только одной компонентой смеси. Эта схема оказывается формально очень удобной для решения статистической

задачи разделения конечных смесей вида (5.1), то есть задачи статистического оценивания параметров этой смеси.

Будем предполагать, что пары значений (x_j, y_j) являются стохастически независимыми реализациями пары случайных величин (X, Y).

Совместную плотность случайных величин X и Y обозначим $f_\theta(x, y)$. По определению случайных величин X и Y, так как дискретная случайная величина Y абсолютно непрерывна относительно считающей меры и принимает значения $i = 1, 2, \ldots, k$, то ее маргинальная плотность равна

$$f_\theta^Y(i) = p_i, \quad i = 1, 2, \ldots, k,$$

в то время как условная плотность случайной величины X при фиксированном значении $Y = i$ равна

$$f_\theta(x|i) = \psi_i(x; t_i).$$

Поэтому, если бы значения $\boldsymbol{y} = (y_1, \ldots, y_n)$ были известны, то логарифм *полной* функции правдоподобия имел бы вид

$$\log L(\theta; \boldsymbol{x}, \boldsymbol{y}) = \log \prod_{j=1}^n f_\theta(x_j, y_j) = \sum_{j=1}^n \log f_\theta(x_j, y_j) =$$

$$= \sum_{j=1}^n \log[f_\theta(x_j|y_j) f_\theta^Y(y_j)] = \sum_{j=1}^n \log[p_{y_j} \psi_{y_j}(x_j; t_{y_j})] =$$

$$= \sum_{j=1}^n \log p_{y_j} + \sum_{j=1}^n \log \psi_{y_j}(x_j; t_{y_j}).$$

Обратим внимание, что в этом представлении совместная плотность распределения случайных величин X и Y относительно произведения меры Лебега и считающей меры (см. главу 2) равна

$$f_\theta(x, y) = p_y \psi_y(x; t_y).$$

Используя соотношение (1.5), мы замечаем, что, с другой стороны,

$$p_y \psi_y(x; t_y) = f_\theta(x, y) = f_\theta(y|x) f_\theta^X(x), \tag{5.2}$$

откуда с учетом вида модели (5.1) мы получаем следующее представление условной плотности ненаблюдаемой случайной величины Y при известном значении x наблюдаемой случайной величины X:

$$f_\theta(y|x) = \frac{f_\theta(x, y)}{f_\theta^X(x)} = \frac{p_y \psi_y(x; t_y)}{\sum_{i=1}^k p_i \psi_i(x; t_i)}. \tag{5.3}$$

Учитывая определение случайной величины Y в рассматриваемой задаче, мы можем интерпретировать правую часть (5.3) как *апостериорную* вероятность того, что наблюдение x было сгенерировано в соответствии с распределением, задаваемым компонентой ψ_i смеси (5.1).

Следовательно, с учетом стохастической независимости реализаций $\{(x_j, y_j)\}_{j=1}^n$ пары случайных величин (X, Y) совместная условная плотность распределения набора $\boldsymbol{y} = (y_1, \ldots, y_n)$ реализаций ненаблюдаемой случайной величины Y при фиксированных значениях $\boldsymbol{x} = (x_1, \ldots, x_n)$ наблюдаемой случайной величины X равна

$$f_\theta(\boldsymbol{y}|\boldsymbol{x}) = \prod_{j=1}^n f_\theta(y_j|x_j) = \prod_{j=1}^n \frac{p_{y_j}\psi_{y_j}(x_j; t_{y_j})}{\sum_{i=1}^k p_i\psi_i(x_j; t_i)}.$$

Подставляя полученное выражение в соотношение (2.2), определяющее функцию $Q(\theta; \theta^{(m)})$ (m – номер итерации EM-алгоритма), и вводя обозначение $\mathcal{Y} = \mathbf{N}_k^n$, где $\mathbf{N}_k = \{1, 2, \ldots, k\}$, а степень n понимается в смысле декартова (прямого) произведения множеств, мы получаем

$$Q(\theta; \theta^{(m)}) = \sum_{\boldsymbol{y}\in\mathcal{Y}} [\log f_\theta(\mathbf{x}, \boldsymbol{y})] f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x}) =$$

$$= \sum_{\boldsymbol{y}\in\mathcal{Y}} \sum_{j=1}^n \log[p_{y_j}\psi_{y_j}(x_j; t_{y_j})] \prod_{r=1}^n f_{\theta^{(m)}}(y_r|x_r) =$$

$$= \sum_{y_1=1}^k \sum_{y_2=1}^k \cdots \sum_{y_n=1}^k \sum_{j=1}^n \log[p_{y_j}\psi_{y_j}(x_j; t_{y_j})] \prod_{r=1}^n f_{\theta^{(m)}}(y_r|x_r) =$$

$$= \sum_{y_1=1}^k \sum_{y_2=1}^k \cdots \sum_{y_n=1}^k \sum_{j=1}^n \sum_{l=1}^k \delta_{l,y_j} \log[p_l\psi_l(x_j; t_l)] \prod_{r=1}^n f_{\theta^{(m)}}(y_r|x_r) =$$

$$= \sum_{l=1}^k \sum_{j=1}^n \log[p_l\psi_l(x_j; t_l)] \sum_{y_1=1}^k \sum_{y_2=1}^k \cdots \sum_{y_n=1}^k \delta_{l,y_j} \prod_{r=1}^n f_{\theta^{(m)}}(y_r|x_r), \qquad (5.4)$$

где $\delta_{a,b}$ – символ Кронекера: $\delta_{a,b} = 1$, если $a = b$, и $\delta_{a,b} = 0$, если $a \neq b$.

Чтобы упростить правую часть (5.4), заметим, что для $l \in \mathbf{N}_k$

$$\sum_{y_1=1}^k \sum_{y_2=1}^k \cdots \sum_{y_n=1}^k \delta_{l,y_j} \prod_{r=1}^n f_{\theta^{(m)}}(y_r|x_r) =$$

$$= \Bigg(\sum_{y_1=1}^{k} \cdots \sum_{y_{j-1}=1}^{k} \sum_{y_{j+1}=1}^{k} \cdots \sum_{y_n=1}^{k} \prod_{\substack{r=1 \\ r\neq j}}^{n} f_{\theta^{(m)}}(y_r|x_r) \Bigg) f_{\theta^{(m)}}(l|x_j) =$$

$$= \prod_{\substack{r=1 \\ r\neq j}}^{n} \Bigg(\sum_{y_r=1}^{k} f_{\theta^{(m)}}(y_r|x_r) \Bigg) f_{\theta^{(m)}}(l|x_j) = f_{\theta^{(m)}}(l|x_j), \qquad (5.5)$$

так как в соответствии с соотношением (5.2)

$$\sum_{i=1}^{k} f_{\theta^{(m)}}(i|x_r) = 1.$$

Используя (5.5), выражение (5.4) для $Q(\theta;\theta^{(m)})$ можно переписать в виде

$$Q(\theta;\theta^{(m)}) = \sum_{l=1}^{k}\sum_{j=1}^{n} f_{\theta^{(m)}}(l|x_j) \log[p_l \psi_l(x_j;t_l)] =$$

$$= \sum_{l=1}^{k}\sum_{j=1}^{n} f_{\theta^{(m)}}(l|x_j) \log p_l + \sum_{l=1}^{k}\sum_{j=1}^{n} f_{\theta^{(m)}}(l|x_j) \log \psi_l(x_j;t_l). \qquad (5.6)$$

Теперь видно, что для отыскания максимума функции $Q(\theta;\theta^{(m)})$ по $\theta = (p_1,\ldots,p_k,t_1,\ldots,t_k)$ можно максимизировать слагаемые в правой части (5.6) независимо друг от друга, так как они зависят от *разных* параметров: первое зависит только от весов $p_1,\ldots,p_k$, а второе – только от параметров $t_1,\ldots,t_k$ компонент смеси.

Найдем значения $p_1,\ldots,p_k$, максимизирующие первое слагаемое в правой части (5.6). С этой целью, учитывая очевидное ограничение

$$\sum_{l=1}^{k} p_l = 1,$$

воспользуемся методом неопределенных множителей Лагранжа, согласно которому искомые значения $p_1,\ldots,p_k$ удовлетворяют уравнениям

$$\frac{\partial}{\partial p_l}\Bigg[\sum_{l=1}^{k}\sum_{j=1}^{n} f_{\theta^{(m)}}(l|x_j) \log p_l + \lambda\Bigg(\sum_{l=1}^{k} p_l - 1 \Bigg)\Bigg] = 0, \quad l = 1,\ldots,k,$$

очевидно, эквивалентным уравнениям

$$\frac{1}{p_l}\sum_{j=1}^{n} f_{\theta^{(m)}}(l|x_j) + \lambda = 0, \quad l = 1,\ldots,k.$$

Просуммировав эти уравнения по l, мы приходим к заключению, что $\lambda = n$, откуда вытекает, что искомые значения весов равны

$$p_l^* = \frac{1}{n} \sum_{j=1}^{n} f_{\theta^{(m)}}(l|x_j). \tag{5.7}$$

Подразумевая, что значение $\theta^{(m)} = (p_1^{(m)}, \ldots, p_k^{(m)}, t_1^{(m)}, \ldots, t_k^{(m)})$ параметра θ на m-ой итерации EM-алгоритма известно, и учитывая (5.3), мы теперь можем выписать значения весов $p_1^{(m+1)}, \ldots, p_k^{(m+1)}$ на $(m+1)$-ой итерации EM-алгоритма. А именно, с учетом (5.7) мы имеем

$$p_l^{(m+1)} = p_l^* = \frac{1}{n} \sum_{j=1}^{n} \frac{p_l^{(m)} \psi_l(x_j; t_l^{(m)})}{\sum_{i=1}^{k} p_i^{(m)} \psi_i(x_j; t_i^{(m)})}, \qquad l = 1, \ldots, k. \tag{5.8}$$

Оптимальные значения параметров $t_1, \ldots, t_k$, доставляющие максимум второго слагаемого в правой части (5.6), зависят от конкретного аналитического выражения для $\psi_i(x; t_i)$, $i = 1, \ldots, k$. В следующем разделе мы подробно рассмотрим ситуацию, в которой $\psi_i(x; t_i)$ – нормальные плотности.

6.

Разделение конечных смесей нормальных распределений с помощью EM-алгоритма

В данном разделе мы рассмотрим EM-алгоритм применительно к анализу специального случая модели (5.1), в котором

$$\psi_i(x;t_i) = \frac{1}{\sigma_i}\phi\Big(\frac{x-a_i}{\sigma_i}\Big), \quad x \in \mathbb{R},$$

где

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{x^2}{2}\Big\}, \quad x \in \mathbb{R},$$

– плотность стандартного нормального распределения вероятностей, $t_i = (a_i, \sigma_i^2)$, $a_i \in \mathbb{R}$, $\sigma_i > 0$, $i = 1, \ldots, k$. Этот случай, пожалуй, чаще всего встречается в прикладных работах, связанных с применением EM-алгоритма.

Для данного случая апостериорная вероятность того, что наблюдение x_j было сгенерировано в соответствии с распределением, задаваемым i-ой компонентой смеси (5.1) (правая часть (5.3)) имеет вид

$$f_\theta(i|x_j) = \frac{\dfrac{p_i}{\sigma_i}\phi_i\Big(\dfrac{x_j-a_i}{\sigma_i}\Big)}{\displaystyle\sum_{r=1}^{k}\frac{p_r}{\sigma_r}\phi_r\Big(\frac{x_j-a_r}{\sigma_r}\Big)}.$$

Пусть значение

$$\theta^{(m)} = (p_1^{(m)}, \ldots, p_k^{(m)}, a_1^{(m)}, \ldots, a_k^{(m)}, \sigma_1^{(m)}, \ldots, \sigma_k^{(m)})$$

параметра θ на m-ой итерации EM-алгоритма известно. Обозначим

$$g_{ij}^{(m)} = f_{\theta^{(m)}}(i|x_j) = \frac{\dfrac{p_i^{(m)}}{\sigma_i^{(m)}}\phi_i\Big(\dfrac{x_j - a_i^{(m)}}{\sigma_i^{(m)}}\Big)}{\sum\limits_{r=1}^{k}\dfrac{p_r^{(m)}}{\sigma_r^{(m)}}\phi_r\Big(\dfrac{x_j - a_r^{(m)}}{\sigma_r^{(m)}}\Big)} =$$

$$= \frac{\dfrac{p_i^{(m)}}{\sigma_i^{(m)}}\exp\left\{-\frac{1}{2}\left(\dfrac{x_j - a_i^{(m)}}{\sigma_i^{(m)}}\right)^2\right\}}{\sum\limits_{r=1}^{k}\dfrac{p_r^{(m)}}{\sigma_r^{(m)}}\exp\left\{-\frac{1}{2}\left(\dfrac{x_j - a_r^{(m)}}{\sigma_r^{(m)}}\right)^2\right\}}.$$

Тогда в соответствии с (5.7) и (5.8) уточненное значение параметра p_i на $(m+1)$-ой итерации EM-алгоритма примет вид

$$p_i^{(m+1)} = \frac{1}{n}\sum_{j=1}^{n} g_{ij}^{(m)}. \tag{6.1}$$

В то же время в рамках рассматриваемой модели конечной смеси нормальных законов второе слагаемое в правой части (5.6) с учетом введенных обозначений оказывается равным

$$\sum_{i=1}^{k}\sum_{j=1}^{n} f_{\theta^{(m)}}(i|x_j)\log\psi_i(x_j;t_i) =$$

$$= -\sum_{i=1}^{k}\sum_{j=1}^{n} g_{ij}^{(m)}\left[\frac{1}{2}\log 2\pi + \log\sigma_i + \frac{1}{2}\left(\frac{x_j - a_i}{\sigma_i}\right)^2\right] =$$

$$= -\frac{1}{2}\log 2\pi\sum_{i=1}^{k}\sum_{j=1}^{n} g_{ij}^{(m)} - \sum_{i=1}^{k}\log\sigma_i\sum_{j=1}^{n} g_{ij}^{(m)} - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} g_{ij}^{(m)}\left(\frac{x_j - a_i}{\sigma_i}\right)^2. \tag{6.2}$$

Заметим, что первое слагаемое в правой части (6.2) не зависит от параметров a_i и σ_i. Обозначим

$$\Psi = \sum_{i=1}^{k}\log\sigma_i\sum_{j=1}^{n} g_{ij}^{(m)} + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n} g_{ij}^{(m)}\left(\frac{x_j - a_i}{\sigma_i}\right)^2.$$

Чтобы найти значения a_i и σ_i, $i = 1, \ldots, k$, которые доставляют максимум второму слагаемому в правой части (5.6), продифференцируем Ψ по a_i

и σ_i, $i = 1, \ldots, k$, приравняем эти частные производные нулю и решим полученные уравнения относительно a_i и σ_i, $i = 1, \ldots, k$. В результате получим

$$a_i^* = \frac{1}{\sum_{j=1}^n g_{ij}^{(m)}} \sum_{j=1}^n g_{ij}^{(m)} x_j,$$
$$\sigma_i^* = \left[\frac{1}{\sum_{j=1}^n g_{ij}^{(m)}} \sum_{j=1}^n g_{ij}^{(m)} (x_j - a_i^*)^2 \right]^{1/2}, \ i = 1, \ldots, k.$$

Таким образом, уточненные значения параметров a_i и σ_i на $(m+1)$-ой итерации EM-алгоритма имеют вид

$$a_i^{(m+1)} = \frac{1}{\sum_{j=1}^n g_{ij}^{(m)}} \sum_{j=1}^n g_{ij}^{(m)} x_j,$$
$$\sigma_i^{(m+1)} = \left[\frac{1}{\sum_{j=1}^n g_{ij}^{(m)}} \sum_{j=1}^n g_{ij}^{(m)} (x_j - a_i^{(m+1)})^2 \right]^{1/2}, \ i = 1, \ldots, k.$$

Эти соотношения в совокупности с (6.1) определяют EM-алгоритм для разделения конечных смесей нормальных законов. Отметим, что фактически указанные рекуррентные соотношения реализуют оба этапа (E-этап и M-этап) EM-алгоритма *одновременно*.

Необходимо отметить, что, хотя в прикладных работах EM-алгоритм чаще всего применяется к исследованию модели (5.1) с нормальными компонентами, именно эта модель не удовлетворяет условиям, гарантирующим правильную работу EM-алгоритма. А именно, во всех упоминавшихся выше работах, в которых исследовалась сходимость EM-алгоритма, его сходимость доказана при обязательном условии ограниченности логарифма функции правдоподобия (не говоря уж о том, что для сходимости проксимальных алгоритмов нужно, чтобы оптимизируемая ими функция обладала свойствами выпуклости (Васильев, 2002)). Для смесей нормальных распределений указанные условия, вообще говоря, не выполняются. Чтобы убедиться в этом, достаточно рассмотреть случай двухкомпонентной смеси, для которой логарифм функции правдоподобия имеет вид

$$\log L(a_1, a_2, \sigma_1, \sigma_2, p; \boldsymbol{x}) =$$
$$= \sum_{j=1}^n \log \left[\frac{p}{\sqrt{2\pi}\sigma_1} \exp \left\{ -\frac{1}{2} \left(\frac{x_j - a_1}{\sigma_1} \right)^2 \right\} + \frac{1-p}{\sqrt{2\pi}\sigma_2} \exp \left\{ -\frac{1}{2} \left(\frac{x_j - a_2}{\sigma_2} \right)^2 \right\} \right].$$

Если $a_i \neq x_j$ для $i = 1, 2$ и $j = 1, \ldots, n$, то, обозначив $z_{ij} = |x_j - a_i|/\sigma_i$, мы будем иметь

$$\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{ -\frac{1}{2}\left(\frac{x_j - a_i}{\sigma_i}\right)^2 \right\} = \frac{z_{ij}}{\sqrt{2\pi}\sigma_i z_{ij}} e^{-z_{ij}^2/2} \leq$$

$$\leq \frac{1}{\sqrt{2\pi}|x_j - a_i|} \max_{z>0} z e^{-z^2/2} = \frac{1}{\sqrt{2\pi e}|x_j - a_i|}.$$

Поэтому равномерно по p, σ_1 и σ_2

$$\log L(a_1, a_2, \sigma_1, \sigma_2, p; \boldsymbol{x}) \leq \sum_{j=1}^{n} \log\left(\frac{1}{\sqrt{2\pi e}|x_j - a_i|}\right) =$$

$$= n - \tfrac{1}{2}[\log(2\pi) + 1] - \sum_{j=1}^{n} \log|x_j - a_i|.$$

Однако, полагая, к примеру, $a_1 = x_1$, мы сразу замечаем, что $\log L(a_1, a_2, \sigma_1, \sigma_2, p; \boldsymbol{x})$ неограниченно возрастает как $-\log\sigma_1$ при $\sigma_1 \to 0$.

Эта особенность в совокупности с наличием большого числа стационарных точек (локальных максимумов) логарифма функции правдоподобия для модели (5.1) с большим ($k \geq 2$) числом нормальных компонент ведет к таким неприятным свойствам EM-алгоритма в рамках задачи разделения смесей нормальных законов как повышенная чувствительность (он находит не "правильные" оценки параметров, но наиболее правдоподобные, причем правдоподобие найденных EM-алгоритмом оценок оказывается заметно выше правдоподобия "правильных" оценок) и большая неустойчивость по отношению к начальному приближению и исходным данным. Эти свойства и некоторые способы противодействия им будут рассмотрены ниже.

В работе (Xu and Jordan, 1996) приведены некоторые теоретические результаты, показывающие (вопреки эвристическому заключению, данному в статье (Redner and Walker, 1984)), что EM-алгоритм оказывается более эффективным средством решения задачи разделения конечной смеси нормальных законов, нежели другие процедуры численной оптимизации, например, такие, как методы градиентного восхождения или методы сопряженных градиентов.

7.

Модификации EM-алгоритма

7.1. Медианные модификации EM-алгоритма

Как было экспериментально установлено, EM-алгоритм обладает сильной неустойчивостью по начальным данным. Например, в случае четырехкомпонентной смеси нормальных законов при объеме выборки 200–300 наблюдений замена лишь одного наблюдения другим может кардинально изменить итоговые оценки, полученные с помощью EM-алгоритма. Возможно, эта неустойчивость обусловлена тем, что стандартные (наиболее правдоподобные для случая нормального распределения) оценки математического ожидания и дисперсии (среднее арифметическое и выборочная дисперсия) при "засорении" (*контаминации*) выборки "посторонними" или "паразитными" наблюдениями становятся заметно менее эффективными по сравнению со, скажем, выборочной медианой. Этот эффект обнаружен Дж. Тьюки (Tukey, 1960) и описан, например, в (Айвазян и др., 1983) и (Королев, 2006). Формально модель контаминации Тьюки сводится к тому, что вместо "чистого" модельного распределения, интерпретируемого как *однородная* модель, в качестве модельного распределения рассматривается неоднородная модель, имеющая вид смеси исходного "чистого" распределения и некоторого другого закона, описывающего "засоряющие" наблюдения. В задаче разделения смесей по самóй сути модели, когда оцениваются параметры одной компоненты смеси, наблюдения с распределениями, соответствующими другим компонентам, являются "загрязняющими". Это обстоятельство может сыграть особенно важную роль при реализации SEM-алгоритма, описываемого ниже.

Для противодействия указанной неустойчивости EM-алгоритма можно использовать медианные модификации EM-алгоритма. Смысл этих модификаций в том, что наиболее "неустойчивые" этапы выполнения EM-алгоритма заменяются более устойчивыми. В частности, на M-этапе неустойчивые моментные оценки наибольшего правдоподобия (которые для нормальных компонент минимизируют квадратичный риск) заменяются более устойчивыми (робастными) оценками медианного типа, оптимальными в смысле среднего абсолютного отклонения.

Опишем две возможные медианные модификации M-этапа EM-алгоритма. В рамках этих модификаций параметры a_i оцениваются одинаково. Различными являются лишь оценки параметров σ_i.

Пусть числа $g_{ij}^{(m)}$ известны. По числам $g_{ij}^{(m)}$ определим "вероятности" $p_{ij}^{(m)}$ по правилу

$$p_{ij}^{(m)} = g_{ij}^{(m)} \left(\sum_{j=1}^{n} g_{ij}^{(m)} \right)^{-1}, \quad i = 1, \ldots, k; \; j = 1, \ldots, n$$

(n – объем выборки, k – число компонент смеси). Пусть $\boldsymbol{x} = (x_1, \ldots, x_n)$ – выборка. Тогда число $p_{ij}^{(m)}$ можно интерпретировать как вероятность того, что наблюдение x_j имеет распределение, определяемое i-й компонентой смеси.

Введем "фиктивные" случайные величины $\xi_i^{(m)}$, $i = 1, \ldots, k$, которые, соответственно, принимают значение x_j с вероятностями $p_{ij}^{(m)}$, $i = 1, \ldots, k$, $j = 1, \ldots, n$ (несложно видеть, что

$$\sum_{j=1}^{n} p_{ij}^{(m)} = 1 \Big).$$

При этом оценка параметра сдвига i-й компоненты смеси на $(m+1)$-й итерации, приведенная в предыдущем разделе, оказывается в точности равной математическому ожиданию случайной величины $\xi_i^{(m)}$:

$$a_i^{(m+1)} = \frac{1}{\sum_{j=1}^{n} g_{ij}^{(m)}} \sum_{j=1}^{n} g_{ij}^{(m)} x_j = \sum_{j=1}^{n} p_{ij}^{(m)} x_j = \mathsf{E}_{\theta^{(m)}} \xi_i^{(m)}.$$

Для того чтобы построить модификацию EM-алгоритма, более устойчивую по отношению к наличию "засоряющих" наблюдений (а при оценивании параметров какой-либо компоненты смеси наблюдения, распределения которых соответствуют другим компонентам, неизбежно будут "засоряющими" по отношению к оцениваемой компоненте), в качестве оценки

параметра a_i на $(m+1)$-й итерации предлагается взять медиану $\operatorname{med}\xi_i^{(m)}$ случайной величины $\xi_i^{(m)}$, которую можно вычислить так. Переупорядочим значения $x_1,\ldots,x_n$ случайной величины $\xi_i^{(m)}$ по неубыванию. Получим вариационный ряд $x_{(1)},\ldots,x_{(n)}$. Ясно, что одно и то же переупорядочение имеет место для значений всех случайных величин $\xi_i^{(m)}$. Одновременно переставятся и вероятности $p_{ij}^{(m)}$, соответствующие значениям каждой случайной величины $\xi_i^{(m)}$. Пусть $\widehat{p}_{ij}^{(m)}$ – это та из вероятностей $p_{ij}^{(m)}$, которая соответствует значению $x_{(j)}$ случайной величины $\xi_i^{(m)}$. Положим

$$J_i=\min\{j:\widehat{p}_{i1}^{(m)}+\widehat{p}_{i2}^{(m)}+\ldots+\widehat{p}_{ij}^{(m)}\geq\tfrac{1}{2}\}.$$

Тогда

$$a_i^{(m+1)}=\operatorname{med}\xi_i^{(m)}=x_{(J_i)}. \tag{7.1}$$

Для оценивания параметра σ_i на $(m+1)$-ой итерации сначала по указанной выше схеме вычислим медиану случайной величины $\bigl|\xi_i^{(m)}-a_i^{(m+1)}\bigr|$,

$$\widehat{m}_i^{(m+1)}=\operatorname{med}\bigl|\xi_i^{(m)}-a_i^{(m+1)}\bigr|.$$

Затем введем "фиктивную" случайную величину $\zeta_i^{(m+1)}$ с функцией распределения

$$\mathsf{P}_{\theta^{(m+1)}}(\zeta_i^{(m+1)}<x)=\Phi\biggl(\frac{x-a_i^{(m+1)}}{\sigma_i^{(m+1)}}\biggr),$$

то есть распределение случайной величины $\zeta_i^{(m+1)}$ является i-й компонентой смеси. "Эмпирическим" аналогом случайной величины $\zeta_i^{(m+1)}$ является случайная величина $\xi_i^{(m)}$, введенная ранее. В идеале (при достаточно большом m и при большом n) должно быть справедливо приближенное равенство $\mathsf{P}_{\theta^{(m+1)}}(\zeta_i^{(m+1)}<x)\approx\mathsf{P}_{\theta^{(m+1)}}(\xi_i^{(m)}<x)$, $-\infty<x<+\infty$.

Таким образом, отыскав эмпирическую медиану $\widehat{m}_i^{(m+1)}$ (то есть медиану случайной величины $\bigl|\xi_i^{(m)}-a_i^{(m+1)}\bigr|$), в соответствии с идеологией метода моментов мы можем сказать, что она близка к медиане $\mu_i^{(m+1)}$ случайной величины $\bigl|\zeta_i^{(m+1)}-a_i^{(m+1)}\bigr|$.

Медиана $\mu_i^{(m+1)}$ случайной величины $\bigl|\zeta_i^{(m+1)}-a_i^{(m+1)}\bigr|$ определяется из условия

$$\mathsf{P}_{\theta^{(m+1)}}\bigl(\bigl|\zeta_i^{(m+1)}-a_i^{(m+1)}\bigr|\leq\mu_i^{(m+1)}\bigr)=\tfrac{1}{2}.$$

Но

$$\mathsf{P}_{\theta^{(m+1)}}\bigl(\bigl|\zeta_i^{(m+1)}-a_i^{(m+1)}\bigr|\leq\mu_i^{(m+1)}\bigr)=$$

$$= \mathsf{P}_{\theta^{(m+1)}}(-\mu_i^{(m+1)} \le \zeta_i^{(m+1)} - a_i^{(m+1)} \le \mu_i^{(m+1)}) =$$
$$= \mathsf{P}_{\theta^{(m+1)}}(a_i^{(m+1)} - \mu_i^{(m+1)} \le \zeta_i^{(m+1)} \le a_i^{(m+1)} + \mu_i^{(m+1)}) =$$
$$= \Phi\left(\frac{(a_i^{(m+1)} + \mu_i^{(m+1)}) - a_i^{(m+1)}}{\sigma_i^{(m+1)}}\right) - \Phi\left(\frac{(a_i^{(m+1)} - \mu_i^{(m+1)}) - a_i^{(m+1)}}{\sigma_i^{(m+1)}}\right) =$$
$$= \Phi\left(\frac{\mu_i^{(m+1)}}{\sigma_i^{(m+1)}}\right) - \Phi\left(-\frac{\mu_i^{(m+1)}}{\sigma_i^{(m+1)}}\right) = 2\Phi\left(\frac{\mu_i^{(m+1)}}{\sigma_i^{(m+1)}}\right) - 1.$$

Следовательно, справедливо соотношение

$$2\Phi\left(\frac{\mu_i^{(m+1)}}{\sigma_i^{(m+1)}}\right) - 1 = \frac{1}{2},$$

то есть

$$\Phi\left(\frac{\mu_i^{(m+1)}}{\sigma_i^{(m+1)}}\right) = \frac{3}{4},$$

что эквивалентно соотношению

$$\frac{\mu_i^{(m+1)}}{\sigma_i^{(m+1)}} = u_{3/4},$$

где $u_{3/4}$ – квантиль порядка $3/4$ стандартного нормального закона. В таблицах находим $u_{3/4} \approx 0.6745$. Следуя идеологии метода моментов, приравняем эмпирическую медиану $\widehat{m}_i^{(m+1)}$ теоретической медиане $\mu_i^{(m+1)}$ и окончательно получим уравнение для оценки параметра σ_i на $(m+1)$-ой итерации:

$$\sigma_i^{(m+1)} = \frac{\widehat{m}_i^{(m+1)}}{u_{3/4}} = 1.4826\,\widehat{m}_i^{(m+1)}. \tag{7.2}$$

Оценки $p_i^{(m+1)}$ весов p_i в модели (5.1) ищутся, как и ранее, по формулам (6.1). Числа же $g_{ij}^{(m+1)}$ на каждой итерации переназначаются так же, как и ранее, а именно,

$$g_{ij}^{(m+1)} = \frac{\dfrac{p_i^{(m+1)}}{\sigma_i^{(m+1)}} \exp\left\{-\dfrac{1}{2}\left(\dfrac{x_j - a_i^{(m+1)}}{\sigma_i^{(m+1)}}\right)^2\right\}}{\displaystyle\sum_{r=1}^{k} \frac{p_r^{(m+1)}}{\sigma_r^{(m+1)}} \exp\left\{-\frac{1}{2}\left(\frac{x_j - a_r^{(m+1)}}{\sigma_r^{(m+1)}}\right)^2\right\}}. \tag{7.3}$$

Итак, соотношения (6.1), (7.1), (7.2) и (7.3) определяют первую медианную модификацию EM-алгоритма.

Вторая медианная модификация EM-алгоритма отличается от первой лишь способом оценивания параметров σ_i. А именно, вычислим $\mathsf{E}_{\theta^{(m+1)}}\bigl|\zeta_i^{(m+1)} - a_i^{(m+1)}\bigr|$. Имеем

$$\mathsf{E}_{\theta^{(m+1)}}\bigl|\zeta_i^{(m+1)} - a_i^{(m+1)}\bigr| = \int\limits_{-\infty}^{\infty} \bigl|x - a_i^{(m+1)}\bigr| d_x\Phi\left(\frac{x - a_i^{(m+1)}}{\sigma_i^{(m+1)}}\right) =$$

$$= 2\int\limits_{0}^{\infty} x d_x\Phi\left(\frac{x}{\sigma_i^{(m+1)}}\right) = \sigma_i^{(m+1)}\sqrt{\frac{2}{\pi}}.$$

Эмпирическим аналогом величины $\mathsf{E}_{\theta^{(m+1)}}\bigl|\zeta_i^{(m+1)} - a_i^{(m+1)}\bigr|$ является величина

$$s_i^{(m+1)} = \mathsf{E}_{\theta^{(m)}}\bigl|\xi_i^{(m)} - a_i^{(m+1)}\bigr| = \sum_{j=1}^{n} p_{ij}^{(m)}\bigl|x_j - a_i^{(m+1)}\bigr|.$$

Реализуя метод моментов и приравнивая величину $\mathsf{E}_{\theta^{(m+1)}}\bigl|\zeta_i^{(m+1)} - a_i^{(m+1)}\bigr|$ ее эмпирическому аналогу, мы получаем еще одну оценку для параметра σ_i на $(m+1)$-ой итерации:

$$\sigma_i^{(m+1)} = \sqrt{\frac{\pi}{2}} \cdot s_i^{(m+1)} = 1.2533 \cdot s_i^{(m+1)}. \tag{7.4}$$

Таким образом, вторая медианная модификация EM-алгоритма определяется соотношениями (6.1), (7.1), (7.4) и (7.3).

Заметим, что вторая модификация более соответствует духу так называемой L_1-теории устойчивого оценивания в силу известного свойства

$$\arg\min_a \mathsf{E}_{\theta^{(m+1)}}\bigl|\zeta_i^{(m+1)} - a\bigr| = \operatorname{med}\zeta_i^{(m+1)} \quad \bigl(\approx \operatorname{med}\xi_i^{(m)} = a_i^{(m+1)}\bigr).$$

7.2. SEM-алгоритм

Как уже отмечалось, классический EM-алгоритм относится к категории так называемых “жадных” алгоритмов (greedy algorithms) в том смысле, что он “бросается” на первый попавшийся локальный максимум. Другими словами, будучи методом *локальной оптимизации*, он приводит не к глобальному максимуму функции правдоподобия, а к тому локальному максимуму, который, грубо говоря, является ближайшим к начальному приближению.

Самый простой способ противодействия этому свойству заключается в том, чтобы, не ограничиваясь единственным начальным приближением и, соответственно, единственной траекторией EM-алгоритма, реализовать несколько траекторий, задавая (например, случайно) нескольких различных начальных приближений, а затем выбрать тот из результатов, для которого правдоподобие является наибольшим среди всех реализованных траекторий EM-алгоритма. Однако при таком подходе остается неясным ответ на вопрос о том, каким механизмом разумнее всего пользоваться при переходе от одного начального приближения к другому. В частности, когда начальное приближение задается случайно, без дополнительной информации нельзя исчерпывающим образом определить распределение вероятностей, в соответствии с которым следует генерировать очередное начальное приближение.

Другой, оказавшийся весьма эффективным, способ заключается как бы в случайном, но целенаправленном "встряхивании" наблюдений (выборки) на каждой итерации. Этот способ лежит в основе SEM-алгоритма, название которого является аббревиатурой термина Stochastic EM-algorithm (стохастический (или случайный) EM-алгоритм). SEM-алгоритм, предложенный в работах (Broniatowski, Celeux and Diebolt, 1983), (Celeux and Diebolt, 1984), (Celeux and Diebolt, 1985), весьма прост.

Чтобы описать SEM-алгоритм, представим ненаблюдаемую информацию в иной форме (однако, по сути, эквивалентной старой форме). А именно, будем считать, что каждому наблюдению x_j соответствует *вектор* $\vec{y}_j = (y_{1j}, y_{2j}, \ldots, y_{kj})$, $j = 1, \ldots, n$, где k – число компонент смеси, n – объем выборки. При этом

$$y_{ij} = \begin{cases} 1, & \text{если наблюдение } x_j \text{ порождено } i\text{-й компонентой смеси,} \\ 0, & \text{в противном случае.} \end{cases}$$

Связь между "старой" y_j и "новой" $\vec{y}_j$ формами представления ненаблюдаемой информации такова: y_j равен такому номеру i, для которого $y_{ij} = 1$. При каждом j единице равна только одна из компонент вектора $\vec{y}_j$, остальные компоненты этого вектора равны нулю.

В терминах величин $\boldsymbol{y} = \{\vec{y}_j = (y_{1j}, y_{2j}, \ldots, y_{kj}),\ j = 1, \ldots, n\}$ логарифм полной функции правдоподобия для модели (5.1) принимает вид

$$\log L(\theta; \boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{n} \sum_{i=1}^{k} y_{ij} \log[p_i \psi_i(x_j; t_i)] =$$

$$= \sum_{i=1}^{k} \log p_i \sum_{j=1}^{n} y_{ij} + \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij} \log \psi_i(x_j; t_i). \tag{7.5}$$

Векторы $\vec{y}_j = (y_{1j}, y_{2j}, \ldots, y_{kj})$, $j = 1, \ldots, n$, разбивают исходную наблюдаемую выборку $\boldsymbol{x}$ на k классов (кластеров) $\mathcal{K}_1, \ldots, \mathcal{K}_k$:

$$\boldsymbol{x} = \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_k.$$

Для каждого $i = 1, \ldots, k$ с формальной точки зрения $\mathcal{K}_i$ – это множество тех наблюдений x_j, каждому из которых соответствует $y_{ij} = 1$. При этом каждое наблюдение x_j входит ровно в один кластер, то есть $\mathcal{K}_i \cap \mathcal{K}_j = \varnothing$ при $i \neq j$. Пусть ν_i – это число наблюдений, попавших в кластер $\mathcal{K}_i$, $i = 1, \ldots, k$,

$$\nu_i = \sum_{j=1}^{n} y_{ij}.$$

Очевидно, что $\nu_1 + \ldots + \nu_k = n$. Тогда, продолжая (7.5), для логарифма полной функции правдоподобия в модели (5.1) мы получаем представление

$$\log L(\theta; \boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{k} \nu_i \log p_i + \sum_{i=1}^{k} \sum_{j:\, x_j \in \mathcal{K}_i} \log \psi_i(x_j; t_i). \tag{7.6}$$

Если бы величины y_{ij} были известны, то искать значение θ, максимизирующее функцию правдоподобия (7.6), можно было бы, максимизируя по θ каждое из слагаемых в правой части (7.6), поскольку эти слагаемые зависят только от "своих" групп параметров. А именно, с помощью метода неопределенных множителей Лагранжа несложно убедиться, что максимум первого слагаемого по набору $p_1, \ldots, p_k$ при очевидном ограничении $p_1 + \ldots + p_k = 1$ достигается при

$$p_i^* = \frac{\nu_i}{n} \tag{7.7}$$

(см. (5.7)). Далее заметим, что

$$\sum_{j:\, x_j \in \mathcal{K}_i} \log \psi_i(x_j; t_i) = \log \prod_{j:\, x_j \in \mathcal{K}_i} \psi_i(x_j; t_i) \equiv \log L_i(t_i; \mathcal{K}_i),$$

где $L_i(t_i; \mathcal{K}_i)$ – это функция правдоподобия параметра t_i, построенная по подвыборке (кластеру) $\mathcal{K}_i$ в предположении, что каждый элемент подвыборки имеет плотность распределения $\psi_i(x; t_i)$. Отсюда видно, что, полагая

$$t_i^* = \arg\max_t L_i(t; \mathcal{K}_i), \qquad i = 1, \ldots, k, \tag{7.8}$$

мы доставляем максимум второму слагаемому в правой части (7.6). Легко видеть, что соотношение (7.8) определяет обычные оценки наибольшего правдоподобия для параметров i-ой компоненты смеси (5.1), построенные по подвыборке наблюдений, распределение которых равно этой компоненте, то есть по кластеру $\mathcal{K}_i$.

Таким образом, если бы величины y_{ij} были известны, то оценки наибольшего правдоподобия параметров модели (5.1) определялись бы соотношениями (7.7) и (7.8). Однако на практике величины y_{ij} не известны. Идея SEM-алгоритма заключается в том, что эти величины определяются с помощью специального имитационного моделирования. Итерационный SEM-алгоритм определяется так.

Предположим, что известны значения $g_{ij}^{(m)}$ апостериорных вероятностей принадлежности наблюдения x_j к кластеру $\mathcal{K}_i$, $i = 1, \ldots, k$; $j = 1, \ldots, n$; m – номер итерации (отметим, что

$$\sum_{i=1}^{k} g_{ij}^{(m)} = 1$$

для каждого j и при каждом m).

На первом этапе SEM-алгоритма (S-этапе, от слов Stochastic или Simulation) для каждого $j = 1, \ldots, n$ генерируются векторы $\vec{y}_j\,^{(m+1)} = (y_{1j}^{(m+1)}, y_{2j}^{(m+1)}, \ldots, y_{kj}^{(m+1)})$ как реализации случайных векторов с полиномиальным распределением с параметрами 1 и $g_{1j}^{(m)}, \ldots, g_{kj}^{(m)}$ ($g_{ij}^{(m)}$ – это вероятность того, что величина $y_{ij}^{(m+1)}$ равна единице). По векторам $\vec{y}_j\,^{(m+1)}$ определяется разбиение выборки $\boldsymbol{x} = (x_1, \ldots, x_n)$ на кластеры $\mathcal{K}_1^{(m+1)}, \ldots, \mathcal{K}_k^{(m+1)}$ и соответствующие числа $\nu_1^{(m+1)}, \ldots, \nu_k^{(m+1)}$ (численности кластеров) на $(m+1)$-й итерации. (Можно сказать, что на S-этапе реализуется случайное "встряхивание" исходной выборки, о котором говорилось выше.)

На втором этапе (M-этапе), этапе максимизации, в соответствии с формулами (7.7) и (7.8) вычисляются оценки максимального правдоподобия компонент параметра θ:

$$p_i^{(m+1)} = \frac{\nu_i^{(m+1)}}{n}, \tag{7.9}$$

$$t_i^{(m+1)} = \arg\max_t L_i(t; \mathcal{K}_i^{(m+1)}), \qquad i = 1, \ldots, k. \tag{7.10}$$

Наконец, на третьем этапе (E-этапе), переназначаются вероятности g_{ij}. Название этого этапа восходит к слову Expectation. Это обусловлено тем,

что, если $\vec{Y}_j{}^{(m+1)} = (Y_{1j}{}^{(m+1)}, \ldots, Y_{kj}{}^{(m+1)})$ – это случайный вектор, реализацией которого является вектор $\vec{y}_j{}^{(m+1)}$, а $\vec{X} = (X_1, \ldots, X_n)$ – это случайный вектор, реализацией которого является выборка $\boldsymbol{x} = (x_1, \ldots, x_n)$, то по определению

$$g_{ij}^{(m+1)} = \mathsf{E}_{\theta^{(m+1)}}(Y_{ij}{}^{(m+1)} | X_j)$$

(очевидно, $Y_{ij}^{(m+1)}$ – это индикатор (случайного) события $\{X_j \in \mathcal{K}_i^{(m+1)}\}$, а математическое ожидание индикатора случайного события равно вероятности этого события). При известном значении $X_j = x_j$ в соответствии с (5.3) мы имеем

$$g_{ij}^{(m+1)} = \frac{p_i^{(m+1)} \psi_i(x_j; t_i^{(m+1)})}{\sum_{r=1}^{k} p_r^{(m+1)} \psi_r(x_j; t_r^{(m+1)})}. \tag{7.11}$$

Для случая смеси нормальных распределений, в которой

$$\psi_i(x; t_i) = \frac{1}{\sigma_i} \phi\Big(\frac{x - a_i}{\sigma_i}\Big), \qquad x \in \mathbb{R},$$

(см. главу 6) SEM-алгоритм выглядит так. Соотношение (7.9) остается без изменений, соотношение (7.10) трансформируется в пару соотношений

$$a_i^{(m+1)} = \frac{1}{\nu_i^{(m+1)}} \sum_{j=1}^{n} y_{ij}^{(m+1)} x_j = \frac{1}{\nu_i^{(m+1)}} \sum_{j:\, x_j \in \mathcal{K}_i^{(m+1)}} x_j, \tag{7.12}$$

$$\sigma_i^{(m+1)} = \left[\frac{1}{\nu_i^{(m+1)}} \sum_{j=1}^{n} y_{ij}^{(m+1)} \Big(x_j - a_i^{(m+1)} \Big)^2 \right]^{1/2} =$$

$$= \left[\frac{1}{\nu_i^{(m+1)}} \sum_{j:\, x_j \in \mathcal{K}_i^{(m+1)}} \Big(x_j - a_i^{(m+1)} \Big)^2 \right]^{1/2}.$$

Соотношение же (7.11) принимает вид (7.3).

Свойства SEM-алгоритма были подвергнуты эмпирическому исследованию в (Celeux and Diebolt, 1984) и теоретическому исследованию в (Diebolt and Celeux, 1993), (Ip, 1994) и (Diebolt and Ip, 1996). В частности, в этих работах для многих достаточно общих конкретных случаев отмечено, что построенная SEM-алгоритмом последовательность $\{\theta^{(m)}\}_{m \geq 1}$, вообще говоря, не сходится с вероятностью единица, но образует цепь Маркова, которая при некоторых дополнительных условиях регулярности довольно быстро сходится к стационарному распределению. Стационарность

достигается после довольно продолжительного периода "приработки" алгоритма. При этом получаемые с помощью SEM-алгоритма оценки параметров смеси являются асимптотически несмещенными в том смысле что оценка максимального правдоподобия параметров смеси является асимптотически эквивалентной математическому ожиданию $\theta^{(m)}$ относительно стационарного распределения. Поэтому в качестве "окончательной" оценки $\widetilde{\theta}^{(m)}$ параметра θ после m итераций SEM-алгоритма в упомянутых работах предлагается использовать "выборочное среднее"

$$\widetilde{\theta}^{(m)} = \widetilde{\theta}^{(m)}(m_0) = \frac{1}{m - m_0} \sum_{r=m_0+1}^{m} \theta^{(m)},$$

где m_0 – настолько большое число, что при $r > m_0$ цепь Маркова $\theta^{(r)}$ близка к стационарному режиму.

Многочисленные реализации SEM-алгоритма показали, что он работает относительно быстро по сравнению с другими методами, результаты его работы практически не зависят от начального приближения, позволяет избегать выхода на неустойчивые локальные максимумы анализируемой функции правдоподобия за счет постоянного случайного "встряхивания" выборки и, более того – как правило, приводит к глобальному максимуму этой функции. Кроме того, SEM-алгоритм легко модифицировать с целью отыскания числа k компонент смеси, если оно заранее не известно (об этом см. ниже).

Так как на каждой итерации SEM-алгоритма в каждый из кластеров $\mathcal{K}_1^{(m+1)}, \ldots, \mathcal{K}_k^{(m+1)}$ могут попасть "лишние" наблюдения, фактически распределенные в соответствии с другими компонентами смеси, то в силу причин, о которых говорилось в предыдущем разделе, устойчивые медианные модификации могут быть весьма перспективными для SEM-алгоритма. Медианные модификации SEM-алгоритма определяются следующим образом.

Упорядочим элементы выборки $\boldsymbol{x} = (x_1, \ldots, x_n)$, попавшие в кластер $\mathcal{K}_i^{(m+1)}$, по неубыванию. Полученный в результате набор обозначим $\mathcal{K}_i^{(m+1)} = \{x_{i,1}^{(m+1)}, \ldots, x_{i,\nu_i}^{(m+1)}\}$. Тогда в случае смеси нормальных компонент вместо (7.12) можно использовать более устойчивую оценку

$$a_i^{(m+1)} = \begin{cases} \frac{1}{2}\left(x_{i,\nu_i^{(m+1)}/2} + x_{i,\nu_i^{(m+1)}/2+1}\right), & \text{если } \nu_i^{(m+1)} - \text{четное}, \\ x_{i,[\nu_i^{(m+1)}/2]+1}, & \text{если } \nu_i^{(m+1)} - \text{нечетное}, \end{cases}$$

где символ $[z]$ обозначает целую часть числа z. Другими словами, в качестве оценки параметра a_i на $(m+1)$-й итерации SEM-алгоритма можно

использовать выборочную медиану кластера $\mathcal{K}_i^{(m+1)}$.

Исходя из тех же рассуждений, что и в предыдущем разделе, в качестве оценки параметра σ_i на $(m+1)$-й итерации SEM-алгоритма можно взять

$$\sigma_i^{(m+1)} = \sqrt{\frac{\pi}{2}} \cdot S_i^{(m+1)} = 1.2533 \cdot S_i^{(m+1)},$$

где

$$S_i^{(m+1)} = \frac{1}{\nu_i^{(m+1)}} \sum_{j=1}^{\nu_i^{(m+1)}} \left| x_{i,j}^{(m+1)} - a_i^{(m+1)} \right|$$

– выборочное среднее абсолютное отклонение, вычисленное для кластера $\mathcal{K}_i^{(m+1)}$.

7.3. CEM-алгоритм

В работах (Celeux and Govaert, 1991), (Celeux and Govaert, 1992) был предложен так называемый *классификационный* EM-алгоритм или CEM-алгоритм (аббревиатура английского названия Classification EM-algorithm). Этот алгоритм совпадает с SEM-алгоритмом за исключением того, что вместо S-этапа для определения величин $\vec{y}_j = (y_{1j}, y_{2j}, \ldots, y_{kj})$ используется детерминированное правило, эквивалентное классификации по принципу максимума апостериорной вероятности, согласно которому для каждого j наблюдение x_j приписывается к тому из кластеров $\mathcal{K}_1, \ldots, \mathcal{K}_k$, вероятность принадлежности x_j к которому является наибольшей. А именно, на $(m+1)$-й итерации CEM-алгоритма равной единице полагается та из компонент вектора $\vec{y}_j{}^{(m+1)} = (y_{1j}^{(m+1)}, y_{2j}^{(m+1)}, \ldots, y_{kj}^{(m+1)})$, номер которой равен номеру наибольшего из чисел $g_{1j}^{(m+1)}, g_{2j}^{(m+1)}, \ldots, g_{kj}^{(m+1)}$.

Формулы для оценок параметров на очередной итерации CEM-алгоритма (в том числе и его медианной модификации) идентичны соответствующим формулам для SEM-алгоритма.

7.4. MCEM и SAEM-алгоритмы

7.4.1. Классический MCEM-алгоритм

При решении некоторых прикладных задач с помощью классического EM-алгоритма вычисление функции $Q(\theta; \theta^{(m)})$ по формуле (2.2) на E-этапе мо-

жет быть весьма трудоемким из-за того, что соответствующий интеграл невозможно привести к замкнутому аналитическому виду в терминах элементарных функций. Чтобы обойти это препятствие, в работе (Wei and Tanner, 1990) предложен MCEM-алгоритм (Monte-Carlo EM-algorithm). Суть этого алгоритма заключается в том, что функция $Q(\theta;\theta^{(m)})$ заменяется ее аналогом $\overline{Q}(\theta;\theta^{(m)})$, вычисленным с помощью имитационного моделирования как

$$\overline{Q}(\theta;\theta^{(m)}) = \frac{1}{N_m}\sum_{r=1}^{N_m}\log f_\theta(\boldsymbol{x},\boldsymbol{y}_r), \tag{7.13}$$

где $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{N_m}$ – независимые псевдослучайные реализации ненаблюдаемой величины $\boldsymbol{y}$, вычисленные (симулированные) в соответствии с условным распределением $f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x})$. При этом в силу закона больших чисел при достаточно больших N_m имеет место приближенное равенство

$$\overline{Q}(\theta;\theta^{(m)}) \approx Q(\theta;\theta^{(m)}). \tag{7.14}$$

Далее на M-этапе $(m+1)$-й итерации MCEM-алгоритма ищется такое значение $\theta^{(m+1)}$ параметра θ, для которого

$$\overline{Q}(\theta^{(m+1)};\theta^{(m)}) \geq \overline{Q}(\theta;\theta^{(m)})$$

для любого $\theta\in\Theta$.

Хотя аппроксимация (7.14) в целом позволяет преодолеть вычислительные сложности на E-этапе, она привносит в метод дополнительную погрешность

$$\varepsilon^{(m)}(N_m) = \left|\int[\log f_\theta(\boldsymbol{x},\boldsymbol{y})]f_{\theta^{(m)}}(\boldsymbol{y}|\boldsymbol{x})\mu_{\mathbf{Y}}(d\boldsymbol{y}) - \frac{1}{N_m}\sum_{r=1}^{N_m}\log f_\theta(\boldsymbol{x},\boldsymbol{y}_r)\right|,$$

величина которой зависит от объема N_m имитируемой выборки значений ненаблюдаемой величины $\boldsymbol{y}$. В принципе, величина $\varepsilon^{(m)}(N_m)$ может быть сделана произвольно малой за счет простого увеличения N_m. Однако на практике вычислительные ресурсы (время, мощность процессора, объем памяти) всегда ограничены. Более того, как отмечено в работе (Wei and Tanner, 1990), для гарантированной сходимости MCEM-алгоритма допустимы умеренные значения N_m на начальных итерациях, но на заключительном этапе необходима все бóльшая и бóльшая точность, то есть все меньшие и меньшие значения $\varepsilon^{(m)}(N_m)$ и, следовательно, все бóльшие и бóльшие значения N_m. Вопросам оптимального распределения объема симуляции и вычислительных ресурсов по итерациям MCEM-алгоритма посвящены работы (Booth and Hobert, 1999), (Levine and Casella, 2001), (Levine and Fan, 2004), (Caffo et al., 2005).

7.4.2. MC-модификация SEM-алгоритма

Легко видеть, что с формальной точки зрения SEM-алгоритм, описанный в разделе 7.2, является специальным MCEM-алгоритмом для решения задачи разделения смесей вероятностных распределений с $N_m = 1$, $m \geq 1$. Однако возможны и нетривиальные (с $N_m > 1$) MC-версии SEM-алгоритма для разделения смесей.

Действительно, в отличие от классического SEM-алгоритма, в рамках которого на $(m+1)$-й итерации генерируется лишь одна реализация $\boldsymbol{y}^{(m+1)} = \{\vec{y}_1\,^{(m+1)}, \ldots, \vec{y}_n\,^{(m+1)}\}$ векторов, определяющих разбиение выборки $\boldsymbol{x}$ на кластеры (см. раздел 7.2), можно на каждой итерации генерировать не одну, но $N_{m+1} > 1$ реализаций указанного вектора $\boldsymbol{y}_1^{(m+1)}, \ldots, \boldsymbol{y}_{N_{m+1}}^{(m+1)}$. Обозначим кластер $\mathcal{K}_i$ при r-й реализации $\boldsymbol{y}_r^{(m+1)}$ указанного вектора на $(m+1)$-й итерации через $\mathcal{K}_{i,r}^{(m+1)}$, а количество элементов выборки $\boldsymbol{x}$, попавших в этот кластер, через $\nu_{i,r}^{(m+1)}$, $r = 1, \ldots, N_{m+1}$. Тогда в качестве оценок весов p_i следует взять

$$p_i^{(m+1)} = \frac{1}{nN_{m+1}} \sum_{r=1}^{N_{m+1}} \nu_{i,r}^{(m+1)}.$$

При этом, если рассматривается модель (5.1) с нормальными компонентами, то в качестве оценок параметров a_i и σ_i на $(m+1)$-й итерации следует взять

$$a_i^{(m+1)} = \frac{1}{N_{m+1}\nu_i^{(m+1)}} \sum_{r=1}^{N_{m+1}} \sum_{j:\, x_j \in \mathcal{K}_{i,r}^{(m+1)}} x_j,$$

$$\sigma_i^{(m+1)} = \left[\frac{1}{N_{m+1}\nu_i^{(m+1)}} \sum_{r=1}^{N_{m+1}} \sum_{j:\, x_j \in \mathcal{K}_{i,r}^{(m+1)}} \left(x_j - a_i^{(m+1)} \right)^2 \right]^{1/2}.$$

Величины же $g_{ij}^{(m+1)}$ переназначаются, как и в классическом SEM-алгоритме, по формуле (7.3).

Асимптотические свойства SEM- и MCEM-алгоритмов описаны в статье (Nielsen, 2000).

7.4.3. SAEM-алгоритм

Существуют алгоритмы, родственные MCEM-алгоритмам, которые, в противоположность классическому MCEM-алгоритму, сходятся при постоянных (не увеличивающихся) и довольно умеренных объемах N_m симуляции

на каждой итерации. Таким свойством, в частности, обладает так называемый SAEM-алгоритм.

Рассмотрим последовательность $\{\gamma_m\}_{m\geq 1}$ положительных чисел такую, что

$$\sum_{m\geq 1}\gamma_m = \infty, \quad \sum_{m\geq 1}\gamma_m^2 < \infty.$$

Определим последовательность функций $W^{(m)}(\theta)$ с помощью рекуррентного соотношения

$$W^{(m)}(\theta) = (1-\gamma_m)W^{(m-1)}(\theta) + \gamma_m \overline{Q}(\theta;\theta^{(m)}) =$$

$$= W^{(m-1)}(\theta) + \gamma_m\left(\frac{1}{N_m}\sum_{r=1}^{N_m}\log f_\theta(\boldsymbol{x},\boldsymbol{y}_r) - W^{(m-1)}(\theta)\right). \tag{7.16}$$

При этом последовательность $\{\theta^{(m)}\}_{m\geq 1}$ формируется в соответствии с правилом

$$\theta^{(m)} = \arg\max_\theta W^{(m)}(\theta), \quad m \geq 1. \tag{7.17}$$

Рекуррентное соотношение (7.16) совпадает с соотношением, определяющим метод оценивания, называемый *стохастической аппроксимацией*, который был предложен в работе (Robbins and Monro, 1951), см. также (Вазан, 1972), (Невельсон и Хасьминский, 1972). Поэтому разновидность EM-алгоритма, определяемая соотношениями (7.16) и (7.17), называется SAEM-алгоритмом (Stochastic Approximation EM-algorithm), см. (Celeux and Diebolt, 1992), (Delyon et al., 1999).

Одна из замечательных особенностей SAEM-алгоритма заключается в том, что он сходится при постоянных значениях N_m (Delyon et al., 1999). Более того, представляется вполне разумным, что при вычислении очередного значения $\theta^{(m)}$ SAEM-алгоритм использует информацию о *всех* симулированных данных на *всех* предшествующих итерациях.

В отличие от MCEM-алгоритма, в котором решение о значении N_m принимается на *каждой* итерации, в рамках SAEM-алгоритма решение о выборе параметров метода $\{\gamma_m\}_{m\geq 1}$ принимается *единовременно*, обычно до начала работы алгоритма. В работе (Polyak and Juditsky, 1992) показано, что скорость сходимости SAEM-алгоритма оптимальна, если шаг метода γ_m выбирается в соответствии с соотношением $\gamma_m \propto 1/m^\alpha$ при $\frac{1}{2} < \alpha \leq 1$. Тем не менее необходимо заметить, что при больших шагах (то есть $\alpha \approx \frac{1}{2}$) SAEM-алгоритм быстро попадает в окрестность решения, однако одновременно возрастает имитационная погрешность $\varepsilon^{(m)}(N_m)$ (см. предыдущий

раздел). С другой стороны, при малых шагах (то есть $\alpha \approx 1$) имитационная погрешность быстро убывает, однако сильно уменьшается скорость сходимости алгоритма (Jank, 2004).

SEM-, SAEM- и MCEM-алгоритмы были подвергнуты систематическому сравнительному анализу в работе (Celeux, Chauveau and Diebolt, 1995). В этой работе отмечено, что многочисленные численные эксперименты показали, что, как правило, SEM-алгоритм оказывался более эффективным. Однако на некоторых смесях ни один из этих алгоритмов не дал разумного решения. Тем не менее, авторы упомянутой работы считают, что и в таких случаях SEM-алгоритм может быть использован в качестве средства разведочного анализа, позволяющего обозначить возможные точки максимума функции правдоподобия с целью их дальнейшего исследования другими средствами.

8.

Приближенное разделение конечных смесей с помощью метода фиксированных компонент для выбора начального приближения EM-алгоритма

8.1. Основная идея метода фиксированных компонент

Идея, которая лежит в основе метода разделения смесей, описываемого в данном разделе, очень близка к идее гармонического анализа, когда периодическая функция раскладывается в ряд Фурье, то есть представляется взвешенной комбинацией (рядом), возможно, бесконечно большого числа синусов и косинусов с различными (но кратными) периодами (то есть с различными параметрами масштаба – частотами). Возможность приближения исходной функции с помощью такого разложения обоснована тем обстоятельством, что семейство синусов и косинусов с указанными свойствами образует базис, то есть полную систему (линейно независимых) функций в пространстве (регулярных) периодических функций.

Аналогия будет видна более отчетливо, если мы рассмотрим семейство чисто масштабных смесей нормальных законов с нулевым сред-

ним. Поскольку при некоторых условиях регулярности каждое распределение вероятностей, сосредоточенное на неотрицательной полуоси, может быть приближено решетчатым распределением с произвольно высокой точностью (скажем, в метрике Леви, метризующей слабую сходимость), мы можем заключить, что семейство нормальных распределений $\mathbf{N}_{\rm rat} = \{\mathcal{N}(0,s) : s$ – рациональное$\}$ образует счетную полную систему функций в пространстве масштабных смесей нормальных законов с нулевым средним. То есть для любой масштабной смеси нормальных законов и произвольно малого $\epsilon > 0$ существует конечная линейная комбинация распределений из семейства $\mathbf{N}_{\rm rat}$ такая, что расстояние (скажем, расстояние Леви) между смесью и линейной комбинацией не превышает $\epsilon > 0$. Поскольку число слагаемых в такой линейной комбинации конечно и все параметры масштаба рациональны, найдется такое (минимальное) значение параметра масштаба, что параметры масштаба всех членов рассматриваемой линейной комбинации будут кратными этому минимальному значению параметра масштаба.

Рассмотрим смесь функций распределения вида

$$F(x) = \sum_{i=1}^{k} p_i \Phi\Big(\frac{x-a_i}{\sigma_i}\Big), \qquad x \in \mathbb{R}, \tag{8.1}$$

где $k \geq 1$ – целое. В классической задаче разделения смесей параметрами, подлежащими статистическому оцениванию, являются тройки (p_i, a_i, σ_i), $i = 1, \ldots, k$, где $a_i \in \mathbb{R}$, $\sigma_i > 0$, $p_i \geq 0$, $p_1 + \ldots + p_k = 1$.

Предположим, что заранее известны числа $\underline{a}$, $\overline{a}$ и $\overline{\sigma}$ такие, что $\underline{a} \leq a_i \leq \overline{a}$ и $\sigma_i \leq \overline{\sigma}$ при всех $i = 1, \ldots, k$. Другими словами, известны диапазоны изменения неизвестных параметров a_i и σ_i.

Идея, лежащая в основе рассматриваемого подхода, заключается в замене *интервалов* $[\underline{a}, \overline{a}]$ и $(0, \overline{\sigma}]$ возможных значений неизвестных параметров масштаба σ_i и сдвига a_i *дискретными* множествами известных точек. Эти точки могут быть выбраны, например, исходя из следующих соображений.

Пусть ε_a и ε_σ – положительные числа, определяющие априорные требования к точности оценивания параметров a_i и σ_i:

$$\max_i |a_i - \widehat{a}_i| \leq \varepsilon_a, \qquad \max_i |\sigma_i - \widehat{\sigma}_i| \leq \varepsilon_\sigma, \tag{8.2}$$

где $\widehat{a}_i$ и $\widehat{\sigma}_i$ – искомые оценки параметров. Числа ε_a и ε_σ также можно интерпретировать как пороги различимости возможных значений параметров:

значения a', a'' и σ', σ'', соответственно, считаются неразличимыми, если

$$|a' - a''| \le \varepsilon_a, \quad |\sigma' - \sigma''| \le \varepsilon_\sigma. \tag{8.3}$$

Положим $k_a = [(\overline{a} - \underline{a})/\varepsilon_a] + 1$, $k_\sigma = [\overline{\sigma}/\varepsilon_\sigma] + 1$, где символ $[z]$ обозначает целую часть числа z. Для $r = 1, 2, \ldots, k_a + 1$ положим $\widetilde{a}_r = \underline{a} + (r-1)\varepsilon_a$. Аналогично, для $l = 1, 2, \ldots, k_\sigma$ положим $\widetilde{\sigma}_l = l\varepsilon_\sigma$. Тогда точки с координатами $(\widetilde{a}_r, \widetilde{\sigma}_l)$ образуют узлы конечной сети, накрывающей прямоугольник $\{(a, \sigma) : \underline{a} \le a \le \overline{a},\, 0 \le \sigma \le \overline{\sigma}\}$, представляющий собой множество возможных значений параметров сдвига и масштаба компонент смеси (8.1) (чтобы избежать возможной некорректности, мы исключили возможность равенства параметра масштаба нулю). Число узлов полученной сети равно $K = (k_a + 1)k_\sigma$. Для удобства записи и упрощения обозначений перенумеруем каким-либо образом узлы указанной сети, вводя *единый* индекс i для координат $(\widetilde{a}_i, \widetilde{\sigma}_i)$ узла с номером i после перенумерации, $i = 1, \ldots, K$.

Базовая посылка рассматриваемого подхода заключается в аппроксимации смеси (8.1) смесью с заведомо бóльшим числом *известных* компонент:

$$F(x) = \sum_{i=1}^{k} p_i \Phi\Big(\frac{x - a_i}{\sigma_i}\Big) \approx \sum_{i=1}^{K} \widetilde{p}_i \Phi\Big(\frac{x - \widetilde{a}_i}{\widetilde{\sigma}_i}\Big) \equiv \widetilde{F}(x), \quad x \in \mathbb{R}. \tag{8.4}$$

Такое приближение практически допустимо, поскольку в силу соотношений (8.2) и (8.3) для любой пары (a_r, σ_r) параметров компоненты смеси (8.1) обязательно найдется практически не отличимая от нее пара $(\widetilde{a}_i, \widetilde{\sigma}_i)$ параметров компоненты смеси $\widetilde{F}(x)$. Веса же остальных компонент смеси $\widetilde{F}(x)$, для параметров которых не найдется "близкой" пары параметров (a_r, σ_r) компоненты смеси (8.1), можно считать равными нулю. Действительно, если бы в соотношении (8.4) вместо *приближенного* было бы *точное* равенство, то в силу идентифицируемости семейства конечных смесей нормальных законов, в полном соответствии с определением идентифицируемости конечных смесей с точностью до переиндексации были бы справедливы равенства:

$$k = K, \quad p_i = \widetilde{p}_i, \quad a_i = \widetilde{a}_i, \quad \sigma_i = \widetilde{\sigma}_i, \quad i = 1, \ldots, k.$$

Заметим, что неизвестными параметрами смеси $\widetilde{F}(x)$ являются *только* веса $\widetilde{p}_1, \ldots, \widetilde{p}_K$.

Пусть $\boldsymbol{x} = (x_1, \ldots, x_n)$ – (независимая) выборка наблюдений, каждое из которых представляет собой реализацию случайной величины с функцией

распределения $F(x)$, задаваемой соотношением (8.1). Пусть $(x_{(1.1)}, \ldots, x_{(n)})$ и $F_n(x)$ – соответственно, вариационный ряд и эмпирическая функция распределения, построенные по выборке $\boldsymbol{x}$,

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}(x_j < x), \quad x \in \mathbb{R}.$$

При этом очевидно, что

$$F_n(x_{(j)}) = \frac{j}{n}, \quad j = 1, \ldots, n. \tag{8.5}$$

В силу теоремы Гливенко при больших n равномерно по $x \in \mathbb{R}$ выполняется соотношение

$$F_n(x) \approx F(x). \tag{8.6}$$

Обозначим

$$\Phi_{ij} = \Phi\Big(\frac{x_{(j)} - \widetilde{a}_i}{\widetilde{\sigma}_i}\Big), \quad j = 1, \ldots, n;\ i = 1, \ldots, K.$$

Отметим, что величины Φ_{ij} *известны*.

Из (8.4) и (8.5) вытекает, что при больших n

$$\widetilde{F}(x) \approx F_n(x), \quad x \in \mathbb{R},$$

откуда с учетом (8.6) мы получаем приближенное соотношение

$$\widetilde{F}(x_{(j)}) = \sum_{i=1}^{K} \widetilde{p}_i \Phi_{ij} \approx \frac{j}{n}, \quad j = 1, \ldots, n. \tag{8.7}$$

Соотношение (8.7) можно использовать для отыскания оценок параметров $\widetilde{p}_1, \ldots, \widetilde{p}_K$ с помощью метода наименьших квадратов и метода наименьших модулей, рассматриваемых в последующих разделах. Реализация этих методов в рассматриваемом случае довольно проста, поскольку $\widetilde{F}(x)$ зависит от параметров $\widetilde{p}_1, \ldots, \widetilde{p}_K$ *линейно*, так что мы не выходим за рамки линейной модели регрессионного анализа.

8.2. Разделение конечных смесей вероятностных распределений с фиксированными компонентами при помощи метода наименьших квадратов

Рассматриваемую задачу удобно записать в векторно-матричном виде. Обозначим

$$\mathbf{p} = \begin{pmatrix} \widetilde{p}_1 \\ \widetilde{p}_2 \\ \cdots \\ \widetilde{p}_K \end{pmatrix}, \quad \mathbf{\Phi} = \begin{pmatrix} \Phi_{11} & \Phi_{21} & \cdots & \Phi_{K1} \\ \Phi_{12} & \Phi_{22} & \cdots & \Phi_{K2} \\ \cdots & \cdots & \cdots & \cdots \\ \Phi_{1n} & \Phi_{2n} & \cdots & \Phi_{Kn} \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} \frac{1}{n} \\ \frac{2}{n} \\ \cdots \\ \frac{n-1}{n} \\ 1 \end{pmatrix}.$$

Тогда соотношение (8.7) можно переписать в матричной форме:

$$\mathbf{\Phi p} \approx \mathbf{u}$$

или

$$\mathbf{\Phi p} + \mathbf{d} = \mathbf{u},$$

где

$$\mathbf{d} = \begin{pmatrix} d_1 \\ d_2 \\ \cdots \\ d_n \end{pmatrix}$$

– вектор-столбец невязок с компонентами

$$d_j = \frac{j}{n} - \sum_{i=1}^{K} \widetilde{p}_i \Phi_{ij}.$$

Предположим, что $n \geq K$. В этом случае ранг случайной (в силу случайности выборки $\boldsymbol{x} = (x_1, \ldots, x_n)$) матрицы $\mathbf{\Phi}$ с вероятностью единица равен K. Тогда из общей теории метода наименьших квадратов для линейных моделей вытекает, что решение

$$\widehat{\mathbf{p}}^* = \begin{pmatrix} \hat{p}_1^* \\ \hat{p}_2^* \\ \cdots \\ \hat{p}_K^* \end{pmatrix}$$

безусловной задачи наименьших квадратов

$$\arg\min_{\mathbf{p}} \mathbf{d}^\top \mathbf{d} = \arg\min_{\mathbf{p}} \sum_{j=1}^{n} \left[\sum_{i=1}^{K} \widetilde{p}_i \Phi_{ij} - \frac{j}{n} \right]^2 \tag{8.8}$$

с вероятностью единица существует, единственно и имеет вид

$$\widehat{\mathbf{p}}^* = (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{u}. \tag{8.9}$$

При этом сумма компонент вектора $\widehat{\mathbf{p}}^*$ отнюдь не обязана быть равной единице. Обозначим

$$\gamma = 1 - \sum_{i=1}^{K} \hat{p}_i^*. \tag{8.10}$$

Пусть $\mathbf{a}$ – K-мерный вектор-столбец, все компоненты которого равны единице:

$$\mathbf{a}^\top = (1,\, 1,\, \ldots,\, 1).$$

Тогда соотношение (8.10) можно записать в виде

$$\gamma = 1 - \mathbf{a}^\top \widehat{\mathbf{p}}^*.$$

Но в рассматриваемом случае компоненты искомого вектора $\mathbf{p}$, будучи вероятностями в дискретном распределении, связаны очевидным условием

$$\widetilde{p}_1 + \ldots + \widetilde{p}_K = 1. \tag{8.11}$$

В терминах вектора $\mathbf{a}$ условие (8.11) запишется как

$$\mathbf{a}^\top \mathbf{p} = 1. \tag{8.12}$$

Хорошо известно, что, если ранг матрицы $\mathbf{\Phi}$ равен K, то решение $\widehat{\mathbf{p}}$ задачи (8.8) при условии (8.12) имеет вид

$$\widehat{\mathbf{p}} = \widehat{\mathbf{p}}^* + (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{a} [\mathbf{a}^\top (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{a}]^{-1} (1 - \mathbf{a}^\top \widehat{\mathbf{p}}^*) \tag{8.13}$$

(см., например, с. 85–86 в (Себер, 1980)), где вектор $\widehat{\mathbf{p}}^*$ определен в (8.9). Несложно видеть, что выражение $\mathbf{a}^\top (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{a}$ равно сумме всех элементов матрицы $(\mathbf{\Phi}^\top \mathbf{\Phi})^{-1}$. Обозначим эту сумму S. Тогда с учетом (8.10) соотношение (8.13) примет вид

$$\widehat{\mathbf{p}} = \widehat{\mathbf{p}}^* + \frac{\gamma}{S} \cdot (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{a}. \tag{8.14}$$

Несложно видеть, что i-ая компонента s_i вектор-столбца $(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{a}$ равна сумме всех элементов i-ой строки матрицы $(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}$. Таким образом, из (8.14) мы окончательно получаем, что

$$\widehat{p}_i = \hat{p}_i^* + \frac{\gamma s_i}{S}, \quad i = 1, \ldots, K. \tag{8.15}$$

Условие $n \geq K$, упрощающее вычисления, определяет выбор числа $k_a + 1$ возможных значений параметров сдвига и числа k_σ возможных значений параметров масштаба компонент исходной смеси (8.1). А именно, чтобы обеспечить существование единственного решения $\widehat{\mathbf{p}}$ условной задачи наименьших квадратов, параметры сетки k_a и k_σ должны быть связаны с объемом выборки n соотношением

$$(k_a + 1)k_\sigma \leq n.$$

Случай $n < K$ нуждается в более тщательном анализе, поскольку в таком случае оценки наименьших квадратов определены неоднозначно.

Если оценки $\widehat{p}_i$ удовлетворяют неравенствам

$$0 \leq \widehat{p}_i \leq 1 \tag{8.16}$$

для всех $i = 1, \ldots, k$, то задача решена. Однако может случиться так, что некоторые из последних неравенств не реализуются на полученном векторе. В таком случае можно идти несколькими путями.

Во-первых, в работе (Waterman, 1974) показано, что задачу наименьших квадратов с ограничениями типа неравенств можно свести к задаче последовательного решения нескольких задач наименьших квадратов без ограничений. Вытекающий из этого результата алгоритм решения рассматриваемой задачи таков.

Если для некоторых $i = 1, \ldots, k$ соотношения (8.16) не выполнены, то минимум функции

$$SS(\mathbf{p}) = \sum_{j=1}^{n} \left[\sum_{i=1}^{K} \widetilde{p}_i \Phi_{ij} - \frac{j}{n} \right]^2$$

на множестве

$$\mathcal{P} = [0,1]^K \bigcap \{\mathbf{p} : \widetilde{p}_1 + \ldots + \widetilde{p}_K = 1\}$$

достигается в одной из граничных точек, в которых хотя бы для одного i выполнено равенство $p_i = 0$ (напомним, что $\widehat{p}_1 + \ldots + \widehat{p}_K = 1$ по построению). Пусть I – некоторое подмножество множества $\{1, 2, \ldots, K\}$. Как

известно, всего таких подмножеств 2^K. Для каждого такого подмножества I мы находим $\widehat{\mathbf{p}}_I$ – точку минимума функции

$$SS_I(\mathbf{p}_I) = \sum_{j=1}^{n} \left[\sum_{i \in I} \widetilde{p}_i \Phi_{ij} - \frac{j}{n} \right]^2$$

при условии $\mathbf{p}_I^\top \mathbf{p}_I = 1$ по формулам, аналогичным (8.14) и (8.15) (здесь $\mathbf{p}_I$ – это вектор $\mathbf{p}$, компоненты которого с номерами, не попавшими в множество I, равны нулю). Если при этом $\widehat{\mathbf{p}}_I \in \mathcal{P}$, то вычисляется значение $SS_I(\widehat{\mathbf{p}}_I)$. Перебрав все возможные подмножества I, можно найти решение исходной задачи – точку

$$\widehat{\mathbf{p}} = \arg\min_I SS_I(\widehat{\mathbf{p}}_I).$$

Однако при таком подходе требуется решить 2^K задач наименьших квадратов с линейным ограничением, что при больших K занимает чрезвычайно много времени.

Во-вторых, в работе (Judge and Takayama, 1966) показано, что задачу наименьших квадратов с ограничениями типа неравенств можно свести к задаче квадратичного программирования, и предложена итерационная процедура – модифицированный симплекс-метод – для ее решения. Эта процедура реализована в виде встроенного средства Optimization Toolbox системы MATLAB и использована, в частности, в работе (Королев, Ломской, Пресняков и Рэй, 2005) для решения исходной задачи наименьших квадратов с ограничениями типа неравенств и равенства. Однако при этом объем вычислений трудно оценить, а статистические свойства оценок, получаемых при таком подходе, чрезвычайно трудно исследовать.

Наконец, исходную задачу наименьших квадратов с ограничениями типа неравенств и равенства можно заменить приближенной, накинув конечную сетку и на множество возможных значений вектора $(p_1, \ldots, p_k)$, сведя таким образом решение задачи к простому перебору.

8.3. Разделение конечных смесей вероятностных распределений с фиксированными компонентами при помощи метода наименьших модулей

8.3.1. Решение в смысле минимума sup-нормы

Соотношение (8.7) допускает различные конкретизации, связанные с различными нормами невязки. Решение задачи по методу наименьших квадратов, рассмотренное в предыдущем разделе, связано с минимизацией обычной евклидовой нормы невязки.

В данном разделе мы рассмотрим решение, связанное с минимизацией sup-нормы вектора невязок

$$\mathbf{d} = \begin{pmatrix} d_1 \\ d_2 \\ \dots \\ d_n \end{pmatrix},$$

где, как и ранее,

$$d_j = \frac{j}{n} - \sum_{i=1}^{K} \widetilde{p}_i \Phi_{ij}.$$

А именно, мы будем искать решение задачи

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \max_{1 \le j \le n} |d_j|. \tag{8.17}$$

Как известно, задача (8.17) может быть сведена (см., например, с. 91 в (Васильев и Иваницкий, 2003) и с. 137 в (Ашманов и Тимохов, 1991)) к задаче линейного программирования вида

$$f(\mathbf{p}, \theta) = \theta \longrightarrow \min_{(\mathbf{p},\theta) \in \mathcal{Q}'}, \tag{8.18}$$

$$\mathcal{Q}' = \Bigg\{ (\mathbf{p}, \theta) \in \mathbb{R}_+^{K+1} : \ \widetilde{p}_1 + \ldots + \widetilde{p}_K \ge 1; \ \widetilde{p}_1 + \ldots + \widetilde{p}_K \le 1;$$

$$\theta \ge \sum_{i=1}^{K} \widetilde{p}_i \Phi_{ij} - \frac{j}{n}; \ \theta \ge -\sum_{i=1}^{K} \widetilde{p}_i \Phi_{ij} + \frac{j}{n}, \ j = 1, \ldots, n \Bigg\}. \tag{8.19}$$

Последняя задача, как известно, решается с помощью стандартного симплекс-метода, заключающегося в направленном переборе угловых точек множества $\mathcal{Q}'$ (см., например, главу 1 в (Васильев и Иваницкий, 2003)).

8.3.2. Решение в смысле минимума $\mathbf{L}_1$-нормы

В данном разделе мы рассмотрим решение, связанное с минимизацией L_1-нормы вектора невязок $\mathbf{d}$. А именно, мы будем искать решение задачи наименьших модулей

$$\mathbf{p}^\star = \arg\min_{\mathbf{p}} \sum_{j=1}^{n} |d_j|. \tag{8.20}$$

Обозначим $\vec{\theta} = (\theta_1, \ldots, \theta_n)^\top$. Как известно, задача (8.20) может быть сведена (см., например, с. 91 в (Васильев и Иваницкий, 2003) и с. 137 в (Ашманов и Тимохов, 1991)) к задаче линейного программирования вида

$$f(\mathbf{p}, \theta) = \sum_{j=1}^{n} \theta_j \longrightarrow \min_{(\mathbf{p}, \vec{\theta}) \in \mathcal{Q}''}, \tag{8.21}$$

$$\mathcal{Q}'' = \Big\{ (\mathbf{p},\, \vec{\theta}) \in \mathbb{R}_+^{K+n} :\ \widetilde{p}_1 + \ldots + \widetilde{p}_K \geq 1;\ \widetilde{p}_1 + \ldots + \widetilde{p}_K \leq 1;$$

$$\theta_j \geq \sum_{i=1}^{K} \widetilde{p}_i \Phi_{ij} - \frac{j}{n};\ \theta_j \geq -\sum_{i=1}^{K} \widetilde{p}_i \Phi_{ij} + \frac{j}{n},\ j = 1, \ldots, n \Big\}. \tag{8.22}$$

Последняя задача также решается с помощью стандартного симплекс-метода, заключающегося в направленном переборе угловых точек множества $\mathcal{Q}''$ (см., например, главу 1 в (Васильев и Иваницкий, 2003)).

Вычислительные реализации задач линейного программирования (8.18), (8.19) и (8.21), (8.22) в популярных системах MATHCAD и MATLAB обладают намного более высоким быстродействием, нежели вычислительные реализации EM-алгоритма.

С содержательной точки зрения, рассмотренные в данном разделе задачи построения оценок вектора весов $\mathbf{p}$ вполне равноправны с задачей, рассмотренной в предыдущем разделе. Более того, как известно, оценки наименьших модулей более устойчивы по отношению к наличию резко выделяющихся наблюдений, нежели оценки наименьших квадратов. Поэтому оценки минимума sup-нормы и минимума L_1-нормы в определенном смысле предпочтительнее оценок наименьших квадратов.

Конечно, рассмотренные методы дают лишь приближенное решение задачи разделения смесей. Однако при его реализации удается избежать итерационных процедур типа EM-алгоритма для поиска решений экстремальных задач. Решение, полученное такими методами, вполне можно использовать как начальное приближение для EM-алгоритма или его модификаций с целью последующего получения более точного решения.

При этом, имея решение (8.15) задачи разделения смесей, начальное приближение EM-алгоритма можно выбирать следующим образом. Упорядочим полученные оценки весов $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_K$ по убыванию и получим набор $\widehat{p}_{i_1}, \widehat{p}_{i_2}, \ldots, \widehat{p}_{i_K}$ такой, что $\widehat{p}_{i_1} \geq \widehat{p}_{i_2} \geq \ldots \geq \widehat{p}_{i_K}$. Выберем порог $\delta > 0$ из тех соображений, что на практике, если $p_{i_j} < \delta$, то мы будем считать, что компонента смеси, соответствующая весу p_{i_j}, отсутствует. Положим

$$k = \max\{j : p_{i_j} \geq \delta\}.$$

Тогда в качестве начального приближения для EM-алгоритма возьмем параметры a_{i_j} и $\sigma^2_{i_j}$, соответствующие весам $\widehat{p}_{i_1}, \widehat{p}_{i_2}, \ldots, \widehat{p}_{i_k}$. Особо следует отметить, что при таком подходе число компонент смеси определяется автоматически по заданному порогу пренебрежимости δ.

Обратим внимание, что в рамках рассмотренного в данном разделе подхода возможно классическое детерминистическое истолкование точности в отношении приближения параметров a_i и σ_i, при котором точность характеризуется *одним* числом, задаваемым шагом дискретной сетки, накидываемой на множества значений указанных параметров. В отношении же весовых параметров $p_1, \ldots, p_k$, чтобы охарактеризовать точность приближения, приходится пользоваться статистическим истолкованием, при котором одного числа недостаточно, а нужно задавать еще и надежность вывода (коэффициент доверия или уровень значимости).

8.4. Разделение конечных смесей вероятностных распределений с фиксированными компонентами при помощи "усеченного" EM-алгоритма

Для решения задачи отыскания оценок весов $\widetilde{p}_i$, $i = 1, \ldots, K$, смеси $\widetilde{F}(x)$ (см. (8.4)) в методе фиксированных компонент можно использовать "усеченный" EM-алгоритм.

В разделе 5 показано, что при применении EM-алгоритма к решению задачи разделения смесей вероятностных распределений (5.1) оценки весов компонент можно искать независимо от оценок параметров самих компонент (см. представление (5.6)). Более того, при известных значениях параметров компонент (или их оценок) оценки весов имеют довольно простой вид (5.8). Это позволяет указать несложную итерационную процедуру оценивания весов $\widetilde{p}_i$ смеси $\widetilde{F}(x)$. Эта процедура по сути представляет собой

"усеченный" вариант EM-алгоритма. В этом варианте за ненадобностью отсутствует шаг, связанный с переназначением оценок параметров компонент смеси, поскольку в методе фиксированных компонент значения параметров компонент смеси $\widetilde{F}(x)$ *известны*.

Пусть, как и ранее, $\mathbf{x} = (x_1, \ldots, x_n)$ – выборка, $(\widetilde{a}_i, \widetilde{\sigma}_i)$ – узлы сетки, накидываемой на множество значений параметров компонент, $i = 1, \ldots, K$. Для удобства обозначим

$$x_{ji} = \frac{x_j - \widetilde{a}_i}{\widetilde{\sigma}_i}, \quad \phi_{ij} = \frac{1}{\widetilde{\sigma}_i}\phi(x_{ji}), \quad j = 1, \ldots, n;\ i = 1, \ldots, K,$$

где, как и ранее, $\phi(x)$ – стандартная нормальная плотность. Тогда обсуждаемый "усеченный" вариант EM-алгоритма для оценивания весов $\widetilde{p}_i$ смеси $\widetilde{F}(x)$ определяется с помощью следующего рекуррентного соотношения, вытекающего из (5.8):

$$\tilde{p}_i^{(m+1)} = \frac{\tilde{p}_i^{(m)}}{n} \sum_{j=1}^{n} \frac{\phi_{ij}}{\sum_{r=1}^{K} \tilde{p}_r^{(m)} \phi_{rj}} \quad j = 1, \ldots, n;\ i = 1, \ldots, K. \tag{8.23}$$

Здесь $m \geq 1$ – номер итерации.

Поскольку функция правдоподобия параметров $\mathbf{p} = (\widetilde{p}_1, \ldots, \widetilde{p}_K)^\top$ регулярна (в частности, обладает свойствами гладкости), то последовательность $\{\tilde{p}_i^{(m)}\}_{m \geq 1}$, определенная соотношением (8.23), сходится к соответствующим оценкам максимального правдоподобия весов в смеси $\widetilde{F}(x)$. Однако, как уже отмечалось выше, функция правдоподобия в исходной задаче разделения смесей $F(x)$ нормальных распределений (см. (8.4)) обладает весьма нерегулярным рельефом. К тому же узлы $(\widetilde{a}_i, \widetilde{\sigma}_i)$ сетки могут не совпасть с "истинными" значениями (a_j, σ_j) параметров компонент исходной смеси. Поэтому сходимость рекурсии (8.23), максимизирующей правдоподобие оценок весов в смеси $\widetilde{F}(x)$, к оценкам максимального правдоподобия параметров p_i в исходной смеси $F(x)$, к сожалению, отнюдь не гарантирована.

8.5. Разделение конечных смесей вероятностных распределений с фиксированными компонентами при помощи байесовской классификации

Если объем выборки n намного превосходит число K компонент смеси $\widetilde{F}(x)$, то для оценивания весов $\widetilde{p}_i$, $i = 1, \ldots, K$, можно использовать совсем простую вычислительную процедуру. Из общей теории байесовских правил классификации (см., например, (Андерсон, 1963)) известно, что оптимальное решение задачи классификации наблюдений $\boldsymbol{x} = (x_1, \ldots, x_n)$ на K классов (совокупностей), определяемых распределениями $\Phi((x-a_i)/\sigma_i)$, $i = 1, \ldots, K$, которое минимизирует максимальную вероятность ошибочной классификации, имеет следующий вид.

Обозначим

$$k(j) = \arg \max_{1 \le i \le K} \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left\{ - \frac{(x_j - a_i)^2}{2\sigma_i^2} \right\}.$$

В соответствии с описываемым правилом классификации наблюдение x_j приписывается к классу с номером $k(j)$, $j = 1, \ldots, n$.

Пусть

$$\nu_i = \sum_{j:\, k(j)=i} 1$$

– количество наблюдений, отнесенных к классу с номером i, $i = 1, \ldots, K$. Тогда в качестве оценки веса $\widetilde{p}_i$ следует взять величину

$$\tilde{p}_i^* = \frac{\nu_i}{n}, \quad i = 1, \ldots, K.$$

9.

Выбор модели (определение типа и числа компонент смеси)

Для конкретных прикладных задач, математически формализованных с помощью смесей вероятностных (в частности, гауссовых) распределений, огромную важность имеет интерпретация полученных результатов, например, возможность отождествления полученных компонент смеси с теми или иными реальными объектами и/или процессами. Очевидно, что такая интерпретация может быть разумной только лишь, если рассматриваемая формальная модель адекватна исследуемому процессу или явлению.

Включая в модель все больше и больше параметров, можно добиться все большего и большего ее согласия с анализируемыми данными, то есть увеличить ее адекватность. Поэтому, казалось бы, при использовании математических моделей типа конечных смесей нормальных законов следует включать в смеси как можно больше компонент, и единственным обстоятельством, определяющим выбор модели, является лишь мощность используемой вычислительной системы (производительность, память и другие ресурсы). Однако на самом деле ситуация намного сложнее.

К примеру, когда исследуются компоненты волатильности хаотических стохастических процессов, в качестве моделей распределений их приращений можно рассматривать два типа смесей: чисто масштабные смеси, в которых смешивание производится только по параметру масштаба, а параметры сдвига равны нулю (например, если исследователь заведомо заинтересован лишь в чисто стохастической интерпретации волатильности), и общие сдвиг-масштабные смеси (которые позволяют выделить также и динамическую составляющую волатильности). Оказывается, что на практике два этих типа смесей принципиально по-разному ведут себя по от-

ношению к увеличению числа их компонент: для чисто масштабных смесей наблюдается *эффект насыщения*, когда заметное повышение согласия модели с данными наблюдается лишь при увеличении числа компонент смеси до определенного порога (как правило, довольно небольшого и равного 4 или 5), а при дальнейшем увеличении числа компонент (повышении "проработанности" модели) изменения согласия незначительны и практически несущественны. Для моделей типа общих сдвиг-масштабных смесей наблюдается принципиально иной *эффект перетекания волатильности* из диффузионной (стохастической) составляющей в динамическую: при малом (2-3) числе компонент диффузионная составляющая превалирует над динамической, а при увеличении числа компонент соотношение этих составляющих принципиально изменяется в пользу динамической. С формальной точки зрения эффект перетекания волатильности сводится к тому, что модели типа смесей превращаются во все более и более точные непараметрические оценки плотности, аналогичные ядерным оценкам. Интерпретация таких оценок в силу их почти непараметрической природы весьма проблематична.

Таким образом, хотелось бы заранее решить, модель какого типа использовать в той или иной конкретной ситуации. В частности, хотелось бы заранее ответить на вопрос, значима ли динамическая составляющая волатильности, или ее присутствие обусловлено лишь чисто стохастическими флуктуациями результатов измерений вокруг их среднего значения. Если отличие оценок параметров сдвига компонент от нуля на самом деле вызвано не систематическими, но чисто стохастическими причинами, то, как следует из сказанного выше, использование формально более точной модели типа сдвиг-масштабной смеси может сильно исказить интерпретацию результатов. Аналогично, к неправильной, искаженной интерпретации результатов может привести и использование формально более точной модели с завышенным числом компонент или неточной модели с заниженным числом компонент.

Некоторые критерии выбора подходящей модели (типа смеси и количества компонент) рассмотрены в данном разделе.

9.1. Некоторые сведения из теории проверки сложных статистических гипотез

Здесь мы приведем некоторые основные определения и нужные нам результаты из теории проверки сложных статистических гипотез.

В математической статистике базовым понятием, формализующим имеющиеся в распоряжении исследователя данные, является понятие *статистической структуры*, под которой подразумевается тройка объектов $(\mathbb{R}^n, \mathcal{B}_n, \mathcal{P})$, где $\mathbb{R}^n$ – это n-мерное евклидово пространство, которому принадлежат возможные значения $\boldsymbol{x} = (x_1, \ldots, x_n)$ наблюдаемого случайного вектора $\mathbf{X} = (X_1, \ldots, X_n)$ (выборки), $\mathcal{B}_n$ – это борелевская σ-алгебра подмножеств n-мерного евклидова пространства, $\mathcal{P}$ – семейство вероятностных мер, содержащее ту, которая задает неизвестное "истинное" распределение наблюдаемого случайного вектора $\mathbf{X}$. Обычно семейство $\mathcal{P}$ задается в виде

$$\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}, \tag{9.1}$$

где Θ – некоторое множество, имеющее, вообще говоря, более простую структуру, нежели $\mathcal{P}$.

Если говорить нестрого, то под *статистической гипотезой* H обычно подразумевают некое утверждение, описывающее неизвестное распределение наблюдаемого случайного вектора $\mathbf{X}$, то есть сужающее множество $\mathcal{P}$. С формальной же точки зрения, статистическая гипотеза является подмножеством множества $\mathcal{P}$. Пусть $\mathcal{P}_0$ – некоторое подмножество множества $\mathcal{P}$ (возможно, совпадающее с $\mathcal{P}$). Объединяя неформальное объяснение с математически корректным формализмом, мы будем обозначать статистическую гипотезу о том, что распределение наблюдаемого случайного вектора $\mathbf{X}$ определяется какой-либо мерой из подмножества $\mathcal{P}_0$, в виде $H\colon\ \mathsf{P} \in \mathcal{P}_0$. Если семейство $\mathcal{P}$ задано в параметрическом виде (9.1), то под статистической гипотезой мы будем подразумевать подмножество Θ_0 множества Θ. Для такой параметрической гипотезы мы будем использовать обозначение $H\colon\ \theta \in \Theta_0$. В дальнейшем мы будем иметь дело только с параметрическими гипотезами. Если гипотеза $H\colon\ \theta \in \Theta_0$ описывает распределение наблюдаемого случайного вектора $\mathbf{X}$ однозначно, то есть множество Θ_0 содержит единственный элемент, то гипотеза $H\colon\ \theta \in \Theta_0$ называется *простой*. Гипотеза, не являющаяся простой, называется *сложной*.

Предположим, что существует σ-конечная мера μ, относительно которой все меры из семейства (9.1) являются абсолютно непрерывными. Соответствующие плотности случайного вектора $\mathbf{X}$ обозначим $f_\theta^{\mathbf{X}}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$. Предположим, что компоненты $X_1, \ldots, X_n$ случайного вектора $\mathbf{X}$ стохастически независимы и имеют одно и то же распределение. Тогда функция правдоподобия параметра θ примет вид

$$L(\theta; \boldsymbol{x}) \equiv f_\theta^{\mathbf{X}}(\boldsymbol{x}) = \prod_{j=1}^{n} f_\theta(x_j), \qquad \theta \in \Theta,$$

где $f_\theta(x)$, $x \in \mathbb{R}$, – плотность распределения случайной величины X_1 (относительно меры μ).

В дальнейшем мы будем предполагать, что $\Theta \subseteq \mathbb{R}^r$ при некотором $r \geq 1$, то есть мы считаем, что $\theta = (\theta_1, \ldots, \theta_r)$, причем $\theta_i \in \mathbb{R}$, $i = 1, \ldots, r$. Нас будут интересовать сложные ("цилиндрические") статистические гипотезы следующего вида. Пусть s – целое число, $1 \leq s < r$. Зафиксируем числа $t_1, \ldots, t_{r-s}$. Назовем статистическую гипотезу

$$H_s\colon\ \theta \in \Theta_s \equiv \{\theta = (\theta_1, \ldots, \theta_r) \in \Theta :\ \theta_i = t_{i-s},\ i = \overline{s+1, r}\}$$

цилиндрической сложной гипотезой размерности s, порожденной числами $t_1, \ldots, t_{r-s}$. Несложно видеть, что $\Theta_s \subset \Theta$.

Обозначим

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; \boldsymbol{x}), \tag{9.2}$$

$$\widehat{\theta}^s = \arg\max_{\theta \in \Theta_s} L(\theta; \boldsymbol{x}). \tag{9.3}$$

Для проверки сложной гипотезы $H_s\colon\ \theta \in \Theta_s$ используется критерий отношения правдоподобия, основанный на статистике

$$\lambda_s(\mathbf{X}) = \frac{L(\widehat{\theta}^s; \mathbf{X})}{L(\widehat{\theta}; \mathbf{X})}. \tag{9.4}$$

Очевидно, что всегда $0 \leq \lambda_s(\boldsymbol{x}) \leq 1$. Чем ближе значение статистики $\lambda_s(\boldsymbol{x})$ к нулю, тем менее правдоподобна гипотеза $H_s\colon\ \theta \in \Theta_s$ по сравнению с гипотезой $H\colon\ \theta \in \Theta$.

Чтобы придать последнему утверждению строгую форму и получить возможность сделать более или менее обоснованный вывод о том, есть ли *систематические* причины отвергнуть гипотезу $H_s\colon\ \theta \in \Theta_s$ или согласиться с ней, можно использовать следующий фундаментальный результат, доказанный С. Уилксом еще в 1938 г. (Wilks, 1938) и развитый в работе (Chernoff, 1954), см. также (Уилкс, 1967), раздел 13.8. Заметим, что стремление к нулю величины $\lambda_s(\boldsymbol{x})$ эквивалентно неограниченному возрастанию величины $-\log \lambda_s(\boldsymbol{x})$.

Теорема 9.1. *Пусть компоненты вектора $\mathbf{X} = (X_1, \ldots, X_n)$ – независимы и одинаково распределены, причем функция правдоподобия дважды непрерывно дифференцируема по $\theta \in \Theta$, а статистика $\lambda_s(\mathbf{X})$ определена в соответствии с* (9.4). *Тогда равномерно по $x \in \mathbb{R}$*

$$\lim_{n\to\infty} \mathsf{P}_\theta\big(-2\log \lambda_s(\mathbf{X}) < x\big) = \frac{1}{2^{(r-s)/2}\Gamma\Big(\dfrac{r-s}{2}\Big)} \int\limits_0^x z^{(r-s)/2} e^{-z/2} dz$$

для любого $\theta \in \Theta_s$, где $\Gamma(\cdot)$ – гамма-функция Эйлера. Другими словами, распределение статистики $-2\log\lambda_s(\mathbf{X})$ при гипотезе $H_s = \Theta_s$ при неограниченно увеличивающемся объеме выборки n стремится к распределению хи-квадрат с числом степеней свободы, равным $(r-s)$.

Основанный на теореме 9.1 критерий проверки сложной цилиндрической гипотезы $H_s\colon\ \theta \in \Theta_s$ против сложной альтернативы $H\colon\ \theta \in \Theta$ выглядит так. Заранее фиксируется малое положительное число α – уровень значимости. Оно характеризует требования к качеству вывода (уровень значимости – это вероятность ошибочно отвергнуть верную гипотезу). На основе имеющейся выборки $\boldsymbol{x}$ по приведенным выше формулам (9.2), (9.3), (9.4) вычисляется значение статистики $\Lambda_s(\boldsymbol{x}) = -2\log\lambda_s(\boldsymbol{x})$. Если полученное значение $\Lambda_s(\boldsymbol{x})$ окажется бóльшим $\chi_{r-s}(1-\alpha)$ – квантили порядка $(1-\alpha)$ распределения хи-квадрат с числом степеней свободы, равным $(r-s)$, – то гипотеза $H_s\colon\ \theta \in \Theta_s$ отвергается. Если же оказывается, что $\Lambda_s(\boldsymbol{x}) \leq \chi_{r-s}(1-\alpha)$, то оснований отвергать гипотезу $H_s\colon\ \theta \in \Theta_s$ нет. При этом вероятность ошибочно отвергнуть гипотезу $H_s\colon\ \theta \in \Theta_s$ в том случае, когда она на самом деле верна, примерно равна α.

9.2. Проверка значимости динамической составляющей волатильности с помощью критерия отношения правдоподобия

Несложно убедиться, что если распределение некоторой случайной величины X имеет вид конечной сдвиг/масштабной смеси нормальных законов, то есть

$$\mathsf{P}(X < x) = \sum_{i=1}^{k} p_i \Phi\Big(\frac{x - a_i}{\sigma_i}\Big), \tag{9.5}$$

где $p_i \geq 0$, $a_i \in \mathbb{R}$, $\sigma_i > 0$, $i = 1, \ldots, k$, $p_1 + \ldots + p_k = 1$, то дисперсия случайной величины X может быть представлена в виде суммы двух слагаемых

$$\mathsf{D}X = \sum_{i=1}^{k} (a_i - \overline{a})^2 p_i + \sum_{i=1}^{k} p_i \sigma_i^2, \tag{9.6}$$

где

$$\overline{a} = \sum_{i=1}^{k} a_i p_i.$$

При этом первое выражение в (9.6) характеризует ту часть дисперсии, которая обусловлена наличием ненулевых сдвигов, то есть "динамическую" составляющую дисперсии (или волатильности), тогда как второе выражение в (9.6) характеризует "чисто диффузионную" составляющую дисперсии.

Как уже говорилось, при анализе волатильности реальных хаотических процессов модели типа чисто масштабных смесей с $a_1 = \ldots = a_k = 0$ и модели общего типа, в которых хотя бы некоторые a_i отличны от нуля, обладают разными свойствами. Более того, интерпретация моделей этих двух разных типов может быть принципиально разной. Поэтому чрезвычайную важность приобретает вопрос о возможности проверки гипотезы о том, что наблюдаемая волатильность имеет чисто стохастическую, диффузионную природу, то есть гипотезы $H_0\colon\ a_1 = \ldots = a_k = 0$. Если эта гипотеза отвергается, то возникает принципиально иное объяснение наблюдаемой изменчивости исследуемого процесса. Приведенная в предыдущем разделе теорема 9.1 позволяет построить вполне разумный критерий для проверки упомянутой гипотезы.

Наряду с моделью (9.5) рассмотрим модель

$$\mathsf{P}(X < x) = \sum_{i=1}^{k} p_j \Phi\Big(\frac{x}{\sigma_i}\Big). \tag{9.7}$$

Пусть в нашем распоряжении имеется выборка $\boldsymbol{x} = (x_1, \ldots, x_n)$ независимых реализаций случайной величины X, в отношении распределения которой имеются два предположения (9.5) и (9.7). В соответствии с моделями (9.5) и (9.7) рассмотрим две функции правдоподобия

$$L_0(p_1, \ldots, p_k, \sigma_1, \ldots, \sigma_k; \boldsymbol{x}) = \prod_{j=1}^{n} \Big[\sum_{i=1}^{k} \frac{p_i}{\sigma_i} \phi\Big(\frac{x_j}{\sigma_i}\Big) \Big]$$

и

$$L(p_1, \ldots, p_k, a_1, \ldots, a_k, \sigma_1, \ldots, \sigma_k; \boldsymbol{x}) = \prod_{j=1}^{n} \Big[\sum_{i=1}^{k} \frac{p_i}{\sigma_i} \phi\Big(\frac{x_j - a_i}{\sigma_i}\Big) \Big],$$

где, как обычно, $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ – стандартная нормальная плотность. Пусть

$$(p_1^\circ, \ldots, p_k^\circ, \sigma_1^\circ, \ldots, \sigma_k^\circ) = \arg \max_{\substack{p_1, \ldots, p_k \\ \sigma_1, \ldots, \sigma_k}} L_0(p_1, \ldots, p_k, \sigma_1, \ldots, \sigma_k; \boldsymbol{x}),$$

$$(\widehat{p}_1, \ldots, \widehat{p}_k, \widehat{a}_1, \ldots, \widehat{a}_k, \widehat{\sigma}_1, \ldots, \widehat{\sigma}_k) =$$

$$= \arg \max_{\substack{p_1,\ldots,p_k \\ a_1,\ldots,a_k \\ \sigma_1,\ldots,\sigma_k}} L(p_1,\ldots,p_k,a_1,\ldots,a_k,\sigma_1,\ldots,\sigma_k;\boldsymbol{x}).$$

Положим

$$\Lambda(\boldsymbol{x}) = -2\log\left(\frac{L_0(p_1^\circ,\ldots,p_k^\circ,\sigma_1^\circ,\ldots,\sigma_k^\circ;\boldsymbol{x})}{L(\widehat{p}_1,\ldots,\widehat{p}_k,\widehat{a}_1,\ldots,\widehat{a}_k,\widehat{\sigma}_1,\ldots,\widehat{\sigma}_k;\boldsymbol{x})}\right).$$

В соответствии с теоремой 9.1, если справедлива гипотеза H_0: $a_1 = \ldots = a_k = 0$, то при $n \to \infty$ распределение случайной величины $\Lambda(\mathbf{X})$ стремится к распределению хи-квадрат с k степенями свободы. Поэтому критерий проверки гипотезы H_0: $a_1 = \ldots = a_k = 0$ устроен так. Пусть α – малое положительное число (уровень значимости), $\chi_k(1-\alpha)$ – $(1-\alpha)$-квантиль распределения хи-квадрат с k степенями свободы. Если

$$\Lambda(\boldsymbol{x}) > \chi_k(1-\alpha),$$

то гипотеза H_0: $a_1 = \ldots = a_k = 0$ отвергается, то есть отличие динамической компоненты волатильности от нуля статистически значимо. Если же

$$\Lambda(\boldsymbol{x}) \leq \chi_k(1-\alpha),$$

то систематических ("системных") оснований отвергнуть гипотезу H_0 в пользу предположения о том, что $a_i \neq 0$ для некоторых $i \in \{1,2,\ldots,k\}$, нет. В таком случае при некоторых дополнительных условиях, смысл которых в том, что оценкам максимального правдоподобия "запрещается" близко приближаться к границе параметрического множества (см. главу 6), при больших n вероятность ошибочно отвергнуть гипотезу о том, что динамическая компонента волатильности равна нулю, должна быть примерно равной α.

9.3. "Последовательный" критерий отношения правдоподобия для определения числа компонент смеси

Во многих реальных прикладных задачах, математическая формализация которых сводится к задаче статистического разделения смесей, число компонент смеси заранее не известно. Существует несколько более или менее обоснованных методов экспериментального отыскания числа компонент. К сожалению, во всех методах экспериментального определения числа компонент смеси в той или иной степени присутствует элемент произвола.

Некоторые из этих методов будут рассмотрены в данном и следующих разделах.

Предположим, что в модели (5.3.15) $\psi_1(x;t) \equiv \ldots \equiv \psi_k(x;t)$, то есть все компоненты смеси принадлежат к одному и тому же аналитическому типу и различаются лишь значением параметра t, так что модель (5.3.15) трансформируется к виду

$$f_\theta^X(x) = \sum_{i=1}^{k} p_i \psi(x;t_i). \tag{9.8}$$

Пусть, как и ранее, $\boldsymbol{x} = (x_1, \ldots, x_n)$ – исходная выборка, являющаяся наблюдаемым значением случайного вектора $\mathbf{X} = (X_1, \ldots, X_n)$, компоненты которого являются независимыми копиями случайной величины, плотность распределения которой имеет вид (9.8). Соответствующая этой модели функция правдоподобия имеет вид

$$L_k(\theta;\boldsymbol{x}) = \prod_{j=1}^{n} \left[\sum_{i=1}^{k} p_i \psi(x_j;t_i) \right], \tag{9.9}$$

где $\theta = (p_1, \ldots, p_k, t_1, \ldots, t_k)$.

Чтобы подчеркнуть зависимость размерности параметра $\theta = (p_1, \ldots, p_k, t_1, \ldots, t_k)$ модели (9.8) от числа компонент k, мы будем этот параметр и его оценку максимального правдоподобия обозначать $\theta(k)$ и $\widehat{\theta}(k)$, соответственно. Предположим, что значения этих оценок для $k = 1, 2, \ldots$ уже найдены. Другими словами, известны значения

$$\widehat{\theta}(k) = \arg\max_{\theta(k)} L_k(\theta(k);\boldsymbol{x}), \quad k = 1, 2, \ldots,$$

где функция правдоподобия $L_k(\theta(k);\boldsymbol{x})$ определена в (9.9).

Для статистического определения числа компонент смеси (9.8) можно воспользоваться теоремой 9.1, согласно которой при некоторых дополнительных условиях регулярности на функцию правдоподобия $L_k(\theta(k);\boldsymbol{x})$ распределение статистики критерия отношения правдоподобия

$$\Lambda_{k+1}(\mathbf{X}) = -2\log\left(\frac{L_k(\widehat{\theta}(k);\mathbf{X})}{L_{k+1}(\widehat{\theta}(k+1);\mathbf{X})}\right) \tag{9.10}$$

при $n \to \infty$ (n – объем выборки), вычисленное при справедливости гипотезы H_k о том, что "истинное" число компонент смеси равно k, стремится к

распределению хи-квадрат с числом степеней свободы, равным $l+1$, где l – размерность параметра t в модели (9.9). Число $l+1$ оказывается равным разности размерностей параметров $\theta(k+1)$ и $\theta(k)$: действительно, если число компонент смеси увеличивается на единицу, то к параметру $\theta(k)$ добавляются l параметров новой компоненты и один весовой параметр p_{k+1}.

Тогда формальная процедура статистического определения числа компонент смеси может быть сведена к последовательной проверке гипотезы H_k против альтернативы H_{k+1} с помощью критерия (9.10) (см., например, (Айвазян и др., 1989)). Зададим заранее уровень значимости α (понимаемый как вероятность ошибочно отвергнуть верную гипотезу H_k). Для $k = 1, 2, \ldots$ последовательно вычисляем значение статистики (9.10) и сравниваем его с критическим значением, равным $(1-\alpha)$-квантили $\chi_{l+1}(1-\alpha)$ распределения хи-квадрат с числом степеней свободы, равным $l+1$. Если статистика (9.10) оказывается большей, чем указанная $(1-\alpha)$-квантиль, то гипотеза H_k отвергается. В качестве итоговой оценки числа k компонент смеси (9.8) принимается такое значение k, при котором гипотеза H_k впервые оказывается не отвергнутой. Обозначим такую оценку $\widehat{k}_n$.

В работе (Орлов, 1983) показано, что построенная указанным образом оценка $\widehat{k}_n$ оказывается несколько завышенной. А именно, асимптотическое (при $n \to \infty$) распределение случайной величины $\widehat{k}_n - k$ (относительно любой из вероятностных мер $\mathsf{P}_{\theta(k)}$, соответствующей "истинному" значению числа компонент k) является геометрическим с параметром α:

$$\lim_{n\to\infty} \mathsf{P}_{\theta(k)}\big(\widehat{k}_n - k < 0\big) = 0,$$

$$\lim_{n\to\infty} \mathsf{P}_{\theta(k)}\big(\widehat{k}_n - k = r\big) = (1-\alpha)\alpha^r, \qquad r = 0, 1, 2, \ldots.$$

Несложно видеть, что при этом

$$\lim_{n\to\infty} \mathsf{E}_{\theta(k)}\widehat{k}_n = k + \frac{\alpha}{1-\alpha},$$

то есть асимптотическое смещение оценки $\widehat{k}_n$ равно $\alpha/(1-\alpha)$.

К сожалению, модели типа конечной смеси нормальных распределений не удовлетворяют условиям регулярности (см. главу 6). Поэтому реальный уровень значимости при проверке гипотезы H_k против альтернативы H_{k+1} может существенно отличаться от α вследствие отличия асимптотического распределения статистики критерия отношения правдоподобия от распределения хи-квадрат с указанным выше числом степеней свободы. Примеры вычисления предельных законов, имеющих вид распределений взвешенных сумм независимых хи-квадрат-распределенных случайных величин,

для статистик модифицированных тестов типа отношения правдоподобия и дальнейшие ссылки можно найти, например, в недавних работах (Lo, Mendell and Rubin, 2001), (Lo, 2005) и (Chen and Kalbfleisch, 2005).

9.4. Определение числа компонент смеси с помощью SEM-алгоритма

Для определения числа компонент смеси может быть приспособлен SEM-алгоритм. Соответствующая его модификация определяется следующим образом.

До начала работы алгоритма фиксируются два числа: K – максимально возможное число компонент и ν_0 – минимально допустимое число наблюдений в одном кластере. По поводу обоснования выбора этих чисел можно заметить следующее. Значение порога K в значительной мере определяется соображениями практической интерпретируемости полученного результата. Порог же ν_0 определяется соображениями минимально допустимой значимости компонент. Действительно, в соответствии с формулой (7.7), вес компоненты оказывается прямо пропорциональным численности соответствующего кластера. Поэтому если из каких-либо соображений нецелесообразно считать значимыми компоненты смеси, веса которых меньше некоторого заранее задаваемого (малого) значения, скажем, α, то ν_0, соответственно, определяется как $\nu_0 = \alpha \cdot n$, где n – объем выборки.

В рассматриваемой версии SEM-алгоритма величина k – число компонент смеси – является параметром, который может изменяться на каждой итерации. Поэтому набор начальных приближений включает и значение $k^{(0)} = K$.

Согласно работе (Celeux and Diebolt, 1984) в рассматриваемой версии SEM-алгоритма S-этап $(m+1)$-й итерации дополняется следующими действиями. Если после симуляции величин $\boldsymbol{y}$ какие-либо из соответственно полученных кластеров оказываются малочисленными, то есть при каких-либо i выполняются неравенства

$$\nu_i^{(m+1)} \leq \nu_0, \tag{9.11}$$

то соответствующий кластер аннулируется. Процедура аннулирования (изъятия) малочисленных кластеров на $(m+1)$-й итерации заключается в том, что:

1°. Переопределяются апостериорные вероятности $g_{ij}^{(m)}$: вводятся обновленные с учетом изъятия малочисленных кластеров величины

$$\widetilde{g}_{ij}^{(m)} = \begin{cases} 0, & \text{если выполнено условие (9.11),} \\ \dfrac{g_{ij}^{(m)}}{\sum\limits_{i \in \mathcal{M}(m+1)} g_{ij}^{(m)}}, & \text{в противном случае,} \end{cases}$$

где $\mathcal{M}(m+1)$ – множество номеров всех "достаточно представительных" кластеров (то есть тех кластеров, для которых условие (9.11) не выполнено) на $(m+1)$-й итерации.

2°. Наблюдения, попавшие в аннулированные кластеры, перераспределяются по "достаточно представительным" кластерам, для чего для каждого из таких наблюдений генерируются векторы $\vec{y}_j{}^{(m+1)} = (y_{1j}^{(m+1)}, y_{2j}^{(m+1)}, \ldots, y_{k^{(m)}j}^{(m+1)})$ как реализации случайных векторов с полиномиальным распределением с параметрами 1 и $\widetilde{g}_{1j}^{(m)}, \ldots, \widetilde{g}_{k^{(m)}j}^{(m)}$. "Непристроенное" наблюдение x_j приписывается к тому из "достаточно представительных" кластеров, номер которого совпадает с номером компоненты вектора $\vec{y}_j{}^{(m+1)}$, равной единице.

3°. Переопределяется число компонент смеси: $k^{(m+1)} = k^{(m)} - l_0^{(m+1)}$, где $l_0^{(m+1)}$ – число малочисленных кластеров (то есть удовлетворяющих условию (9.11)) на $(m+1)$-й итерации.

Таким образом, в соответствии с описанной процедурой аннулирования малочисленных кластеров, в ходе работы данной версии SEM-алгоритма число компонент смеси может только уменьшиться.

9.5. Информационные критерии выбора модели (числа компонент и типа смеси)

9.5.1. Информационный критерий Акаике (AIC)

Этот критерий предложен Хиротугу Акаике в 1971 г., впервые описан в работе (Akaike, 1973) и исследован в работах (Akaike, 1974), (Акаике, 1983), (Bozdogan, 1987) и многих других. Хотя в работе (Akaike, 1974) автор пояснил, что предложенное им название этого критерия AIC является аббревиатурой английского термина "An Information Criterion", в работах других авторов это название расшифровывается как "Akaike Information Criterion".

Пусть $\boldsymbol{x} = (x_1, \ldots, x_n)$ – исходная наблюдаемая выборка, являющаяся реализацией случайного вектора $\mathbf{X} = (X_1, \ldots, X_n)$, компоненты которого независимы и одинаково распределены. Пусть $f^{\mathbf{X}}(\boldsymbol{x})$ – неизвестная "истинная" плотность распределения случайного вектора $\mathbf{X}$. Для удобства обозначений в дальнейшем мы будем опускать верхний индекс у плотности и вместо $f^{\mathbf{X}}(\boldsymbol{x})$ будем писать просто $f(\boldsymbol{x})$.

Пусть $f_M(\boldsymbol{x}; \theta)$ – модельная (гипотетическая) плотность распределения случайного вектора $\mathbf{X}$ (рассматриваемые плотности понимаются относительно некоторой σ-конечной меры μ). Предположим, что размерность параметра θ равна d, то есть $\theta = (\theta_1, \ldots, \theta_d)$. Пусть $\widehat{\theta} = \widehat{\theta}(\boldsymbol{x})$ – оценка максимального правдоподобия параметра θ, построенная по выборке $\boldsymbol{x}$.

В качестве показателя качества модели Х. Акаике предложил рассматривать расстояние Кульбака–Лейблера между "истинной" $f(\cdot)$ и наиболее правдоподобной "модельной" $f_M(\cdot; \widehat{\theta})$ плотностями (см. раздел 4.2). Модель тем лучше, чем меньше указанное расстояние.

При фиксированном значении $\widehat{\theta} = \widehat{\theta}(\boldsymbol{x})$ это расстояние записывается в следующем виде. Пусть $\mathbf{Z}$ – "фиктивный" случайный вектор, имеющий такую же плотность распределения $f(\boldsymbol{x})$, как и вектор $\mathbf{X}$. Тогда

$$\mathcal{D}_{KL}[f(\cdot); f_M(\cdot; \widehat{\theta}(\boldsymbol{x}))] = \mathsf{E}\left[\log \frac{f(\mathbf{Z})}{f_M(\mathbf{Z}; \widehat{\theta}(\mathbf{X}))} \middle| \mathbf{X} = \boldsymbol{x} \right] =$$

$$= \mathsf{E}[\log f(\mathbf{Z}) | \mathbf{X} = \boldsymbol{x}] - \mathsf{E}[\log f_M(\mathbf{Z}; \widehat{\theta}(\mathbf{X})) | \mathbf{X} = \boldsymbol{x}], \qquad (9.12)$$

где математическое ожидание берется по "истинному" распределению $f(\boldsymbol{x})$. Расстояние Кульбака–Лейблера по форме тесно связано с таким понятием теории информации, как энтропия. Теоретико-информационное название расстояния Кульбака–Лейблера (9.12) – *негэнтропия* (отрицательная энтропия) распределения $f(\cdot)$ по отношению к распределению $f_M(\cdot; \widehat{\theta}(\boldsymbol{x}))$. Негэнтропию также называют *информацией по Кульбаку*, см., например, (Акаике, 1983), (Кульбак, 1967). Эти обстоятельства обусловили терминологию, согласно которой обсуждаемые критерии качества модели называют *информационными*.

Обратим внимание, что первое слагаемое в правой части (9.12) не зависит от гипотетической модели, и потому всегда постоянно, так что качество модели, понимаемое в смысле расстояния Кульбака–Лейблера между модельным и истинным распределениями, характеризуется вторым слагаемым в правой части (9.12). Однако практическое вычисление второго

слагаемого

$$-\mathsf{E}[\log f_M(\mathbf{Z};\widehat{\theta}(\mathbf{X}))|\mathbf{X}=\boldsymbol{x}] = -\int \log f_M(\mathbf{z};\widehat{\theta}(\boldsymbol{x}))f(\mathbf{z})\mu(d\mathbf{z}) \tag{9.13}$$

невозможно, поскольку "истинное" распределение $f(\mathbf{z})$ не известно.

Вместо неизвестной величины (9.13) можно использовать ее статистическую оценку $-\log f_M(\boldsymbol{x};\widehat{\theta}(\boldsymbol{x}))$. Однако при этом серьезную помеху представляет то обстоятельство, что статистика $-\log f_M(\mathbf{X};\widehat{\theta}(\mathbf{X}))$ не является несмещенной оценкой величины (9.13). Действительно, с одной стороны, в соответствии с теоремой 9.1, при некоторых условиях регулярности, обеспечивающих асимптотическую нормальность оценки максимального правдоподобия параметра θ и существование такого θ, при котором $f(\boldsymbol{x}) \equiv f_M(\boldsymbol{x};\theta)$, при $n \to \infty$ справедливо соотношение

$$2\log f_M(\mathbf{X};\widehat{\theta}(\mathbf{X})) - 2\log f(\mathbf{X}) \;\sim\; \chi^2_d, \tag{9.14}$$

где запись $Y \sim \chi^2_d$ означает, что случайная величина Y имеет распределение хи-квадрат с d степенями свободы.

С другой стороны, при условиях теоремы 9.1 распределение оценки максимального правдоподобия, вычисленное при "истинном" значении параметра θ, является асимптотически нормальным с математическим ожиданием, равным "истинному" значению параметра, и ковариационной матрицей, выражающейся в терминах фишеровской информации. Х. Акаике заметил, что этому утверждению можно придать такой вид: при $n \to \infty$

$$\mathsf{E}\big[2\log f(\mathbf{Z}) - 2\log f_M(\mathbf{Z};\widehat{\theta}(\mathbf{X}))|\mathbf{X}\big] \;\sim\; \chi^2_d, \tag{9.15}$$

где математическое ожидание берется в соответствии с "истинным" распределением $f(\boldsymbol{x})$, а $\mathbf{Z}$ – введенный выше случайный вектор, плотность распределения которого совпадает с $f(\boldsymbol{x})$ (см. (Akaike, 1973), (Акаике, 1983), (Cavanaugh, 1997)).

Возьмем теперь математические ожидания обеих частей в соотношениях (9.14) и (9.15). Соответственно, получим асимптотические (при $n \to \infty$) равенства

$$2\mathsf{E}\log f_M(\mathbf{X};\widehat{\theta}(\mathbf{X})) - 2\mathsf{E}\log f(\mathbf{X}) \;=\; d \tag{9.16}$$

и, с учетом равенства $\mathsf{EE}[\log f(\mathbf{Z})|\mathbf{X}] = \mathsf{E}\log f(\mathbf{X})$,

$$2\mathsf{E}\log f(\mathbf{X}) - 2\mathsf{EE}[\log f_M(\mathbf{Z};\widehat{\theta}(\mathbf{X}))|\mathbf{X}] \;=\; d, \tag{9.17}$$

Сложив (9.16) и (9.17), мы окончательно получим асимптотическое (при $n \to \infty$) равенство

$$\mathsf{E}\{-2\mathsf{E}[\log f_M(\mathbf{Z}; \widehat{\theta}(\mathbf{X}))|\mathbf{X}]\} = \mathsf{E}\{-2\log f_M(\mathbf{X}; \widehat{\theta}(\mathbf{X})) + 2d\}.$$

Таким образом, статистика

$$AIC = -2\log f_M(\mathbf{X}; \widehat{\theta}(\mathbf{X})) + 2d \tag{9.18}$$

является асимптотически несмещенной оценкой удвоенной величины (9.13) в том смысле, что математическое ожидание разности этих двух величин, вычисленное в соответствии с "истинным" распределением $f(\boldsymbol{x})$, стремится к нулю при $n \to \infty$ при указанных выше условиях регулярности.

Приведенные выше рассуждения дают основания считать величину AIC (9.18) критерием качества модели $f_M(\boldsymbol{x}; \theta)$, точнее, мерой "несогласия" модели и реальных данных: чем меньше AIC, тем лучше модель.

При этом ясно, что включение в модель дополнительных параметров может только увеличить правдоподобие модели и, стало быть, уменьшить первое слагаемое в AIC. Однако при этом увеличивается второе слагаемое, играющее роль "штрафа" за использование дополнительных параметров.

Пусть теперь $\boldsymbol{x} = (x_1, \ldots, x_n)$ – исходная наблюдаемая выборка, элементы которой являются независимыми реализациями случайной величины, плотность распределения которой имеет вид

$$f(x) = \sum_{i=1}^{k} \frac{p_i}{\sigma_i} \phi\Big(\frac{x - a_i}{\sigma_i}\Big), \quad x \in \mathbb{R}, \tag{9.19}$$

где $\phi(x)$ – стандартная нормальная плотность. Соответствующая этой модели функция правдоподобия имеет вид

$$L_k(\theta; \boldsymbol{x}) = \prod_{j=1}^{n} \Big[\sum_{i=1}^{k} \frac{p_i}{\sigma_i} \phi\Big(\frac{x_j - a_i}{\sigma_i}\Big) \Big],$$

где $\theta = (p_1, \ldots, p_k, a_1, \ldots, a_k, \sigma_1, \ldots, \sigma_k)$. Размерность параметра θ равна $d = 3k - 1$. Поэтому с использованием критерия Акаике ответ на вопрос о том, какое значение числа k компонент смеси является "оптимальным", имеет вид

$$k_{\text{opt}} = \arg\min_k \{-\log L_k(\widehat{\theta}(k); \boldsymbol{x}) + 3k - 1\},$$

где $\widehat{\theta}(k)$ – оценка максимального правдоподобия параметра θ, построенная в соответствии с моделью (9.19).

Точно так же, в соответствии с критерием Акаике вопрос о том, какая из моделей смесей – чисто масштабная или общая сдвиг-масштабная – лучше, при фиксированном k решается путем сравнения двух значений:

$$\tfrac{1}{2}AIC_{sl} + 1 = -\log L_k(\widehat{\theta}(k); \boldsymbol{x}) + 3k,$$

соответствующего модели, имеющей вид общей сдвиг-масштабной смеси (9.19), и

$$\tfrac{1}{2}AIC_{s} + 1 = -\log L^{\circ}_k(\widehat{\theta}^{\circ}(k); \boldsymbol{x}) + 2k,$$

соответствующего модели

$$f(x) = \sum_{i=1}^{k} \frac{p_i}{\sigma_i}\phi\Big(\frac{x}{\sigma_i}\Big), \quad x \in \mathbb{R},$$

где

$$L^{\circ}_k(\theta^{\circ}; \boldsymbol{x}) = \prod_{j=1}^{n}\Big[\sum_{i=1}^{k} \frac{p_i}{\sigma_i}\phi\Big(\frac{x_j}{\sigma_i}\Big)\Big], \quad \theta^{\circ} = (p_1, \ldots, p_k, \sigma_1, \ldots, \sigma_k),$$

$$\widehat{\theta}^{\circ}(k) = \arg\max_{\theta^{\circ}} L^{\circ}_k(\theta^{\circ}; \boldsymbol{x}).$$

При этом следует выбрать ту модель, которая соответствует меньшему из чисел AIC_{sl} и AIC_s.

Известны многие модификации критерия Акаике. В частности, в работе (Bozdogan, 1987) предложен *состоятельный* информационный критерий Акаике ($CAIC$) вида

$$CAIC = -2\log f_M(\boldsymbol{x}; \widehat{\theta}) + d(\log n + 1), \tag{9.20}$$

где n – объем выборки. Состоятельность такого критерия, по аналогии с классической трактовкой этого свойства в математической статистике, заключается в том, что вычисленная в соответствии с "истинной" моделью вероятность правильного выбора модели с помощью этого критерия при упоминавшихся выше условиях регулярности, обеспечивающих асимптотическую нормальность оценки максимального правдоподобия параметра θ и существование такого θ, при котором $f(\boldsymbol{x}) \equiv f_M(\boldsymbol{x}; \theta)$, стремится к единице при $n \to \infty$. В работе (Bozdogan, 1994) также рассматривались модифицированные информационные критерии Акаике вида

$$AIC_3 = -2\log f_M(\boldsymbol{x}; \widehat{\theta}) + 3d$$

и

$$AIC_4 = -2\log f_M(\boldsymbol{x}; \widehat{\theta}) + 4d.$$

Известны и другие имеющие простой вид модификации критерия Акаике: критерий Акаике для малых выборок (Corrected Akaike Information Criterion)

$$AIC_c = AIC + \frac{2d(d+1)}{n-d+1}$$

(Sugiura, 1978), (Hurvich and Tsai, 1989), (Hurvich and Tsai, 1995), (Cavanaugh, 1997); критерий Хэннана–Куинна

$$HQ = -2\log f_M(\boldsymbol{x};\widehat{\theta}) + 2d\log(\log n)$$

(Hannan and Quinn, 1979); критерии минимальной длины описания модели (Minimum Description Length)

$$MDL_2 = -2\log f_M(\boldsymbol{x};\widehat{\theta}) + 2d\log n,$$

$$MDL_5 = -2\log f_M(\boldsymbol{x};\widehat{\theta}) + 5d\log n$$

(Liang et al., 1992) (по поводу принципа минимальной длины описания модели см. работу (Rissanen, 1976)). В некоторых модификациях критерия Акаике используются более сложные штрафные функции, зависящие от фишеровской информации (Takeuchi, 1976), (Bozdogan, 1987), (Bozdogan, 1994) и других характеристик (Ishiguro et al., 1997).

Из всех формальных модификаций критерия Акаике наиболее используемой является байесовский информационный критерий, рассматриваемый в следующем разделе.

9.5.2. Байесовский информационный критерий (BIC)

Пусть, как и ранее, $\boldsymbol{x} = (x_1, \ldots, x_n)$ – исходная наблюдаемая выборка, элементы которой являются независимыми реализациями случайной величины, плотность распределения которой имеет вид (5.1). Соответствующая этой модели функция правдоподобия имеет вид

$$L_k(\theta;\boldsymbol{x}) = \prod_{j=1}^{n}\left[\sum_{i=1}^{k} p_i\psi_i(x_j;t_i)\right], \tag{9.21}$$

где $\theta = (p_1, \ldots, p_k, t_1, \ldots, t_k)$. В соответствии с принятой нами терминологией такая функция правдоподобия является “неполной”.

В работе (Schwartz, 1978), посвященной оцениванию размерности модели (то есть числа задействованных в ней параметров) предложен следующий критерий, названный *байесовским информационным критерием*. В соответствии с этим критерием следует вычислить величину

$$BIC(k,n) = 2\log L_k(\widehat{\theta};\boldsymbol{x}) - M_k \log n,$$

где $\widehat{\theta}$ – оценка максимального правдоподобия параметра θ, функция $L_k(\widehat{\theta};\boldsymbol{x})$ определена соотношением (9.21), M_k – число независимых параметров модели (5.1) при фиксированном k, причем само число компонент смеси не является независимым параметром, n – объем выборки. Как видно, байесовский информационный критерий асимптотически (при $n \to \infty$) аналогичен состоятельному информационному критерию Акаике (9.20). Однако при обосновании величины BIC использована байесовская идеология (также см. (Akaike, 1978), (Акаике, 1983)). Если априори все возможные значения k одинаково правдоподобны, то величина $BIC(k,n)$ с точностью до аддитивной константы пропорциональна апостериорной вероятности того, что наблюдения $\boldsymbol{x}$ соответствуют модели (5.1) с данным k. Поэтому, чем больше значение $BIC(k,n)$, тем больше правдоподобие модели (5.1) с соответствующим k.

Как мы уже отмечали, согласие модели (5.1) с данными $\boldsymbol{x}$ может только возрасти, если в модель будут добавлены дополнительные параметры. Следовательно, сама функция правдоподобия $L_k(\widehat{\theta};\boldsymbol{x})$ не может быть использована в качестве критерия адекватности модели. Поэтому, как и в информационном критерии Акаике, в байесовском информационном критерии, определяемом величиной $BIC(k,n)$, к функции правдоподобия добавляется второе слагаемое, интерпретируемое как "штраф" или наказание за использование дополнительных параметров.

Как отмечено в (Schwartz, 1978), критерий $BIC(k,n)$ можно использовать не только для выбора k, но и для сравнения разных моделей типа (5.1), определяемых разными параметризациями.

Размерность M_k параметра θ в модели (9.19) равна $3k-1$. Поэтому с использованием байесовского информационного критерия решение задачи об "оптимальном" значении числа k компонент смеси имеет вид

$$k_{\text{opt}} = \arg\max_k \{2\log L_k(\widehat{\theta}(k);\boldsymbol{x}) - (3k-1)\log n\},$$

где $\widehat{\theta}(k)$ – оценка максимального правдоподобия параметра θ, построенная в соответствии с моделью (9.19), по выборке $\boldsymbol{x} = (x_1,\ldots,x_n)$.

Точно так же, в соответствии с байесовским информационным критерием вопрос о том, какая из моделей смесей – чисто масштабная или общая сдвиг-масштабная – лучше, при фиксированном k решается путем сравнения двух значений:

$$BIC_{sl} = 2 \log L_k(\widehat{\theta}(k); \boldsymbol{x}) - (3k - 1) \log n,$$

соответствующего модели, имеющей вид общей сдвиг-масштабной смеси (9.19), и

$$BIC_s = 2 \log L_k^{\circ}(\widehat{\theta}^{\circ}(k); \boldsymbol{x}) - (2k - 1) \log n,$$

соответствующего модели (9.19), в которой $a_1 = \ldots = a_k = 0$. При этом следует выбрать ту модель, которая соответствует большему из чисел BIC_{sl} и BIC_s.

Общепринятая шкала значений критерия $BIC(k, n)$ такова (см., например, (Kass and Raftery, 1995), (Fraley and Raftery, 1998a), (Fraley and Raftery, 1998b).

Если $BIC(k, n) < 2$, то согласие модели и данных плохое.

Если $2 \leq BIC(k, n) < 6$, то согласие модели и данных удовлетворительное.

Если $6 \leq BIC(k, n) < 10$, то согласие модели и данных хорошее.

Если $BIC(k, n) \geq 10$, то согласие модели и данных отличное.

Список литературы

1. С. А. Айвазян, И. С. Енюков и Л. Д. Мешалкин. *Прикладная статистика. Основы моделирования и первичная обработка данных.* "Финансы и статистика", Москва, 1983.

2. С. А. Айвазян, В. М. Бухштабер, И. С. Енюков и Л. Д. Мешалкин. *Прикладная статистика. Классификация и снижение размерности.* "Финансы и статистика", Москва, 1989.

3. Х. Акаике. Развитие статистических методов. – в кн.: П. Эйкхофф (ред.) *Современные методы идентификации систем.* "Мир", Москва, 1983, с. 148-176.

4. Т. Андерсон. *Введение в многомерный статистический анализ.* ГИФМЛ, Москва, 1963.

5. С. А. Ашманов и А. В. Тимохов. *Теория оптимизации в задачах и упражнениях.* "Наука", Москва, 1991.

6. Д. А. Батракова и В. Ю. Королев. Вероятностно-статистический анализ хаотических информационных потоков в телекоммуникационных сетях с помощью метода скользящего разделения смесей. – в сб. *Системы и средства информатики. Специальный выпуск.* ИПИРАН, Москва, 2006, с. 183-209.

7. Д. А. Батракова, В. Ю. Королев и С. Я. Шоргин. Новый метод вероятностно-статистического анализа информационных потоков в телекоммуникационных сетях. – *Информатика и ее применения*, 2007, т. 1, вып. 1, в печати.

8. М. Вазан. *Стохастическая аппроксимация.* "Мир", Москва, 1972.

9. Ф. П. Васильев. *Методы оптимизации.* Факториал Пресс, Москва, 2002.

10. Ф. П. Васильев и А. Ю. Иваницкий. *Линейное программирование*. Факториал Пресс, Москва, 2003.

11. О. К. Исаенко и В. Ю. Урбах. Разделение смесей распределений вероятностей на их составляющие. – в сб.: *Итоги науки и техники. Теория вероятностей, математическая статистика и теоретическая кибернетика*. Издательство ВИНИТИ, Москва, 1976, с. 37-58.

12. В. Ю. Королев. *Теория вероятностей и математическая статистика*. "Проспект", Москва, 2006.

13. В. Ю. Королев. Новый подход к определению и анализу компонент волатильности финансовых индексов. – *Актуарий*, 2007, т. 1, вып. 1, с. 47-49.

14. В. Ю. Королев, В. А. Ломской, Р. Р. Пресняков и М. Рэй. Анализ компонент волатильности с помощью метода скользящего разделения смесей. – в сб. *Системы и средства информатики. Специальный выпуск*. ИПИРАН, Москва, 2005, с. 180-206.

15. В. Ю. Королев и Н. Н. Скворцова. Новый метод вероятностно-статистического анализа процессов плазменной турбулентности. – *Системы и средства информатики*, ИПИ РАН, Москва, 2005, специальный выпуск, с. 126-179.

16. В. М. Круглов. Смеси распределений вероятностей. – *Вестник Московского университета, Серия 15 Вычислительная математика и кибернетика*, 1991, № 2, с. 3-15.

17. С. Кульбак. *Теория информации и статистика*. "Наука", Москва, 1967.

18. Р. Дж. А. Литтл и Д. Б. Рубин. *Статистический анализ данных с пропусками*. "Финансы и статистика", Москва, 1991.

19. М. Б. Невельсон и Р. З. Хасьминский. *Стохастическая аппроксимация и рекуррентное оценивание*. "Наука", Москва, 1972.

20. А. И. Орлов. Некоторые вероятностные вопросы теории классификации. – в кн. *Прикладная статистика*. "Наука", Москва, 1983, с. 166-179.

21. Дж. Себер. *Линейный регрессионный анализ*. "Мир", Москва, 1980.

22. С. Уилкс. *Математическая статистика*. "Наука", Москва, 1967.

23. П. Халмош. *Теория меры*. Изд-во иностранной литературы, Москва, 1953.

24. М. И. Шлезингер. О самопроизвольном распознавании образов. – в сб.: *Читающие автоматы*. "Наукова думка", Киев, 1965.

25. М. И. Шлезингер. Взаимосвязь обучения и самообучения в распознавании образов. – *Кибернетика*, 1968, № 2, с. 81-88.

26. H. Akaike. Information theory and an extension of the maximum likelihood principle. – in: B. N. Petrov and F. Csake (eds.) *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest, 1973, p. 267-281.

27. H. Akaike. A new look at the statistical model identification. – *IEEE Transactions on Automatic Control, AC-19*, 1974, p. 716-723.

28. H. Akaike. A Bayesian analysis of the minimum AIC procedure. – *Ann. Inst. Statist. Math.*, 1978, vol. 30A, p. 9-14.

29. J. A. Bilmes. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. Technical report TR-97-021. International Computer Science Institute, Berkeley, CA, 1998. (Avaliable at: http://ssli.ee.washington.edu/people/bulyko/papers/em.pdf)

30. J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. – *Journal of the Royal Statistical Society, Series B*, 1999, vol. 61, p. 265-285.

31. H. Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. – *Psychometrika*, 1987, vol. 52(3), p. 345-370.

32. H. Bozdogan. Mixture-model cluster analysis using model selection criteria and a new information measure of complexity. – in: H. Bozdogan (Ed.) *Proceedings of the First US–Japan Conference on the Frontiers of Statistical Modeling: An Information Approach*. Vol. 2. Kluwer Academic Publishers, Boston, 1994, p. 69-113.

33. R. A. Boyles. On the convergence of the EM-algorithm. – *Journal of the Royal Statistical Society, Series B*, 1983, vol. 45, p. 47-50.

34. M. Broniatowski, G. Celeux and J. Diebolt. Reconnaissance de mélanges de densités par un algorithme d'apprentisage probabiliste. – in: E. Diday, M. Jambu, L. Lebart, J.-P. Pagès and R. Tomasone (Eds.) *Data Analysis and Informatics*, III, North Holland, Amsterdam, 1983, p. 359-373.

35. B. S. Caffo, W. S. Jank and G. L. Jones. Ascent-Based Monte Carlo EM. – *Journal of the Royal Statistical Society, Series B*, 2005, vol. 67, p. 235-252.

36. J. Cavanaugh. Unifying the derivations of the Akaike and corrected Akaike information criteria. – *Statistics and Probability Letters*, 1997, vol. 31, p. 201-208.

37. G. Celeux and J. Diebolt. *Reconnaissance de mélanges de densité et classification. Un algorithme d'apprentisage probabiliste: l'algorithme SEM.* Rapport de Recherche de l'INRIA RR-0349. Centre de Rocquencourt. 1984.

38. G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. – *Computational Statistics Quarterly*, 1985, vol. 2, No. 1, p. 73-82.

39. G. Celeux and G. Govaert. *A Classification EM Algorithm for Clustering and Two Stochastic Versions.* Rapport de Recherche de l'INRIA RR-1364. Centre de Rocquencourt. 1991.

40. G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. – *Computational Statistics and Data Analysis*, 1992, vol. 14, p. 315-332.

41. G. Celeux and J. Diebolt. A stochastic approximation type algorithm for the mixture problem. – *Stochastics and Stochastics Reports*, 1992, vol. 41, p. 119-134.

42. G. Celeux, D. Chauveau and J. Diebolt. *On Stochastic Versions of the EM-algorithm.* Rapport de Recherche de l'INRIA RR-2514. Centre de Rocquencourt. 1995.

43. J. Chen and J. D. Kalbfleisch. Modified likelihood ratio test in finite mixture models with a structural parameter. – *Journal of Statistical Planning and Inference*, 2005, vol. 129, p. 93-107.

44. H. Chernoff. On the distribution of the likelihood ratio. – *Annals of Mathematical Statistics*, 1954, vol. 25, p. 573-578.

45. S. Chrétien and A. Hero. Kullback proximal algorithms for maximum likelihood estimation. – *IEEE Transactions on Information Theory*, 2000, vol. 46, No. 5, p. 1800-1810.

46. T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, New York, 1991.

47. N. E. Day. Divisive cluster analysis and test for multivariate normality. – *Session of the ISI*, London, 1969.

48. N. E. Day. Estimating the components of a mixture of normal distributions. – *Biometrika*, 1969, vol. 56, No. 3, p. 463-474.

49. B. Delyon, M. Lavielle and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. – *The Annals of Statistics*, 1999, vol. 27, p. 94-128.

50. A. Dempster, N. Laird and D. Rubin. Maximum likelihood estimation from incomplete data. – *Journal of the Royal Statistical Society, Series B*, 1977, vol. 39, p. 1-38.

51. J. Diebolt and G. Celeux. Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Communications in Statistics B: Stochastic Models*, 1993, vol. 9, No 4, p. 599-613.

52. J. Diebolt and E. H. S. Ip. Stochastic EM: method and application. – in: W. R. Gilks, S. Richardson and D. J. Spiegelhalter (Eds.) *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996, p. 259-273.

53. B. S. Everitt and D. J. Hand. *Finite Mixture Distributions*. Chapman and Hall, 1981.

54. M. A. T. Figueiredo. *Lecture Notes on the EM Algorithm*, 2004. Available at: http://www.stat.duke.edu/courses/Spring06/sta376/Support/EM/EM.Mixtures.Figueiredo.2004.pdf

55. C. Fraley and A. E. Raftery. *How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis*. Technical Report No. 329, Department of Statistics, University of Washington, Seattle, 1998. Also available at http://www.ics.uci.edu/ smyth/ courses/ics274/fraley__raftery.pdf

56. C. Fraley and A. E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. – *Computer Journal*, 1998, vol. 41, p. 578-588.

57. E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. – *Journal of the Royal Statistical Society*, 1979, vol. B41, p. 190-195.

58. M. J. R. Healy and M. H. Westmacott. Missing values in experiments analyzed on automatic computers. – *Applied Statistics*, 1956, vol. 5, p. 203-206.

59. C. M. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. – *Biometrika*, 1989, vol. 76, p. 297-307.

60. C. M. Hurvich and C.-L. Tsai. Model selection for extended quasi-likelihood models in small samples. – *Biometrics*, 1995, vol. 51, p. 1077-1084.

61. E. H. Ip. *A Stochastic EM Estimator in the Presence of Missing Data – Theory and Practice*. PhD Dissertation, Stanford University, 1994, 127 p.

62. M. Ishiguro, Y. Sakamoto and G. Kitagawa. Bootstrapping log likelihood and EIC, an extension of AIC. – *Annals of the Institute of Statistical Mathematics*, 1997, vol. 49, p. 411-434.

63. W. Jank. *Implementing and Diagnosing the Stochastic Approximation EM Algorithm.* Technical Report, University of Maryland, 2004. Also available at: http:// www.smith.umd.edu/faculty/wjank/ monitorSAEM.pdf

64. H. Jeffreys. *Theory of Probability.* 3rd edition. Clarendon, 1961.

65. M. Jordan and L. Xu. Convergence results for the EM approach to mixtures of experts architectures. – *Neural Networks*, 1996, vol. 8, p. 1409-1431.

66. G. G. Judge and T. Takayama. Inequality restrictions in regression analysis. – *Journal of American Statistical Association*, 1966, vol. 61, No. 1, p. 166-181.

67. R. E. Kass and A. E. Raftery. Bayes factors. – *Journal of the American Statistical Society*, 1995, vol. 90, p. 773-795.

68. J. Kazakos. Recursive estimation of prior probabilities using a mixture. – *IEEE Transactions on Information Theory*, 1977, vol. 23, No. 2, p. 203-210.

69. V. Yu. Korolev and M. Rey. Statistical analysis of volatility of financial time series and turbulent plasmas by the method of moving separation of mixtures. – in: V. Yu. Korolev and N. N. Skvortsova (Eds.) *Stochastic Models of Plasma Turbulence* VSP, Utrecht, 2005.

70. S. Kullback and R. A. Leibler. On information and sufficiency. – *Annals of Mathematical Statistics*, 1951, vol. 22, p. 79-86.

71. R. A. Levine and G. Casella. Implementations of the Monte Carlo EM algorithm. – *Journal of Computational and Graphical Statistics*, 2001, vol. 10, p. 422-439.

72. R. F. Levine and J. Fan. An automated (Markov Chain) Monte Carlo EM algorithm. – *Journal of Statistical Computation and Simulation*, 2004, vol. 74, p. 349-359.

73. Z. Liang, R. J. Jaszczak, R. E. Coleman. Parameter estimation of finite mixtures using the EM algorithm and information criteria with applications to medical image processing. – *IEEE Transactions Nuclear Science*, 1992, vol. 39, No. 4, p. 901-913.

74. Y. Lo, N. R. Mendell and D. B. Rubin. Testing the number of components in a normal mixture. – *Biometrika*, 2001, vol. 88, No. 3, p. 767-778.

75. Y. Lo. Likelihood ratio tests of the number of components in a normal mixture with unequal variances. – *Statistics and Probability Letters*, 2005, vol. 71, p. 225-235.

76. B. Martinet. Regularisation d'inéquations variationnelles par approximations successives. – *Revue Francaise d'Informatique et le Recherche Operationelle*, 1970, vol. 3, p. 154-170.

77. A. G. McKendrick. Applications of mathematics to medical problems. – *Proceedings of the Edinburgh Mathematical Society*, 1926, vol. 44, p. 98-130.

78. G. McLachlan and K. Basford. *Mixture Models: Inference and Application to Clustering*. Marcel Dekker, New York, 1988.

79. G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley and Sons, New York, 1997.

80. G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley and Sons, New York, 2000.

81. P. Medgyessy. *Decomposition of Superpositions of Distribution Functions*. Publishing House of the Hungarian Academy of Sciences, Budapest, 1961.

82. R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. – in: M. I. Jordan (Ed.). *Learning in Graphical Models*. Kluwer Academic Publishers, 1998, p. 355-368.

83. S. F. Nielsen. The stochastic EM algorithm: estimation and asymptotic results. – *Bernoulli*, 2000, vol. 6, No. 3, p. 457-489.

84. B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. – *SIAM Journal of Control and Optimization*, 1992, vol. 30, p. 838-855.

85. B. L. S. Prakasa Rao. *Identifiability in Stochastic Models*. Academic Press, Boston–San Diego–New York–London–Sydney–Tokyo–Toronto, 1992.

86. R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. – *SIAM Review*, 1984, vol. 26, No. 2, p. 195-239.

87. J. Rissanen. Minimax entropy estimation of models for vector processes. – in: R. K. Mehra and D. G. Lainiotis (Eds.) *System Identification. Advances and Case Studies*. Academic Press, New York, 1976, p. 97-119.

88. H. Robbins and S. Monro. A stochastic approximation method. – *The Annals of Mathematical Statistics*, 1951, vol. 22, p. 400-407.

89. R. Rockafellar. Monotone operators and the proximal point algorithm. – *SIAM Journal on Control and Optimization*, 1976, vol. 14, p. 877-898.

90. G. Schwartz. Estimating the dimension of a model. – *The Anaals of Statistics*, 1978, vol. 6, p. 461-464.

91. N. Sugiura. Further analysis of the data by Akaike's information criterion and the finite corrections. – *Communications in Statistics. Theory and Methods*, 1978, vol. A7, p. 13-26.

92. K. Takeuchi. Distribution of informational statistics and a criterion of model fitting. – *Suri-Kagaku* (*Mathematical Sciences*), 1976, vol. 153, p. 12-18 (in Japanese).

93. G. M. Tallis. The identifiability of mixtures of distributions. – *Journal of Applied Probability*, 1969, vol. 6, No. 2, p. 389-398.

94. M. Tanner. *Tools for Statistical Inference*. Springer-Verlag, New York, 1993.

95. H. Teicher. Identifiability of mixtures. – *Ann. Math. Stat.*, 1961, vol. 32, p. 244-248.

96. H. Teicher. Identifiability of finite mixtures. – *Ann. Math. Stat.*, 1963, vol. 34, No. 4, p. 1265-1269.

97. D. M. Titterington, A. F. Smith and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, Chichester–New York–Brisbane–Toronto–Singapore, 1987.

98. J. W. Tukey. A survey of sampling from contaminated distributions. – in: I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann (Eds.) *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*. Stanford University Press, Stanford, CA, 1960, p. 448-485.

99. M. S. Waterman. A restricted least squares problem. – *Technometrics*, 1974, vol. 16, No. 1, p. 135-136.

100. G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. – Journal of the American Statistical Association, 1990, vol. 85, p. 699-704.

101. S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. – *The Annals of Mathematical Statistics*, 1938, vol. 9, p. 60-62.

102. J. H. Wolfe. Pattern clustering by multivariate mixture analysis. – *Multivariate Behavioral Research*, 1970, vol. 5, p. 329-350.

103. C. F. Wu. On the convergence properties of the EM-algorithm. – *The Annals of Statistics*, 1983, vol. 11, No. 1, p. 95-103.

104. L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. – *Neural Computation*, 1996, vol. 8, p. 129-151.

105. S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. – *Ann. Math. Statistics*, 1968, vol. 39, p. 209-214.