

УДК 519.24

А.В. ПОПОВ, И.Н. КОЛЕСНИК

Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Украина

РАЗРАБОТКА МЕТОДА ПОСТРОЕНИЯ НЕГАУССОВСКИХ СТАТИСТИЧЕСКИХ МОДЕЛЕЙ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Построение адекватных статистических моделей выборочных данных, имеющих существенно негауссовское распределение вероятностей, является актуальной задачей экспериментальных исследований. Предложено для описания негауссовских экспериментальных данных использовать систему распределений Джонсона, которая позволяет описывать практически любые унимодальные и широкий класс бимодальных распределений. Предложен метод нахождения оценок параметров законов распределения Джонсона по выборочным данным, основанный на методе моментов и использующий оптимизационные процедуры при нелинейной целевой функции с ограничениями. Эффективность и устойчивость метода исследована путем статистического моделирования. Представлены результаты апробации разработанного метода на реальных данных дистанционного зондирования.

Ключевые слова: статистическая модель, негауссовское распределение, метод моментов, оптимизация.

Введение

Одной из задач экспериментальных исследований является построение статистических моделей измеряемых параметров сигналов, технологических процессов, диагностических параметров, классификационных признаков и т.д. При этом, как правило, измеряемые величины являются многомерными случайными процессами, взаимосвязанными между собой, и могут иметь различные законы распределения вероятностей. Особенно актуальной данная проблема является для многоканальных радиотехнических систем дистанционного зондирования [1], в которых параметры сигналов, несущие информацию об объектах дистанционного наблюдения, во многих случаях являются негауссовскими [2].

В качестве примера на рис. 1, 2 представлены гистограммы поляризационных параметров (коэффициентов асимметрии и анизотропии [3]) радиолокационного объекта – лесного массива, являющихся информативными в задачах экологического мониторинга, оценки состояния лесных ресурсов и распознавания радиолокационных объектов [3]. Построение даже одномерной статистической модели таких данных вызывает затруднения. Гистограмма, представленная на рис. 1, с достаточной степенью точности может быть описана полинормальным распределением [2], а для описания данных, представленных на рис. 2, необходимо использовать составные законы распределения [4], т.е. существующие методы построения статистических моделей существенно негауссовских данных, к сожалению, не обладают универсальностью.

Известен ряд универсальных статистических моделей, таких как система распределений Пирсона и система распределений Джонсона [5], обеспечивающих аппроксимацию практически любых одно-модальных и широкого спектра бимодальных распределений вероятностей. Наиболее привлекательной для использования в качестве универсальной статистической модели является система распределений Джонсона, которая состоит из 3-х распределений, в отличие от системы 12-ти распределений Пирсона. Более того, для распределений Джонсона известна многомерная форма [6], позволяющая описывать векторные данные с различными законами распределения компонент с учетом их корреляции.

Недостатком данной системы распределений является отсутствие прямых методов оценки параметров распределений по выборочным данным. Известные методы [5] их расчета являются итерационными и не всегда обеспечивают сходимость.

Постановка задачи исследования

В данной работе предлагается использование распределений Джонсона в качестве универсальных статистических моделей экспериментальных данных и разрабатывается метод оценки их параметров по выборочным данным на основе известного метода моментов [5, 7]. Задача оценки параметров распределения формулируется в виде оптимизационной задачи с ограничениями и нелинейной целевой функцией.

Целью работы является разработка устойчивого алгоритма оценки параметров распределений Джонсона по выборочным данным, имеющим суще-

ственно негауссовское распределение вероятностей, исследование эффективности метода и его апробация на реальных данных.

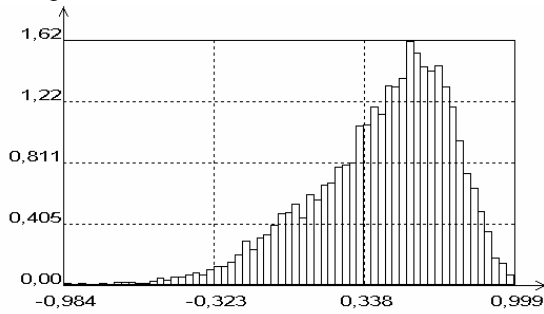


Рис. 1. Гистограмма коэффициента асимметрии объекта дистанционного зондирования (реальные данные)

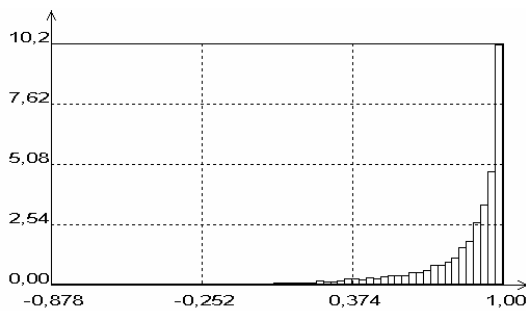


Рис. 2. Гистограмма коэффициента анизотропии объекта дистанционного зондирования (реальные данные)

1. Методология выбора статистической модели экспериментальных данных

Одной из задач статистической обработки данных является построение статистической модели и оценка ее параметров. Для решения данной задачи применяются типовые этапы статистического анализа наблюдений [7], основными из которых являются:

- нахождение выборочного среднего значения (оценка математического ожидания)

$$\hat{x} = \frac{1}{N} \cdot \sum_{i=1}^N x_i, \quad (1)$$

где N – количество отсчетов;

x_i – значения отсчетов, $i = 1, N$;

- оценка выборочного среднеквадратического отклонения

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^N (x_i - \hat{x})^2}{N-1}}; \quad (2)$$

- оценка коэффициента асимметрии распределения

$$\hat{a} = \frac{1}{N} \cdot \frac{\sum_{i=1}^N (x_i - \hat{x})^3}{\hat{\sigma}^3}; \quad (3)$$

- оценка коэффициента эксцесса выборочных данных

$$\hat{e} = \frac{1}{N} \cdot \frac{\sum_{i=1}^N (x_i - \hat{x})^4}{\hat{\sigma}^4}. \quad (4)$$

Два последних коэффициента используются для определения типа закона распределения экспериментальных данных. Если их теоретические значения для различных законов распределения нанести на плоскость в системе координат $\{a, e\}$, как показано на рис. 3, то может быть получена диаграмма, называемая плоскостью моментов [5].

Найденные по выборочным данным оценки коэффициентов асимметрии (3) и эксцесса (4) также наносятся на плоскость моментов. Положение точки, определяемое этими коэффициентами, сравнивается с положением точек, соответствующих известным законам распределения. В качестве статистической модели экспериментальных данных принимается ближайший закон распределения.

2. Распределение Джонсона

Одной из достаточно универсальных статистических моделей является система распределений Джонсона, которая занимает на плоскости моментов большие области (рис. 4.), что позволяет описывать различные законы распределения экспериментальных данных [5], кроме мультимодальных, попадающих в т.н. критическую область (см. рис. 4).

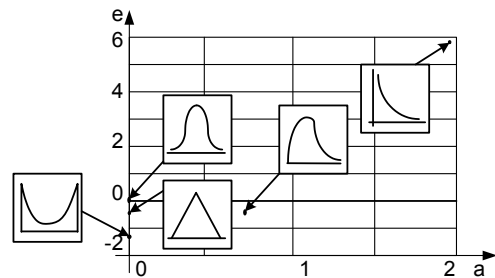


Рис. 3. Положение различных законов распределения на плоскости моментов

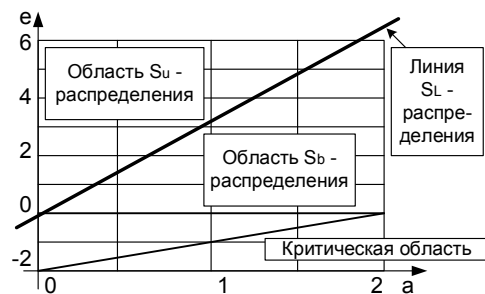


Рис. 4. Положение распределений Джонсона на плоскости моментов

Джонсон предложил находить эмпирические распределения путём нелинейных преобразований:

$$Z = \gamma + \eta \cdot \tau(x, \varepsilon, \lambda), \quad (5)$$

где $\eta > 0$, $\gamma \in (-\infty, +\infty)$, $\varepsilon \in (-\infty, +\infty)$, $\lambda > 0$ – параметры закона распределения Джонсона, обеспечивающие преобразование случайной величины x в случайную величину Z , распределённую по нормальному закону с нулевым средним и единичной дисперсией [5].

Система нелинейных преобразований Джонсона (5) определяет систему законов распределений Джонсона, состоящую в отличие от системы распределений Пирсона, всего из трёх типов распределений S_L , S_B и S_U , и имеет два параметра формы γ и η , один параметр, характеризующий центр распределения ε и один параметр масштаба λ .

S_L – распределение Джонсона основано на преобразовании

$$\tau_1(x, \varepsilon, \lambda) = \ln\left(\frac{x - \varepsilon}{\varepsilon - x + \lambda}\right), \quad (6)$$

плотность распределения случайной величины x имеет вид:

$$f_1(x) = -\frac{\eta}{\sqrt{2\pi(x - \varepsilon)}} \times \exp\left\{-\frac{1}{2} \cdot \eta^2 \left[\frac{\gamma}{\eta} + \ln(x - \varepsilon)\right]^2\right\}, \quad (7)$$

где $x > \varepsilon$, $\eta > 0$, $\gamma \in (-\infty, +\infty)$, $\varepsilon \in (-\infty, +\infty)$.

S_B – распределение Джонсона:

$$\tau_2(x, \varepsilon, \lambda) = \ln\left(\frac{x - \varepsilon}{\lambda + \varepsilon - x}\right), \quad (8)$$

$$f_2(x) = \frac{\eta}{\sqrt{2\pi}} \cdot \frac{\lambda}{(x - \varepsilon) \cdot (\lambda - x + \varepsilon)} \times \exp\left\{-\frac{1}{2} \left[\gamma + \eta \ln\left(\frac{x - \varepsilon}{\lambda + \varepsilon - x}\right)\right]^2\right\}, \quad (9)$$

где $\varepsilon \leq x \leq \varepsilon + \lambda$, $\eta > 0$, $\lambda > 0$.

На рис. 5 представлен вид (9) при различных значениях η , γ .

S_U – распределение:

$$\tau_3(x; \varepsilon, \lambda) = \text{Arsh}\left(\frac{x - \varepsilon}{\lambda}\right), \quad (10)$$

где $x \in (+\infty; -\infty)$;

$$f_3(x) = \frac{\eta}{\sqrt{2\pi}} \frac{1}{\sqrt{(x - \varepsilon)^2 + \lambda^2}} \times \exp\left\{-\frac{1}{2} \left[\gamma + \eta \ln\left\{\left(\frac{x - \varepsilon}{\lambda}\right) + \sqrt{\left(\frac{x - \varepsilon}{\lambda}\right)^2 + 1}\right\}\right]^2\right\}. \quad (11)$$

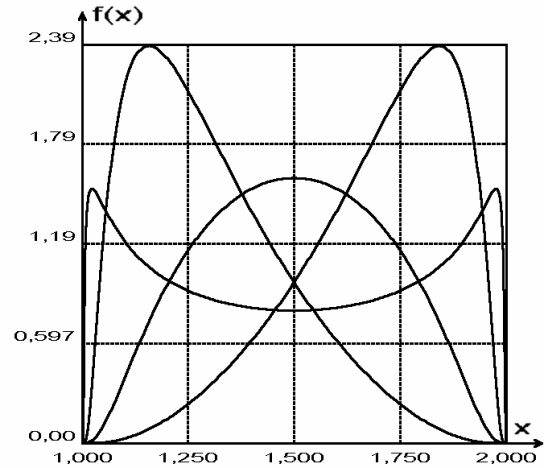


Рис. 5. Форма S_B – распределения Джонсона при различных значениях η и γ

Известным методом нахождения оценок параметров ε , λ , η , γ распределений (7), (9), (11) является метод процентилей, рассмотренный в [5].

Процентиль x_α определяется решением интегрального уравнения

$$\alpha = \int_{-\infty}^{x_\alpha} f(x; \varepsilon, \lambda, \eta, \gamma) dx \quad (12)$$

при заданной вероятности α . Для нахождения четырех параметров задаются 4 значения α . Для них вычисляются процентиля нормированного нормального распределения z , которые связаны с эмпирическими процентилями \hat{x} соотношениями (6), (8) или (10). За счет приравнивания соответствующих процентилей получают четыре уравнения вида

$$z = \gamma + \eta \cdot \tau(\hat{x}, \varepsilon, \lambda), \quad (13)$$

решение которых дает оценки искомых параметров.

Уравнения (13) нелинейны и должны решаться численными методами. Поскольку процентиля в (13) должны определяться из решения уравнения (12), алгоритм решения системы уравнений (13) получается достаточно сложным и, как показали проведенные исследования, не всегда обеспечивает сходимость.

3. Метод моментов

Для определения параметров распределений Джонсона (7), (9), (11) можно использовать метод моментов [7], который основывается на нахождении следующих числовых характеристик случайной величины x по ее плотности распределения $f(x)$:

– момент первого порядка – математическое ожидание:

$$m_x = \int_{-\infty}^{\infty} x \cdot f(x) dx ; \quad (14)$$

– момент второго порядка – дисперсия:

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - m_x)^2 \cdot f(x) dx ; \quad (15)$$

– коэффициент асимметрии:

$$a_x = \int_{-\infty}^{\infty} (x - m_x)^3 \cdot f(x) dx ; \quad (16)$$

– коэффициент эксцесса:

$$e_x = \int_{-\infty}^{\infty} (x - m_x)^4 \cdot f(x) dx . \quad (17)$$

При использовании этого метода получается система уравнений, которая имеет следующий вид:

$$\begin{cases} \int_{-\infty}^{\infty} x \cdot f(x; \varepsilon, \lambda, \eta, \gamma) dx = \hat{x} \\ \int_{-\infty}^{\infty} (x - m_x)^2 \cdot f(x; \varepsilon, \lambda, \eta, \gamma) dx = \hat{\sigma}^2 \\ \int_{-\infty}^{\infty} (x - m_x)^3 \cdot f(x; \varepsilon, \lambda, \eta, \gamma) dx = \hat{a} \\ \int_{-\infty}^{\infty} (x - m_x)^4 \cdot f(x; \varepsilon, \lambda, \eta, \gamma) dx = \hat{e}, \end{cases} \quad (18)$$

где \hat{x} , $\hat{\sigma}$, \hat{a} , \hat{e} – выборочные оценки, которые находятся по экспериментальным данным согласно (1) – (4).

Моменты распределений Джонсона (14) – (17) существенно зависят от всех четырех параметров плотности распределения $f(x; \varepsilon, \lambda, \eta, \gamma)$. Вид этих зависимостей представлен на рис. 6 – 9.

Анализ рис. 6 – 9 показывает, что система уравнений (18) так же является нелинейной, однако может быть однозначно решена относительно искомых параметров ε , λ , η , γ численными методами, поскольку не содержит переменных пределов интегрирования.

4. Предлагаемый метод нахождения оценок параметров закона распределения Джонсона

Задачу нахождения параметров закона распределения Джонсона путем решения системы уравнений (18) можно сформулировать как оптимизационную. Поскольку в точке $(\varepsilon, \lambda, \eta, \gamma)$, соответствующей решению системы уравнений (18), правые части уравнений должны быть равны левым, то целевая

функция, построенная по методу наименьших квадратов [8], будет иметь вид

$$Q(\varepsilon, \lambda, \eta, \gamma) = (m_x - \hat{x})^2 + (\sigma_x - \hat{\sigma})^2 + (a_x - \hat{a})^2 + (e_x - \hat{e})^2 \quad (19)$$

и должна быть минимальна в искомой точке в пространстве варьируемых переменных $(\varepsilon, \lambda, \eta, \gamma)$, т.е. целевая функция (19) формулирует оптимизационную задачу на минимум.

Для параметров распределения рассматриваемого закона имеется ряд ограничений, которые необходимо учесть в целевой функции (19). В частности $\lambda > 0$, $\eta > 0$. Данные ограничивающие неравенства могут быть учтены с помощью штрафной функции, которая имеет вид:

$$\psi(\lambda, \eta) = \psi_\lambda(\lambda) + \psi_\eta(\eta), \quad (20)$$

где

$$\psi_\lambda(\lambda) = \begin{cases} 0, \lambda \geq 0 \\ |\lambda| \cdot K_1 + K_0, \lambda \leq 0, \end{cases}$$

$$\psi_\eta(\eta) = \begin{cases} 0, \eta \geq 0 \\ |\eta| \cdot K_1 + K_0, \eta \leq 0. \end{cases}$$

Если варьируемый параметр находится в пределах области допустимых значений, то штрафная функция (20) равна нулю, иначе ее значение больше 0 и растет пропорционально отклонению за допустимые пределы, что определяется коэффициентом K_1 . Поскольку для параметров λ , η значение нуль также не разрешено, то в штрафную функцию вводится линейное смещение K_0 .

Результирующая целевая функция имеет вид:

$$Q(\varepsilon, \lambda, \eta, \gamma) = (m(\varepsilon, \lambda, \eta, \gamma) - \hat{x})^2 + (\sigma(\varepsilon, \lambda, \eta, \gamma) - \hat{\sigma})^2 + (a(\varepsilon, \lambda, \eta, \gamma) - \hat{a})^2 + (e(\varepsilon, \lambda, \eta, \gamma) - \hat{e})^2 + \psi(\lambda, \eta). \quad (21)$$

Анализ целевой функции (21) показывает, что в окрестности минимума она является гладкой и унимодальной, как показано на рис. 10, но имеет изломы за счет ограничений на параметры λ , η . В связи с этим для решения оптимизационной задачи был выбран метод Розенброка [8], обеспечивающий минимизацию унимодальных функций с разрывами производных.

Таким образом, по выборочным данным найдутся оценки значений \hat{m} , $\hat{\sigma}$, \hat{a} , \hat{e} согласно (1) – (4), которые используются в качестве исходных данных в целевой функции вида (21), минимум которой соответствует решению системы уравнений (18).

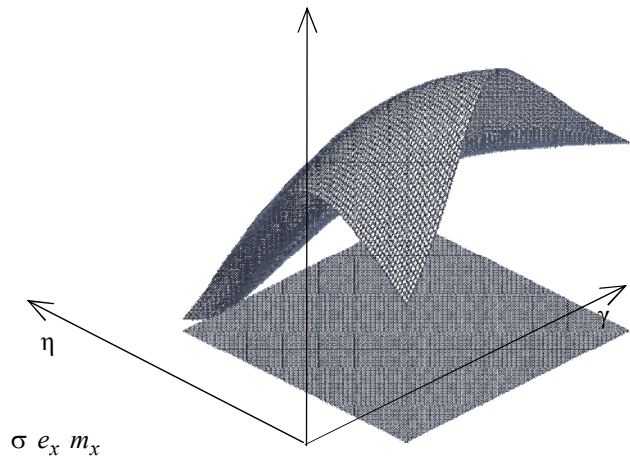


Рис. 6. Зависимость математического ожидания Sb распределения Джонсона от параметров η, γ при фиксированных ε, λ

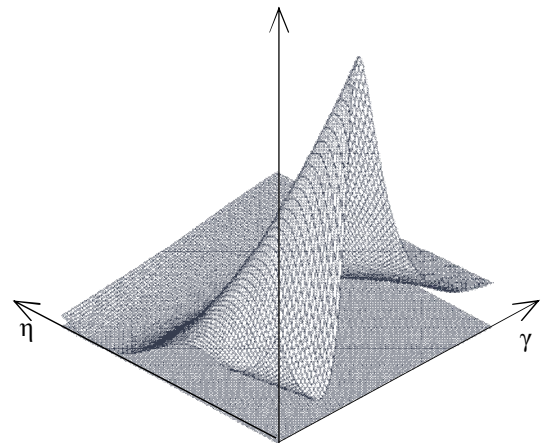


Рис. 9. Зависимость эксцесса Sb распределения Джонсона от параметров η, γ при фиксированных ε, λ

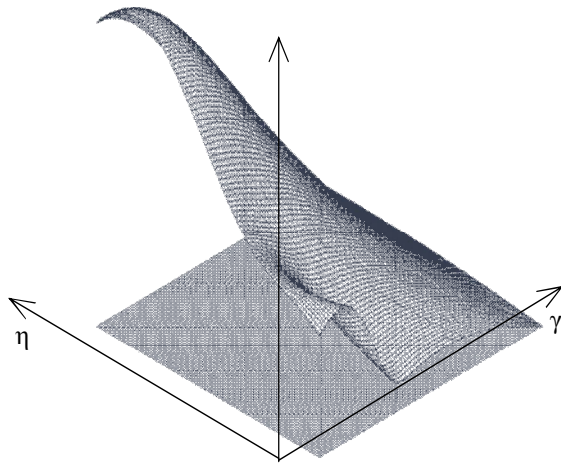


Рис. 7. Зависимость среднеквадратического отклонения Sb распределения Джонсона от параметров η, γ при фиксированных ε, λ

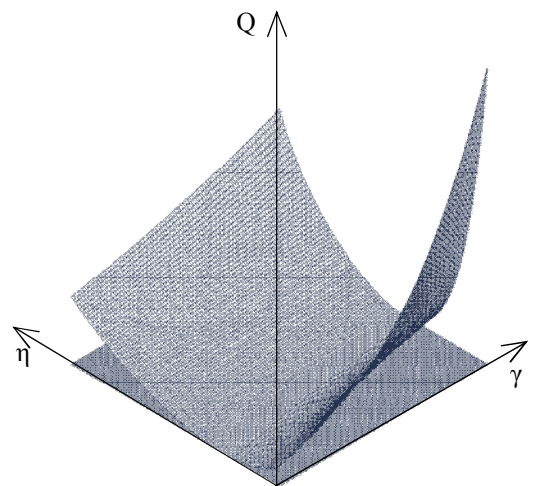


Рис. 10. Графическое представление целевой функции

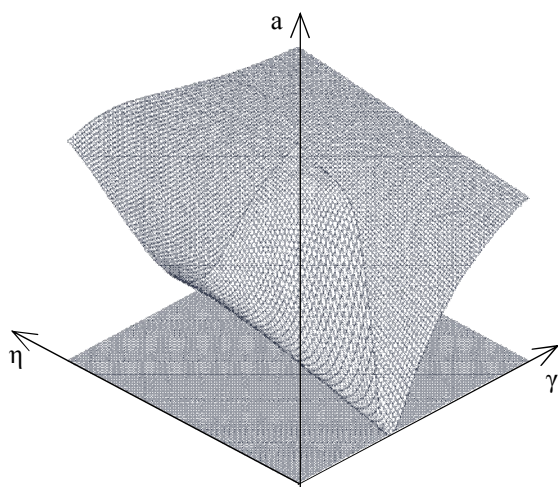


Рис. 8. Зависимость асимметрии Sb распределения Джонсона от параметров η, γ при фиксированных ε, λ

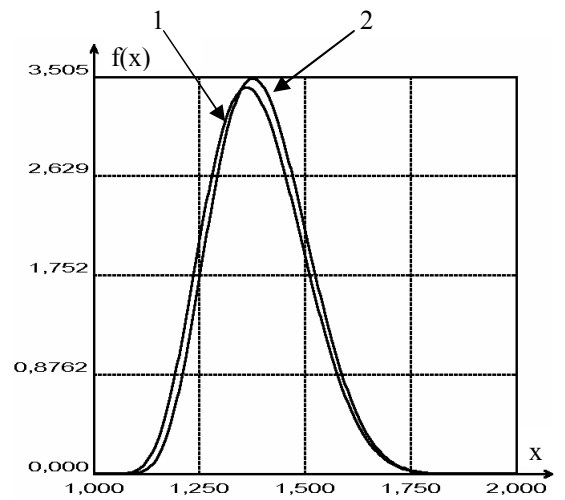


Рис. 11. Закон распределения Джонсона: 1 – с тестовыми значениями параметров; 2 – с найденными значениями параметров

5. Анализ эффективности предлагаемого метода

Проверка разработанного метода проводилась с использованием тестовых данных при различных известных значениях параметров $\varepsilon, \lambda, \eta, \gamma$ S_B – распределения Джонсона (9). Так, например, для тестовой выборки с параметрами $\varepsilon=1, \lambda=1, \eta=1, \gamma=1$ были найдены согласно (1) – (4) значения $\hat{x}=1,384, \hat{\sigma}=0,112, \hat{a}=0,287, \hat{e}=-0,225$, которые использовались в качестве эталонных данных для целевой функции. После проведения оптимизации были получены значения $\varepsilon=1,05, \lambda=0,98, \eta=1,07, \gamma=0,92$. Расчет моментов распределения (9) согласно (14) – (17) при полученных значениях искомых параметров дает значения $m_x=1,384, \sigma_x=0,112, a_x=0,264, e_x=-0,346$. Различия между исходными и полученными данными объясняются пологостью целевой функции в точке минимума (рис. 10), что требует высокой точности вычислений, повышение которой существенно увеличивает количество итераций, необходимых для решения системы уравнений (18). Так, например, при точности нахождения параметров $\varepsilon, \lambda, \eta, \gamma 10^{-3}$ требуется 20...50 итераций, а при точности $10^{-6} - 9...15$ тысяч.

Графики закона распределения Джонсона с тестовыми значениями и полученными в результате оценки параметров по выборочным данным практически совпадают и представлены на рис. 11.

На рис. 12. представлены результаты обработки экспериментальных данных (коэффициент асимметрии [3] объекта дистанционного зондирования – лесного массива), гистограмма которых приведена на рис. 1. Получена статистическая модель вида (9) с параметрами: $\varepsilon=-0,999; \lambda=1,98; \eta=1,29; \gamma=-1,35$. Как следует из рис. 12, полученную аппроксимацию гистограммы экспериментальных данных можно считать вполне адекватной. Для построения статистической модели потребовалось 527 итераций.

На рис. 13 представлены результаты построения статистической модели коэффициента анизотропии [3] объекта дистанционного зондирования (лесного массива) по реальным радиолокационным данным. Гистограмма, показанная на рис. 2, не аппроксимируется ни одним из известных авторам законом распределения. Как следует из рис. 13, S_B – распределение Джонсона с параметрами $\varepsilon=-0,74; \lambda=1,7; \eta=0,529; \gamma=-1,55$ с высокой точностью описывает данный закон распределения, для построения модели которого потребовалось всего 56 итераций.

Разработанный метод применялся для построения статистических моделей экспериментальных

данных, полученных с помощью радиолокационного поляриметра [3]. При обработке более 2000 выборок данных в 90% случаев аппроксимация гистограмм была признана адекватной. Опыт применения разработанного метода показывает, что построение статистической модели будет успешным, если закон распределения экспериментальных данных не попадает в критическую область (см. рис. 4).

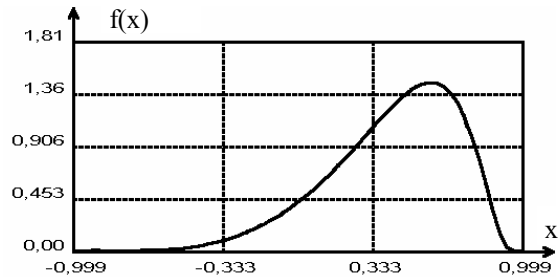


Рис. 12. Статистическая модель экспериментальных данных, представленных на рис. 1.

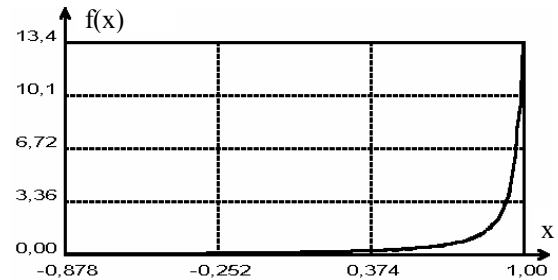


Рис. 13. Статистическая модель экспериментальных данных, представленных на рис. 2.

Заключение

Построение статистических моделей измеряемых параметров сигналов, технологических процессов, диагностических параметров, классификационных признаков и т.д. является одной из важнейших задач экспериментальных исследований, т.к. позволяет обобщить полученные опытным путем данные. В случае негауссовских законов распределения данных использование разнородных статистических моделей усложняет их дальнейшую обработку.

Применение распределений Джонсона позволяет получать универсальное вероятностное описание практически любых унимодальных и широкого спектра бимодальных распределений. Предложен метод нахождения оценок параметров законов распределения Джонсона по выборочным данным, основанный на методе моментов и использующий оптимизационную процедуру для нелинейной целевой функции с ограничениями.

Исследования методом статистического моделирования показали эффективность и устойчивость предложенного метода. Результаты апробации раз-

рабочего метода на реальных данных дистанционного зондирования подтверждают его эффективность для построения статистических моделей экспериментальных данных с существенно негауссовскими законами распределения.

Литература

1. Радиолокационные методы и средства оперативного дистанционного зондирования Земли с аэрокосмических носителей / Под ред. С.Н. Конюхова, В.И. Драновского, В.Н. Цимбала. – К.: НАНУ, 2007. – 440 с.

2. Popov A.V. Image clustering algorithm using polynormal distribution / A.V. Popov, O.B. Pogrebnyak, A.N. Brashevan // *Mathematical Methods in Pattern and Image Analysis: Proc. SPIE*. – 2005. – V. 5916. – P. 341-349.

3. Бабаков М.Ф. Применение поляризационно-модулированных сигналов для селекции и распознавания радиолокационных объектов / М.Ф. Бабаков, А.В. Попов // *Зарубежная радиоэлектроника. Успехи современной радиоэлектроники*. – 1999. – № 11. – С. 42–45.

4. Storvik B. On the Combination of Multisensor Data Using Meta-Gaussian Distributions / B. Storvik, G. Storvik, R. Fjortoft // *IEEE Transactions on Geoscience and Remote Sensing*, 2009. – V. 47, Issue 7. – P. 2372 - 2379.

5. Хан Г. Статистические модели в инженерных задачах: пер. с англ. / Г. Хан, С. Шапиро. – М.: Мир, 1969. – 395 с.

6. Бабаков М.Ф. Об одном способе аппроксимации распределений многомерных поляриметрических характеристик / М.Ф. Бабаков // *Автоматизированные системы управления: сб. науч. тр. Харьк. авиац. ин-та им. Н.Е. Жуковского*. – Вып. 3. – X., 1981. – С. 166-167.

7. Бендат Дж. Прикладной анализ случайных данных: пер. с англ. / Дж. Бендат, А. Пирсол. – М.: Мир, 1989. – 540 с.

8. Шуп Т. Решение инженерных задач на ЭВМ: практическое руководство: пер. с англ. / Т. Шуп. – М.: Мир, 1982. – 238 с.

Поступила в редакцию 28.08.2009

Рецензент: д-р техн. наук, проф., проф. кафедры производства радиоэлектронных систем Г.Я. Красовский, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина.

РОЗРОБКА МЕТОДУ ПОБУДОВИ НЕГАУСОВСЬКИХ СТАТИСТИЧНИХ МОДЕЛЕЙ ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ

А.В. Попов, І.М. Колесник

Побудова адекватних статистичних моделей вибіркових даних, що мають істотно негаусовський розподіл імовірностей, є актуальною задачею експериментальних досліджень. Запропоновано для опису негаусовських експериментальних даних використовувати систему розподілів Джонсона, що дозволяє описувати практично будь-які унімодальні та широкий клас бімодальних розподілів. Запропоновано метод отримання оцінок параметрів законів розподілу Джонсона по вибіркоvim даним, заснований на методі моментів і що використовує оптимізаційні процедури при нелінійній цільовій функції з обмеженнями. Ефективність і стабільність методу досліджена шляхом статистичного моделювання. Представлено результати апробації розробленого методу на реальних даних дистанційного зондування.

Ключові слова: статистична модель, негаусовський розподіл, метод моментів, оптимізація.

THE DEVELOPMENT OF A METHOD FOR NON-GAUSSIAN STATISTICAL MODELS CONSTRUCTING IN APPLICATION TO EXPERIMENTAL DATA

A.V. Popov, I.N. Kolesnik

Constructing of adequate statistical models for sampled data that have substantially non-Gaussian probability distributions is a problem of current importance in experimental researches. For non-Gaussian experimental data description it is suggested to use the system of Johnson's distributions that allows to describe almost every unimodal and a broad class of bimodal distributions. A method for distribution law parameters estimation that is based on the method of moments and uses optimization techniques for non-linear goal function with limitations is suggested. The efficiency and the stability of the method are investigated using statistical modeling. Approbation results of the suggested method on real remote sensing data are presented.

Key words: statistical model, non-Gaussian distribution, method of moments, optimization.

Попов Анатолий Владиславович – канд. техн. наук, доцент, доцент кафедры производства радиоэлектронных систем, Национальный аэрокосмический университет им. Н.Е. Жуковского «Харьковский авиационный институт», Харьков, Украина, E-mail: a.v.popov@inbox.ru.

Колесник Инна Николаевна – студентка кафедры производства радиоэлектронных систем, Национальный аэрокосмический университет им. Н.Е. Жуковского «Харьковский авиационный институт», Харьков, Украина.