ОЦЕНКА КАЧЕСТВА АППРОКСИМАЦИИ ПЛОТНОСТЕЙ РАСПРЕДЕЛЕНИЯ ПЛОТНОСТЯМИ ИЗ ЭКСПОНЕНЦИАЛЬНОГО СЕМЕЙСТВА РАСПРЕДЕЛЕНИЙ

М.С. Лохвицкий, *к.т.н., доцент, Московский технический университет связи и информатики.*

УДК 621.396

Аннотация. Рассматриваются вопросы оценки качества аппроксимации неизвестных плотностей распределения плотностями из экспоненциального семейства распределений и полигауссовыми распределениями. В качестве критериев качества используются как статистические (критерий *Х*и-квадрат), так и критерии близости функций плотностей распределения в пространстве *L2*.

Ключевые слова: аппроксимация неизвестных плотностей; экспоненциальное семейство распределений; критерии качества; критерий *Х*и-квадрат; пространство *L2*.

QUALITY ASSESSMENT OF APPROXIMATION DISTRIBUTION DENSITIES BY THE DENSITIES FROM DISTRIBUTION EXPONENTIAL FAMILY

Mikhail Lokhvitsky, *candidate of technical science, associate professor, Moscow technical university of communications and informatics.*

Annotation. The problems of assessing the quality of approximation of unknown distribution densities by densities from the distribution exponential family and poly-Gaussian distributions are considered. As the quality criteria both statistical (Chi-square criterion) and the proximity criteria of the distribution density functions in the space are used.

Keywords: approximation of unknown densities; distribution exponential family; quality criteria; Chi-square criterion; L2 space.

Введение

При приеме сигнала в современных системах телекоммуникаций часто приходится работать с плотностями распределений вероятностей, имеющих нестандартный вид [1, 2]. Это определяется тем, что передаваемая информация обладает различными статистическими свойствами. В общем случае, аналитическое выражение для плотности (в дискретном случае – для вероятности) можно получить по формуле полной вероятности [1]. Как правило, это будет многомодальное распределение. Аналогичные ситуации возникают и при анализе новых физических явлений. Один из способов решения таких задач – это аппроксимация плотностей распределения плотностями из некоторого семейства распределений. В качества такого семейства в [1, 2] предлагается использовать эспоненциальное семейство, а в работах [3, 4] полигауссовые распределения. Плотности распределений из этих семейств в качестве аппроксимирующих обладают рядом полезных свойств.

Выбор экспоненциального семейства в качестве аппроксимирующего связан с тем, что только плотности из этого семейства распределений обладают достаточными статистиками [2], что позволяет существенно сократить число вычислений при обработке выборок большого объема. Кроме этого, плотностями из этого семейства хорошо аппроксимируется широкий класс плотностей. Само экспоненциальное распределение возникает при разложении в ряд логарифма плотности распределения. Плотность распределения из этого семейства имеет вид:

$$p(x\,|\,\vec{\alpha}) = \exp\left\{\sum_{i=1}^{N}\alpha_i\varphi_i\left(x\right) + C\left(\vec{\alpha}\right) + \varphi_0\right\} \qquad , \qquad (1)$$

где: α_i – неопределенные параметры, которые оцениваются на первоначальном этапе работы системы в режиме обучения по выборке;

$\varphi_i, i = \overline{1,N}$ – первые N - базисные функции полной системы функций на пространстве входных реализаций X ;

$C(\vec{\alpha})$ – функция, определяемая из условия нормировки плотности.

$$\begin{aligned}
&\int_X \exp\left(C(\alpha)\right)\cdot\exp\left(\sum_{i=1}^{N}\alpha_i\varphi_i(x)\right)dx = 1\\
&\exp\left(C(\alpha)\right) = \frac{1}{\int_X \exp\left\{\sum_{i=1}^{N}\alpha_i\varphi_i(x)\right\}dx}\quad .\\
&C(\alpha) = -\ln\int_X \exp\left\{\sum_{i=1}^{N}\alpha_i\varphi_i(x)\right\}dx
\end{aligned} \tag{2}$$

Для экспоненциального семейства распределения необходимые и достаточные статистики будут иметь вид:

$$\begin{aligned}
&S_i\left(x[1],\ldots,x[n]\right) = \sum_{k=1}^{n}\varphi\left(x[k]\right)\\
&\left(i = \overline{1,N}\right).
\end{aligned} \tag{3}$$

Если в качестве функции $\varphi_i = x^i$ используются степенные функции, то достаточные статистики имеют вид:

$$\begin{aligned}
&S_i\left(x[1],\ldots,x[n]\right) = \sum_{k=1}^{n}x^i[k]\\
&\left(i = \overline{1,N}\right).
\end{aligned} \tag{4}$$

К недостаткам выбора экспоненциального распределения следует отнести невозможность, в общем случае, получить в явном виде аналитическое выражение для функции нормировки (2).

Оценки параметров плотностей из экспоненциального семейства $\vec{\alpha}$ для больших *n*, используя метод максимального правдоподобия, можно получить из рекуррентных формул [1, 2]:

$$\alpha_i[n] = \alpha_i[n-1] + \frac{1}{n}\sum_{j=1}^{N}b_{ij}\left\{x^i[n] - \frac{1}{n-1}\sum_{k=1}^{n-1}x^i[k]\right\}, i = \overline{1,N}\ . \tag{5}$$

Здесь коэффициенты b_{ij} являются элементами матрицы (6):

$$B = \left\|b_{ij}\right\| = \left\|-\frac{\partial^2 C}{\partial\alpha_i\partial\alpha_j}\left(\vec{\alpha}[n-1]\right)\right\|^{-1}. \tag{6}$$

При малых n оценки параметров $\vec{\alpha}$ определяются по формулам (7). Это связано с тем, что при малых *n* матрица, элементы которой используются в формуле (6), будет вырожденной.

$$\alpha_i[n] = \alpha_i[n-1] + \sum_{j=1}^{N} \gamma_{ij}[n] \left\{ \frac{\partial C}{\partial \alpha_j} \left(\vec{\alpha}[n-1] \right) + x^j[n] \right\}. \qquad (7)$$

Элементы матрицы γ_{ij} для формулы (7) можно определить [1, 2] так:

$$\gamma[n] = \frac{1}{n}\|b_{ij}\| = \begin{Vmatrix} \frac{a}{n} & 0 & \dots & 0 \\ 0 & \frac{a}{n} & \cdots & 0 \\ 0 & 0 & \dots & \frac{a}{n} \end{Vmatrix}, a > 0. \qquad (8)$$

Алгоритм нахождения оценок параметров, используя метод максимального правдоподобия (МММ), показан на рис. 1.

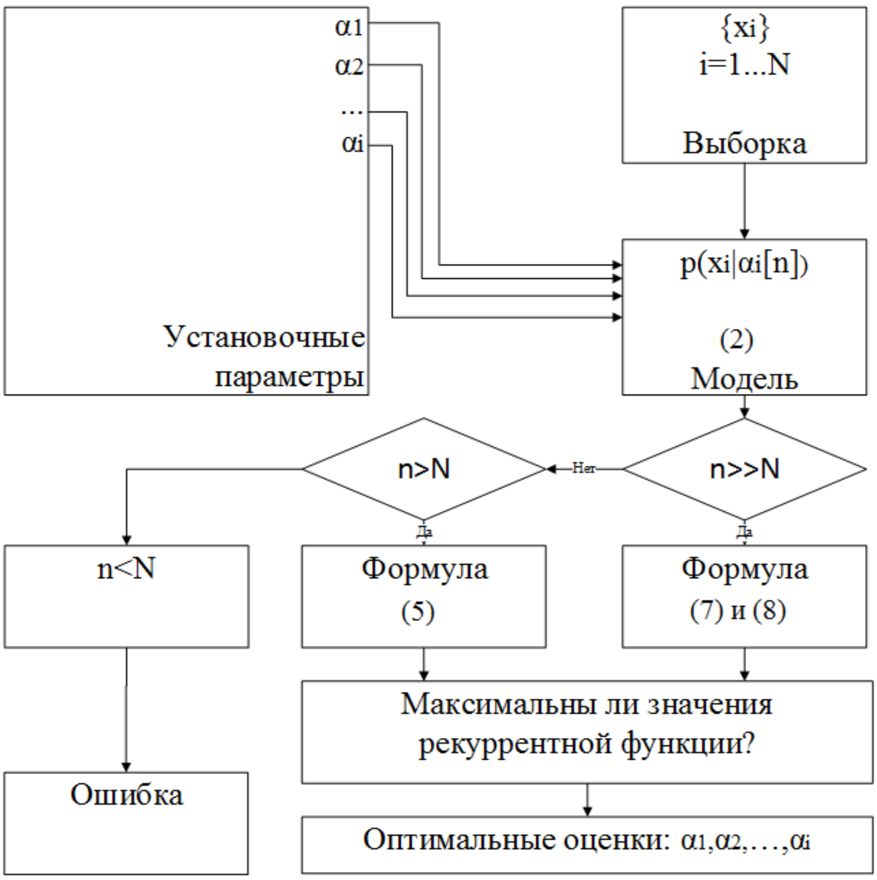

Рисунок 1

Для сравнения различных методов оценивания параметров распределения использовался так же и метод моментов (ММ).

Все выше изложенные идеи были проверены для случая аппроксимации плотностью из экспоненциального семейства смеси двух нормальных законов с различными параметрами (табл. 1, 2). Во всех рассмотренных примерах параметры распределений были: $m_1 = 0, m_2 = 4, \sigma_1^2 = 1, \sigma_2^2 = 1, \alpha = 0{,}5$.

Последний параметр – это весовой коэффициент в смеси плотностей.
Таблица 1

№	Объем выборки	Оценка m_1	Оценка дисперсии1	Оценка m_2	Оценка дисперсии2	Оценка весового коэф.
1	50	0,114	0,845	4,115	0,731	0,501
2	200	0,065	0,944	4,09	0,953	0,509
3	10000	-0,015	1,008	3,982	1,003	0,497

Таблица 2

№	Объем выборки	Оценка m_1	Оценка дисперсии1	Оценка m_2	Оценка дисперсии2	Оценка весового коэф.
1	50	0,171	0,731	3,957	1,176	0,462
2	200	-0,106	0,662	4,021	0,938	0,485
3	10000	0,081	1,142	3,993	1,015	0,503

В задачах аппроксимации плотностей естественно важным является качество аппроксимации. При этом можно и нужно использовать различные критерии качества. Первый класс критериев – это статистические. Самые универсальные и популярные из них – это критерии согласия χ^2.

Основная гипотеза: $H_0:\ p(x) = \hat{p}(x)$. Альтернативная гипотеза : $H_1:\ p(x) \neq \hat{p}(x)$.

Т. е. проверяется гипотеза о том, что статистические данные (выборка) соответствуют полученной аппроксимирующей плотности. Чтобы проверить основную гипотезу будем использовать критерий согласия Фишера. Суть критерия: числовая ось разбивается на l интервалов. Затем вычисляются относительные частоты O_l, вероятности E_i попадания в интервалы и статистика критерия:

$$\chi_н^2 = \sum_{i=1}^{k} \frac{(O_l - E_l)^2}{E_l}. \qquad (9)$$

По таблице χ^2 находится порог, и значение статистики (9) сравнивается с величиной этого порога. Если порог превышен, то гипотеза отвергается. В нашем случае это показывает, что качество аппроксимации неудовлетворительное. В противном случае – нет оснований отвергнуть основную гипотезу.

Второй класс критериев – это критерии, основанные на оценке близости графиков функций. Естественно такие критерии можно использовать только тогда, когда нам известна плотность, которую мы хотим заменить на близкую к ней плотность из экспоненциального семейства. Для нас эти критерии интересны тем, что, используя их, можно оценить эффективность самих методов аппроксимации. Рассмотрим два таких критерия. В первом критерии вычисляется модуль максимального отклонения теоретической плотности от ее оценки:

$$\max\left| p(x) - \hat{p}(x) \right|. \qquad (10)$$

Второй критерий основан на оценке близости функций в пространстве L_2:

$$\int_a^b \left(p(x)-\hat{p}(x)\right)^2 dx \quad . \qquad (11)$$

Рассмотрим качество оценивания для примеров вычисления, приведенных в табл. 1и 2. Начнем с метода максимального правдоподобия (табл. 3), затем рассмотрим метод моментов (табл. 4).

Таблица 3

№	Максимум модуля разности	Отклонение в пространстве L_2	Статистический критерий согласия
1	0,024	0,018	Гипотеза H_0
2	0,009	0,015	Гипотеза H_0
3	0,001	0,0087	Гипотеза H_0

Таблица 4

№	Максимум модуля разности	Отклонение в пространстве L_2	Статистический критерий согласия
1	0,0022	0,019	Гипотеза H_0
2	0,005	0,016	Гипотеза H_0
3	0,002	0,0089	Гипотеза H_0

Заключение

Если параметры аппроксимирующей плотности были хорошо оценены (т.е. близки к теоретическим значениям параметров), то и качество аппроксимации плотностей по всем критериям будет хорошим.

Литература

1. Лохвицкий М.С. Качество аппроксимации плотностей распределения плотностями из экспоненциального семейства распределений // Мобильный бизнес: перспективы развития и реализации систем радиосвязи в России и за рубежом. Сборник материалов 44-ой Международной конференции РАЕН, 2019. – С. 8-10.
2. Шахгильдян В.В., Лохвицкий М.С. Методы адаптивного приема сигналов. – М.: Изд-во: Связь, 1974. – С. 3-158.
3. Застела М.Ю. Основы радиоэлектроники. – М., Юрайт, 2019.
4. Матвеев В.А., Сикарев А.А., Фалько А.И. Подавление сосредоточенных помех. Радиэлектроника, – Т. XX. – № 4. 1977.
5. Лохвицкий М.С., Сорокин А.С., Шорин О.А. Мобильная связь: стандарты, структуры, планирование. – М.: Изд-во Горячая линия - Телеком, 2018. – С. 264.
6. Дынкин Е.Б. Необходимые и достаточные статистики для семейства распределений вероятностей. – «Успехи математических наук», 1951. – Вып. 1. – С. 68-90.
7. Большаков И.А., Гуткин Л.С., Левин Б.Р., Стратонович Р.Л. Математические основы современной радиоэлектроники. – М.: Изд-во «Советское радио», 1972. – С. 447.