ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 7 за 2016 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Кирина-Лилинская Е.П., Орлов Ю.Н., Федоров С.Л.

Метод базисных паттернов в анализе нестационарных временных рядов

Рекомендуемая форма библиографической ссылки: Кирина-Лилинская Е.П., Орлов Ю.Н., Федоров С.Л. Метод базисных паттернов в анализе нестационарных временных рядов // Препринты ИПМ им. М.В.Келдыша. 2016. № 7. 20 с. doi:10.20948/prepr-2016-7
URL: http://library.keldysh.ru/preprint.asp?id=2016-7

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Е.П. Кирина-Лилинская,
Ю.Н. Орлов, С.Л. Федоров

Метод базисных паттернов в анализе нестационарных временных рядов

Москва — 2016

Кирина-Лилинская Е.П., Орлов Ю.Н., Федоров С.Л.

Метод базисных паттернов в анализе нестационарных временных рядов

Даны определения и практические примеры приближенного разложения выборочной плотности вероятности по системе базисных паттернов для нестационарного временного ряда. Описана методика построения базисных паттернов для решения задачи статистического распознавания определенного типа состояния фрагмента временного ряда. Построена статистика, имеющая свойства предиктора разладки трендового состояния.

Ключевые слова: нестационарный временной ряд, базисные паттерны, согласованный уровень стационарности, распознавание образов

Kirina-Lilinskaya E.P., Orlov Yu.N., Fedorov S.L.

The basis patterns method for non-stationary time-series analysis

The definitions of approximate expansion of empirical distribution function density over special basis patterns are proposed for non-stationary time series. The method of statistical recognition of the local state in terms of distribution function is formulated. The statistics of trend prediction is constructed.

Key words: non-stationary time series, basis patterns, self-consistent stationary level, pattern recognition

Работа выполнена при поддержке гранта РФФИ, проект № 16-01-00342.

Содержание

Введение

Многие встречающиеся на практике временные ряды, нестационарные в смысле своей функции распределения, обладают тем не менее некоторыми стационарными свойствами. Во-первых, наблюдаемая в течение конечного промежутка времени последовательность значений случайной величины ограничена, поэтому из практических соображений удобно считать случайную величину равномерно ограниченной по времени, что выполняется во всяком случае на некотором горизонте прогнозирования. При этом выборочные моменты всех порядков, естественно, конечны. Во-вторых, зачастую наблюдаемые значения генерируются некоторой физической системой, для которой характерно пребывать в определенных состояниях, интересных для наблюдателя. Таковыми являются, например, метеорологические ряды данных (температура больше или меньше нуля, давление, характерное для циклона или антициклона), ряды приростов цен на бирже (тренды вверх или вниз) и т.п. Эти состояния, трактуемые как состояния с определенным распределением наблюдаемого параметра, могут быть описаны в терминах выборочной функции распределения этого параметра, а переход от одного состояния к другому – как разладка или эволюция нестационарного распределения.

В работе [1] было исследовано предположение о том, что для рядов вышеуказанного типа выборочные распределения, отвечающие явно выделяемым состояниям, таким как тренд вверх (вниз), кластеризуются. Эта гипотеза подтвердилась на достаточно большом количестве примеров. Не претендуя на всеобщность, можно во всяком случае сказать, что на практике существуют ряды, состояния которых в смысле фрагмента их траекторий достаточно хорошо соответствуют определенным эталонам в терминах функций распределения или их плотностей. Эталон (или базисный паттерн) распределения временного ряда представляет собой средневзвешенное состояние распределений фрагментов траектории случайного процесса, входящих в выделенный кластер. В таком подходе эталоны в виде функций распределения характеризуют типовые состояния изучаемой системы.

Текущее состояние, распознаваемое по близости выборочного распределения к эталонному в определенной норме, может относиться как к локально установившемуся эталонному состоянию, так и к переходному состоянию. Это последнее, в свою очередь, может быть близким к некому новому эталону, а может представляться в виде линейной комбинации уже имеющихся эталонов. Тем самым возникает задача оптимального, т.е. с наименьшей ошибкой, разложения текущего состояния по базисным паттернам. Представляет интерес ситуация, когда любое выборочное состояние может быть с заданной точностью представлено как линейная комбинация базисных паттернов. Эволюция коэффициентов разложения по такому базису заменяет эволюцию собственно функции распределения.

В данной работе изучаются аспекты приближенного разложения вектора состояния в виде дискретного набора вероятностей по паттернам, составляющим в определенном смысле базис состояний временного ряда.

1. Оптимальное разложение вектора по системе паттернов

Следуя [2], введем обозначения:

$f_N(x,t)$ – выборочная плотность функции распределения (ВПФР) по выборке длины N значений x в момент времени t;

$F_N(x,t)=\int_0^x f_N(z,t)dz$ – выборочная функция распределения (ВФР);

$\varphi_i(x)$ – i-й базисный паттерн в смысле плотности распределения.

В силу равномерной ограниченности случайной величины считаем, что $x\in[0;1]$. При численном анализе этот отрезок разбивается на n промежутков. Для краткости вероятность попадания в k-й промежуток обозначается так же, как и плотность распределения, т.е. $f_N(k,t)$. Если рассматривается выборка в фиксированный момент времени, то плотность обозначается $f_N(k)$.

Базисные паттерны корректно определены в том случае, если так называемый согласованный уровень стационарности (СУС, [2, 3]) фрагментов, образующих i-й базисный кластер, меньше, чем расстояние между разными паттернами. В норме L1 СУС $s(N)$ выборочных плотностей, построенных по выборкам длины N, определяется формулой

$$\int_0^{s(N)} g_N(\rho)d\rho=1-s(N)/2, \qquad (1.1)$$

где $g_N(\rho)$ есть плотность распределения расстояний между выборками длины N в данной норме. Если расстояние между распределениями вычисляется в гистограммной норме L1 через плотности распределений, то для построения соответствующей гистограммы выбирается количество классовых интервалов, оптимальным образом отражающее особенности рассматриваемого выборочного распределения. Число n интервалов разбиения гистограммы представляет собой размерность вектора состояния. При оптимальном разбиении величина $1/n$ совпадает со статистической неопределенностью оценки вероятности по выборке заданной длины N. В работе [4] показано, что для стационарного ряда $s(N)=2/n$. При больших длинах выборки наблюдается зависимость $n=AN^{1/3}$, где A есть эмпирический коэффициент порядка двойки.

Здесь следует сделать уточнение, касающееся представления базисного паттерна в виде гистограммы плотности вероятностей. Паттерн образован выборками, совокупная длина которых значительно больше, чем длина N анализируемого фрагмента, для которого и строится ВПФР $f_N(x,t)$. Поэтому паттерны $\varphi_i(x,t)$ определены значительно точнее в статистическом смысле, чем анализируемый фрагмент, длина выборки которого определяется

практическими соображениями: максимально допустимым запаздыванием в принятии решения, квазипериодичностью (возможной) функционирования системы и т.п. Если паттерн представляет функцию распределения, то мелкость разбиения области изменения случайной величины значения не имеет, и тогда совпадение размерностей векторов базиса и фрагмента реализуется автоматически в соответствии с требованиями к точности вычислений. Если же сравниваются плотности распределений, то паттерны должны быть представлены в укрупненном виде в соответствии с числом классовых интервалов, на которые оптимально разбит изучаемый фрагмент.

Векторы, расстояние между которыми меньше СУС, считаются с принятой точностью совпадающими. Это обстоятельство приводит к необходимости дать подходящее определение приближенного разложения вектора по базису, если сами базисные состояния известны не абсолютно точно. Поскольку рассматриваемые нами векторы представляют собой вероятности попадания значений случайной величины в заданные классовые интервалы, то координаты этих векторов неотрицательны, а их сумма равна единице. Если какой-либо такой вектор состояния разложен в линейную комбинацию других векторов состояния, то сумма коэффициентов разложения также равна единице. В случае если эти численно определяемые коэффициенты неотрицательны, их можно трактовать как вероятности реализации данных базисных состояний применительно к текущему состоянию. Именно в таких случаях разложение текущего выборочного состояния по системе паттернов имеет смысл. Если же какие-либо из коэффициентов оказались отрицательны, то выбранный базис не полон и требуется провести дальнейшую кластеризацию состояний. В дальнейшем для краткости, если это не препятствует пониманию, индекс N у вектора состояния $f_N(k,t)$ будем опускать.

Определение 1. Вектор состояния $\mathbf{f} \in R^n$ *считается* α*-разложенным по системе векторов* $\{\varphi_1,...,\varphi_p\}$, $\varphi_i \in R^n$, *если существуют такие неотрицательные числа* $y_1,...,y_p$, *что*

$$\left\| \mathbf{f} - \sum_{k=1}^{p} y_k \varphi_k \right\| \le \alpha, \quad \sum_{k=1}^{p} y_k = 1. \qquad (1.2)$$

В частном случае равномерного оптимального разбиения гистограммы на n классовых интервалов $\alpha \le 2/n$. При неравномерном разбиении под точностью понимается статистическая неопределенность гистограммы, построенной по выборке длины N.

Определение 2. Совокупность векторов $\{\varphi_1,...,\varphi_p\}$, $\varphi_i \in R^n$ *будем называть* α*-независимой (или* α*-базисом), если ни один из векторов этой совокупности не может быть* α*-разложенным по оставшейся системе векторов в смысле определения 1. В противном случае совокупность называется* α*-зависимой.*

Например, в [1] была выбрана типовая система паттернов плотностей функций распределения, отвечающая движениям траектории временного ряда, идентифицируемым как up (тренд вверх), down (тренд вниз) и flat (боковик).

Расстояния между каждой парой из этих трех паттернов в два-три раза превосходили СУС расстояний между фрагментами, образующими каждый из этих трех кластеров. Однако полусумма паттернов up и down с точностью порядка выбранного значения α давала паттерн flat, так что эти три паттерна оказались α-зависимы на данном уровне значимости.

Термин α-зависимости относится только к тем разложениям, которые имеют вероятностную интерпретацию, т.е. когда сумма коэффициентов разложения равна единице и сами коэффициенты неотрицательны. В противном случае, даже когда норма невязки мала, векторы не считаются зависимыми. Такая трактовка принята в силу специфики рассматриваемой задачи, когда раскладываются не произвольные векторы из некоторого евклидового пространства, а векторы, в виде которых представлены гистограммы вероятностей состояний изучаемой системы. Именно для таких векторов из определений 1 и 2 следует условие α-зависимости пары векторов: два вектора $\mathbf{f},\mathbf{g}\in R^n$ α-зависимы тогда и только тогда, когда $\|\mathbf{f}-\mathbf{g}\|\le\alpha$.

Заметим, что если установлена α-зависимость между плотностями, т.е. разложение $\mathbf{f}=\sum_{k=1}^{p}y_k\varphi_k+\mathbf{r}$ имеет норму невязки $\|\mathbf{r}\|\le\alpha$, то можно рассмотреть и разложение между функциями распределения, после чего использовать для установления величины невязки другую норму. Это может быть полезно для определения наилучшей нормы в смысле распознавания кластерной принадлежности текущего фрагмента.

В общем виде задача разложения вектора $\mathbf{f}\in R^n$ по заданному набору линейно независимых векторов $\{\varphi_1,...,\varphi_p\}$, $\varphi_i\in R^n$ сводится к нахождению вектор-строки $\mathbf{y}^T=(y_1,...,y_p)$, $p\le n$, минимизирующей в смысле 2-нормы функционал $\|\mathbf{f}-\Phi\mathbf{y}\|$, где $\Phi_{n\times p}$ есть матрица, столбцы которой составляют векторы φ_i. Минимизация этого функционала осуществляется ортогональным проектированием вектора $\mathbf{f}$ на p-мерное подпространство, натянутое на векторы $\{\varphi_1,...,\varphi_p\}$. Это проектирование представляет собой так называемое QR-разложение матрицы Φ в произведение специальной матрицы $Q_{n\times p}$, такой, что $Q^TQ=I_{p\times p}$ и верхней треугольной матрицы $R_{p\times p}$, что эквивалентно процессу ортогонализации Грама-Шмидта. В результате такого разложения получаем:

$$\mathbf{f}-\Phi\mathbf{y}=I\mathbf{f}-QR\mathbf{y}=\left(I-QQ^T+QQ^T\right)\mathbf{f}-QR\mathbf{y}=Q\left(Q^T\mathbf{f}-R\mathbf{y}\right)+\left(I-QQ^T\right)\mathbf{f}. \qquad (1.3)$$

Векторы, в виде суммы которых в последнем равенстве (1.3) представлено данное разложение, ортогональны:

$$\left(R\mathbf{y}-Q^T\mathbf{f}\right)^TQ^T\left(I-QQ^T\right)\mathbf{f}=\left(R\mathbf{y}-Q^T\mathbf{f}\right)^T\left(Q^T{}_{p\times n}I_{n\times n}-I_{p\times p}Q^T{}_{p\times n}\right)\mathbf{f}=$$

$$=\left(R\mathbf{y}-Q^T\mathbf{f}\right)^TO_{p\times n}\mathbf{f}=\mathbf{o}.$$

Второе слагаемое в (1.3) не зависит от $\mathbf{y}$. Следовательно, с учетом ортогональности указанных слагаемых, минимальное по $\mathbf{y}$ значение нормы $\|\mathbf{f}-\Phi\mathbf{y}\|$ равно норме этого второго слагаемого и достигается тогда, когда первое слагаемое равно нулю: $R\mathbf{y}-Q^T\mathbf{f}=\mathbf{o}$.

Итак, оптимальное разложение определяется вектором

$$\mathbf{y}_{opt}=R^{-1}Q^T\mathbf{f}. \tag{1.4}$$

Величина

$$\mathbf{r}=\mathbf{f}-\Phi\mathbf{y}_{opt}=\left(I-QQ^T\right)\mathbf{f} \tag{1.5}$$

есть невязка разложения (1.3). Ошибкой разложения считается 2-норма невязки, т.е. величина $\delta=\|\mathbf{r}\|=\left\|\left(I-QQ^T\right)\mathbf{f}\right\|$. Относительная ошибка определяется как

$$\varepsilon=\frac{\delta}{\|\mathbf{f}\|}=\frac{\left\|\left(I-QQ^T\right)\mathbf{f}\right\|}{\|\mathbf{f}\|}. \tag{1.6}$$

Пусть теперь, как это и бывает на практике, вектор текущего состояния $\mathbf{f}$ и матрица Φ базисных паттернов известны неточно. Неточность здесь имеет не измерительную, а статистическую природу, поскольку вместо генеральных совокупностей приходится иметь дело с выборочными распределениями. Возникает вопрос: как эта неточность повлияет на вычисление оптимального разложения, насколько эта процедура устойчива к малым (и не очень) возмущениям, какова в этом случае невязка? Положим

$$\xi=\max\left(\frac{\|\Delta\Phi\|}{\|\Phi\|},\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}\right) \tag{1.7}$$

и введем число обусловленности $\kappa(\Phi)$ матрицы Φ в смысле 2-нормы как отношение наибольшего и наименьшего ее сингулярных чисел. Поскольку матрица Φ по построению имеет полный столбцевой ранг, ее наименьшее сингулярное число строго больше нуля. Однако если базисные векторы оказываются близкими, то число обусловленности может быть очень большим. Согласно [5], 2-норма относительной вариации оптимального разложения оценивается сверху следующим образом:

$$\frac{\|\Delta\mathbf{y}\|}{\|\mathbf{y}\|}\le\xi\cdot\left(\frac{2\kappa(\Phi)}{\cos\theta}+\kappa^2(\Phi)tg\theta\right)+O\left(\xi^2\right), \tag{1.8}$$

где $\sin\theta=\varepsilon$ есть синус угла между раскладываемым вектором $\mathbf{f}$ и вектором $\Phi\mathbf{y}_{opt}$ в соответствии с (1.6). В результате может оказаться так, что разложение, например, в двумерное подпространство является более точным, чем в трехмерное. Это связано с тем, что если раскладывается n-мерный вектор, лежащий в плоскости двух базисных паттернов, но определенный с ошибкой порядка ξ (1.7), то возможен вычислительный артефакт нахождения большой проекции на третий паттерн согласно (1.8). Такой пример будет приведен ниже.

Описанный общий метод разложения вектора по заданному базису модифицируется применительно к рассматриваемым векторам, которые суть

гистограммные оценки плотности вероятности. Число базисных паттернов не может быть больше, чем число n классовых интервалов в их гистограммном представлении, поскольку оно заведомо не превосходит теоретического максимального расстояния между двумя плотностями в норме L1, которое равно двум, деленного на точность оценки гистограммы, равной по построению $2/n$. Но на практике максимальные расстояния между выборками, длина N которых не очень велика (например, $N=100$), составляют величину порядка 0,9, а СУС выборок таких длин близок к 0,3. Это означает, что актуальным является проецирование состояний на двумерный или трехмерный базис.

Выпишем поэтому конкретные формулы проецирования вектора состояния на два и три базисных паттерна. Рассмотрим сначала проекцию на два паттерна.

Итак, ищется оптимальное в смысле 2-нормы разложение вектора $\mathbf{f} \in R^n$ по базисным паттернам φ_1 и φ_2 следующего специального вида:

$$\mathbf{f} = y\varphi_1 + (1-y)\varphi_2. \tag{1.9}$$

Из-за условия равенства единице суммы коэффициентов разложения получается эффективное разложение не по двум, а всего лишь по одному вектору, который представляет собой разность двух базисных паттернов:

$$\mathbf{f} - \varphi_2 \equiv \psi = y(\varphi_1 - \varphi_2) \equiv y\varphi. \tag{1.10}$$

Условие минимума по y величины $\|\psi - y\varphi\|^2$ дает ортогональную проекцию вектора ψ на базисный вектор φ:

$$y_{opt} = \frac{(\psi, \varphi)}{\|\varphi\|^2}. \tag{1.11}$$

Тогда из (1.9) и (1.10) следует, что аппроксимация $\tilde{\mathbf{f}}$ состояния $\mathbf{f}$ имеет вид

$$\tilde{\mathbf{f}} = \frac{(\mathbf{f} - \varphi_2, \varphi_1 - \varphi_2)}{\|\varphi_1 - \varphi_2\|^2}\varphi_1 + \left(1 - \frac{(\mathbf{f} - \varphi_2, \varphi_1 - \varphi_2)}{\|\varphi_1 - \varphi_2\|^2}\right)\varphi_2. \tag{1.12}$$

Квадрат нормы невязки равен

$$\delta^2 = \|\mathbf{r}\|^2 = \|\mathbf{f} - \varphi_2\|^2 \sin^2(\mathbf{f} - \varphi_2, \varphi_1 - \varphi_2), \tag{1.13}$$

а относительная невязка равна

$$\varepsilon = \frac{\delta}{\|\mathbf{f}\|} = \frac{\|\mathbf{f} - \varphi_2\|}{\|\mathbf{f}\|} \cdot |\sin(\mathbf{f} - \varphi_2, \varphi_1 - \varphi_2)|. \tag{1.14}$$

Если полученная относительная ошибка не превосходит точность, с которой определено состояние $\mathbf{f}$, т.е. величину $2/n$, то формула (1.12) дает α-разложение этого состояния в смысле определения 1. Исходя из того, что минимум 2-нормы вектора состояния достигается на равномерном распределении и равен $1/\sqrt{n}$, а максимум нормы разности не превосходит $\sqrt{2}$, получаем, что ошибка аппроксимации имеет следующую оценку сверху:

$$\varepsilon^2 \le 2n\sin^2(\mathbf{f} - \varphi_2, \varphi_1 - \varphi_2). \tag{1.15}$$

Эффективность разложения определяется тем, насколько ε меньше, чем неопределенность ξ вектора состояния. Если $\varepsilon \le \xi$, разложение приемлемо. Поскольку неопределенность ξ имеет порядок $2/n$, то условие $\varepsilon \le 2/n$ означает, что квадрат синуса угла между векторами в (1.15) должен быть не больше, чем $2/n^3$. Исходя же из того, что число классовых интервалов, т.е. размерность вектора состояния, пропорциональна корню третьей степени из длины выборки, формирующей текущий фрагмент, получаем, что для хорошей аппроксимации требуется уменьшение квадрата синуса угла пропорционально обратной длине выборки, т.е. $\sin(\mathbf{f}_N - \varphi_2, \varphi_1 - \varphi_2) \propto 1/\sqrt{N}$.

Если же $\varepsilon > \xi$, то возможны несколько трактовок такого явления. Во-первых, это может считаться индикацией разладки системы в целом, когда состояние значительно выходит из плоскости базисных паттернов. Во-вторых, это может означать, что изначально был использован неполный базис состояний, и потому требуется построение еще одного вектора-паттерна. Наконец, возможен вариант, что данный подход не является адекватным для рассматриваемого случайного процесса.

Пусть формула (1.12) дает α-разложение состояния $\mathbf{f}$ по двум базисным паттернам φ_1 и φ_2. Тогда вероятность того, что это состояние является на самом деле состоянием φ_1, равна y_{opt} (1.11), а состоянием φ_2 – соответственно $1 - y_{opt}$. С точки зрения максимального правдоподобия состояние $\mathbf{f}$ следует распознавать как реализацию более вероятного паттерна. Легко показать, что в этом случае расстояние между вектором состояния и более вероятным паттерном наименьшее. Рассмотрим $\rho_i = \|\mathbf{f} - \varphi_i\|$, $i = 1, 2$. Пусть, например, $y_1 = y_{opt} > y_2 = 1 - y_{opt}$, т.е. $y_1 > 1/2$, и, следовательно, первый паттерн более вероятен как кандидат на идентификацию текущего состояния, чем второй. Тогда из (1.9) следует, что

$$\|\mathbf{f} - \varphi_1\| = (1 - y_{opt})\|\varphi_1 - \varphi_2\| < y_{opt}\|\varphi_1 - \varphi_2\| = \|\mathbf{f} - \varphi_2\|. \qquad (1.16)$$

Подчеркнем, что этот результат не зависит от нормы, в которой проводится идентификация текущего состояния.

Рассмотрим теперь проекцию на три паттерна. Ищется разложение вида

$$\mathbf{f} = y_1\varphi_1 + y_2\varphi_2 + (1 - y_1 - y_2)\varphi_3. \qquad (1.17)$$

Обозначим для краткости $\varphi_{ij} = \varphi_i - \varphi_j$. Тогда перепишем (1.17) в виде

$$\mathbf{f} - \varphi_3 \equiv \psi = y_1\varphi_{13} + y_2\varphi_{23}. \qquad (1.18)$$

Минимизируя по y_1, y_2 квадрат 2-нормы $\|\psi - y_1\varphi_{13} - y_2\varphi_{23}\|^2 \to \min$, получаем систему уравнений для определения коэффициентов y_1, y_2:

$$\begin{cases} y_1|\varphi_{13}|^2 + y_2(\varphi_{13}, \varphi_{23}) = (\mathbf{f} - \varphi_3, \varphi_{13}); \\ y_1(\varphi_{13}, \varphi_{23}) + y_2|\varphi_{23}|^2 = (\mathbf{f} - \varphi_3, \varphi_{23}). \end{cases} \qquad (1.19)$$

Матрица системы (1.19) есть матрица Грама α-независимой системы векторов $\{\varphi_{13}, \varphi_{23}\}$, поэтому ее определитель строго положителен:

$$G=\begin{vmatrix}(\varphi_{13},\varphi_{13}) & (\varphi_{13},\varphi_{23})\\ (\varphi_{13},\varphi_{23}) & (\varphi_{23},\varphi_{23})\end{vmatrix}=|\varphi_{13}|^2|\varphi_{23}|^2\sin^2(\varphi_{13},\varphi_{23})>0. \qquad (1.20)$$

Собственные значения матрицы Грама суть сингулярные числа $\sigma_{1,2}$ матрицы Φ, меньшее из которых может быть весьма мало, если угол между векторами в (1.20) находится на пределе разделимости состояний, определяемом СУС. Эти числа равны

$$\sigma_{1,2}=\frac{1}{2}\left(|\varphi_{13}|^2+|\varphi_{23}|^2\pm\sqrt{|\varphi_{13}|^4+|\varphi_{23}|^4+2|\varphi_{13}|^2|\varphi_{23}|^2\cos^2 2(\varphi_{13},\varphi_{23})}\right). \qquad (1.21)$$

Оптимальное разложение (1.17) определяется коэффициентами

$$y_{1opt}=\frac{|\varphi_{23}|^2(\mathbf{f}-\varphi_3,\varphi_{13})-(\varphi_{13},\varphi_{23})(\mathbf{f}-\varphi_3,\varphi_{23})}{G},$$
$$y_{2opt}=\frac{|\varphi_{13}|^2(\mathbf{f}-\varphi_3,\varphi_{23})-(\varphi_{13},\varphi_{23})(\mathbf{f}-\varphi_3,\varphi_{13})}{G}. \qquad (1.22)$$

В общем случае эти коэффициенты не обязаны быть неотрицательными числами, меньшими единицы. Если же вектор $\mathbf{f}$ оказался α-разложенным по базисным паттернам, то, как и в двумерном случае, базисный паттерн, к которому ближе всего вектор состояния, имеет наибольшую вероятность при распознавании, трактуемую как значение коэффициентов разложения.

2. Пример разложения состояния по двумерному базису

Проиллюстрируем описанный подход разложением фрагмента траектории конкретного временного ряда (биржевого индекса с минутным шагом по времени) по базису из трендовых паттернов, построенных для этого ряда.

Фрагмент изучаемого временного ряда показан на рис. 1. Для удобства проведено обезразмеривание данных на значение индекса в начальный момент. На этом рисунке видны участки преимущественного движения траектории вверх (промежутки с номерами по оси абсцисс 150-200, 370-420, 500-560) и вниз (соответственно, 240-350, 560-620), отбираемые экспертным образом. Отобрав достаточное количество фрагментов ряда для создания паттернов плотности распределения, отвечающих таким движениям, получаем распределения, приведенные на рис. 2.

Экспертный отбор фрагментов определенного типа вызван тем, что реализация безошибочного алгоритмического поиска нужной траектории весьма затруднительна. Предлагаемый в работе метод сравнения распределений с экспертно отобранными эталонами как раз и представляет один из способов распознавания требуемых образов после соответствующего «обучения» алгоритма.

Рис. 1 – Фрагмент изучаемого временного ряда

Эталоны плотностей распределений трендов вверх (up) и вниз (down) приведены на рис. 2. Для тренда вниз превалируют отрицательные приросты значений ряда, а для тренда вверх – положительные. Совокупная длина фрагментов для построения базовых паттернов составляет 3000 данных для каждого типа тренда. Расстояние между паттернами в норме L1 равно 0,62, а в норме L2 равно 0,24. Длина паттернов в норме L1 равна единице, а в норме L2 равна 0,37 для паттерна down (базисный вектор φ_1) и 0,36 для паттерна up (базисный вектор φ_2), т.е. 2-нормы паттернов примерно одинаковы. Косинус угла между паттернами равен $\cos(\varphi_1,\varphi_2)=0{,}78$.

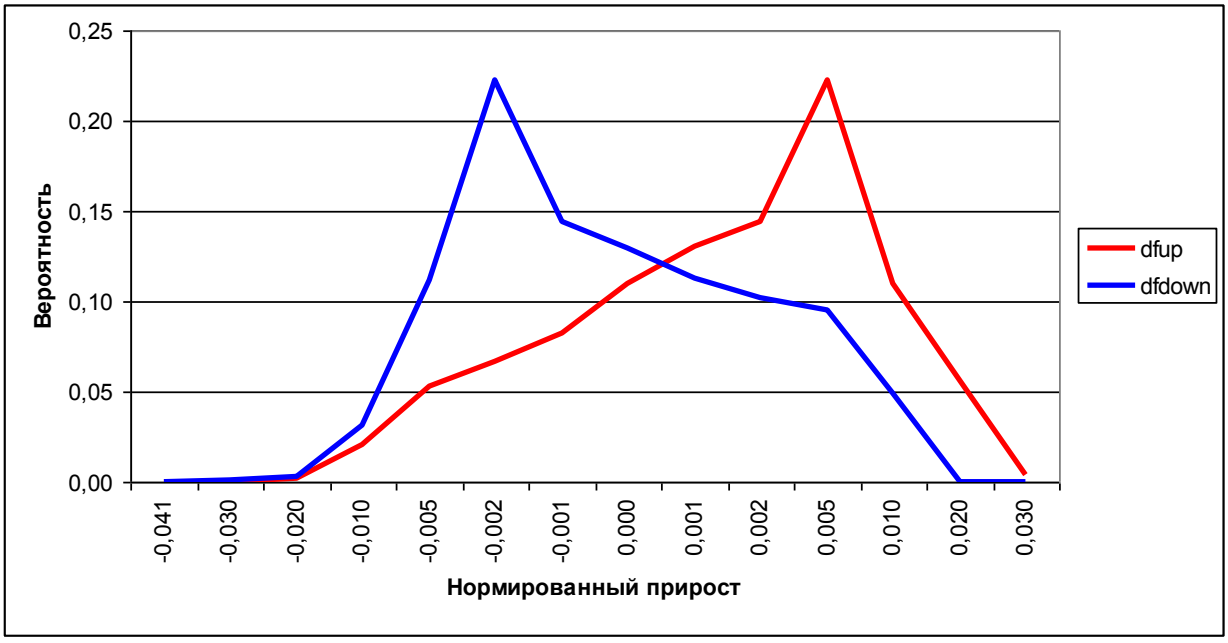

Рис. 2 – Базисные паттерны трендов вверх и вниз

При построении паттернов было проведено неравномерное разбиение гистограммы на классовые интервалы общим числом $n=15$. Малые приросты значений ряда весьма вероятны и потому требуют достаточно мелкой сетки, а большие приросты, встречающиеся относительно редко, можно относить к сравнительно крупным классовым интервалам.

Статистическая неопределенность выборочной плотности оценивается согласно методике, разработанной в [4]. Рассматривается *t*-статистика Стьюдента для выборки длины N с выборочным распределением $f_N(j)$ попадания значения ряда в j-й классовый интервал:

$$t=\sqrt{N-1}\frac{\left|f_N(j)-f^*(j)\right|}{s_N(j)}, \qquad (2.1)$$

где $f^*(j)$ есть гипотетическая генеральная вероятность, если бы процесс был стационарным, а

$$s_N^2(j)=f_N(j)\cdot\left(1-f_N(j)\right). \qquad (2.2)$$

На уровне значимости α выражение $\left|f_N(j)-f^*(j)\right|$ не превосходит величины $t_{1-\alpha/2}(N-1)s_N(j)/\sqrt{N-1}$, где $t_\alpha(N-1)$ есть α-квантиль распределения Стьюдента с $N-1$ степенью свободы. При больших N можно считать $N-1\approx N$ и число степеней свободы в квантиле распределения Стьюдента взять для простоты бесконечным (тогда это распределение совпадает с нормальным). Потребовав одновременно с критерием Стьюдента выполнения условия $\sum_{j=1}^{n}\left|f_N(j)-f^*(j)\right|\le\alpha$, получаем на уровне значимости α оценку

$$\frac{t_{1-\alpha/2}}{\alpha}\le\frac{\sqrt{N}}{\Sigma_N(n)}, \qquad (2.3)$$

где

$$\Sigma_N(n)=\sum_{i=1}^{n}s_N(j)=\sum_{i=1}^{n}\sqrt{f_N(i)\left(1-f_N(i)\right)}. \qquad (2.4)$$

Вычислив по конкретной выборке правую часть в (2.3), по таблице квантилей распределения Стьюдента (см., напр., [6]) определяем точность α оценки распределения по гистограмме. В [4] эта оценка точности применялась для построения равномерного разбиения на классовые интервалы. Здесь же мы применяем ее для произвольного разбиения. В частности, для фрагмента ряда длины 750, представленного на рис. 1, соответствующая ВПФР приведена на рис. 3. Для нее сумма в (2.4) приблизительно равна $\Sigma_{750}(13)=3$, так что правая часть в (1.25) равна 9, что дает оценку для точности $\alpha=0{,}12$.

Рис. 3 – Пример плотности распределения приростов временного ряда

Проведем теперь разложение этого вектора состояния по двум базисным паттернам, векторы которых представлены на рис. 2. Согласно формулам (1.9)-(1.14) получаем, что $y_{opt} = 0{,}24$, а относительная погрешность $\varepsilon = 0{,}45$. Эта погрешность существенно больше статистической неопределенности гистограммы данного состояния, равной, как было найдено выше, $\alpha = 0{,}12$. Таким образом, если не обращать внимания на точность, то с вероятностью $1 - y_{opt} = 0{,}76$ это состояние является трендом вверх, что в целом соответствует рис. 1. Но, строго говоря, полученный результат не является α-разложением вектора состояния. Поэтому значение вероятности тренда вверх на уровне 0,76 не является вполне обоснованным.

Чтобы оценить нижнюю границу максимального из коэффициентов y_{opt} или $1 - y_{opt}$, поступим следующим образом. Будем трактовать погрешность α как неопределенность в коэффициентах разложения и определим максимальную вариацию δy, при которой относительная вариация вектора разложения будет находиться в этих пределах. Для этого наряду с вектором разложения по формуле (1.12)

$$\tilde{\mathbf{f}} = y_{opt}\varphi_1 + (1 - y_{opt})\varphi_2$$

рассмотрим вектор

$$\delta\tilde{\mathbf{f}} = \delta y \cdot \varphi_{12} \tag{2.5}$$

и определим, при каких δy будет выполняться $\delta\tilde{\mathbf{f}} / \|\tilde{\mathbf{f}}\| \le \alpha$. Максимально допустимая вариация коэффициентов разложения определяется условием

$$|\delta y|_{\max} = \frac{\alpha|\tilde{\mathbf{f}}|}{|\varphi_{12}|}. \tag{2.6}$$

Если выяснится, что $y_{opt} - |\delta y|_{\max} < 1/2$ (если этот коэффициент в исходном разложении был наибольшим) или $1 - y_{opt} - |\delta y|_{\max} < 1/2$, то вывод о преимущественном тренде в соответствии с разложением (1.12) не является достоверным. В нашем примере $\varepsilon = 0{,}12, \left|\tilde{\mathbf{f}}\right| = 0{,}34, |\varphi_{12}| = 0{,}24$. В результате из (2.6) находим, что $|\delta y|_{\max} = 0{,}21$, так что с точностью α определения плотности вероятности тренд вверх может иметь вероятность всего лишь на уровне 0,55.

Следовательно, чтобы делать более обоснованные выводы, надо ввести еще один базисный паттерн или, возможно, несколько паттернов. Для отбора соответствующих фрагментов временного ряда надо выяснить, в какие моменты точность разложения выборки по двум паттернам наихудшая, после чего отобрать эти выборки и провести их кластеризацию. Рассмотрим поэтому более детально точность разложения состояния по двум паттернам.

3. Точность разложения выборочных состояний временного ряда

Поскольку рассматриваемый временной ряд нестационарный, точность разложения зависит от двух моментов времени – начала и конца выборки, что в наших терминах означает номер t текущего значения элемента ряда и длину N выборки назад от этого номера. Тогда $y_{opt} = y_{opt}(N,t)$ есть оптимальный коэффициент (1.11) разложения (1.10) состояния по двумерному базису.

На рис. 4 приведена поверхность $y_{opt}(N,t)$ для образца ряда длиной 7000 точек (за 10 торговых дней), а на рис. 5 – точность $\varepsilon(N,t)$ соответствующего разложения согласно (1.14).

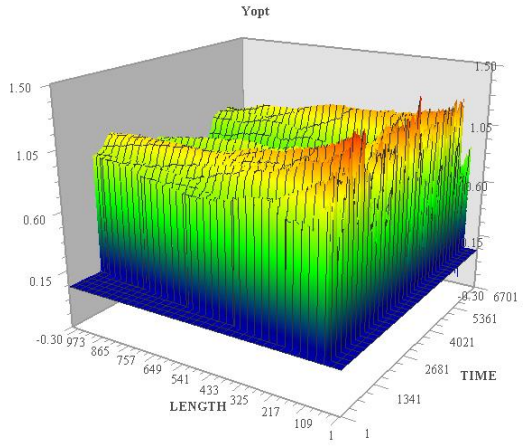

Рис. 4 – Коэффициент разложения состояния по двум паттернам в зависимости от времени и длины выборки

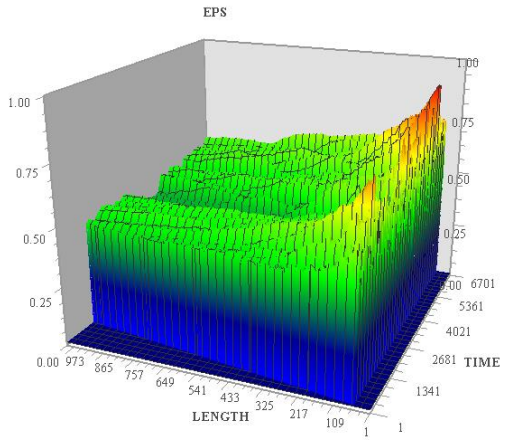

Рис. 5 – Норма относительной невязки разложения состояния по двум паттернам в зависимости от времени и длины выборки

Из рис. 5 видно, что в целом имеется тенденция снижения нормы невязки с увеличением длины выборки. Здесь следует подчеркнуть, что на практике часто требуется идентификация выборок относительно малых длин, чтобы принять управляющее решение с минимальным запаздыванием. Но из рис. 4 видно, что для таких длин (порядка 100 точек) разложение не только имеет большую невязку, но и в отдельных случаях не допускает вероятностной интерпретации в силу превышения коэффициентом $y_{opt}(N,t)$ значения единицы или снижения до отрицательных величин.

Для наглядности сечение поверхности $y_{opt}(N,t)$ плоскостью $N=100$ показано на рис. 6. Видно, что весьма часто наблюдаются ситуации, когда $y_{opt}(100,t)>1$. Но разложение устойчиво, т.е. нет хаотических всплесков в окрестности значений $y_{opt}(100,t)=1$. Это означает, что потеря вероятностной трактовки разложения связана не с вычислительной процедурой, а с тем, что состояние значительно удалено от плоскости базисных паттернов.

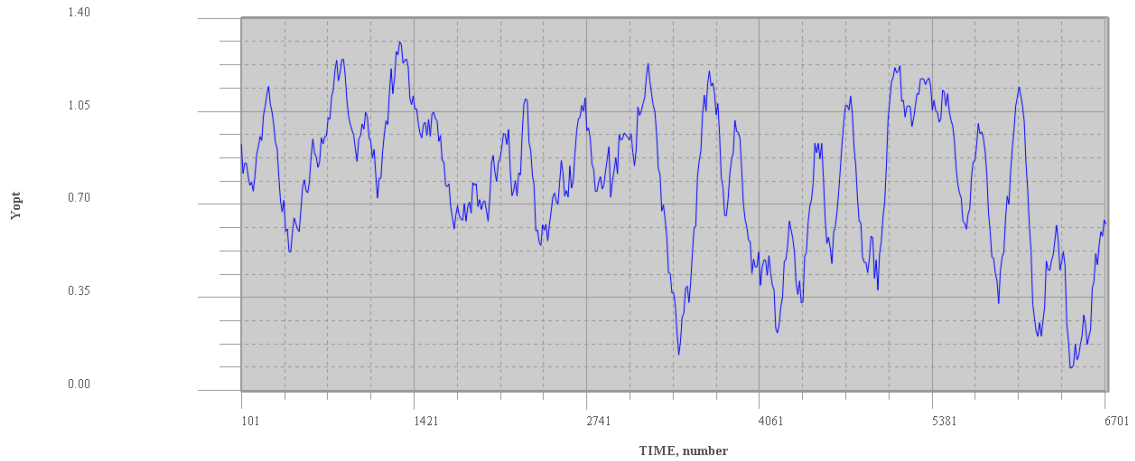

Рис. 6 – Коэффициент разложения состояния по двум паттернам для выборки длины 100

4. Разложение по трем базисным паттернам

Примеры, когда коэффициенты разложения становятся больше единицы или отрицательны, показывают, что двумерный базис может быть неадекватен. В частности, на рис. 1 таким состояниям отвечают выборки с номерами элементов 1-50 точек и 600-750. Видно, что это состояния с очень малой волатильностью, когда почти все распределение попадает в 5-6 классовых интервалов вблизи нуля. Кроме них, плохо идентифицируются состояния с аномально большой волатильностью, т.е. распределения с «толстыми хвостами». Низковолатильные состояния достаточно легко выделяются в виде паттерна (рис. 7), тогда как состояния с высокой волатильностью не образуют кластер в том смысле, что между распределениями этих фрагментов расстояния примерно такие же, как и между паттернами, что не позволяет уверенно распознавать такие фрагменты.

Рис. 7 – Паттерн «низковолатильного» состояния

Состояния с низкой или высокой волатильностью также являются важными для пользователя, но они не имеют «трендового» качества, т.е. могут принадлежать трендовым состояниям любой направленности. Выделяя такие фрагменты и объединяя получившийся паттерн с системой трендовых паттернов, мы получаем некоторый комбинированный базис, который, возможно, более точно может давать идентификацию текущего состояния. Такая гипотеза возникает при сравнении распределения фрагмента на рис. 3, имеющего три выраженных моды, и паттернов (рис. 2 и рис. 7), которые отвечают этим модам: левой моде отвечает паттерн down, правой – паттерн up, а центральной – паттерн низковолатильного состояния.

Разложим тогда вектор состояния по трем указанным паттернам согласно формулам (1.17)-(1.22). Для сравнения с разложением по двум паттернам

возьмем выборку длины $N=100$ в скользящем временном окне. Выборочные плотности распределения (т.е. векторы состояния) показаны на рис. 8.

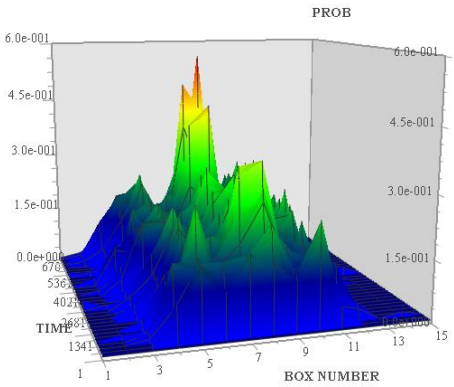

Рис. 8 – Выборочные плотности $f_{100}(k,t)$

При расчетах оказалось, что среднее значение ошибки разложения (1.6) практически не изменилось (уменьшилось с 0,53 для двух паттернов до 0,52 для трех). Но качество вероятностной интерпретации разложения заметно ухудшилось, что можно наблюдать на рис. 9-10, где показаны коэффициенты разложения $y_1(100,t)$ (аналог коэффициента $y_{opt}(100,t)$ на рис. 6) и $y_3(100,t)$.

Сравнивая эти графики, видим, что значения коэффициентов по совокупности гораздо чаще выходят за пределы промежутка $[0;1]$. Именно, в двумерном базисе лишь четверть (0,24) состояний не имела вероятностной трактовки, тогда как в трехмерном таких состояний стало почти в два раза больше (0,42), что не отвечает реальности. Это связано с тем, что при невысокой точности определения гистограммы, равной в данном случае 0,2 согласно (2.3), разложение по базису из близких векторов приводит к увеличению невязки, ибо число обусловленности для матрицы (1.20) равно 32.

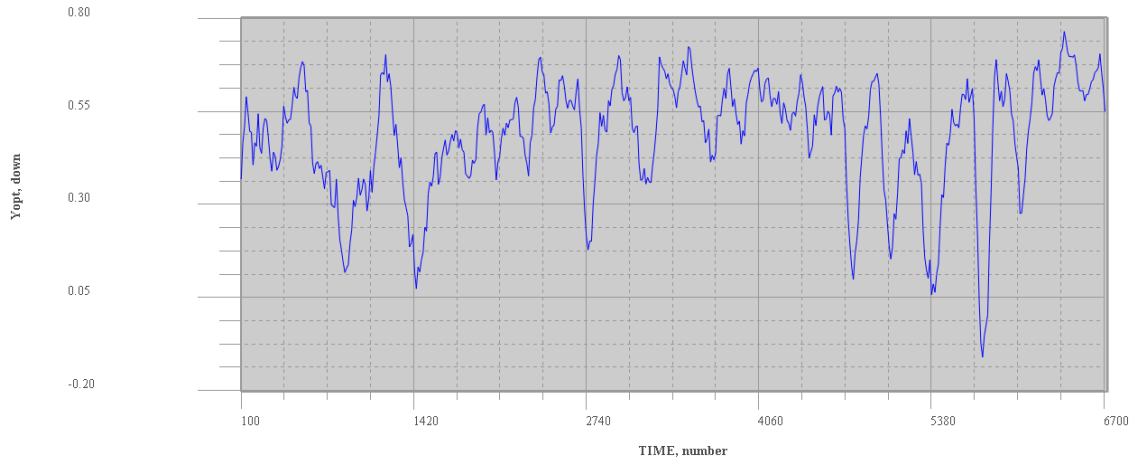

Рис. 9 – Коэффициент разложения состояния по трем паттернам (паттерн down) для выборки длины 100

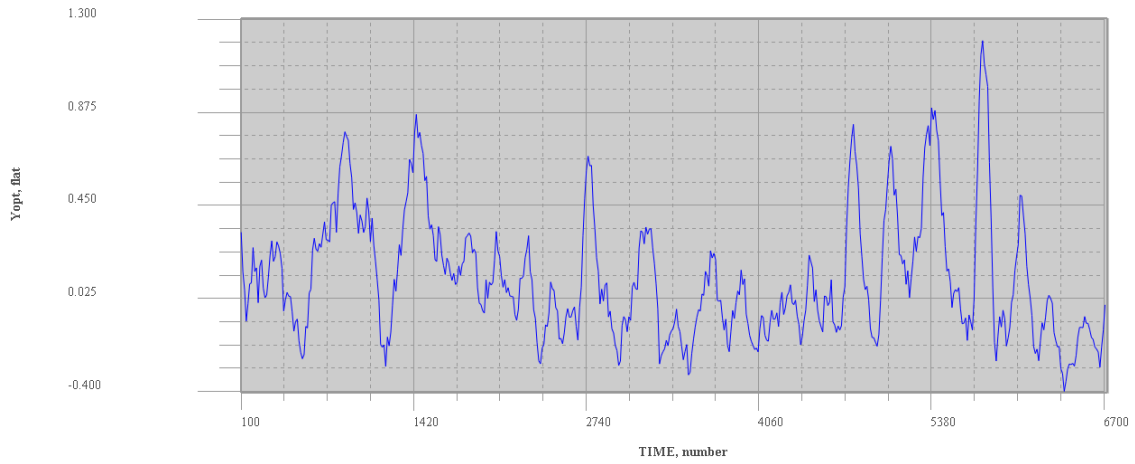

Рис. 10 – Коэффициент разложения состояния по трем паттернам (паттерн flat) для выборки длины 100

Таким образом, хотя типичные состояния рассматриваемого временного ряда кластеризуются, они не образуют α-базиса в смысле определений 1 и 2. Поэтому идентификацию состояний лучше проводить путем нахождения ближайшего паттерна в той или иной норме, как это сделано в [7], а не разлагать выборочную плотность вероятности по неполной системе паттернов. Тем не менее, сами коэффициенты разложения могут иметь определенный предсказательный смысл, для выяснения которого надо обратиться к анализу траектории, порожденной коэффициентами разложения выборочной функции распределения.

5. Динамика компонент разложения вектора по паттернам как предиктор состояния

Рассмотрим ряд из значений коэффициента $y_{opt}(N,t)$, который получается при разложении текущего состояния (1.9) по двум паттернам в скользящем окне длины N. Для $N=100$ такой пример приведен выше на рис. 6. Далее для определенности рассматривается именно этот ряд, так что соответствующие коэффициенты обозначаются просто $y(t)$.

Из вида траектории $y(t)$ следует, что она достаточно гладкая, с преимущественным сохранением направления движения. Расстояние между двумя последовательными пересечениями уровня «безразличия» $y=0{,}5$ меняется от 150 до 230 для скользящего окна длины 100, так что смена индикации тренда в данном подходе не является хаотическим процессом. Отметим, что выход траектории за пределы отрезка $[0;1]$ также не полностью случаен, а происходит по достижении предельных значений коэффициента $y(t)$. Следовательно, разладка трендового характера ряда может быть опознана с приемлемым уровнем задержки. Как было указано выше, доля нетипичных состояний составила 0,24 от общего числа, а на рис. 6 видно, что количество участков траектории вне отрезка $[0;1]$ равно 12 (разумеется, для

рассматриваемого временного ряда). Тогда одно небазисное состояние имеет в среднем длину 0,02 от исходного фрагмента в 7000 данных, т.е. составляет 140 минут. Поэтому в скользящем окне длины 100 можно по наблюдению за точкой поворота траектории $y(t)$ с достаточным запасом по времени предсказать смену тренда. Для этого следует построить удобную аппроксимацию траектории. Вид аппроксимации следует из рассмотрения совместного распределения значений $\{y(t), y(t+1)\}$, представленного на рис. 11. Траектория оказалась сильно коррелированной (носитель распределения сосредоточен на диагонали), так что связь между соседними по времени значениями линейна. Поэтому модель траектории можно приближенно задать в виде кусочно-треугольной периодической функции. В данном примере таких треугольников 12 (см. рис. 6). Предсказание состояния со сменой тренда будет вытекать из наблюдения в реальном времени за движением коэффициента $y(t)$ и оценкой параметров (длина, высота) такого треугольника.

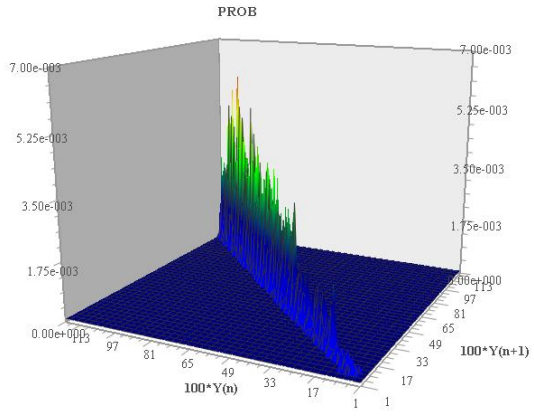

Рис. 11 – Совместное распределение вероятностей $\{y(t), y(t+1)\}$

Таким образом, исследование возможности разложения вектора состояния ВПФР по базисным паттернам привело к двум «новостям»: плохой и хорошей. Отрицательный результат состоит в том, что получить точное α-разложение в смысле определений 1 и 2 для биржевых рядов не удается, так как в тех случаях, когда коэффициенты разложения имеют вероятностную интерпретацию, невязка с реальным состоянием слишком велика. Увеличение базисных паттернов ухудшает точность идентификации состояния. Но есть и положительный результат: статистика коэффициентов разложения обладает выраженным квазипериодическим характером и позволяет оценить типичные периоды смены трендов при сканировании ряда выборкой определенной длины.

Описанный подход может быть применен для построения предиктора смены тренда, что представляет очевидную практическую важность.

Литература

1. Босов А.Д., Орлов Ю.Н., Федоров С.Л. О распределении рядов абсолютных приростов цен на финансовых рынках // Препринты ИПМ им. М.В. Келдыша. 2014. № 96. С. 1-15.
URL: http://library.keldysh.ru/preprint.asp?id=2014-96
2. Орлов Ю.Н. Кинетические методы исследования нестационарных временных рядов. – М.: МФТИ, 2014. – 276 с.
3. Орлов Ю.Н., Шагов Д.О. Индикативные статистики для нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша. 2011. № 53. С. 1-20.
URL: http://library.keldysh.ru/preprint.asp?id=2011-53
4. Орлов Ю.Н. Оптимальное разбиение гистограммы для оценивания выборочной плотности распределения нестационарного временного ряда // Препринты ИПМ им. М.В. Келдыша. 2013. № 14. С. 1-26
URL: http://library.keldysh.ru/preprint.asp?id=2013-14
5. Деммель Дж. Вычислительная линейная алгебра. Теория и приложения (пер. с англ.). – М.: Мир, 2001. – 436 с.
6. Гнеденко Б.В. Курс теории вероятностей. – М.: Физматлит, 1961. – 406 с.
7. Власюк А.А., Орлов Ю.Н. Точность идентификации выборочных распределений временных рядов в зависимости от типа распределения, нормы и длины выборки // Препринты ИПМ им. М.В. Келдыша. 2015. № 17. С. 1- 25.
URL: http://library.keldysh.ru/preprint.asp?id=2015-17