**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2012 Т. 4 № 1 С. 45–53**

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.254

О применении асимптотических критериев для определения числа компонент смеси вероятностных распределений

А. К. Горшенин

Институт проблем информатики Российской академии наук,
Россия, 119333, Москва, ул. Вавилова, д. 44, кор. 2

E-mail: agorshenin@ipiran.ru

Получено 13 февраля 2012 г.

В статье демонстрируется практическая эффективность применения асимптотически наиболее мощных критериев проверки гипотез о числе компонент смеси в моделях добавления и расщепления компонент. Тестовые данные представляют собой выборки из различных конечных смесей нормальных законов. Проводится сравнение результатов для разнообразных уровней значимости и весов.

Ключевые слова: конечная смесь нормальных распределений, асимптотически наиболее мощный критерий

On application of the asymptotic tests for estimating the number of mixture distribution components

A. K. Gorshenin

*Institute of Informatics Problems, Russian Academy of Sciences,
Vavilova street, 44/2, Moscow, 119333, Russia*

Abstract. — The paper demonstrates the efficiency of asymptotically most powerful test of statistical hypotheses about the number of mixture components in the adding and splitting component models. Test data are the samples from different finite normal mixtures. The results are compared for various significance levels and weights.

Keywords: finite mixture of normal distributions, asymptotically most powerful test

Citation: *Computer Research and Modeling*, 2012, vol. 4, no. 1, pp. 45–53 (Russian).

Введение

Для адекватного описания сложных стохастических систем (например, финансовых рынков, турбулентной плазмы и так далее) часто используется математическая модель, основанная на конечной смеси вероятностных распределений, то есть

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

где $k \geqslant 1$ — 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 $p_i \geqslant 0$, $\sum\limits_{i=1}^{k} p_i = 1$; t_i, $i = 1, \ldots, k$ — многомерные параметры. Плотности $\psi_1, \ldots, \psi_k$ обычно называют компонентами смеси (1), а параметры $p_1, \ldots, p_k$ — весами соответствующих компонент.

На практике крайне важна корректная интерпретация полученных результатов (например, в физике турбулентной плазмы необходимо соотносить полученные компоненты смеси с наблюдаемыми в плазме процессами). Поэтому возникает задача определения неизвестных параметров смеси на основе анализа данных, для которых строится модель. Для ее решения принято использовать так называемые алгоритмы EM-типа (EM, SEM, MCEM-алгоритмы и их всевозможные модификации [Dempster, Laird, Rubin, 1977; Broniatowski, Celeux, Diebolt, 1984; Горшенин, Королев, Турсунбаев, 2008]). Однако данные алгоритмы при определении параметров смеси используют заданное число компонент и в процессе итерационной процедуры не могут менять это заданное число. При этом обычно число компонент также является неизвестным параметром. Существуют предназначенные для статистического определения числа компонент так называемые информационные критерии, основанные на функции правдоподобия: критерий Акаике [Akaike, 1973], байесовский критерий [Schwartz, 1978], критерий Ло [Lo, Mendell, Rubin, 2001]. Однако в ряде практически значимых моделей нарушаются условия регулярности (например, для конечной смеси нормальных законов), что приводит к необходимости накладывать дополнительные искусственные технические условия для корректности использования таких критериев. Более того, критерии Акаике и байесовский применимы только при очень больших объемах выборок, а распределение статистики критерия Ло на практике исключительно сложно вычисляется.

Для преодоления указанных недостатков в работах [Бенинг, Горшенин, Королев, 2011; Горшенин, 2011a; Горшенин, 2011b] были предложены асимптотически наиболее мощные критерии проверки гипотез о числе компонент смеси. Для формализации задачи использовались две модели конечных сдвиг-масштабных смесей произвольных абсолютно непрерывных распределений: добавления и расщепления компоненты. При этом такая формализация хорошо согласуется с наиболее часто встречающимися на практике случаями.

Настоящая статья посвящена исследованию практической целесообразности использования полученных в работах [Бенинг, Горшенин, Королев, 2011; Горшенин, 2011a; Горшенин, 2011b] статистических критериев, изучению их эффективности в зависимости от задаваемого уровня значимости, а также от объема выборки.

Модели добавления и расщепления компоненты

Рассмотрим каждую из моделей, а также укажем статистики, на основе которых построены упоминавшиеся выше асимптотически наиболее мощные критерии.

Модель добавления компоненты формализуется следующим образом. Предполагается, что каждое из независимых наблюдений $(X_1, \ldots, X_n)$ имеет плотность, представимую в виде

конечной k-компонентной смеси некоторых законов распределения вида ($\theta \in [0,1]$):

$$p(x,\theta) = (1-\theta)\cdot\sum_{i=1}^{k} p_i\psi_i(x) + \theta\cdot\psi_{k+1}(x) = (1-\theta)f(x) + \theta g(x), \quad \sum_{i=1}^{k} p_i = 1. \tag{2}$$

Отметим, что рассматриваются такие версии плотностей $\psi_i(x)$, $i = 1,\ldots,k$, что функция $f(x)$ строго положительна. В этом случае критерий проверки гипотез о числе компонент смеси основан на статистике (функции $f(x)$ и $g(x)$ определены в соотношении (2))

$$T_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{g(X_i)}{f(X_i)} - 1\right). \tag{3}$$

Данная модель ориентируется на проверку значимости произвольной компоненты с возможным малым весом. Формулировка и доказательство теоремы об асимптотических свойствах критерия, основанного на статистике T_1, приводятся в статье [Бенинг, Горшенин, Королев, 2011].

Решая задачу об уменьшении числа компонент в подгоняемой модели смеси вероятностных распределений, важно не исключить из рассмотрения практически важные компоненты, ошибочно объединив их в одну компоненту. Это означает, что возможна ситуация, когда в смеси присутствуют компоненты с близкими значениями параметров, в том числе и весов. Этому случаю соответствует модель расщепления компоненты. Предполагаем, что каждое из независимых наблюдений имеет плотность вида (для некоторого $\theta \in [0,1]$):

$$p(x,\theta) = \sum_{i=1}^{k-1} p_i\psi_i(x) + (p_k - \theta)\psi_k(x) + \theta\psi(x) = f(x) + \theta g(x), \tag{4}$$

где

$$f(x) = \sum_{i=1}^{k} p_i\psi_i(x),\ \sum_{i=1}^{k} p_i = 1,\ p_i \geqslant 0,\ g(x) = \psi(x) - \psi_k(x),\ 0 \leqslant \theta \leqslant p_k,$$

функция $\psi(x)$ является плотностью из того же семейства распределений, что и все $\psi_i(x)$. Рассматриваются такие версии плотностей ψ_i, $i = 1,\ldots,k$, что функция $f(x)$ строго положительна. Отметим, что в этом случае функция $g(x)$, вообще говоря, не является плотностью какого-либо распределения. Критерий проверки гипотез о числе компонент смеси в модели расщепления компоненты основан на статистике (функции $f(x)$ и $g(x)$ определены в соотношении (4))

$$T_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)}. \tag{5}$$

Формулировка и доказательство теоремы об асимптотических свойствах критерия, основанного на статистике T_2, приводятся, например, в статье [Горшенин, 2011а].

Тестирование критериев T_1 и T_2

В данном разделе рассмотрим эффективность практического применения критериев, основанных на статистиках T_1 и T_2, определяемых соотношениями (3) и (5) соответственно, для выборок конечного объема для моделей добавления и расщепления компонент на примере серий различных тестовых выборок.

Для проверки эффективности каждого из методов был проведен статистический эксперимент. Выбирались различные функции распределения, имеющие вид конечных смесей нормальных законов (рассматривались сдвиговые, масштабные и сдвиг-масштабные смеси). Затем с помощью метода обратных функций моделировались выборки из данного распределения: по 200 выборок объемом 1 000, 5 000 и 10 000 элементов. Затем к каждой из выборок применялись оба критерия. Результаты сравнивались с истинным количеством компонент в смеси (то есть отвергают ли критерии компоненту с «малым» весом). При этом в целях более подробного исследования свойств критериев в качестве проверяемых весов выбирались как действительно малые значения (0.01, 0.05), так и достаточно большие (0.1, 0.25, 1/3, 0.5). Также рассматривались различные уровни значимости (для единообразия приведены результаты для значений 0.1, 0.15, 0.2, 0.3, хотя для больших весов проверки проводились и для уровней 0.05 и 0.01, однако для сопоставимых значений весов и уровней значимости такое сравнение было бы некорректно).

Отметим, что для нахождения критического значения в обеих моделях предполагается знание значения фишеровской информации, определяемой выражениями

$$I_1 = \int\limits_{-\infty}^{\infty} \frac{g^2(x)}{f(x)}\,dx - 1 \tag{6}$$

в модели добавления компоненты (функции $f(x)$ и $g(x)$ определяются из соотношения (2)) и

$$I_2 = \int\limits_{-\infty}^{\infty} \frac{g^2(x)}{f(x)}\,dx \tag{7}$$

в модели расщепления компоненты (функции $f(x)$ и $g(x)$ определяются из соотношения (4)).

Однако, за исключением случаев проверки гипотез вида «нормальное распределение» против альтернативы «смесь двух нормальных распределений», для которого соответствующий интеграл можно найти явно, уже для случаев проверки гипотез вида «смесь содержит 2 компоненты» против альтернативы вида «смесь содержит 3 компоненты» соответствующие интегралы (6) и (7) не вычисляются в элементарных функциях. Поэтому для нахождения значения фишеровской информации используется численное интегрирование по методу Гаусса–Кронрода (см., например, [Кронрод, 1964]). Более того, интегралы в формулах (6) и (7) берутся по бесконечной области. Для перехода к конечным пределам интегрирования воспользуемся следующими соображениями. Пусть $\sigma_{j_0} = \max\limits_{1 \leqslant j \leqslant k} \sigma_j$. Тогда

$$\frac{g^2(x)}{f(x)} \leqslant \frac{\sigma_{j_0}}{p_{j_0}\sigma_{k+1}\sqrt{2\pi}}\,\exp\left\{-\frac{(x-A)^2}{2B^2} + C\right\},$$

где

$$A = \left(\frac{a_{k+1}}{\sigma_{k+1}^2} - \frac{a_{j_0}}{2\sigma_{j_0}^2}\right)\left(\frac{1}{\sigma_{k+1}^2} - \frac{1}{2\sigma_{j_0}^2}\right)^{-1},\ 2B^2 = \left(\frac{1}{\sigma_{k+1}^2} - \frac{1}{2\sigma_{j_0}^2}\right)^{-1},$$

$$C = \frac{A^2}{2B^2} - \frac{a_{k+1}^2}{\sigma_{k+1}^2} + \frac{a_{j_0}^2}{2\sigma_{j_0}^2}.$$

Окончательно получаем

$$\frac{1}{B\sqrt{2\pi}} \int \exp\left\{-\frac{(x-A)^2}{2B^2}\right\} dx < \frac{\varepsilon}{2C_0},$$

где

$$C_0 = \frac{2Be^C\sigma_{j_0}}{p_{j_0}\sigma_{k+1}^2}.$$

Таким образом, для того чтобы приблизить интеграл по бесконечной области интегралом по конечной, необходимо найти квантили нормального распределения с параметрами A и B^2 уровней $\frac{\varepsilon}{2C_0}$ и $1 - \frac{\varepsilon}{2C_0}$. Обозначим через a ближайшее целое число снизу для первой квантили и через b ближайшее целое число сверху для второй квантили. Тогда

$$\left|\int_{-\infty}^{+\infty}\frac{g^2(x)}{f(x)}\,dx - \int_a^b\frac{g^2(x)}{f(x)}\,dx\right| = \left|\int_{-\infty}^{a}\frac{g^2(x)}{f(x)}\,dx + \int_b^{+\infty}\frac{g^2(x)}{f(x)}\,dx\right| \leqslant$$

$$\leqslant \left|\int_{-\infty}^{a}\frac{g^2(x)}{f(x)}\,dx\right| + \left|\int_b^{+\infty}\frac{g^2(x)}{f(x)}\,dx\right| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Оба критерия были реализованы на встроенном языке программирования пакета MATLAB. В качестве точности приближения интеграла по бесконечному интервалу интегралом по конечному множеству была выбрана величина $\varepsilon = 10^{-12}$.

Результаты тестирования критериев представлены на рис. 1 и 2. На оси абсцисс отмечены значения проверяемых весов, в скобках указан объем выборки, на котором был достигнут подобный результат. На оси ординат отмечено число правильных решений для серий выборок в процентах для каждого из критериев. Если на каком-то объеме выборки успешность теста значимости компоненты с некоторым весом составляла 100% или была близка к этому значению, данный объем выборки признавался достаточным для правильного распознавания данного веса. Столбцы каждого цвета соответствуют различным уровням значимости критерия.

Отметим, что так как оба критерия являются асимптотическими, увеличение объема выборки положительно сказывается на успешности проведения теста. Так, на рис. 3 и 4 для обеих моделей можно проследить изменение результатов для малого веса 0.01 с изменением размера выборки с 1 000 элементов до 5 000 и 10 000 элементов. При заданных уровнях значимости критерия, при объеме выборки в 1 000 элементов каждый из критериев ошибается значительно в большем числе случаев, чем допускает заданный уровень. На выборках размером в 5 000 элементов ситуация улучшается: на уровнях 0.2 и 0.3 ошибок меньше уровня значимости. А уже на выборках в 10 000 элементов хороший уровень успешности тестов достигается для всех рассматриваемых уровней. Таким образом, если на практике есть необходимость точного различения малых весов, желательно использовать выборки тем большего размера, чем более маленький вес нужно учитывать. При этом в большинстве стандартных случаев для правильного различения весов вплоть до 0.01 достаточно объема выборки в 10 000 элементов.

Однако в ситуации, когда исходная смесь состоит из большого числа компонент и они достаточно близки по всем параметрам, для лучшей эффективности работы критериев необходимо использовать выборки по возможности большего объема, чтобы гарантировать правильность статистического определения числа компонент в смеси, особенно если необходимо правильно различать небольшие веса.

Таким образом, можно сформулировать алгоритм для практического применения критериев, основанных на статистиках (3) и (5). Если нет дополнительной информации о минимальном числе компонент исходя из особенностей области получения данных, для которых строится модель смеси вероятностных распределений, то следует последовательно проверять гипотезу

$$H_0 : \ K = k$$

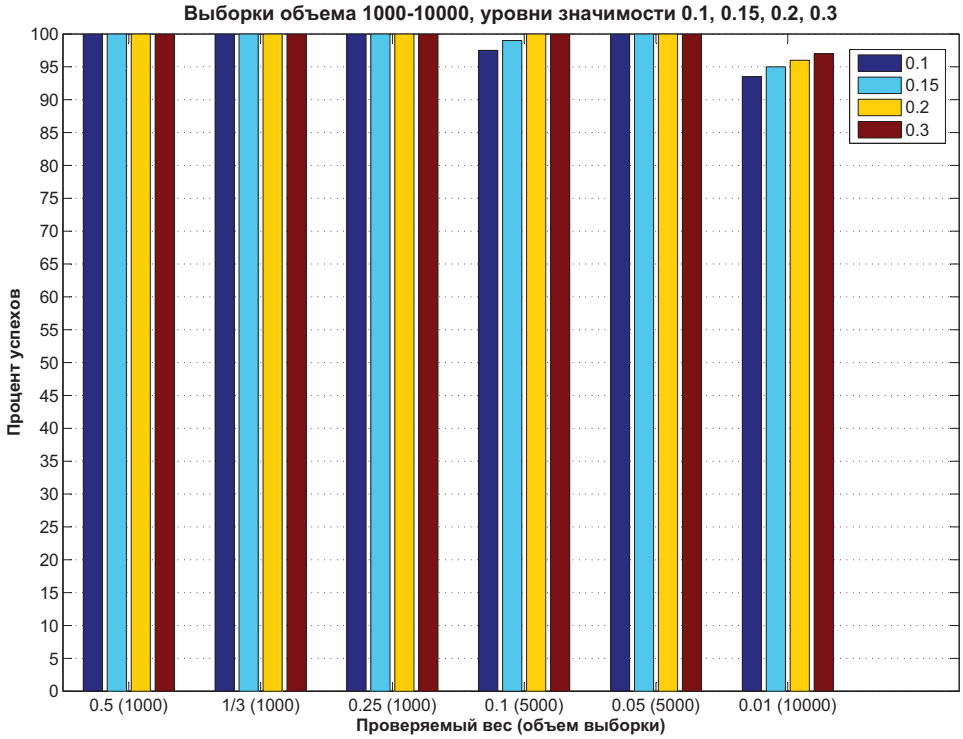

Рис. 1. Результаты применения критерия (модель добавления компоненты)

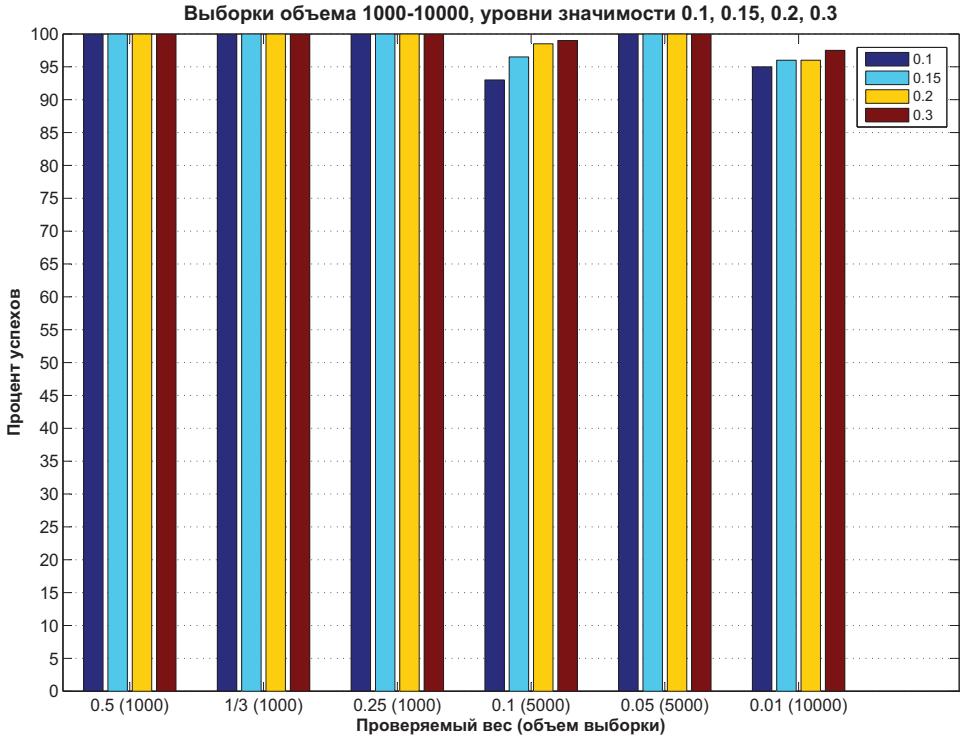

Рис. 2. Результаты применения критерия (модель расщепления компоненты)

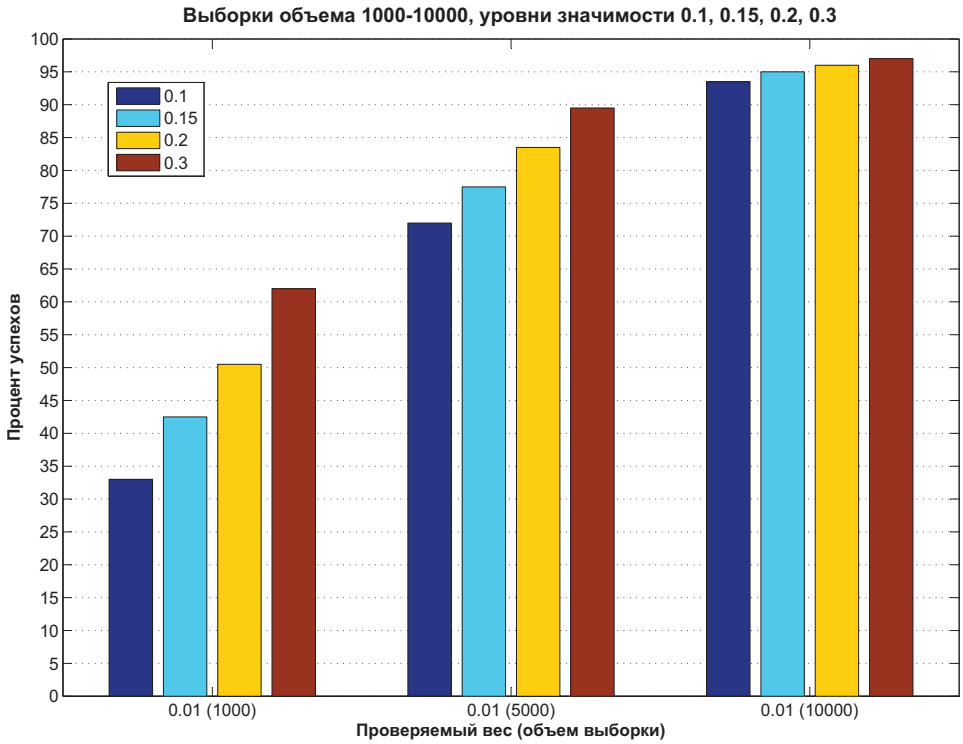

Рис. 3. Асимптотический характер поведения критерия (модель добавления компоненты)

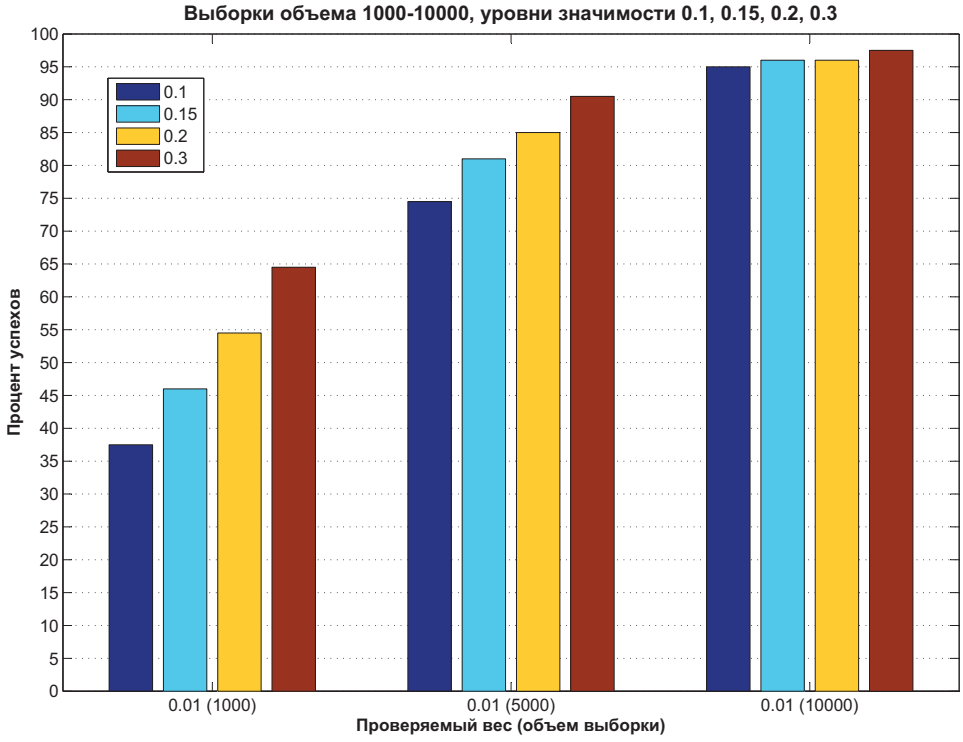

Рис. 4. Асимптотический характер поведения критерия (модель расщепления компоненты)

против альтернативы

$$H_1 : K = k + 1$$

для всех возможных значений k, начиная со значения $k = 1$, до тех пор, пока критерием не будет установлена справедливость гипотезы. Полученное значение k следует рассматривать как истинное значения числа компонент. Если же известны ограничения снизу на число компонент, то начинать тестирование следует именно с этого значения. Основываясь на результатах тестирования критериев на модельных данных, можно утверждать, что если используемый метод оценивания параметров смеси достаточно точно приближает параметры компонент исходной смеси, то при анализе выборки достаточного объема предложенные асимптотически наиболее мощные критерии правильно оценят наличие компонент с малым весом. При этом разумно использовать методы с фиксированным числом компонент, без дополнительных изменений, связанных с возможностью удалять некие компоненты на этапе итерационных шагов (так как сами по себе методы зачастую ошибаются в оценке числа компонент, предпочтительнее использовать возможности, предоставляемые найденными асимптотически наиболее мощными критериями).

Заключение

Результаты тестирования означают, что предложенные критерии могут использоваться и при умеренных объемах выборок без существенных ограничений в большинстве ситуаций. В то же время, в случае необходимости проверки статистической значимости компонент с действительно малыми весами для обеспечения максимальной корректности результатов теста необходимо выбирать выборку максимально возможного объема.

Отметим относительную простоту практического использования критериев, основанных на статистиках (3) и (5). Дело в том, что критическое значение находится как квантиль стандартного нормального распределения соответствующего уровня, а данная задача решается стандартными статистическими процедурами в различных программных пакетах (или же с использованием таблиц). В то же время статистика упомянутого выше критерия Ло, который может быть использован для решения аналогичных задач, в случае справедливости нулевой гипотезы имеет асимптотическое распределение, являющееся взвешенной суммой χ^2-распределений. Функция распределения предельного закона не выражается в элементарных функциях, более того, подынтегральное выражение имеет весьма сложный вид, а также зависит от собственных значений специальной матрицы, формирование которой требует нахождения для каждого элемента выборки значений матрицы размера $(3k-1)\times(3k-1)$. Уже для умеренных значений k, например, $k = 3$, и размера выборки в 1 000 элементов скорость работы, по сравнению с предложенными в этой главе критериями, различается на несколько порядков. Эта разница возрастает с увеличением объема выборки.

Модель расщепления компоненты предназначена для некоторой специфической ситуации, в то время как модель добавления компоненты является вполне универсальной и подходит для большинства случаев. Проведенное сравнение показывает, что статистические критерии для каждой модели являются весьма эффективными, простыми для реализации различными программными средствами и удобными для применения на практике в силу высокой точности и скорости своей работы. Поэтому во многих ситуациях использование данных критериев представляется более предпочтительным, нежели применение методов, основанных на информационных критериях.

Список литературы

Бенинг В. Е., Горшенин А. К., Королев В. Ю. Асимптотически оптимальный критерий проверки гипотез о числе компонент смеси вероятностных распределений // Информатика и ее применения. — 2001. — Т. 5, вып. 3. — С. 4–16.

Горшенин А. К. Проверка статистических гипотез в модели расщепления компоненты // Вестник Московского университета. Сер. 15. Вычисл. матем. и киберн. — 2011. — № 4. — С. 26–32.

Горшенин А. К., Королев В. Ю., Турсунбаев А. М. Медианные модификации EM- и SEM-алгоритмов для разделения смесей вероятностных распределений и их применение к декомпозиции волатильности финансовых временных рядов // Информатика и ее применения. — 2008. — Т. 2, вып. 4. — С. 12–47.

Кронрод А. С. Узлы и веса квадратурных формул. — М.: Наука, 1964.

Akaike H. Information theory and an extension of the maximum likelihood principle.// In: B. N. Petrov and F. Csake (eds.) Second International Symposium on Information Theory. — Budapest, 1973. P. 267–281.

Broniatowski M., Celeux G. and Diebolt J. Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste // Data Analysis and Informatics. — 1984. — Vol. 3. — P. 359–373.

Dempster A., Laird N. and Rubin D. Maximum likelihood estimation from incompleted data // Journal of the Royal Statistical Society, 1977. Series B. — Vol. 39(1). — P. 1–38.

Gorshenin A. K. Testing of statistical hypotheses in the splitting component model // Moscow University Computational Mathematics and Cybernetics. — 2011. — Vol. 35, No. 4. — P. 176–183. DOI: 10.3103/S0278641911040054.

Lo Y., Mendell N. R. and Rubin D. B. Testing the number of components in a normal mixture // Biometrika. — 2001. — Vol. 88, №. 3. — P. 767–778.

Schwartz G. Estimating the dimension of a model // The Annals of Statistics. — 1978. — Vol. 6. — P. 461–464.