УДК 65.011.56

А.Е. Примак, А.Г. Шумихин, С.И. Сташков

Пермский национальный исследовательский
политехнический университет

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ СПЕКТРАЛЬНОГО АНАЛИЗА В ОБУЧАЮЩЕЙ ВЫБОРКЕ ДЛЯ СОЗДАНИЯ МОДЕЛЕЙ ДЛЯ ПОТОЧНОГО АНАЛИЗАТОРА СВЕТЛЫХ НЕФТЕПРОДУКТОВ

Приведены результаты статистического анализа, на основе метода главных компонент, для снижения уровня шума обучающей выборки математических моделей, используемых в спектральном анализе светлых нефтепродуктов.

Переход на новые стандарты в химической и нефтеперерабатывающей промышленности, который наблюдается в настоящее время, ведет к ужесточению требований к значениям показателей качества продуктов и их полуфабрикатов.

Сложность измерения качества нефтепродуктов обусловлена многими причинами, ряд которых требует применения математических моделей связи значений показателей качества со значениями естественных сигналов поточных анализаторов.

Еще до начала выбора математических моделей необходимо убедиться, что набор данных содержит достаточное количество информации о физико-химических свойствах смеси, которые в конечном итоге и определяют качество продукции. Использование спектрального анализа, хотя и является относительно быстрым и недорогим способом получения данных, порождает несколько проблем. К ним можно отнести сложность оборудования для проведения анализа, требовательного к точной калибровке, а также большой объем избыточных данных, не связанных с интересующим исследователя показателем качества. В первом случае существует вероятность появления инструментальной погрешности, приводящей к получению обучающей выборки, представители которой не отражают реальные значения. Модели, обученные на такой выборке, будут обладать неснижаемой систематической погреш-

ностью. Компенсация такой погрешности усложняет алгоритмы измерения показателей качества.

Избыточность данных ведет к значительному зашумлению данных обучающей выборки и создает дополнительные сложности для выявления ошибок при измерении и сбоях оборудования. Поэтому для получения обучающей выборки, пригодной для создания математических моделей, необходимо до начала процедуры обучения провести контроль данных. Для решения этой задачи нами разработана методика, опирающаяся на совокупность методов обработки и анализа данных.

При выборе метода анализа исходят из того, что набор данных является многомерным, в данном случае – двухмерным. Необходимо учесть, что набор данных содержит большое количество избыточной информации, может содержать ошибки измерения анализатора и лабораторного контроля. При подготовке моделей дополнительные трудности создает задача выделения из общего ансамбля данных, связанных с необходимым показателем качества. При спектральном анализе связь между измеренными значениями и показателем качества является нелинейной. Поэтому метод отбора должен обладать возможностью понижать размерность массива данных при сохранении его информативности, а также выявлять скрытые структуры в данных, несущие необходимую информацию. С учетом этих условий был рассмотрен метод главных компонент (МГК), а при анализе использованы опирающиеся на него алгоритмы.

При применении метода главных компонент данные записываются в виде матрицы $\boldsymbol{X}$ – прямоугольной таблицы чисел с размерностью $i \times j$:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1j} \\ x_{21} & x_{22} & \ldots & x_{2j} \\ \vdots & \vdots & & \vdots \\ x_{i1} & x_{i2} & \ldots & x_{ij} \end{pmatrix} \qquad (1)$$

Строки данной матрицы называются образцами. Они нумеруются индексом i, меняющимся от 1 до I. Столбцы называются переменными и нумеруются индексом $j = 1, \ldots, J$.

Цель МГК – извлечение из этих данных нужной информации, определяемой сутью решаемой задачи. Данные могут содержать и избыточную информацию или же не содержать полезную информацию. Данные всегда (или почти всегда) содержат в себе нежелательную бес-

полезную составляющую, называемую шумом, природа которого может быть различной. Что считать шумом, а что полезной информацией, решается с учетом поставленной цели и методов, используемых для ее достижения.

Шум и избыточность в данных проявляют себя через корреляционные связи между переменными. Погрешности в данных приводят к случайным составляющим в связях между переменными. Понятие скрытых, латентных переменных является важнейшим понятием в МГК [1].

Перед применением МГК осуществляется центрирование и нормирование набора данных. Центрирование переменных – это вычитание из каждого элемента столбца среднего по столбцу значения. Центрирование позволяет исключить из модели МГК свободный член. Нормирование выравнивает вклад разных переменных в МГК модель. При этом преобразовании каждый элемент столбца делится на свое стандартное отклонение, т.е.

$$\overset{o}{x}_{ij} = \frac{x_{ij} - m_j}{S_j}; \tag{2}$$

$$m_j = \frac{1}{I}\sum_{i=1}^{I} x_{ij}; \tag{3}$$

$$S_j = \frac{1}{I-1}\sum_{i=1}^{I}\left(x_{ij} - m_j\right); \tag{4}$$

$$x_{ij} := \overset{o}{x}_{ij},$$

где $\overset{o}{x}_{ij}$ – нормированное значение x_{ij}; m_j – среднее значение для j-го столбца; S_j – среднеквадратичное отклонение (стандарт j-го столбца).

После нормирования можно приступать к процедуре разделения данных на «информацию» и «шум». Для усиления влияния переменных, коррелированных с измеренными значениями, используется метод проекции на латентные структуры (ПЛС). При использовании ПЛС производится совместная декомпозиция матрицы $\boldsymbol{X}$ переменных и вектора $\boldsymbol{Y}$ измеренных значений таким образом, что вклад переменных, коррелированных с измеренными значениями, увеличивается, а некоррелированных уменьшается, что позволяет дополнительно понизить их влияние на модель.

При определении размерности модели нами используется анализ квадрата матрицы дисперсий-ковариаций [2]

$$\boldsymbol{C} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X} \tag{5}$$

на наибольшее собственное значение

$$\lambda = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}w = \boldsymbol{t}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{t},$$

где $\boldsymbol{Y}$ – матрица (вектор) откликов; $\boldsymbol{w}$ – матрица взвешенных нагрузок; $\boldsymbol{\lambda}$ – собственные значения; $\boldsymbol{t}$ – матрица счетов.

Зависимость значений λ и его верхнего доверительного предела от размерности модели (числа ГК) представлена на рис. 1. Верхний доверительный предел рассчитывался по формуле

$$E(\lambda) \cong \frac{\mathrm{tr}(t^{\mathrm{T}}t)\cdot \mathrm{tr}(Y^{\mathrm{T}}Y)}{N} + 1{,}645\left(\frac{2\left(t^{\mathrm{T}}t\right)^2 \cdot \mathrm{tr}\left(\left[Y^{\mathrm{T}}Y\right]^2\right)}{N^2}\right)^{\frac{1}{2}}, \tag{6}$$

где tr (.) – след матрицы (.).

На рис. 1 линии пересекаются при размерности, равной 4. Это означает, что оптимальной для построения модели будет размерность, равная 3.

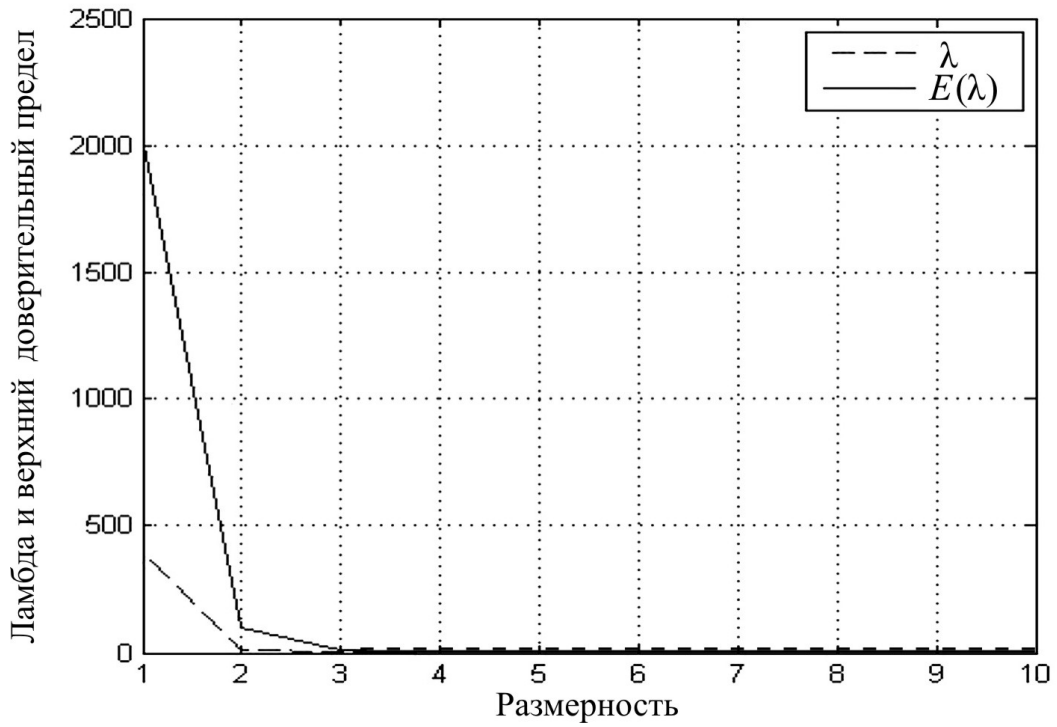

Рис. 1. Зависимость значений λ и его верхнего доверительного предела от числа ГК: – – – график λ; —— график E (λ)

Приведенный выше расчет соответствует модели для ИК-спектрофотометра по показателю качества бензина «октановое число (моторный метод)».

Определив размерность модели, необходимо очистить данные от переменных, не имеющих связи с показателем качества.

Для определения значащих переменных был использован анализ диагональных элементов квадрата матрицы дисперсий-ковариаций (5). На рис. 2 представлена зависимость значений диагональных элементов матрицы ***C*** от длины волны для показателя «октановое число (исследовательский метод)».

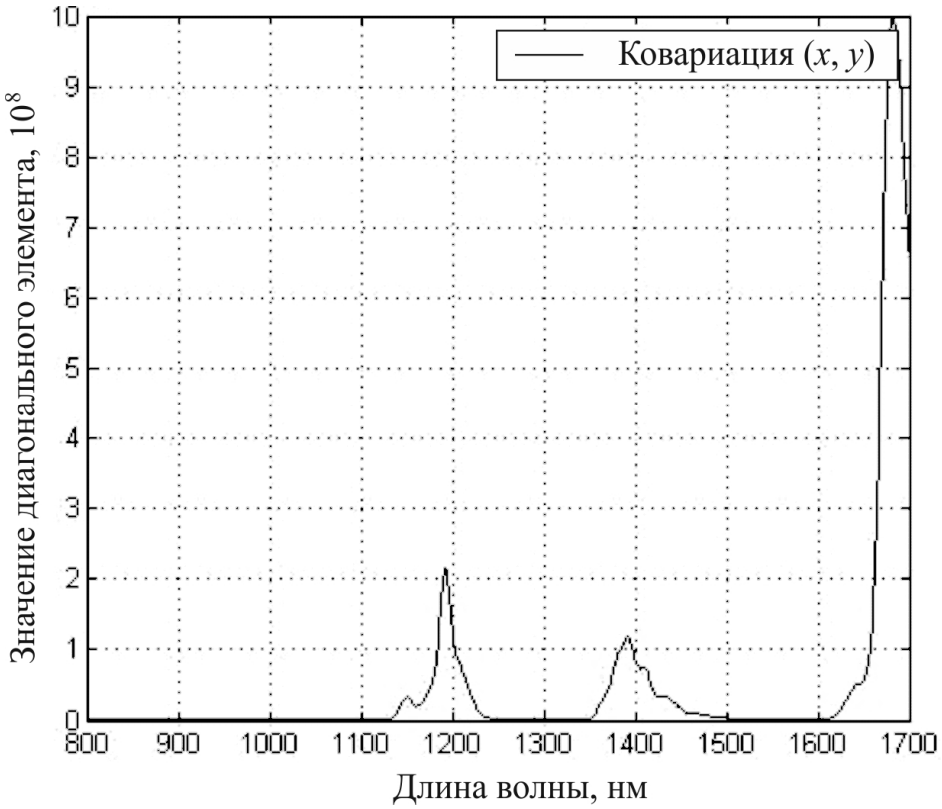

Рис. 2. Зависимость значений диагональных элементов матрицы ***C*** от длины волны для показателя «октановое число (исследовательский метод)»

На рис. 2 видно, что большая часть переменных не связана с показателями качества. Поэтому при построении модели их следует удалить из общей выборки.

Оставшиеся данные анализируются в пространстве главных компонент. Для рассмотрения структуры данных строятся двумерные проекции переменных в пространстве главных компонент, так называемый график счетов [3].

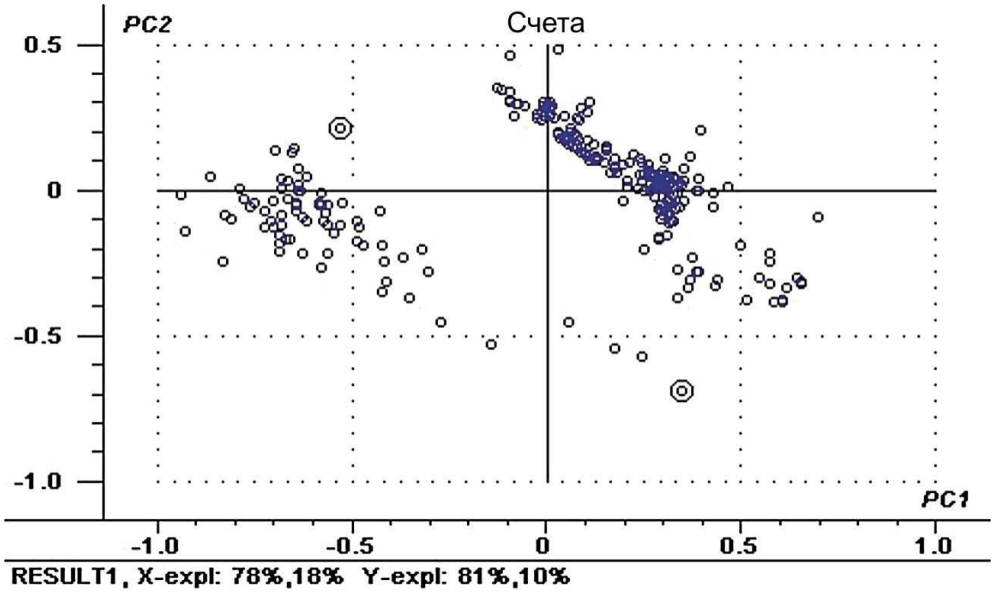

Рис. 3. График счетов для показателя
«октановое число (исследовательский метод)»

Анализ графика счетов позволяет на первоначальном этапе выделить образцы, нетипичные для данной выборки и требующие дополнительной проверки.

Дополнительно к графику счетов строится график «влияние – полнота описания», представленный на рис. 4. Он позволяет оценить степень влияния и полноту описания каждого образца полученной моделью. Вдоль оси абсцисс откладывается размах – вклад этого образца

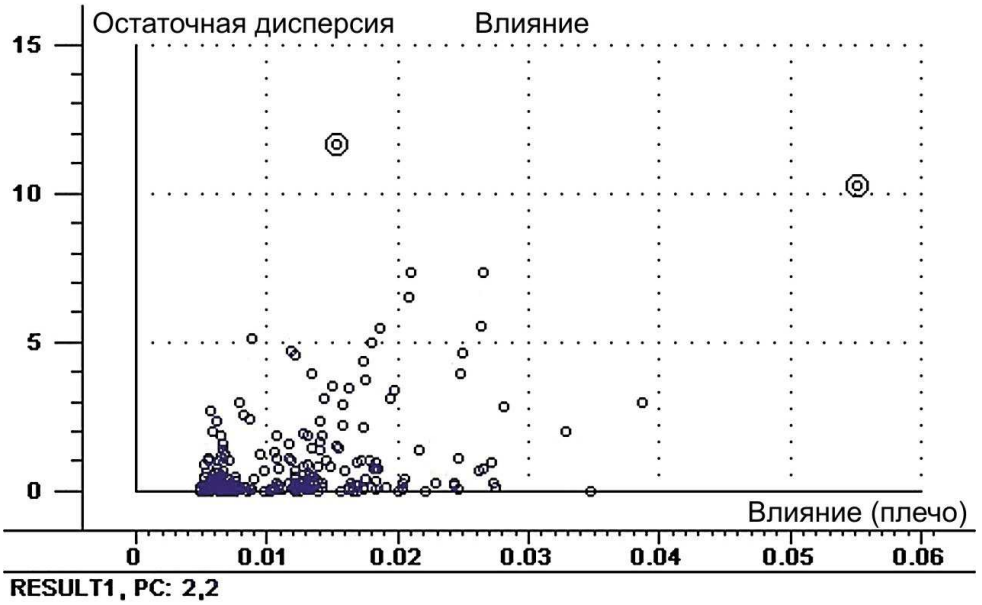

Рис. 4. График «влияние – полнота описания» для показателя
«октановое число (исследовательский метод)»

в общую модель. Вдоль оси ординат откладываются значения остаточной дисперсии, и чем больше их значения, тем хуже данный образец описывается моделью. Таким образом, образцы, расположенные в верхнем правом углу графика влияний, можно сразу исключать из обучающей выборки, поскольку они плохо описываются моделью и имеют на нее большое влияние.

Заключительным этапом обработки является оценка переменных по критериям максимально и минимально возможного.

Можно полагать, что результаты измерений $\boldsymbol{X}$ распределены по нормальному закону с генеральными параметрами m_x и σ_x^2. Выборка имеет объем n, а подозрительный выброс в данной серии измерений равен $x_{\max}$.

Пусть гипотеза H_0 является предположением, что значение $x_{\max}$ принадлежит к той же генеральной совокупности, что и другие $n-1$ измерения этой серии, т.е. $x_{\max}$ не является грубой ошибкой. Альтернативная сложная гипотеза H_1 принимается, а H_0 отклоняется, если при сравнении $x_{\max}$ с некоторым критическим значением $x = \nu_n(q)$ значение $x_{\max}$ попадает в критическое множество при заранее заданном уровне значимости q.

При выборке конечного объема n вместо теоретических параметров m_x и σ_x^2 можно вычислить лишь соответствующие выборочные оценки $\overline{x}$ и S_x^2. Тогда граничные значения x можно записать в виде

$$x_{\max} = \overline{x} + \nu_n(q) S_x, \quad x_{\min} = \overline{x} - \nu_n(q) S_x. \tag{7}$$

Значение $\nu_n(q)$ для уровня значимости q и числа измерений n можно найти из табл. 1 квантилей распределения величины $\nu = (x_{\max} - \overline{x}) / S_x$ или $\nu = (\overline{x} - x_{\min}) / S_x$.

Эти данные получены из исследования вероятности

$$\text{Вер}\{(x_{\max} - \overline{x}) / S \prec \nu\},$$

где $x_{\max}$, $\overline{x}$ и S определены по выборке объема n из нормальной совокупности.

Зависимость $\nu(n)$ для уровня значимости $q = 0{,}1$, построенная по данным табл. 1, представлена на рис. 5.

Таблица 1

Квантили распределения величины $\nu = (x_{\max} - \bar{x}) / S_x$ или $\nu = (\bar{x} - x_{\min}) / S_x$

n	q				n	q			
	0,10	0,05	0,025	0,01		0,10	0,05	0,025	0,01
3	1,406	1,412	1,414	1,414	15	2,326	2,493	2,638	2,800
4	1,645	1,689	1,710	1,723	16	2,354	2,523	2,670	2,837
5	1,791	1,869	1,917	1,955	17	2,380	2,551	2,701	2,871
6	1,894	1,996	2,067	2,130	18	2,404	2,577	2,728	2,903
7	1,974	2,093	2,182	2,265	19	2,426	2,600	2,754	2,932
8	2,041	2,172	2,273	2,374	20	2,447	2,623	2,778	2,959
9	2,097	2,237	2,349	2,464	21	2,467	2,644	2,801	2,984
10	2,146	2,294	2,414	2,540	22	2,486	2,664	2,823	3,008
11	2,190	2,343	2,470	2,606	23	2,504	2,683	2,843	3,030
12	2,229	2,387	2,519	2,663	24	2,520	2,701	2,862	3,051
13	2,264	2,426	2,562	2,714	25	2,537	2,717	2,880	3,071
14	2,297	2,461	2,602	2,759	…	…	…	…	…

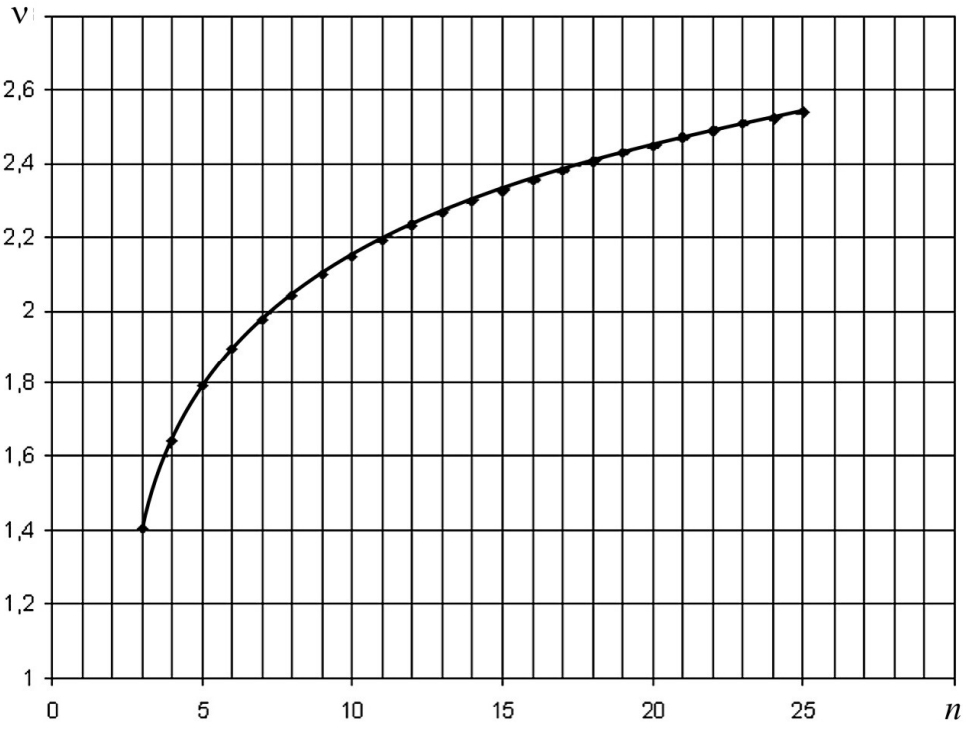

Рис. 5. Зависимость ν (n) для q = 0,1

Экстраполяция данных табл. 1 для интервала значений $n = \overline{15; 25}$ на значения $n = \overline{26; 250}$ осуществлена для различных уровней значимости q с помощью рекуррентной формулы

$$\nu_n(q) = \nu_{n-1}(q) + \frac{1}{2{,}35n}, \qquad (8)$$

где первое значение $\nu_{n-1}(q)$ для различных уровней значимости соответствует значениям, приведенным в табл. 1 при $n = 15$, т.е. $\nu_{n-1}(q) = \nu_{15}(q)$.

В табл. 2 представлен пример квантилей распределения величины $\nu = (x_{\max} - \overline{x}) / S_x$ или $\nu = (\overline{x} - x_{\min}) / S_x$ для $q = 0{,}1$ на интервале значений $n = \overline{3; 250}$.

Таблица 2

Фрагмент квантилей распределения величины $\nu = (x_{\max} - \overline{x}) / S_x$ или $\nu = (\overline{x} - x_{\min}) / S_x$ для $n = \overline{3; 250}$

n	ν	n	ν	n	ν	n	ν
3	1,406	65	3,217	127	4,271	189	5,325
4	1,645	66	3,234	128	4,288	190	5,342
5	1,791	67	3,251	129	4,305	191	5,359
…	…	…	…	…	…	…	…
64	3,2	126	4,254	188	5,308	250	6,362

Зависимость $\nu(n)$ для уровня значимости $q = 0{,}1$, построенная в соответствии с (8), представлена на рис. 6.

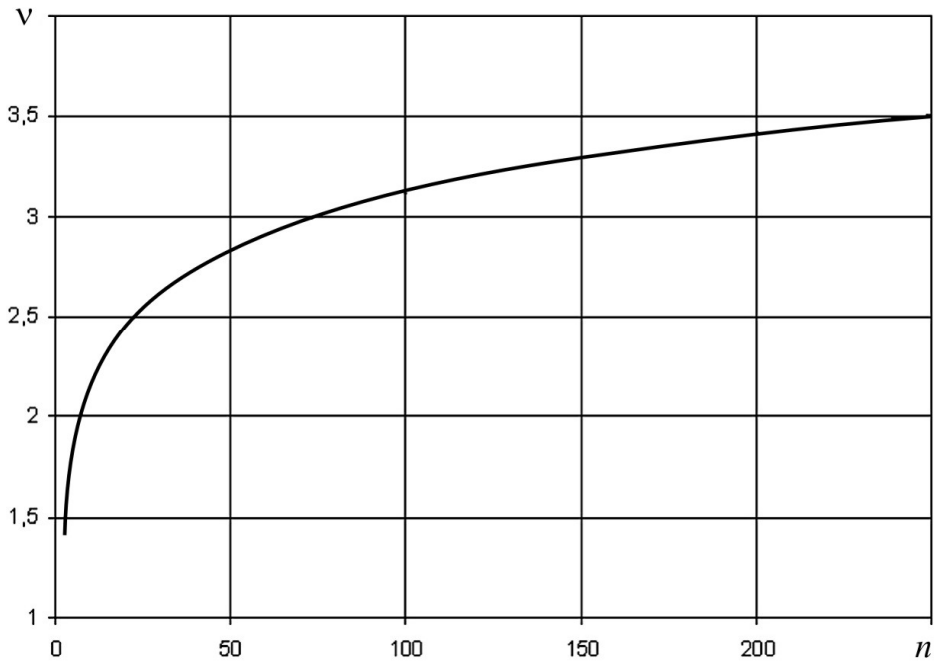

Рис. 6. Зависимость $\nu(n)$ для $q = 0{,}1$ на интервале значений $n = \overline{3; 250}$

После окончания всех вышеперечисленных процедур обучающая выборка несет в себе минимум шумовой составляющей (из набора данных удалена большая часть некоррелированных переменных) и не содержит грубых ошибок инструментального и лабораторного измерений, т.е. подготовлена для обучения моделей.

Список литературы

1. Померанцев А.Л. Метод главных компонент (PCA) / Российское хемометрическое общество [Электронный ресурс]. – URL: http: // www.chemometrics.ru/materials/textbooks/pca.htm

2. Höskuldsson A. PLS Regression and the Covariance // Journ. of Chemometrics. – 2006. – Vol. 20, Is. 8–10. – P. 376–385.

3. Лавренчик В.Н. Постановка физического эксперимента и статистическая обработка его результатов: учеб. пособие для вузов. – М.: Энергоатомиздат, 1986. – 72 с.

Получено 20.06.2012