УДК 519.25+330.46

Жгун Т.В., Липатов А.В., Чалов Г.А.

Новгородский государственный университет им. Ярослава Мудрого, г. Великий Новгород, Россия

ОПРЕДЕЛЕНИЕ ИНФОРМАТИВНОСТИ ИНТЕГРАЛЬНОЙ ХАРАКТЕРИСТИКИ ИЗМЕНЕНИЯ КАЧЕСТВА СИСТЕМЫ*

АННОТАЦИЯ

В статье предлагается подход к определению информативности характеристики изменения качества системы по зашумлённым данным при использовании ОСШ-алгоритма. Алгоритм на основе метода главных компонент определяет эмпирические главные компоненты согласованием направлений собственных векторов для различных наблюдений и выбор действующих переменных этих компонент на основе принятого отношения сигнал/шум. При разделении данных на полезный сигнал и шум критерий информативности, основывающийся на необходимой доле объяснённой дисперсии, теряет смысл. Поэтому определение информативности без учета характеристики шума в рассматриваемых данных бессмысленно и требуется его переопределить с учетом свойств шума. В статье построены априорные и апостериорные оценки информативности решения задачи построения интегральной характеристики изменения качества на основании задаваемого отношения сигнал/шум. Полученные подходы в построении оценок информативности проиллюстрированы на примере вычисления интегрального индикатора качества жизни.

КЛЮЧЕВЫЕ СЛОВА

Интегральная характеристика качества, интегральные индикаторы качества жизни, шум в измеряемых данных, отношение сигнал/шум, метод главных компонент, информативность метода главных компонент.

Tatyana Zhgun, Alexander Lipatov, German Chalov

Novgorod State University a. Yaroslav the Wise, Veliky Novgorod, Russia

DEFINITION OF INFORMATIVITY DURING CALCULATION THE INTEGRAL CHARACTERISTICS OF CHANGES OF THE QUALITY SYSTEM

ABSTRACT

The article offers an approach to the estimation of the informativity of the integral characteristics of changes of the quality system, calculated by the measured data containing noise, when using the SNR-algorithm. The algorithm determines the empirical principal components, choosing the directions of eigenvectors for different observations and determines the operating variables of these components on the basis of the signal-to-noise ratio. The criterion of informativity based on the required proportion of the explained variance, when splitting the data on the signal and noise, is meaningless. Therefore, the a priori definition of informativity without taking into account the characteristics of the noise in the data in question in this case is meaningless and should be overridden based on the properties of noise. The article is constructed a priori and apostore estimating the informativity of solving the problem of constructing the integral characteristics of quality change on the basis of the specified signal-to-noise ratio. The resulting approaches in the construction of a priori and a posteriori estimates of the information content is illustrated by the calculation of the integral indicator quality of life.

KEYWORDS

Integral characteristic quality, integral indicator quality of life, noise in the measured data, the signal-noise ratio, principal component analysis, informativity of the principal component analysis.

* **Труды XI Международной научно-практической конференции «Современные информационные технологии и ИТ-образование» (SITITO'2016), Москва, Россия, 25-26 ноября, 2016**

Введение

Одной из центральных проблем при решении задач обработки информации является проблема выбора информативного подмножества признаков и оценки его пригодности. Все чаще встречаются реальные задачи (например, в генетике), в которых небольшое число (десятки) объектов выборки описывается очень большим числом характеристик (десятками тысяч). При решении этой проблемы возникают вопросы: как организовать выбор наиболее характерных признаков, по каким критериям оценивать информативность выбранной подсистемы признаков. Большой интерес к этим проблемам в различных областях науки и техники обусловлен многообразием прикладных задач, в которых используются результаты. Необходимость в обработке и анализе данных возникает при распознавании объектов, в обработке изображений, при изучении природных ресурсов Земли из космоса, в управлении движущимися объектами, при количественной оценке параметров объектов и т. п.

Большинство систем анализа данных основывается на методах построения пространства признаков меньшей размерности. Задача снижения размерности важна еще и потому, что сложность большинства алгоритмов экспоненциально возрастает с увеличением размерности изображений, а практическая реализация таких алгоритмов требует мощных вычислительных средств. Одним из широко распространенных методов сокращения размерности изображений является метод главных компонент (МГК). В настоящее время для решения задачи поиска и распознавания предлагаются множество алгоритмов, использующих МГК.

Информативность в методе главных компонент

Метод главных компонент – один из способов понижения размерности, состоящий в переходе к новому ортогональному базису, оси которого ориентированы по направлениям максимальной дисперсии набора входных данных. Вдоль первой оси нового базиса дисперсия максимальна, вторая ось максимизирует дисперсию при условии ортогональности первой оси, и т.д., последняя ось имеет минимальную дисперсию из всех возможных. Однако направления, максимизирующие дисперсию, далеко не всегда максимизируют информативность. Может случиться, что именно младшие главные компоненты несут необходимую смысловую нагрузку. Например, при создании цифровой модели рельефа, именно восьмая и девятая главные компоненты дают искомый рельеф, а главные компоненты 12 и 13 в методе «Гусеница» свидетельствует о наличии в анализируемых данных периодики с дробным периодом [1]. На странице сайта *Alglib* [2] приводится пример, когда переменная с максимальной дисперсией не несет почти никакой информации, в то время как переменная с минимальной дисперсией позволяет полностью разделить классы.

Информативность в методе главных компонент основывается на дисперсионном критерии системы признаков, с помощью которого происходит отбор числа главных компонент [3]. Распространённый способ выбора числа главных компонент – оставить число главных компонент, которые объясняют заданный процент общей дисперсии – параметр информативности θ:

$$\gamma_l = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_l}{\lambda_1 + \lambda_2 + \ldots + \lambda_n} \geq \theta \,. \qquad (1)$$

В работах Айвазяна [4] МГК применяется для вычисления интегральной характеристики качества жизни, порог информативности θ выбирается 55%. В **а**нализе пожарной безопасности зданий величина пожарного риска в зданиях, ограничиваются первыми главными компонентами, которые объясняют более 80–90 % дисперсии [5]. При исследовании показателей сердечного ритма исключают только те главные компоненты, которые учитывают менее 5% суммарной дисперсии переменных [6].

Определение информативности МГК при работе с зашумленными данными

Статистические данные неизбежно содержат погрешность измерения и, возможно, иные случайные ошибки. Любой результат, полученный на основании этих данных, будет дублировать эти ошибки. Переход к другому моменту времени означает изменение данных, которое при неизменной структуре системы вызвано как изменением ситуации, так и случайными ошибками. Метод главных компонент на основании различных для разных моментов значений собственных векторов и собственных значений описывает неизменную структуру системы. Следовательно, именно значения собственных чисел и собственных векторов будут тем сигналом, который нужно распознать, т.е. выделить сигнал из зашумленных данных по имеющимся реализациям. Эта задача аналогична задаче восстановления цифровых изображений, искаженных белым гауссовским

шумом. МГК позволяет выделить структуру в многомерном массиве данных и с успехом применяется для распознавания изображений и для шумоподавления.

Современные технические системы и человеческий глаз уверенно выделяют сигнал из шума, если уровень отношения сигнала (а если точнее, суммы сигнала и шума) к шуму (ОСШ, англ. *signal-to-noise ratio*, сокр. *SNR*) в системе составляет около 2,2 единиц. В частности, именно такое пороговое значение используется в фотометрии слабых объектов: при регистрации сигнала от тусклых звезд необходимо, чтобы отношение сигнал/шум превышало 2,2.

Подходы к оценке числа главных компонент по необходимой доле объяснённой дисперсии формально применимы всегда, однако неявно они предполагают, что нет разделения на «сигнал» и «шум», и любая заранее заданная точность имеет смысл. При разделении данных на полезный сигнал и шум задаваемая точность теряет смысл и требуется переопределить понятие информативности. Рассмотрим подход к определению информативности при построения интегральной характеристики изменения качества системы. Определяемая интегральная характеристика есть слабый полезный сигнал, который нужно распознать в зашумленных данных (в данном случае – в статистических данных). Решение получаем с помощью с помощью алгоритма на основе метода главных компонент, учитывающего отношение сигнала к шуму используемых данных – ОСШ-алгоритма [7-9]. Интегральная оценка системы из *m* объектов, каждый из которых характеризуется *n* признаками для момента t имеет вид:

$$q^t = A^t \cdot W^*, \tag{2}$$

где $q^t = \left\langle q^t{}_1, q^t{}_2, \ldots, q^t{}_m \right\rangle^T$ – вектор интегральных индикаторов момента *t*, A^t – матрица предобработанных данных для момента *t*, веса показателей $W^* = \left\langle w^*{}_1, w^*{}_2, \ldots, w^*{}_n \right\rangle$ определены с помощью ОСШ-алгоритма [7-9] путем суммирования выбранного числа эмпирических главных компонент.

Выбор порогового значения ОСШ определяет выбор параметра информативности θ, определяющего относительную доля разброса γ, приходящуюся на первые главные компоненты (1). Если информативность γ выражена в долях единицы, величину отношения сигнал/шум можно представить как отношение полезной части используемой информации γ к неиспользуемой информации 1- γ и тогда справедливо:

$$SNR = \frac{\gamma}{1-\gamma}. \tag{3}$$

Если рассматриваемое значение отношения сигнал/шум не менее заданного порогового значения информативности θ $SNR \geq \theta$, то и $\frac{\gamma}{1-\gamma} \geq \theta$, и тогда справедлива оценка информативности

$$\gamma \geq \frac{\theta}{\theta+1}. \tag{4}$$

Соотношение (4) даёт априорную оценку информативности выбранной системы признаков через используемое пороговое отношение сигнал/шум. Эта оценка и будет априорной оценкой *SNR* - информативности решения. В таблице 1 представлены некоторые значения, связывающие рассматриваемые показатели.

Таблица 1– Связь SNR – информативности с используемым пороговым значением ОСШ

SNR	1.2	1.5	2.0	2.2	3.0	4.0	5.7	9.0	19.0
Информативность, γ	0.550	0.600	0.667	0.688	0.750	0.800	0.851	0.900	0.950

Используемое в ОСШ-алгоритме пороговое значение *SNR* = 2,2 соответствует информативности около 70%. Увеличивая используемое пороговое значение ОСШ, можно надеяться, что информативность будет выше, при этом необходимое число используемых эмпирических главных компонент для вычисления интегральной характеристики увеличивается.

Увеличение порогового значения сигнал/шум при определении эмпирических главных компонент возможность повышения информативности делает иллюзорной. Значение *SNR*=2,2 является оптимистичной величиной для статистических данных, и увеличение этого значения хотя бы до трех единиц не позволит получить достаточного числа эмпирических главных компонент, большая их часть окажутся просто нулевыми. В таблице 2 представлены варианты определения пятой и шестой эмпирических главных компонент при вычислении интегральной характеристики изменения качества. Действующие переменные, у которых вычисленное значение сигнал/шум

(отношение среднего *m* к среднеквадратичному отклонению *s*) превосходит пороговое значение *SNR*=2,2, выделены темным цветом. Факторные нагрузки действующих переменных учитываются при вычислении интегральной характеристики, незначимые переменные обнуляются. Кажется очевидным, что пятая компонента информативнее шестой, так как там действующими оказались 8 из 14 переменных, а в шестой компоненте – только одна действующая переменная из 14. При увеличении порогового значения *SNR* до 3 единиц, в пятой эмпирической главной компоненте останется половина действующих переменных, а в шестой их не останется вовсе, что очевидно информативность решения не увеличит, а уменьшит.

Предложим способ определения информативности интегральной характеристики изменения качества, использующей для построения метод главных компонент, для случая, когда присутствует разделение данных на «сигнал» и «шум». В этом случае заранее заданная точность не имеет смысла, и оценка числа главных компонент по необходимой доле объяснённой дисперсии неприменима. В каждой из эмпирических главных компонент, полученных согласованием направлений собственных векторов для разных наблюдений, определим действующие переменные (для которых выполняется заданное отношение сигнал/шум) и вычислим сумму рассматриваемых величин ОСШ у действующих переменных, и сумму ОСШ у всех переменных эмпирической главной компоненты.

Таблица 2 – Определение эмпирических главных компонент

5	Переменные													
ГК	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2007	0.15	-0.23	0.03	0.26	0.02	0.03	-0.73	0.12	0.06	0.02	-0.34	0.02	-0.25	0.36
2008	-0.04	-0.41	0.08	0.19	-0.03	0.05	-0.68	-0.03	-0.04	-0.13	-0.40	0.12	-0.08	0.34
2009	-0.01	-0.51	0.10	0.22	-0.03	0.05	-0.45	-0.10	0.11	-0.09	-0.51	0.16	-0.19	0.35
2010	0.07	-0.05	0.09	0.24	0.05	0.08	-0.65	0.06	0.27	0.00	-0.61	0.09	-0.11	0.17
2011	0.05	-0.29	0.10	0.23	0.03	0.05	-0.50	-0.03	0.12	-0.06	-0.67	0.13	-0.19	0.26
2012	0.05	-0.30	0.12	0.32	0.04	0.07	-0.34	0.00	0.04	0.00	-0.68	0.14	-0.28	0.32
m	0.04	-0.30	**0.09**	**0.24**	0.01	**0.05**	**-0.56**	0.00	0.09	-0.04	**-0.54**	**0.11**	**-0.18**	**0.30**
s	0.07	0.16	0.03	0.04	0.04	0.02	0.15	0.08	0.10	0.06	0.14	0.05	0.08	0.07
ОСШ	0.68	1.90	**2.75**	**5.47**	0.38	**2.90**	3.69	0.02	0.89	0.76	**3.76**	**2.20**	**2.31**	**4.27**
													Сумма ОСШ по строке	31.99
													Сумма ОСШ действующих переменных	27.36
6	Переменные													
ГК	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2007	0.09	0.25	0.04	-0.29	0.06	0.01	-0.20	0.33	-0.19	-0.12	0.06	-0.77	-0.12	-0.14
2008	-0.19	-0.56	0.09	-0.08	0.04	0.02	0.21	0.01	-0.37	-0.04	0.45	-0.32	-0.28	0.27
2009	0.24	0.58	-0.05	-0.01	0.09	0.03	-0.50	0.16	0.08	0.08	-0.44	-0.08	0.10	-0.31
2010	0.01	0.52	0.04	-0.45	0.12	0.09	-0.31	0.13	0.14	0.04	0.06	-0.54	-0.02	-0.28
2011	0.13	0.58	-0.13	-0.24	0.09	0.21	-0.03	-0.13	0.42	0.10	-0.40	0.02	0.34	-0.20
2012	0.21	0.56	-0.03	-0.33	0.10	0.13	-0.36	0.00	0.15	0.06	-0.41	-0.30	0.09	-0.28
m	0.08	0.32	-0.01	-0.23	**0.09**	0.08	-0.20	0.09	0.04	0.02	-0.11	-0.33	0.02	-0.16
s	0.16	0.45	0.08	0.17	0.03	0.08	0.25	0.16	0.28	0.09	0.36	0.29	0.21	0.22
ОСШ	0.53	0.71	0.09	1.42	**2.87**	1.08	0.78	0.54	0.13	0.23	0.31	1.14	0.08	0.71
													Сумма ОСШ по строке	10.63
													Сумма ОСШ действующих переменных	2.87

Аналогично дисперсионной информативности согласно (1), можно определить *SNR*-информативность для выбранного числа эмпирических главных компонент *N*:

$$\gamma_{SNR} = \frac{S_{11} + S_{12} + \ldots + S_{1N}}{S_{21} + S_{22} + \ldots + S_{2N}}, \qquad (5)$$

где S_{1k} – сумма величин ОСШ у действующих переменных *k*-ой ЭГК, S_{2k} – сумма ОСШ всех переменных *k*-ой ЭГК. В отличие от дисперсионной информативности, *SNR*-информативность не может достигать 100% по логике построения.

Таблица 3 – SNR-информативность Блока 1: «Благосостояние населения» при вычислении интегрального индикатора качества жизни

№ ЭГК	Действующие ОСШ ЭГК	Накопленные действующие ОСШ ЭГК	Все ОСШ ЭГК	Накопленные ОСШ ЭГК	SNR-информативность
1	101.42	101.42	102.23	102.23	0.99
2	54.99	156.41	57.56	159.79	0.98
3	15.41	171.82	19.73	179.52	0.96
4	4.25	176.07	11.33	190.85	0.92
5	3.42	179.49	7.92	198.76	0.90

6	4.90	184.39	10.66	209.42	0.88
7	16.49	200.89	20.94	230.36	0.87
8	10.84	211.73	16.51	246.87	0.86
9	95.67	307.40	100.48	347.35	0.88

Проиллюстрируем дальнейшие рассуждения на примере вычисления интегральной характеристики качества жизни. Для этого воспользуемся переменными из исследования [10]. Рассматриваются 3 блока переменных:

- Блок 1: Уровень благосостояния населения (9 переменных);
- Блок 2: Качество населения (14 переменных);
- Блок 3: Качество социальной сферы (14 переменных).

В таблице 3 приводится пример определения *SNR*-информативности первого блока «Благосостояние населения» при вычислении интегрального индикатора качества жизни. При использовании всех девяти эмпирических главных компонент (ЭГК) *SNR*-информативность вычисляемой интегральной характеристики «Благосостояния населения» составит около 88%.

SNR–информативность убывает при увеличении числа выбираемых эмпирических главных компонент, а дисперсионная информативность – возрастает. На рис.1 приводится определение дисперсионной и *SNR*-информативности при вычислении интегральной характеристики демографического развития России на основании 85 переменных. Рассматриваются все показатели, фиксируемые Росстатом на всем интервале наблюдения.

Рис.1 Определение дисперсионной и SNR – информативности при вычислении интегральной характеристики демографического развития России

Общая информативность выбранной системы признаков представляет собой компромисс между этими значениями. и определяется двумя этими параметрами – дисперсионной и *SNR* - информативностью:

$$\gamma = \gamma_{\sigma} \cdot \gamma_{SNR} \quad . \tag{6}$$

Число эмпирических главных компонент выбирается таким образом, чтобы общая информативность была максимальной. Например, при вычислении интегральной характеристики демографического развития России максимальная информативность составляет 60,52% и будет достигаться при использовании девяти ЭГК (рис.2). При этом *SNR*-информативность составит около 71%, что согласуется с оценкой (4).

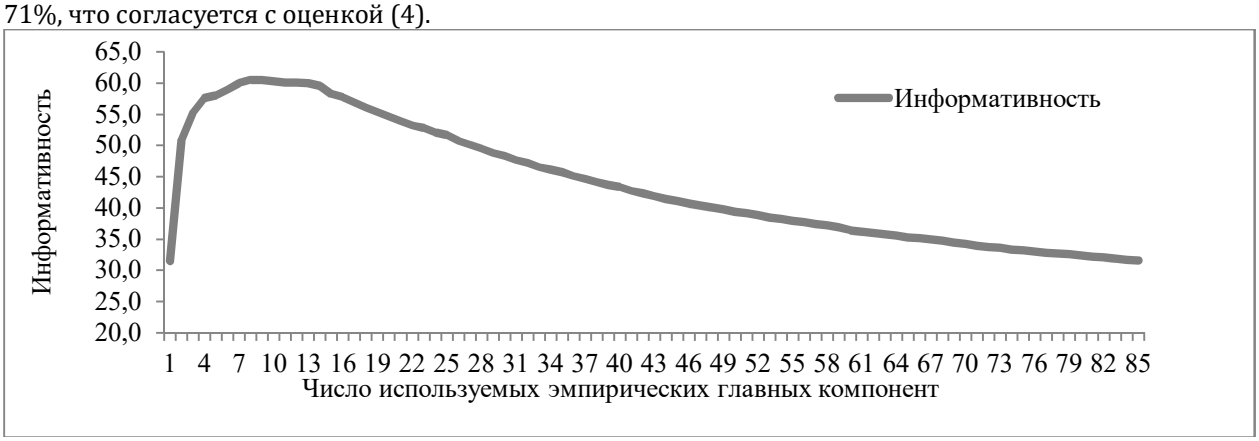

Рис.2 Определение общей информативности при вычислении интегральной характеристики демографического развития России

Пример определения числа эмпирических главных компонент, определяющую

максимальную информативность второго блока, приведен также в таблице 4. Из 14 ЭГК выбирается 11, и общая информативность интегральной характеристики второго блока «Качество населения» составит около 79%. Суммарная информативность первого блока «Благосостояние населения» максимальна при рассмотрении всех 9 ЭГК и составит около 88%. Поэтому для вычисления интегрального показателя первого блока используем все эмпирические главные компоненты. Чем больше переменных описывают систему, тем меньшее их относительное количество участвует в построении композитного индекса. Для второго блока выбираем 11 из 14 ЭГК, для третьего – 10 из 14 ЭГК. В системе, которую описывают 51 переменных, было выбрано 21 ЭГК, в приведенном выше примере из 85 возможных было выбрано 9 ЭГК. В таблице 4 можно сравнить оценки информативности по дисперсионному и *SNR*-критерию. Оценка информативности по *SNR*-критерию при выборе одиннадцати ЭГК выглядит менее оптимистичной, чем по дисперсионному критерию.

Таблица 4 – Информативность Блока 2 «Качество населения» при вычислении интегрального индикатора качества жизни

	Номер эмпирической главной компоненты													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Сумма ОСШ *k*-ой ЭГК	190,1	115,0	37,0	33,2	21,4	25,1	12,9	13,6	15,2	10,8	16,9	15,3	17,2	20,1
Сумма действующих ОСШ *k* - ой ЭГК	186,4	111,7	26,9	21,7	12,5	17,5	3,2	5,0	2,7	3,2	7,0	4,1	6,6	6,2
Накопленные ОСШ *k* - ой ЭГК	190	305	342	375	397	422	435	448	463	474	491	506	524	544
Накопленные действ, ОСШ *k*-ой ЭГК	186	298	325	347	359	377	380	385	387	391	398	402	408	415
SNR - информативность	0,98	0,98	0,95	0,92	0,91	0,89	0,87	0,86	0,84	0,82	0,81	0,79	0,78	0,76
Эмпирические собственны числа (ЭСЧ)	4,5	2,6	1,5	1,2	1,0	0,8	0,7	0,5	0,4	0,3	0,2	0,1	0,1	0,1
Накопленные ЭСЧ	4,5	7,1	8,6	9,8	10,7	11,6	12,2	12,7	13,1	13,4	13,6	13,8	13,9	13,9
Дисперсионная информативность	0,32	0,50	0,61	0,70	0,77	0,83	0,87	0,91	0,94	0,96	0,97	0,98	0,99	0,99
Информативность, %	31,2	49,3	58,2	64,6	69,5	73,7	76,2	78,0	78,3	78,90	**78,93**	78,0	77,2	75,8

Приведем в таблице 5 теперь определение характеристик информативности для третьего блока переменных, характеризующих «Качество социальной сферы». Количество переменных этого блока совпадает с количеством переменных второго блока, однако у него совсем другие численные характеристики – и сумма всех значений ОСШ (395 здесь и 544 для второго блока), и сумма ОСШ у действующих переменных (294 и 415 соответственно). Хотя общая информативность выбираемого числа эмпирических компонент здесь сравнима с общей информативностью второго блока – 77,4% и 78,9% – сами блоки вносят разный вклад в определяемое по трем блокам значение интегрального показателя. Вклад второго блока значительно выше, чем третьего, так как сигнал этого блока значительно «слышнее».

Таблица 5 – Информативность 3 блока «Качество социальной сферы» при вычислении интегрального индикатора качества жизни

	Номер эмпирической главной компоненты													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	11	12	**13**	**14**
Сумма ОСШ *k*-ой ЭГК	159	44,7	51,1	12,8	7,8	7,4	16,5	-	10,2	11,1	12,1	16,2	23,0	23,0
Сумма действующих ОСШ *k* - ой ЭГК	158	38,0	42,6	3,7	2,5	3,0	8,5	-	3,3	3,5	2,5	5,2	6,7	16,2
Накопленные ОСШ *k* - ой ЭГК	159	204	255	268	275	283	299	299	309	321	333	349	372	395
Накопленные действ ОСШ *k*-ой ЭГК	158	196	239	242	245	248	257	257	260	263	266	271	278	294
SNR - информативнос	1,00	0,96	0,94	0,91	0,89	0,88	0,86	0,86	0,84	0,82	0,80	0,78	0,75	0,74
Эмпирические собст-венные числа (ЭСЧ)	4,38	2,09	1,60	1,13	0,97	0,86	0,69	0,60	0,49	0,39	0,32	0,26	0,16	0,07
Накопленные ЭСЧ	4,4	6,5	8,1	9,2	10,2	11,0	11,7	12,3	12,8	13,2	13,5	13,8	13,9	14,0
Дисперсионная информативность	0,31	0,46	0,58	0,66	0,73	0,79	0,84	0,88	0,91	0,94	0,97	0,98	1,00	1,00
Информативность, %	31,1	44,5	54,0	59,6	64,6	69,1	71,8	75,4	76,8	**77,4**	77,15	76,4	74,3	74,4

Следовательно, при вычислении интегральной характеристики системы общую характеристику системы следует определять с учетом весов блоков, определяемых пропорционально вычисленным значениям суммы отношений сигнал/шум у действующих переменных, что аналогично определению веса блока пропорционально силе принятого сигнала. В таблице 6 представлены вычисленные значения весов блоков в зависимости от вычисленных характеристик значений суммы отношений сигнал/шум у действующих переменных. Блок 3 «Качество социальной сферы», в котором 14 переменных, оказался менее значим, чем блок 1 «Уровень благосостояния населения», в котором 9 переменных. А самым значимым оказался второй блок «Качество населения».

SNR-информативность решения задачи вычисления интегрального индикатора качества жизни определяется суммами отношений сигнал/шум у действующих переменных и суммой всех вычисленных отношений сигнал/шум. В рассматриваемом примере она составит около 82%. Полученное значение согласуется с априорной оценкой (4). Однако вовсе не следует, что и при исследовании систем, характеризующихся значительным числом переменных, всегда следует использовать все эмпирические главные компоненты, доставляющие максимум общей информативности. Увеличение числа эмпирических главных компонент будет вносить вклад в вычисляемую интегральную характеристику, пока величина используемых эмпирических чисел превосходит неустранимую ошибку исходных данных.

Таблица 6 – Определение вклада разных блоков в вычисляемое значение интегрального показателя на основе SNR- информативности

Блок	**1**	**2**	**3**	По всем блокам
Сумма действующих ОСШ	307,4	397,7	254,3	959,4
Сумма ОСШ	347,4	491,14	330	1168,3
Вес блока	0,32	0,41	0,27	

Приводимые в статистических справочниках величины имеют зачастую всего три (или даже менее) верные значащие цифры в представлении величин (процентов и пр.). Вычисляемый результат не может иметь точность большую, чем исходные данные, в частности, количество верно значащих цифр результата также составляет три верно значащие цифры (или менее). В вычисляемой интегральной характеристике это соответствует одному знаку до запятой и двум знакам после. Т.е. абсолютная погрешность вычисляемой интегральной характеристики $\Delta y^* \geq 10^{-2}$. Следовательно, эмпирическая главная компонента не изменит величину вычисляемой характеристики, если соответствующее эмпирическое собственное число, участвующее в определении нагрузок эмпирических главных компонент, имеет значение меньшее, чем минимально возможная абсолютная ошибка вычисляемой интегральной характеристики: $\sqrt{\lambda_i} \leq 0.5 \cdot 10^{-2}$. Т.е., стоит отбрасывать все те ЭГК, для которых значение эмпирического собственного числа имеют порядок $\lambda_i \approx \cdot 10^{-5}$ и менее. В рассматриваемом примере при вычислении интегральной характеристики жизни населения рассматривались блоки с девятью и четырнадцатью переменными, и при этом для всех трех блоков минимальное из эмпирических собственных чисел $\lambda_{min} \approx \cdot 10^{-2}$. Т.е. в этом случае даже минимальное эмпирическое собственное число дает ощутимый вклад в вычисляемую характеристику. При рассмотрении демографического состояния России рассматривались 85 переменных в одном блоке. Здесь, начиная в 75-го эмпирического собственного числа, значения ЭСЧ чрезвычайно малы: $\lambda_{75} \approx \cdot 10^{-5}$ и менее. Т.е. использование эмпирических главных компонент, начиная с 75-ой, смысла не имеет. Но в этом случае было достаточно использовать всего девять ЭГК.

Заключение

Традиционные подходы к оценке числа главных компонент по необходимой доле объяснённой дисперсии предполагают, что нет разделения данных на «сигнал» и «шум», и тогда любая заранее заданная точность имеет смысл. При разделении данных на полезный сигнал и шум задаваемая точность бессмысленна и требуется переопределить понятие информативности. В работе предложено определение информативности метода главных компонент для построения интегральной характеристики изменения качества системы с учетом наличия шума в измеряемых данных. Построены априорные и апостериорные оценки информативности, предложен алгоритм определения весов подсистем при вычислении интегральной характеристики с использованием

апостериорных оценок информативности этих подсистем. Приведен пример определения этих характеристик при вычислении интегральной характеристики качества жизни.

Работа выполнена при финансовой поддержке проектной части государственного задания в сфере научной активности Министерства образования и науки Российской Федерации, проект № 1.949.2014/К.

Литература

1. Голяндина Н.Э., Усевич К.Д., Флоринский И.В. Анализ сингулярного спектра для фильтрации цифровых моделей // Геодезия и картография. – 2008. - №5. – Сс. 21-28.
2. Линейный дискриминантный анализ. Alglib. Open sourse [Электронный ресурс]. – URL: http://alglib,sources,ru/dataanalysis/lineardiscriminantanalysis,php – Загл. с экрана (дата обращения: 13.08.2016)
3. Rencher A.C. Methods of multivariate analysis. Wiley. – 2002. – 732 p.
4. Айвазян С.А. К методологии измерения синтетических категорий качества жизни населения // Экономика и математические методы Т. 39. – 2003.– № 2. – Сс. 33-53.
5. Зикратов И.А., Техтереков С.А., Чижов В.А. Методика выбора информативных признаков для классификации объектов на основе метода главных компонент // Вестник Санкт-Петербургсого Университета МЧС России» [Электронный ресурс]– URL: http://vestnik,igps,ru/wp-content/uploads/V63/8,pdf – Загл. с экрана (дата обращения: 13.08.2016),
6. Машин В.А. Методические вопросы использования факторного анализа на примере спектральных показателей сердечного ритма // Экспериментальная психология. – 2010. – Т.3. – №4. – С.119–138.
7. Жгун Т.В. Построение интегральной характеристики изменения качества системы на основании статистических данных как решение задачи выделения сигнала в условиях априорной неопределенности // Вестн. Новг. гос. ун-та. Сер.: Технические науки. – 2014. – № 81. – С. 10-16.
8. Жгун Т.В. Построения интегральной характеристики демографического развития территорий на примере муниципальных образований Новгородской области //Региональная экономика: теория и практика. – 2013. № 36(315),сентябрь. – С. 2-12.
9. Жгун Т.В. Вычисление интегрального показателя эффективности функционирования динамической системы на примере интегральной оценки демографического развития муниципальных образований Новгородской области // Вестн. Новг. Гос. ун-та. Сер.: Физико-математические науки. – 2013. № 75. – Т.2. С. 11-16.
10. Исакин М.А. Модификация метода k -средних с неизвестным числом классов // Прикладная эконометрика. 2006. – Выпуск № 4. – С. 62-70,

References

1. Golyandina N.E., Usevich K.D., Florinsky I.V. The Analysis of a singular range for a filtration of digital models / Geodesy and cartography. 2008, No 5. Pp. 21-28. (In Russ.)
2. Linear discriminant analysis. Alglib. Open sourse . ALGLIB - numerical analysis library, 1999-2015. (In Russ.) Available at: http://alglib,sources,ru/dataanalysis/lineardiscriminantanalysis,php, (acessed 13,01,2016),
3. Rencher A.C. Methods of multivariate analysis. Wiley. 2002. 732 p.
4. Ayvazyan S.A. To methodology of measurement of synthetic categories of quality of life of the population//Economy and mathematical methods. T. 39. 2003.No2. Page 33-53,
5. Zikratov I.A., Tehterekov S.A., Chizhov V.A. The methods of selecting informative features for classification of objects based on the method of principal components/ Vestnik of Saint-Peterburgskogo University of EMERCOM of Russia" (In Russ.) Available at: http://vestnik,igps,ru/wp-content/uploads/V63/8,pdf – (acessed 13.01.2016)
6. Mashin V.A. Methodological issues the use of factor analysis on the example of the spectral indices of cardiac rhythm // Experimental psychology. 2010. Vol. 3. No 4. Pp. 119-138.
7. Zhgun T.V. Creation of the integrated characteristic of change of quality of system on the basis of statistical data as the solution of a problem of allocation of a signal in the conditions of aprioristic uncertainty// Vestn. Novg. the state. un-that. It is gray.: Technical science. 2014. No.81. Page 10-16,
8. Zhgun T.V. Creation of the integrated characteristic of demographic development of territories on the example of municipalities of the Novgorod region / Regional economy: theory and practice. No. 36(315). September. 2013. Page 2-12,
9. Zhgun T.V. Calculation of an integrated indicator of efficiency of functioning of dynamic system on the example of an integrated assessment of demographic development of municipalities of the Novgorod region// Vestn. Novg. the state. un-that. It is gray.: Physical and mathematical sciences., 2013. No. 75, T.2. Page 11-16.
10. Isakin M.A. Modification of a Method to k-Means with Unknown Number of Classes /. Applied Econometrics. 2006. Release no. 4. pp. 62-70. (in Russian)

Поступила: 5.10.2016

Об авторах:

Жгун Татьяна Валентиновна, доцент кафедры прикладной математики и информатики Новгородского государственного университета им, Ярослава Мудрого, кандидат физико-математических наук, zhtv@mail,ru;

Липатов Александр Владимирович, аспирант Новгородского государственного университета им, Ярослава Мудрого;

Чалов Герман Александрович, аспирант Новгородского государственного университета им, Ярослава Мудрого.