Визуальный анализ и обработка многомерных данных

А.Е. Бондарев, В.А. Галактионов, Л.З. Шапиро
bond@keldysh.ru|vlgal@gin.keldysh.ru|pls@gin.keldysh.ru
Институт прикладной математики им. М.В.Келдыша РАН, Москва, Россия

В работе рассматриваются проблемы визуального анализа многомерных наборов данных. Для визуального анализа применяется подход, основанный на построении упругих карт. Для анализа кластерных структур в исходном объеме данных упругие карты используются в качестве методов отображения исходных точек данных в замкнутые многообразия с меньшей размерностью. Уменьшая параметры упругости, можно проектировать поверхность карты, которая намного лучше аппроксимирует многомерный набор данных. Точки исследуемого объема данных проецируются на карту. Расширение построенной карты на плоскость вкупе с отображением в пространство первых трех главных компонент позволяет получить представление о кластерной структуре многомерного набора данных. Построение упругих карт не требует априорной информации о данных и не зависит от характера данных, происхождения данных и т. д. Применение метода «квази-зум» позволяет существенно улучшить результаты в области сгущения точек изучаемого многомерного пространства. В работе приведены результаты применения упругих карт для визуального анализа различных многомерных наборов данных.

Ключевые слова: *многомерные данные, визуальный анализ, упругие карты, кластерные структуры.*

Visual analysis and processing of multidimensional datasets

A.E. Bondarev, V.A. Galaktionov, L.Z. Shapiro
bond@keldysh.ru|vlgal@gin.keldysh.ru|pls@gin.keldysh.ru
Keldysh Institute of Applied Mathematics RAS, Moscow, Russia

The article considers the problems of visual analysis of multidimensional datasets. For visual analysis an approach based on the construction of elastic maps is applied. To analyse clusters in original data volume the elastic maps are used as the methods of original data points mapping to enclosed manifolds having less dimensionality. Diminishing the elasticity parameters one can design map surface which approximates the multidimensional dataset in question much better. The points of dataset in question are projected onto the map. The extension of designed map to a flat plane with mapping into the space of the first three main components allows one to get an insight about the cluster structure of multidimensional dataset. The construction of elastic maps does not require a priori information about the data and does not depend on the nature of the data, the origin of the data, etc. The application of the "quasi-Zoom" method allows one to get the significantly improved results in the area of condensation of the points of the multidimensional space under study. The paper presents the results of applying elastic maps for visual analysis of various multidimensional datasets.

Keywords: *multidimensional data, visual analysis, elastic maps, cluster structures.*

1. Введение

В анализе многомерных данных особое место занимают задачи классификации. При классификации объема многомерных данных может решаться как задача разделения исследуемой совокупности явлений на классы, так и отнесения одного или нескольких явлений к уже существующим классам. Для решения подобных задач используются методы кластерного анализа. Методов и алгоритмов кластерного анализа на современном этапе существует очень много, они постоянно развиваются и отличаются большим разнообразием. Многообразие алгоритмов кластерного анализа обусловлено множеством различных критериев, выражающих те или иные аспекты качества автоматического группирования. При решении задач классификации весьма полезными оказываются подходы визуальной аналитики, являющиеся синтезом нескольких алгоритмов понижения размерности и визуального представления многомерных данных во вложенных в исходный объем многообразиях меньшей размерности.

К таким алгоритмам можно отнести отображение исходного многомерного объема в упругих картах (Elastic Maps) [5,6,8,9] с разными свойствами упругости или эластичности. Эти методы позволяют тем или иным образом выделить из исходного многомерного объема данных содержащуюся в нем кластерную структуру.

Следует заметить, что интерес к упругим картам появился у нас при разработке вычислительной технологии для построения, обработки, анализа и визуального представления многомерных параметрических решений задач газовой динамики. Вычислительная технология реализована как единая технологическая цепочка алгоритмов производства, обработки, визуализации и анализа многомерных данных. Такая технологическая цепочка может рассматриваться как прототип обобщенного вычислительного эксперимента для нестационарных задач вычислительной газовой динамики. Схема реализации подобного обобщенного вычислительного эксперимента представлена на рисунке 1.

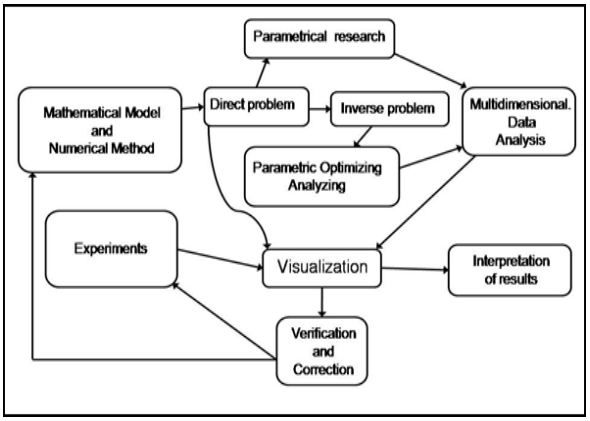

Рис. 1. Схема организации обобщенного вычислительного эксперимента.

Подобный обобщенный вычислительный эксперимент неявно предполагает наличие надежной математической

модели, численного метода для ее решения и набор экспериментальных результатов для верификации. В процессе вычислений необходимо реализовать организацию постоянного сравнения с экспериментальными данными при наличии такой возможности. Набор используемых методов должен включать в себя решение обратных и оптимизационных задач. Будучи реализованными с помощью описанных ранее параллельных интерфейсов, эти методы позволят получать решения задач параметрического исследования и оптимизационного анализа в виде многомерных объемов данных.

Для обработки этих объемов и выявления скрытых взаимозависимостей между изучаемыми в объеме параметрами необходимо интегрировать в общий набор алгоритмов методы анализа многомерных данных и их визуального представления. В итоге подобный обобщенный вычислительный эксперимент позволит получать решение не одной, отдельно взятой, задачи, а решение для целого класса задач, задаваемого диапазонами изменения определяющих параметров. Также следует отметить универсальность подобного обобщенного вычислительного эксперимента. Он может быть применен к широкому кругу задач математического моделирования нестационарных процессов. Практическая реализация подобного обобщенного эксперимента может обеспечивать организацию крупномасштабных промышленных расчетов. Описание элементов реализованной вычислительной технологии приведено в работах [2-4].

На практике упругие карты оказались полезным и достаточно универсальным инструментом, что позволило применять их к многомерным объемам данных разного типа. Например, данный подход был применен к задачам анализа текстовой информации, где в качестве числовых характеристик выступали частоты употребления слов [1].

2. Построение упругих карт

Идеология и алгоритмы реализации построения упругих карт подробно представлены в работах [5,8]. Подобная карта представляет собой систему упругих пружин, вложенную в многомерное пространство данных. Этот подход основывается на аналогии с задачами механики: главное многообразие, проходящее через «середину» данных, может быть представлено как упругая мембрана или пластинка. Метод упругих карт формулируется как оптимизационная задача, предполагающая оптимизацию заданного функционала от взаимного расположения карты и данных.

Согласно [5,6] основой для построения упругой карты является двумерная прямоугольная сетка G, вложенная в многомерное пространство, которая аппроксимирует данные и обладает регулируемыми свойствами упругости по отношению к растяжению и изгибу. Расположение узлов сетки ищется в результате решения оптимизационной задачи на нахождение минимума функционала:

$$D = \frac{D_1}{|X|} + \lambda\frac{D_2}{m} + \mu\frac{D_3}{m} \to min\,,$$

где $|X|$ - число точек в многомерном объеме данных X; m - число узлов сетки, λ, μ - коэффициенты упругости, отвечающие за растяжение и изогнутость сетки соответственно; D_1, D_2, D_3 - слагаемые, отвечающие за свойства сетки.

$$D_1 = \sum_{ij}\sum_{x\in K_{ij}} \left\|x - r^{ij}\right\|^2.$$

D_1 является мерой близости расположения узлов сетки к данным.

Здесь K_{ij} - подмножества точек из X, для которых узел сетки r^{ij} является ближайшим:

$$x \overset{\text{п}}{\to} r^{ij}, \qquad \left\|x - r^{ij}\right\|^2 \to \min, \qquad K_{ij} = \left\{x \in X, x \overset{\text{п}}{\to} r^{ij}\right\},$$

Слагаемое D_2 представляет меру растянутости сетки:

$$D_2 = \sum_{ij}\left\|r^{ij} - r^{i,j+1}\right\|^2 + \sum_{ij}\left\|r^{ij} - r^{i+1,j}\right\|^2.$$

Слагаемое D_3 представляет меру изогнутости (кривизны) сетки:

$$D_3 = \sum_{ij}\left\|2r^{ij} - r^{i,j-1} - r^{i,j+1}\right\|^2 + \\ + \sum_{ij}\left\|2r^{ij} - r^{i-1,j} - r^{i+1,j}\right\|^2.$$

Варьирование параметров упругости заключается в построении упругих карт с последовательным уменьшением коэффициентов упругости, в силу чего карта становится более мягкой и гибкой, наиболее оптимальным образом подстраиваясь к точкам исходного многомерного объема данных. После построения упругую карту можно развернуть в плоскость для наблюдения кластерной структуры в изучаемом объеме данных. Применение упругих карт позволяет более точно и четко определять кластерную структуру изучаемых многомерных объемов данных.

Следует отметить, что при построении упругих карт в многомерном облаке данных, состоящем из сгущений и отдельных отдаленных точек, возникает проблема масштабируемости. Упругая карта будет пытаться подстроиться под рассматриваемый объем в целом – как к отдаленным точкам, так и к областям сгущения, что, естественно, не может получиться одинаково хорошо. Для того чтобы решить эту проблему и обеспечить четкое представление о данных в области сгущений в работе [1] был предложен подход, названный «квази-зум» (quasi-Zoom), заключающийся в вырезании области сгущения из рассматриваемого облака многомерных данных и построения для вырезанной области упругой карты заново.

3. Примеры построения упругих карт

Рассмотрим пример построения упругих карт для широко известного тестового объема многомерных данных IRIS [5]. Данный объем представляет собой набор данных, основанных на измерениях характеристик цветков ириса. Набор данных описывает три сорта ирисов и состоит из 150 точек в четырехмерном пространстве признаков. На рисунке 2 представлена «мягкая» упругая карта для этого набора данных с раскраской по плотности данных. Здесь и далее для построения и визуализации упругих карт применен программный комплекс ViDaExpert, подробно описанный в [8]. На рисунке 3 представлена та же самая карта, развернутая на плоскость. В таком виде карта дает достаточно четкое представление о разделении многомерного объема данных на три класса.

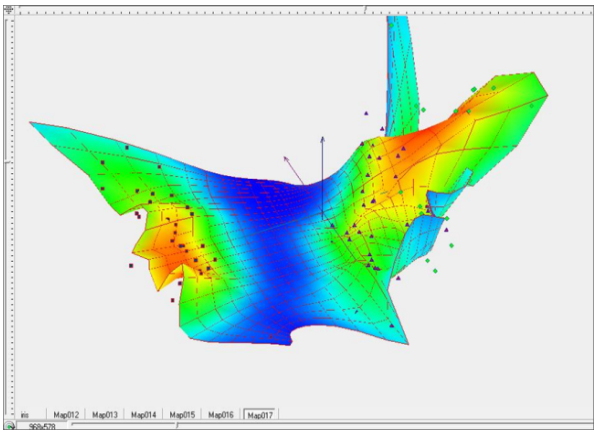

Рис. 2. «Мягкая» упругая карта в пространстве первых трех главных компонент с раскраской по плотности данных.

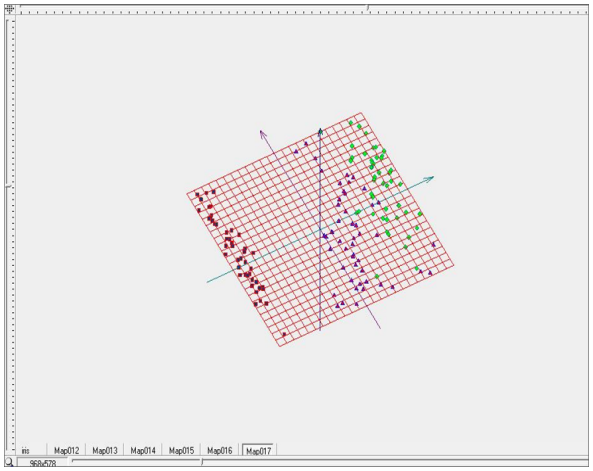

Рис. 3. Развертка «мягкой» упругой карты на плоскость.

Далее рассмотрим результаты применения подхода построения упругих карт к задаче анализа текстовой информации. С точки зрения построения упругих карт исходный многомерный объем является совершенно стандартным. Рассмотрим ниже результаты построения упругих карт для тестового объема [1]. Для первичных тестов было отобрано около 100 глаголов со 155 наиболее связанными с ними существительными. Полученные таким образом данные далее рассматриваются как многомерный объем данных, представляющий собой 100 точек в 155-мерном пространстве. Числовые значения получающейся в результате матрицы определяются как частоты совместного употребления.

Следует отметить, что среди отобранных глаголов содержался ряд пар, представляющих схожие глаголы совершенного и несовершенного вида. Это было сделано для дополнительного контроля в силу предположения, что точки, соответствующие подобным парам, должны находиться недалеко друг от друга на результирующем изображении. Пример «мягкой» упругой карты приведен на рисунке 4. На рисунке 5 представлена та же самая карта в развернутом виде. Здесь видно, что изучаемый объем данных, содержит область высокой плотности данных и точки, достаточно далеко отстоящие от этой области.

Именно в таких случаях возникает проблема масштабируемости, описанная в предыдущем разделе. Для решения этой проблемы был разработан подход «квази-зум» (quasi-Zoom), представленный в работе [1].

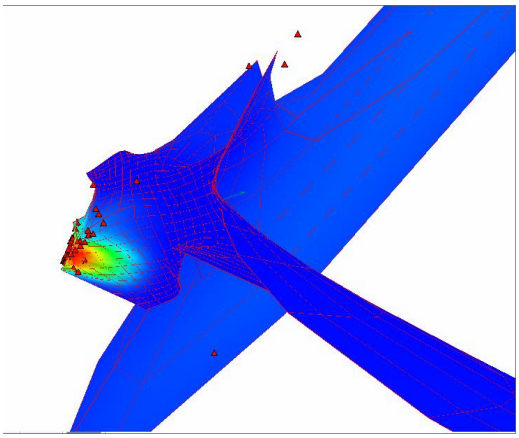

Рис. 4. Упругая карта с раскраской по плотности в применении к объему частот совместного употребления слов [1].

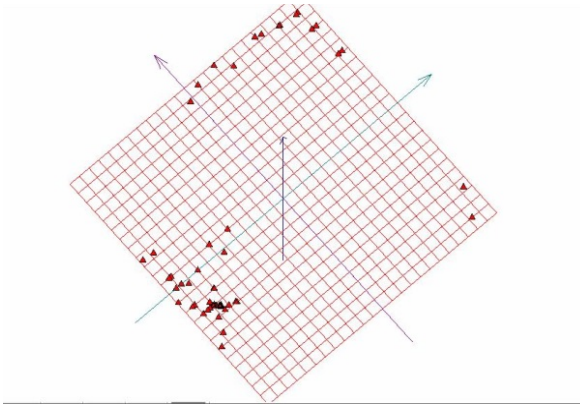

Рис. 5. Построение развертки предыдущей упругой карты на плоскость первых двух главных компонент [1].

При исследовании объема частот совместного употребления глаголов и существительных [1] практическая задача ставилась следующим образом. Нужно было максимально разделить «слипшиеся точки. Для этой цели был использован подход «квази-зум», который позволил решить эту задачу. Суть этого технологического приема заключается в том, что для более тонкой подстройки необходимо выделять большие кластеры в исследуемом объеме многомерных данных и проводить построение упругих карт для выделенных кластеров отдельно, организуя тем самым эффект, подобный функции «zoom» в современной фототехнике. Результаты применения подхода представлены на рисунке 6.

Применение технологий построения упругих карт для решения задач кластерного анализа не предполагает никакой априорной информации об изучаемых данных. Это дает возможность применять их к анализу данных самого различного типа вне зависимости от природы их происхождения. Подобное абстрагирование метода от типа и происхождения данных делает используемый подход построения упругих карт в достаточной степени универсальным инструментом анализа многомерных объемов данных.

Приведем пример применения построения упругих карт к анализу биомедицинских данных. Для этой цели были использованы данные работы [7]. Этот набор данных содержит значения шести биомеханических признаков, используемых для классификации ортопедических пациентов на 2 класса (нормальный «normal», или с отклонением от нормы «abnormal»). Каждый пациент

представлен в наборе данных шестью биомеханическими признаками, полученными из формы и ориентации таза и поясничного отдела позвоночника: тазовое опускание, тазовый наклон, угол поясничного лордоза, крестцовый наклон, радиус таза и степень спондилолистеза.

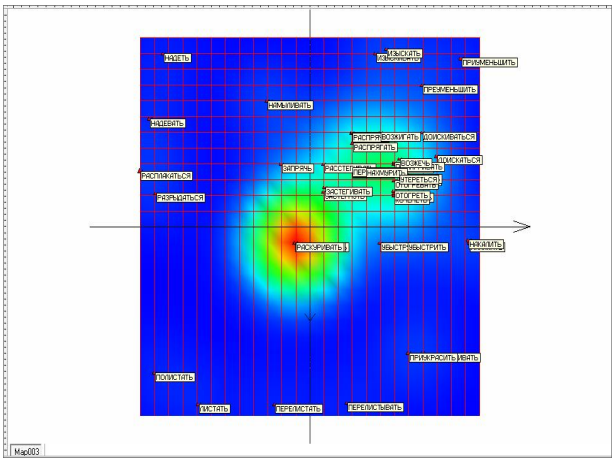

Рис. 6. Результаты применения подхода «quasi-Zoom» для разделения слипшихся точек (развертка упругой карты с раскраской по плотности данных) .

Набор данных содержит 310 точек в 6-мерном пространстве. Рассматривая данный набор как многомерный объем в 6-мерном пространстве признаков, можно применять подход построения упругих карт, варьировать коэффициенты упругости карты для достижения наилучшего результата, далее проецировать точки многомерного объема на полученную карту и строить ее развертку в плоскости двух первых главных компонент.

На рисунке 7 представлена «мягкая» упругая карта для данного набора данных. Зеленые точки соответствуют категории «normal», красные точки соответствуют категории «abnormal». На рисунке 8 представлена развернутая карта с раскраской по плотности данных.

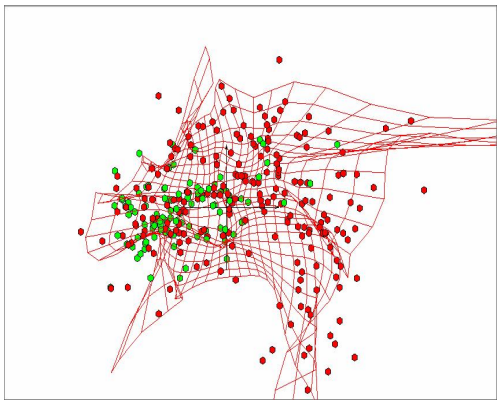

Рис. 7. Пример построения упругой карты для набора биомедицинских данных.

Полученные результаты дают возможность получить представление о взаимном расположении основных классов в изучаемом многомерном объеме. Однако на рисунке 8 видна область смешения данных из двух категорий. Для улучшения разделения необходимо в дальнейшем усовершенствовать алгоритм построения упругих карт за счет возможности сгущения исходной сетки карты в областях повышенной плотности данных. Подобная возможность широко используется в задачах математического моделирования сплошных сред. Автоматическое сгущение сетки в области высоких градиентов в вычислительной механике жидкости и газа является отдельным направлением, где реализовано большое количество эффективных и апробированных на практике алгоритмов.

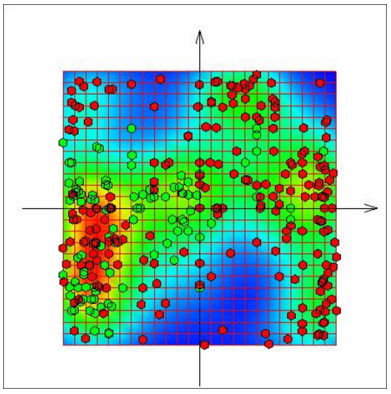

Рис. 8. Развертка предыдущей упругой карты на плоскость первых двух главных компонент с раскраской по плотности данных.

4. Заключение

Для анализа кластерных структур в многомерном объеме данных предлагается использовать технологии построения упругих карт, представляющие собой методы отображения точек исходного многомерного пространства на вложенные в это пространство многообразия меньшей размерности. Варьируя поверхность упругой карты за счет последовательного уменьшения коэффициентов упругости, можно добиться лучшего соответствия подстраивания карты под многомерное облако данных. После уменьшения коэффициентов изгиба и растянутости упругой карты, она становится более мягкой и гибкой, наиболее оптимальным образом подстраиваясь к точкам исходного многомерного объема данных. Применение технологий построения упругих карт для решения задач кластерного анализа не предполагает никакой априорной информации об изучаемых данных и не зависит от их природы, происхождения и т.п. Эти свойства позволяют применить технологии построения упругих карт для выявления кластерных структур. Для решения проблемы масштабируемости, когда упругая карта подстраивается как под область сгущения точек данных, так и к отдельно расположенным точкам облака данных, применяется подход «квази-зум» (quasi-Zoom). Суть подхода заключается в том, что для более тонкой подстройки в многомерном облаке данных выделяются большие кластеры, после чего проводится отдельное построение упругих карт для выделенных кластеров. Приведены примеры построения упругих карт для задач анализа текстов и для медицинских данных. Дальнейшее развитие используемого подхода представляется в данный момент в реализации такого алгоритма при построении упругих карт, где область высокой плотности данных рассматривалась бы аналогично области высоких градиентов в задачах математической физики, и предусматривала бы автоматическое сгущение сетки карты в подобных областях.

5. Благодарности

Данная работа выполнена при поддержке грантов Российского фонда фундаментальных исследований (проекты 16-01-00553а и 17-01-00444а).

6. Литература

[1] Bondarev A.E., Bondarenko A.V., Galaktionov V.A., Klyshinsky E.S. Visual analysis of clusters for a multidimensional textual dataset / Scientific Visualization. V.8, № 3, pp.1-24, 2016, URL: http://sv-journal.org/2016-3/index.php?lang=en

[2] Bondarev A.E., Galaktionov V.A. Analysis of Space-Time Structures Appearance for Non-Stationary CFD Problems // Proceedings of 15-th International Conference On Computational Science ICCS 2015 Rejkjavik, Iceland, June 01-03 2015, Procedia Computer Science, Volume 51, 2015, Pages 1801–1810.

[3] Bondarev A.E., Galaktionov V.A. Multidimensional data analysis and visualization for time-dependent CFD problems // Programming and Computer Software, 2015, Vol. 41, No. 5, pp. 247–252, DOI: 10.1134/S0361768815050023.

[4] Bondarev A.E. Visual analysis and processing of clusters structures in multidimensional datasets // Proceedings of the 2nd International ISPRS Workshop on PSBB, 15–17 May 2017, Moscow, Russia, ISPRS Archives, Volume XLII-2/W4, 2017, pp.151-154. http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W4/151/2017/

[5] Gorban A. et al. Principal Manifolds for Data Visualisation and Dimension Reduction, LNCSE 58, Springer, Berlin – Heidelberg – New York, 2007.

[6] Gorban A., Zinovyev A. Principal manifolds and graphs in practice: from molecular biology to dynamical systems International Journal of Neural Systems, Vol. 20, No. 3 (2010) 219–232.

[7] Rocha Neto A., Barreto G., 2009. On the Application of Ensembles of Classifiers to the Diagnosis of Pathologies of the Vertebral Column: A Comparative Analysis, IEEE Latin America Transactions, 7(4):487-496.

[8] Zinovyev A. Vizualizacija mnogomernyh dannyh [Visualization of multidimensional data]. Krasnoyarsk, publ. NGTU. 2000. 180 p. [In Russian]

[9] Zinovyev A. Data visualization in political and social sciences, In: SAGE «International Encyclopedia of Political Science», Badie, B., Berg-Schlosser, D., Morlino, L. A. (Eds.), 2011.

Об авторах

Бондарев Александр Евгеньевич, к.ф.-м.н., старший научный сотрудник ИПМ им. М.В. Келдыша РАН. Его e-mail: bond@keldysh.ru.

Галактионов Владимир Александрович, д.ф.-м.н., профессор, заведующий отделом компьютерной графики и вычислительной оптики ИПМ им. М.В. Келдыша РАН. Его e-mail: vlgal@gin.keldysh.ru.

Шапиро Лев Залманович, к.ф.-м.н., старший научный сотрудник ИПМ им. М.В. Келдыша РАН. Его e-mail: pls@gin.keldysh.ru.