

*Московский государственный университет имени М.В. Ломоносова*

*Факультет вычислительной математики и кибернетики*

*Кафедра математической статистики*

**Алгоритм Скользящего Разделения Смесей  
 и его применение в анализе реальных данных.**

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**Выполнил:**

Студент 4 курса, ВВО

Птицын Евгений Генрихович

**Научный руководитель:**

д.ф.-м.н., заслуженный профессор

Королёв В.Ю.

Оглавление

[1 Введение 4](#_Toc9502221)

[2 Смесь нормальных законов распределения 5](#_Toc9502222)

[2.1 Определение 5](#_Toc9502223)

[2.2 Идентифицируемость смесей вероятностных распределений 9](#_Toc9502224)

[3 Волатильность 14](#_Toc9502225)

[3.1 Возможные подходы к определению и интерпретации волатильности 17](#_Toc9502226)

[3.1.1 Теоретико-вероятностные характеристики изменчивости случайного процесса 17](#_Toc9502227)

[3.2 Подход, основанный на модели смеси распределений вероятностей 19](#_Toc9502228)

[3.2.1 Декомпозиция волатильности 19](#_Toc9502229)

[3.2.2 Разложение классической волатильности на динамическую и диффузионную компоненты 22](#_Toc9502230)

[3.2.3 Преобразование волатильности при временном скейлинге 24](#_Toc9502231)

[4 Нахождение оценок максимального правдоподобия 27](#_Toc9502232)

[4.1 Метод максимального правдоподобия 27](#_Toc9502233)

[4.2 ЕМ-алгоритм 29](#_Toc9502234)

[4.3 Описание ЕМ алгоритма 29](#_Toc9502235)

[4.3.1 Е- и М- шаги 29](#_Toc9502236)

[4.3.2 Монотонность ЕМ-алгоритма 30](#_Toc9502237)

[4.3.3 ЕМ алгоритм как проксимальный 31](#_Toc9502238)

[4.3.4 Скорость сходимости 33](#_Toc9502239)

[4.3.5 Выбор начального приближения 34](#_Toc9502240)

[4.3.6 Правила остановки 35](#_Toc9502241)

[4.4 Применение ЕМ алгоритма 36](#_Toc9502242)

[4.4.1 Разделение конечных смесей нормальных распределений 36](#_Toc9502243)

[4.5 Модификации ЕМ алгоритма 39](#_Toc9502244)

[4.5.1 Медианные модификации ЕМ-алгоритма 39](#_Toc9502245)

[4.5.2 Первая медианная модификация 40](#_Toc9502246)

[4.5.3 Вторая медианная модификация 40](#_Toc9502247)

[4.5.4 Stochastic EM(SEM) алгоритм 40](#_Toc9502248)

[5 Скользящее (динамическое) разделение смесей вероятностных распределений. СРС-метод 42](#_Toc9502249)

[5.1 Общая идея метода, скользящего разделения смесей 42](#_Toc9502250)

[5.2 Оценивание предполагаемого диффузионного спектра с помощью CРС-метода 43](#_Toc9502251)

[5.3 СРС-метод с оцениванием значений компонент волатильности 45](#_Toc9502252)

[5.3.1 Критерий остановки 45](#_Toc9502253)

[5.3.2 Зависимость СРС-метода от выбора начальных приближений 45](#_Toc9502254)

[5.3.3 О влиянии ширины окна на результат применения СРС-метода 46](#_Toc9502255)

[6 Применение метода к анализу финансовых индексов 46](#_Toc9502256)

[6.1 Погода 59](#_Toc9502257)

[6.1.1 О влиянии масштаба (длины интервала времени между наблюдениями) на результат применения СРС-метода 63](#_Toc9502258)

[7 Реализация метода СРС с применением CUDA. 65](#_Toc9502259)

[8 Заключение 67](#_Toc9502260)

[9 Список источников 68](#_Toc9502261)

# Введение

Анализ данных в современном мире занимает все более решающую роль. Он позволяет получить дополнительную, а зачастую и основную информацию необходимой для принятия важных решений, таких как, например покупка акций.

Журнал Экономист выпустил статью *«Самый ценный ресурс больше не нефть, а данные»* (Economist, 2017), так же как нефть совершила индустриальную революцию и полностью изменила жизнь людей так же и анализ данных в XXI веке полностью меняет нашу жизнь. Появившиеся методы анализа данных позволяют производить анализ огромных массивов информации и извлекать все больше информации из имеющихся данных.

В данной работе рассматривается один из методов анализа данных, с помощью разложения волатильности на компоненты в модели смеси нормальных распределений — метод скользящего разделения смесей (СРС). Этот метод применим к стационарным процессам, таким как цены на финансовые инструменты на бирже.

Современные финансовые рынки представляют из себя очень сложные системы, состояние которых зависит от огромного количества факторов и параметров. Изменение цен на акции, фьючерсы и другие инструменты, равно как и составные инструменты индексы подвержено влиянию таких факторов, как например состояние экономики разных стран, политический климат, различные микро- и макроэкономические показатели, а также, не в последнюю очередь состояние самого рынка.

В отличии от волатильности как скалярного параметра, метод скользящего разделения смесей позволяет представить волатильность в виде разложения на динамическую и диффузионную составляющие. То есть предлагается метод разложения волатильности скалярного процесса в нетривиальную и существенно многомерную информативную картину.

Как будет показано на примере исследования конкретных временных рядов, описывающих эволюцию конкретных финансовых индексов, динамическая и диффузионные составляющие вносят в итоговую волатильность примерно одинаковый вклад, причем в разные моменты времени доминировать могут разные компоненты.

Целью данной работы является обзор применения метода скользящего разделения смесей в анализе данных, а именно получение практических результатов применения метода скользящего разделения смесей на различных данных.

В первой части работы дается теоретическое обоснование метода и его основных составляющих.

Во второй части работы приводится практическая сравнение свойств метода с использованием реальных финансовых данных (Индексы МосБиржи, NASDAQ и т.д.).

Основная идея и теоретическое обоснование метода скользящего разделения смесей было представлено в работе [Королёв, 2011]. Основной инструмент, использующийся для разложения — ЕМ алгоритм был описан в работе [Dempster, Laird, Rubin, 1977]

# Смесь нормальных законов распределения

## Определение

Прежде всего нам потребуется ввести понятие смеси нормальных законов распределения. Рассмотрим функцию вида , при этом множество имеет борелевскую -алгебру . Более того предположим что для каждого фиксированного функция представляет из себя функцию распределения по , а при любом фиксированном функция является измеримой по , то есть для любого выполняется условие . Пусть Q – вероятностная мера, определенная на измеримом пространстве . Тогда функцию распределения

называют смесью функции распределения по относительно Q. Распределение называется смешиваемым, тогда как мера Q задает смешивающее распределение. Если **Y –** m-мерная тождественная случайная величина (т.е. ), определенная на вероятностном пространстве , тогда функцию распределения можно записать в виде

Если — плотность распределения, соответствующая функции распределения ,

тогда смеси соответствует плотность

Если случайный вектор **Y** имеет дискретное распределение и принимает значения с вероятностями соответственно, то получится смесь вида

называемая дискретной. В этом случае функции распределения называются компонентами смеси , а числа называются весам соответствующих компонент,   
. Ели же в дискретной смеси число ненулевых весов конечно (то есть случайный вектор принимает конечное число значений), то дискретную смесь называют конечной.

Если функции распределения соответствует плотность , то дискретной смеси соответствует плотность

Особо подчеркнем, что при разных значениях параметра ***y*** функции распределения могут относиться к различным типам распределения. К примеру, если смесь дискретна, то есть мера Q приписывает вероятность точкам , причем , где — функции распределения, возможно, соответствующие разным типам при разных *j*, то

Более того, если в таком случае компоненты абсолютно непрерывны и имеют плотности , то смесь так же будет абсолютно непрерывной с плотностью

Особую роль играют сдвиг/масштабные смеси. Формально их можно определить следующим образом. Пусть в определении, сформулированном выше, *m =* 2. Предположим, что вектор ***y*** имеетвид.  
где и , так что функция распределения допускает представление

Тогда — это положительная полуплоскость, то есть , и функция распределения

называется *сдвиг/масштабной смесью* распределения *F* относительно Q. Здесь *u* — это параметр масштаба, а *v* — параметр сдвига (положения). Если функция распределения *F* имеет плотность *f*, то функция распределения соответствует плотность  
так что смеси соответствует плотность (2.1.7)

Если и *—* случайные величины, заданные на одном и том же достаточно богатом вероятностном пространстве, так что при каждом фиксированном значении пары случайных величин случайная величина имеет функцию распределения , то смесь (2.1.7) может быть записана в виде

Более того, в таком случае из теоремы Фубини вытекает, что функция распределения соответствует случайной величине , где случайная величина и случайный вектор стохастически независимы. Кстати, легко убедиться, что в таком контексте чисто сдвиговая смесь является не чем иным, как функцией распределения суммы двух независимых случайных величин и , то есть сверткой их функций распределения. В то же время, чисто масштабная смесь  
 является ничем иным, как функцией распределения произведения двух независимых случайных величин и .

Что бы определить дискретную сдвиг/масштабную смесь функции распределения , положим , где , и в качестве специального случая получим

Если при этом , то мы получаем чисто масштабную конечную смесь

Если функция распределения абсолютно непрерывна и имеет плотность , то смеси (2.1.10) функций распределения соответствует смесь плотностей

Физический смысл понятия смеси вероятностных распределений может быть проиллюстрирован на примере дискретной смеси. Рассмотрим некоторую популяцию, которая не является однородной и, в свою очередь, состоит из некоторого числа, скажем, субпопуляций. Предположим, что наблюдаемый признак (или наблюдаемая характеристика) внутри *j*-й субпопуляции распределен в соответствии с функцией распределения , которую можно интерпретировать как условную вероятность того, что значение наблюдаемого признака у случайно выбранного индивидуума будет меньше, чем *x*, при условии, что случайно выбранный индивидуум является представителем *j*-й субпопуляции. Пусть вероятность того, что при случайном выборе индивидуума из всей (генеральной) популяции будет выбран представитель именно *j*-й субпопуляции, равна . Тогда по формуле полной вероятности безусловная вероятность того, что значение наблюдаемого признака у индивидуума, случайно выбранного из всей генеральной популяции, будет меньше, чем , окажется равной

В определенном смысле операция смешивания вероятностных распределений обеспечивает возможность формально интерпретировать популяции, реально являющиеся неоднородными, как однородные. Очень часто нельзя непосредственно определить тип субпопуляции, к которой принадлежит очередное наблюдение. Вследствие этого вся (генеральная) популяция вынужденно считается однородной, хотя на самом деле она таковой не является и содержит индивидуумов, принадлежащих к существенно различным типам. Именно такая ситуация типична для анализа финансовых временных рядов. Поэтому чрезвычайно важно иметь возможность осуществить операцию, в некотором смысле обратную операции смешивания, а именно, операцию разделения (расщепления) смесей. Статистические процедуры, реализующие эту операцию, описываются ниже. Эти процедуры в значительной степени зависят от свойства идентифицируемости смесей вероятностных распределений.

## Идентифицируемость смесей вероятностных распределений

Понятие идентифицируемой смеси интенсивно используется в прикладных задачах, связанных с декомпозицией (разделением, разложением, расщеплением) совокупностей (популяций). В качестве примеров можно упомянуть задачи классификации, распознавания образов или идентификации вероятностных распределений. Определение идентифицируемых семейств смесей распределений вероятностей было предложено в работе (Teicher, 1961).

Пусть функция определена на множестве . Для простоты предположим, что при некотором и множество снабжено борелевской -алгеброй . Как и ранее, будем предполагать, что функция измерима по при каждом фиксированном x и является функцией распределения как функция аргумента x при каждом фиксированном . Пусть – семейство случайных величин, принимающих значения во множестве . Обозначим

Семейство , определяемое ядром и множеством , называется идентифицируемым, если из равенства

с , вытекает, что (здесь и далее символ  обозначает равенство по распределению, то есть совпадение распределений).

К примеру, идентифицируемыми являются конечные смеси нормальных распределений, показательных распределений, пуассоновских распределений и распределений Коши. Однако легко привести очень простые примеры неидентифицируемых семейств. В частности, в качестве ядра рассмотрим равномерное распределение и связанные с ним смеси

Здесь

В данной работе главным образом рассматриваются сдвиг/масштабные смеси, в которых , . Поэтому все, что на самом деле нам нужно, – это результаты об идентифицируемости семейств сдвиг/масштабных смесей одномерных распределений.

Если и – случайные величины, определенные на одном и том же достаточно богатом вероятностном пространстве так, что (i) случайная величина X стохастически независима от пары и (ii) для любых фиксированных значений случайной величины и случайной величины случайная величина имеет функцию распределения , то приведенное выше определение идентифицируемости применительно к сдвиг/масштабным смесям сводится к следующему. Сдвиг/масштабная смесь , порожденная ядром (сдвиг/масштабная смесь функции распределения ) идентифицируема, если из соотношения

где , — тройки случайных величин, обладающих точно такими же свойствами, что присущи описанным выше случайным величинам , вытекает, что

Сужение определения идентифицируемости на класс конечных сдвиг/масштабных смесей сводится к следующему.

Семейство смесей

порожденное ядром , идентифицируемо, если из равенства

вытекает, что

1. ;
2. для каждого индекса существует индекс такой, что

Семейство функций распределения называется аддитивно замкнутым, если для любых справедливо соотношение.

Здесь символ обозначает свертку распределений: если и – функции распределения, то

Иногда свойство (2.2.7) семейств распределений вероятностей называется воспроизводимостью по параметру .

Семейство нормальных законов с нулевым математическим ожиданием  
 является очевидным примером аддитивно замкнутого семейства (относительно дисперсии s).

Следующие результаты принадлежат Г. Тейчеру (Teicher, 1961).

*Теорема 2.2.1. Семейство смесей (2.2.1) функций распределения из аддитивно замкнутого семейства является идентифицируемым.*

Отсюда немедленно вытекает, что семейство масштабных смесей нормальных законов с нулевым средним идентифицируемо.

Смеси, порождаемые ядрами из аддитивно замкнутых семейств, конечно же, не исчерпывают все примеры идентифицируемых смесей. Рассмотрим масштабные смеси распределений, сосредоточенных на неотрицательной полупрямой.

*Теорема 2.2.2. Пусть , где – функция распределения такая, что . Предположим, что преобразование Фурье функции , нигде не обращается в нуль. Пусть – множество всех случайных величин таких, что . Тогда семейство смесей*

*идентифицируемо.*

Аналогичное свойство присуще некоторым семействам сдвиговых смесей распределений вероятностей.

*Теорема 2.2.3. Пусть Q – множество всех случайных величин. Семейство сдвиговых смесей*

*идентифицируемо, если характеристическая функция, соответствующая функции распределения , нигде не обращается в нуль.*

Теорема 2.2.1 гарантирует, что наряду со смесями нормальных законов с нулевым средним, идентифицируемыми являются семейства масштабных смесей любых строго устойчивых законов.

Теорема 2.2.3 гарантирует идентифицируемость семейств сдвиговых смесей любых безгранично делимых (в том числе устойчивых) законов.

Здесь упомянуты только те идентифицируемые семейства, которые так или иначе рассматриваются в данной работе в рамках описания СРС-метода.

К сожалению, примеры ядер, порождающих идентифицируемые семейства сдвиг/масштабных смесей в общей ситуации (то есть без каких-либо дополнительных условий на смешивающие распределения) не известны. Такие примеры известны лишь для дискретных смешивающих законов. В частности, справедливо следующее утверждение, доказанное Г. Тейчером (Teicher, 1963).

*Теорема 2.2.4. Семейство конечных сдвиг/масштабных смесей нормальных законов идентифицируемо.*

Покажем, что семейство сдвиг/масштабных смесей нормальных законов при произвольном (двумерном) смешивающем законе не является идентифицируемым. Пусть и – независимые случайные величины такие, что и имеют одно и то же стандартное нормальное распределение, , а случайная величина не вырождена, то есть ни при каком u не имеет места соотношение . Более того, предположим, что . Тогда распределение случайной величины  
 может быть записано как в виде

так и виде

Легко видеть, что смешивающие двумерные распределения в представлениях (3.2.3) и (3.2.4) различны. Таким образом, можно заключить, что в общем случае задача разделения сдвиг/масштабных смесей, то есть задача реконструкции двумерного смешивающего распределения по распределению смеси является некорректной, так как она может иметь несколько различных решений. Это справедливо даже для случая сдвиг/масштабных смесей нормальных законов. Идея, лежащая в основе описываемого СРС-метода, заключается в замене общей некорректно поставленной задачи аналогичной очень близкой корректной (см. теорему 3.2.4) задачей разделения конечных сдвиг/масштабных смесей нормальных законов. Этот подход, в свою очередь, основан на том обстоятельстве, что любое распределение вероятностей может быть как угодно точно приближено конечным дискретным распределением.

В терминах примера с популяцией, состоящей из нескольких (возможно, бесконечно большого числа) субпопуляций, идея данного подхода может быть пояснена следующим образом. Область значений наблюдаемых характеристик (возможно, континуальная) разбивается на счетное (возможно, конечное) множество групп так, что каждая группа соответствует близким (в определенном смысле) субпопуляциям. Другими словами, все субпопуляции разбиваются на счетное (возможно, конечное) число классов (кластеров). Таким образом, все субпопуляции, попавшие в один класс (кластер) считаются неразличимыми. Другими словами, выделяется счетное (возможно, конечное) множество типических субпопуляций. К примеру, все люди, населяющие нашу планету, различны. Более того, численность населения Земли такова, что эта общая популяция может считаться континуальной (каких бы двух людей с характеризующими их признаками мы ни взяли, практически всегда найдется третий человек с промежуточными признаками). Однако, как известно, все люди подразделяются на конечное число рас.

Идея замены исходной некорректной задачи корректной задачей разделения конечных смесей будет работать на практике, только если решение “суженной” корректной задачи будет близким к множеству решений исходной общей задачи. Другими словами, для того чтобы такой подход привел к разумным результатам, требуется, чтобы семейство сдвиг/масштабных смесей обладало свойством устойчивости относительно смешивающего распределения.

# Волатильность

Масса исследований посвящено моделированию и прогнозу волатильности финансовых инструментов, но в то же время очень мало теоретических моделей, объясняющих откуда вообще берется волатильность. Ролл (Roll, 1984) показывает, что волатильность зависит от микроструктуры рынка. Глостен и Милгром (Glosten, Milgrom 1985) показывает, что по крайней мере один источник волатильности может быть объяснен процессом обеспечения ликвидности.

Ни одно понятие финансовой математики в частности и теории финансов вообще не обладает столь противоречивыми толкованиями, как волатильность. Этот термин возникает в многочисленных контекстах и всегда означает нечто, связанное с неопределенностью, изменчивостью или риском. Однако единого универсального определения волатильности, справедливого во всех обстоятельствах, не существует. Разные источники по-разному трактуют волатильность. Для сравнения рассмотрим определения, приведенные в некоторых источниках.

Джон Халл (Hull, 2003) определяет волатильность как меру неопределенности дохода, реализуемого на данном активе (a measure of the uncertainty of the return realized on an asset).

На латыни слово *volatilis* означает крылатый, летающий, быстрый, быстротечный, мимолетный, временный, преходящий, эфемерный.

Англо-русский словарь В. К. Мюллера (Мюллер, 1967) дает следующие переводы слова “volatility”: 1) (хим.) летучесть; 2) изменчивость, непостоянство.

Словарь “Random House Websters’s concise dictionary” (Random House, New York, 1993) приводит такие интерпретации термина “volatile”: 1)evaporating rapidly; 2) tending or threatening to erupt in violence; explosive; 3) changeable; unstable, что дословно переводится на русский язык как 1) быстро испаряющийся; 2) стремящийся или грозящий излиться или извергнуться со стремительностью или неистовством; взрывоопасный; 3) изменчивый; неустойчивый. Эти интерпретации близки к оригинальным латинским корням слова “volatility”.

В англо-русском словаре математических терминов под редакцией П. С. Александрова, Л. Н. Большева и др. (Александров – ред., 1994) слово “volatility” трактуется как физический термин и переводится как летучесть. Интересно, что в этом словаре слово “volatile” связывается с “математическими машинами”, причем словосочетание “volatile memory” переводится как оперативная память.

Между тем, ни в русско-английском словаре математических терминов Ловатера (Boas, 1990), ни в англо-русском и русско-английском словаре по теории вероятностей, математической статистике и комбинаторике К. А. Боровкова (Боровков, 1994) слов “volatility” и “волатильность” вообще нет. Отсюда можно сделать очевидный вывод, что термин “волатильность” традиционно не считается математическим.

Тем не менее в финансовой математике в рамках некоторых специальных моделей эволюции финансовых индексов, таких как, скажем, геометрическое броуновское движение, волатильность отождествляется с коэффициентом диффузии, в свою очередь, отождествляемым со стандартным отклонением приращений соответствующих случайных процессов на временных интервалах единичной длины. В некоторых случаях этот коэффициент даже предполагается бесконечным, несмотря на то что в реальном мире бесконечно большие перемещения не осуществимы. Ниже мы остановимся на такой интерпретации волатильности подробнее. В качестве примера приведем ссылку на Wikipedia.org (март 2019), Волатильность является важнейшим финансовым показателем и понятием в управлении финансовыми рисками, где представляет собой меру риска использования финансового инструмента за заданный промежуток времени. Тут отмечается еще одна особенность: волатильность так же оказывается зависящей от временного масштаба.

В теории финансов существуют как три конкретных понятия волатильности: историческая волатильность, неявная или подразумеваемая волатильность и вариационная волатильность.

Существует немного методов оценивания исторической волатильности. Рассмотрим лишь некоторые из них. Само название историческая, явно указывает на то, что все эти методы применяют исторические, то есть статистические данные, представляющие собой временной ряд последовательно наблюдавшихся значений рассматриваемого процесса. Обычно, эти методы используются для описания настоящего или прошлого состояния рынка, но также могут быть использованы для прогнозирования будущего состояния.

Неявная (подразумеваемая) волатильность сильно отличается от исторической: она выражает ожидание рынка. Неявной волатильностью обычно называют такое значение волатильности исходных активов, при котором теоретически справедливая цена европейского опциона (вычисляемая по формуле Блэка–Шоулза) совпадает с его рыночной ценой. Неявная волатильность не только отражает ожидаемую волатильность процесса эволюции исходного актива, но и описывает предполагаемую природу развития процесса.

Ясно, что представление о волатильности, как о мере изменчивости процесса, неразрывно связано с выбором масштаба времени, поскольку процесс может развиваться с разной интенсивностью на временных отрезках разной длины. К примеру, если известны значения процесса

и на всех последующих интервалах единичной длины процесс развивается аналогичным образом, то волатильность на единичном интервале равна нулю, так как итоговое приращение процесса за единицу времени равно нулю (P (1) − P (0) = 0), в то время как волатильность на промежутках времени длины 0.1 отнюдь не равна нулю (приращения процесса на таких интервалах равны либо ±2, либо ±1). Понятие вариационной волатильности тесно связано с понятием вариации функции и несет в себе информацию о связи между скоростью изменения рассматриваемого процесса и длиной интервалов времени между последовательными наблюдениями (измерениями, отсчетами).

В данном разделе будут приведены и проинтерпретированы некоторые конкретные определения понятия волатильности с целью сравнения традиционных подходов с тем, который вытекает из рассматриваемой модели. В рамках данного подхода будет предложено более наглядное толкование волатильности по сравнению с традиционными определениями. Как будет показано, величине, обычно трактуемой как коэффициент диффузии, может быть придана многомерная форма, позволяющая получить разложение волатильности на динамическую и стохастическую (трендовую и диффузионную) составляющие. Это позволяет получить более подробную информацию о природе эволюции рассматриваемого процесса.

Интерес к понятию волатильности и важность этого понятия в теории финансов являются очевидными проявлениями следующего хорошо известного принципа: новая информация может быть получена только от изменяющихся объектов (в качестве примера еще одного проявления этой закономерности можно привести принцип достаточности в математической статистике). Таким образом, изменчивость (или волатильность) процесса содержит много информации (если не всю информацию) о его развитии. На упомянутом принципе базируются эффективные методы статистического анализа данных, например, дисперсионный анализ.

Обычно при классическом подходе к анализу эволюции финансовых индексов тренды и волатильности считаются в некотором смысле противоположными понятиями и анализируются отдельно. Ниже будет продемонстрировано, что на самом деле эти понятия нельзя противопоставлять друг другу. Будет показано, что, если существуют нетривиальные тренды, то неизбежно существует компонента волатильности, обусловленная этими трендами. Более того, существует явно выраженная взаимосвязь трендов и волатильностей, которая может быть довольно разумно описана с помощью аппарата теории вероятностей математической статистики.

## Возможные подходы к определению и интерпретации волатильности

### Теоретико-вероятностные характеристики изменчивости случайного процесса

Итак, собрав воедино приведенные выше толкования термина «волатильность», можно заключить, что ближе всего по смыслу к нему, по-видимому, такие термины как «изменчивость» или «подверженность изменениям». При этом, говоря, что один процесс более волатилен, чем другой, мы обычно подразумеваем, что значения одного процесса сильнее разбросаны вокруг своего тренда, нежели значения другого – вокруг своего, или, что то же самое, траектория одного процесса сильнее флуктуирует вокруг некоторого тренда, нежели траектория другого процесса. В теории вероятностей наиболее употребительной мерой разбросанности значений случайной величины является дисперсия. Эта характеристика разброса проста, ее определение прозрачно и хорошо согласуется со здравым смыслом. Дисперсия обладает удобными аналитическими свойствами, среди которых в качестве основного можно упомянуть аддитивность по отношению к стохастически независимым случайным величинам. Переход к той же размерности, что и наблюдаемые процессы, при использовании дисперсии как меры разброса также не вызывает затруднений и вполне естественно сводится к извлечению квадратного корня из дисперсии, называемого среднеквадратичным отклонением. Самый простой подход к строгому определению волатильности случайного процесса средствами теории вероятностей поэтому вполне естественно сводится к отождествлению волатильности процесса с его дисперсией (или, что практически то же самое, с его среднеквадратическим отклонением). Такое определение волатильности будем называть классическим. Однако, так как дисперсия существует далеко не для всех процессов, это определение формально применимо далеко не всегда. Например, для устойчивых процессов Леви, являющихся довольно популярными моделями динамики финансовых индексов, дисперсия не определена (бесконечна).

Для описания степени изменчивости (более точно, для описания ее зависимости от времени) случайных процессов в тех ситуациях, когда возможно отсутствие дисперсии, в физике часто используется понятие “ширины диффузионного пакета”. Чтобы его определить, рассмотрим такую меру рассеяния случайной величины как интерквартильный размах (interquartile range). Пусть для символ обозначает квантиль случайной величины порядка (-квантиль случайной величины ), то есть число, одновременно удовлетворяющее неравенствам

Если функция распределения случайной величины непрерывна, то -квантиль случайной величины удовлетворяет уравнению

Интерквартильный размах случайной величины определяется как

В прикладных исследованиях такая мера рассеяния часто используется в качестве альтернативы среднеквадратичному отклонению, в том числе в ситуациях, когда последнее не определено из-за наличия столь тяжелых хвостов, что второй момент отсутствует. Можно сказать, что интерквартильный размах в той же мере характеризует степень разброса возможных значений случайной величины, в какой медиана характеризует ее центральное значение: внутрь интервала случайная величина попадает с вероятностью и с такой же вероятностью она в этот интервал не попадает. При этом среди всех возможных интервалов с таким свойством (легко видеть, что все такие интервалы имеют своими концами квантили случайной величины порядков и при интервал занимает “срединное” положение, поскольку вероятность того, что случайная величина попадет левее этого интервала равна вероятности того, что случайная величина попадет правее этого интервала (при этом обе эти вероятности равны ). Поэтому длина этого интервала вполне может быть принята в качестве меры разброса случайной величины.

Для стандартно нормально распределенной случайной величины

В общем случае квантили определены неоднозначно. Однако можно считать, что для любой случайной величины и для любых , и в том смысле, что при любом наборе чисел всегда найдутся квантили случайных величин и порядка ,удовлетворяющие указанному равенству. Поэтому

Другими словами, интерквартильный размах пропорционален параметру масштаба. К примеру, если случайная величина имеет нормальное распределение с параметрами  
 и , то , где .

Можно показать, что коэффициент пропорциональности в соотношении , где – случайная величина с конечным вторым моментом, в общем случае не превосходит . А именно, справедливоn следующее утверждение.

/\*\*\*\*\*\*\*/

## Подход, основанный на модели смеси распределений вероятностей

### Декомпозиция волатильности

В рамках нового подхода традиционная интерпретация исторической (статистической) волатильности будет дополнена новой информацией о трендах и диффузиях. Более того, эта новая информация окажется двоякой: с одной стороны, появится возможность учитывать направление и величину трендов, а также интенсивность диффузий. С другой стороны, станет возможным учитывать важность имеющихся трендов и диффузий. Сам факт, что о трендах и диффузиях говорится во множественном числе, указывает на нечто принципиально новое в методах анализа волатильности. А именно, для описания волатильности будет использоваться многомерное представление в терминах компонент волатильности и их весов.

До сих пор волатильность описывалась с помощью скалярной величины, интерпретируемой либо как коэффициент диффузии базового процесса, либо, что ближе к реалиям, как среднее (в определенном смысле) значение абсолютной величины приращения базового процесса. После того, как в главе 2 было дано обоснование целесообразности моделирования распределений приращений рассматриваемых хаотических процессов с помощью смесей нормальных законов, можно сказать, что каждое приращение представляет собой реализацию “случайной” случайной величины. При этом смешивающее распределение задает вероятности, с какими приращение может оказаться реализацией той или иной случайной величины. Каждая из этих случайных величин имеет свое математическое ожидание, свое стандартное отклонение и соответствующий вес – меру “важности”. Таким образом, описание волатильности базового процесса сводится к описанию упомянутых параметров.

Напомним, что распределение некоторой случайной величины является сдвиг/масштабной смесью нормальных законов, если

где и – некоторые случайные величины, или конечной смесью нормальных законов, если

где – натуральное число,

Если распределение приращений индекса описывается общей сдвиг/масштабной смесью нормальных законов, то описание волатильности соответствует

заданию распределения параметров сдвига и масштаба. Однако, как уже отмечалось в разделе 4.2.1, в таком случае интерпретация волатильности затруднена, так как та характеристика, которая традиционно считается скалярной, становится функциональной. Хотя их трактовка остается, в принципе, той же самой, конечные смеси более наглядно интерпретируемы: если распределение приращений индекса описывается конечной сдвиг/масштабной смесью нормальных законов, то волатильность можно задать с помощью вектор-функции с тремя типами компонент: первый тип соответствует весам, два других содержат, соответственно, параметры сдвига и параметры диффузии смешиваемых нормальных законов.

Из сказанного в главах 1 и 2 вытекает, что если в качестве модели неоднородного случайного блуждания, описывающего рассматриваемый процесс эволюции финансового индекса или плазменную турбулентность, используются обобщенные процессы Кокса, то вид смешивающего распределения полностью определяется поведением накопленных интенсивностей и, следовательно, статистическими закономерностями изменений мгновенных (локальных) интенсивностей элементарных (допредельных) процессов. Мгновенные интенсивности случайных блужданий с непрерывным временем тесно взаимосвязаны с мгновенными коэффициентами диффузии и мгновенными трендами. В свою очередь, мгновенные коэффициенты диффузии и мгновенные тренды характеризуют типы “механизмов” (циклических, периодических, тенденциозных или с более сложным поведением), которые явно или неявно определяют функционирование финансового рынка (следует особо отметить, что под “механизмом” здесь подразумевается не физическая или экономическая структура, но абстрактная закономерность или подпроцесс). Напротив, каждый тип “механизма” характеризуется своей собственной мгновенной интенсивностью изменений. Можно считать, что в дискретных смесях вида (3.1.3) параметры и определяют -й тип “механизмов”, в то время как вес определяет долю “механизмов” -го типа в общей структуре рынка и может считаться характеристикой силы влияния “механизмов” -го типа на поведение рынка в целом по сравнению с другими “механизмами”.

Таким образом, статистическое оценивание параметров , и дает возможность, во-первых, выделить типические “механизмы” или структуры в функционировании анализируемой сложной системы и, во-вторых, оценить их относительную важность (или силу влияния).

### Разложение классической волатильности на динамическую и диффузионную компоненты

Предположим, не вдаваясь в более подробную детализацию, что распределение некоторой случайной величины представляет собой сдвиг/масштабную смесь некоторой функции распределения, скажем, . Это означает, что

где и – некоторые случайные величины. Наша цель – разложить стандартное отклонение этой случайной величины на компоненты, характеризуемые случайным сдвигом и случайным масштабом, соответственно.

Основываясь на теореме Фубини, можно утверждать, что смесь (4.3.1) – это распределение случайной величины , где случайная величина имеет распределение и стохастически независима от пары . При этих условиях справедлив следующий результат о разложении дисперсии случайной величины .

*Лемма 4.3.1. Предположим, что имеет место представление , в котором случайная величина стохастически независима от пары , причем случайные величины и обладают конечными моментами второго порядка. Тогда*

*Доказательство*. В силу независимости случайной величины и пары утверждение леммы непосредственно вытекает из следующих несложных выкладок:

Но для справедливо представление

Аналогично

что и завершает доказательство.

Искомое разложение содержится в следующем следствии.

Следствие 4.3.1. Пусть в дополнение к условиям Леммы 4.3.1 справедливы соотношения и . Тогда

Этот результат представляет собой обобщение основного соотношения дисперсионного анализа, в соответствии с которым полная дисперсия раскладывается на сумму внутригрупповой и межгрупповой дисперсий.

Конечная сдвиг/масштабная смесь нормальных законов представима в виде

где пара случайных величин имеет дискретное распределение

так что в соотношении (4.3.2)

где

Первое выражение в (4.3.3) зависит только от весов и параметров положения (сдвига) компонент и потому характеризует ту часть волатильности, которая обусловлена наличием локальных трендов, то есть “динамическую” компоненту волатильности, тогда как второе выражение в (4.3.3) зависит только от весов и параметров масштаба (“коэффициентов диффузии”) компонент и потому характеризует “чисто диффузионную” компоненту волатильности.

Если вспомнить традиционное одномерное представление о волатильности как о стандартном отклонении приращения процесса, то можно заметить, что следствие 4.3.1 уточняет это представление: волатильность процесса представляет собой корень квадратный из суммы двух компонент, первая из которых является характеристикой разбросанности локальных трендов, а вторая характеризует диффузию процесса. Если локальные тренды отсутствуют, то классическая волатильность равна корню квадратному из взвешенной суммы квадратов волатильностей компонент, причем веса компонент показывают важность соответствующей диффузионной компоненты.

Таким образом, обеспечена возможность получать больше информации по сравнению с исследованием обычной (одномерной) волатильности.

### Преобразование волатильности при временном скейлинге

Подобно классической волатильности, многомерная волатильность зависит от длины интервала времени между последовательными отсчетами. Поэтому чрезвычайно важно иметь возможность пересчета волатильности при изменении масштаба времени, чтобы получать дополнительную информацию о поведении изучаемого процесса при больших или меньших временных масштабах, не проводя дополнительных экспериментов и/или измерений.

Предложенная выше модель, основанная на представлении распределения приращений процесса в виде конечной сдвиг/масштабной смеси нормальных законов, позволяет это сделать.

*Лемма 4.3.2. Пусть X и Y – две независимые случайные величины с функциями распределения*

*и*

*Тогда сумма этих случайных величин имеет распределение*

*где параметры , и удовлетворяют следующим соотношениям:*

*если и , то*

*При этом динамическая и диффузионная компоненты волатильности композиции (свертки), соответственно, имеют вид*

*и*

*Где*

Д о к а з а т е л ь с т в о. Достаточно заметить, что характеристические функции случайных величин и , соответственно, имеют вид

и

в силу независимости слагаемых.

Если случайные величины и независимы и одинаково распределены, то мы получаем следующее утверждение.

*Следствие 4.3.2. В дополнение к условиям Леммы 4.3.2 предположим, что случайные величины и одинаково распределены. Тогда композиция (свертка) их распределений имеет вид*

*Параметры , и удовлетворяют следующим соотношениям: если*

*и , то*

*В частности, для совпадающих r и j имеем:*

*При этом динамическая и диффузионная компоненты волатильности композиции (свертки), соответственно, вдвое больше динамической и диффузионной компонент волатильности слагаемых.*

Утверждения, приведенные выше, позволяют пересчитывать «портреты волатильности» с меньших временных масштабов на большие. Обратную операцию, к сожалению, не всегда можно реализовать с абсолютной точностью. Действительно, легко видеть, что если число компонент смеси равно, скажем, k, то число компонент смеси, являющейся сверткой двух исходных, равно

Поэтому, если к экспериментальным данным удалось подогнать, скажем, n-компонентную смесь, то выполнить “развертку” смеси, то есть операцию, обратную вышеописанной свертке, можно как минимум, только если для этого n существует целое k, удовлетворяющее (4.3.4). Более того, по сути, “развертка” заключается в отыскании характеристической функции вида

обеспечивающей равенство

Где – характеристическая функция подогнанной смеси. Но функция, в точности равная квадратному корню из , может не быть характеристической, даже если соотношение (4.3.4) выполнено. Поэтому на практике “развертка” сводится к отысканию псевдорешения уравнения (4.3.6), например, по следующему правилу.

По заданному n определяется

А затем ищется функция вида (4.3.5), исходя из условия

где – какая-либо мера близости характеристических функций, например, сумма квадратов разностей их значений в каких-либо точках .

# Нахождение оценок максимального правдоподобия

## Метод максимального правдоподобия

Оценки максимального правдоподобия и основанные на функции правдоподобия выводы являются важнейшими в статистической теории и анализе данных. Метод максимального правдоподобия является базовым методом с хорошими свойствами. Это наиболее часто используемая методика в частотном анализе, и она может быть в равной степени применима к поиску апостериорного распределения в Байесовском анализе. (C. P. Robert, Bayesian Computational Methods)

Зачастую Байесовские решения оправдывают с помощью функции правдоподобия и оценок максимального правдоподобия (MLE), и Байесовские решения похожи на оценки правдоподобия со штрафом. Оценка максимального правдоподобия является универсальным методом и широко используется в любой области, где используются статистические методы.

Предположим, что наблюдаемые данные имеет плотность , где вектор, содержащий неизвестные параметры в постулированной форме для плотности **.** Наша цель максимизировать функцию правдоподобия как функцию , над параметрическим пространством . Так, что,

Или эквивалентно для логарифма правдоподобия,

Цель ОМП - оценить так, чтобы определить последовательность корней (2.1) согласованную и асимптотически эффективную. Подобная последовательность существует при подходящих регулярных условиях (Cramér 1946). Почти наверно, эти корни соответствуют локальному максимуму на . Для оценочных моделей в общем, функция правдоподобия имеет глобальный максимум на . Типичная последовательность корней, (2.1) с желаемыми асимптотическими свойствами, получается взяв как корень, который глобально максимизирует ; в этом случае, является Оценкой Максимального Правдоподобия. Далее будем называть – ОМП даже в ситуациях, когда она может не давать глобальный максимум правдоподобия. Несомненно, в некоторых примерах на смесях (McLachlan and Peel 2000, Chap. 3), функция правдоподобия не ограниченна. Однако, для этих моделей, при обычных условиях регулярности, могут существовать последовательности корней (2.1) со свойствами согласованности, эффективности и асимптотический нормальности. (McLachlan and Basford 1988, Chap. 12).

Когда функция правдоподобия или логарифмическая функция правдоподобия являются квадратичными в параметрах, как в случае независимых нормально распределенных наблюдений, ее максимум может быть получен решением системы линейных уравнений с параметрами. Однако, часто на практике функция правдоподобия не является квадратичной, что приводит к нелинейным проблемам в оценке максимального правдоподобия. Например: (а) модели приводят к математическим ожиданиям нелинейным в параметрах; (б) несмотря на возможную линейную структуру, функция правдоподобия не квадратичная по параметрам по причине не «нормальных» ошибок, отсутствующих данных или зависимостях.

Традиционно ОМП в этих ситуациях производится численно итерационным методом, решая уравнения, например методом Ньютона-Рафсона (NR) и его вариантами, такими как метод Фишера. При разумных предположениях на и достаточно точных начальные значениях, последовательность полученная методом Ньютона-Рафсона имеет локальную квадратичную сходимость к решению для (2.1). Квадратичная сходимость рассматривается как основное преимущество метода Ньютона-Рафсона. Но в применении эти методы могут быть утомительны аналитически и численно даже в сравнительно простых случаях. Смотри McLachlan and Krishnan (2008, Sect. 1.3) и Meng and van Dyk (1997). ЕМ-алгоритм предлагает лучшую альтернативу по многим критериям. Сейчас это популярный инструмент для итерационного нахождения ОМП в различных проблемах с отсутствующими или неполными данными.

## ЕМ-алгоритм

## Описание ЕМ алгоритма

### Е- и М- шаги

В результате действия ЕМ-алгоритма, представляющего собой итерационную процедуру, вычисляется последовательность значений параметра . Если задано некоторое значение , то вычисление следующего значения можно условно подразделить на два этапа, аббревиатура наименования которых и дала название всей процедуре. Опишем эти этапы.

**E-шаг.** Вычисление математического ожидания (Expectations). Вычисляется значение вектора скрытых переменных по текущему приближению вектора.

Определим функцию как условное математическое ожидание логарифма полной функции правдоподобия при известном значении наблюдаемой компоненты **X*:***

В этом определении *θ* являетсяаргументом функции , **X** и – параметры, так что соотношении (3.1) символ означает усреднение по **Y** относительно меры .

При известном значении **X = *x*** функцию можно вычислить по формуле

**М-шаг**. Производится максимизация функции правдоподобия и находится следующее приближение вектора θ.

Итерационный процесс останавливается в соответствии с заранее согласованным критерием остановки. Например, заранее выбирается какая-нибудь метрика и фиксируется малое положительное число ε. Процесс останавливается на m-ом шаге, если .

Заметим, что иногда название «ЕМ-алгоритм», данное в бумаге Dempster et al. (1977), объясняют не упомянутой выше аббревиатурой английских слов *Expectation-Maximization,* но возводят к термину *Estimation-Maximization.* (например (Айвазян и др., 1989)). По всей вероятности, первый термин все же лучше отражает суть ЕМ-алгоритма.

### Монотонность ЕМ-алгоритма

Свойство монотонности ЕМ-алгоритма было впервые установлено в работе (Шлезингер, 1965). Впоследствии это свойство обобщенных и модифицированных версий ЕМ-алгоритма систематически исследовалось в работах (Dempster, Laird and Rubin, 1977), (Everitt and Hand, 1981), (Boyles,1983), (Wu, 1983), (Redner and Walker, 1984), (Jordan and Xu, 1996), (Xuand Jordan, 1996).

Недавно в работах (Neal and Hinton, 1998) и (Chretien and Hero, 2000) было замечено, что ЕМ-алгоритм принадлежит к классу так называемых проксимальных алгоритмов (или PP-алгоритмов, от английского термина Proximal Point algorithms) (в статье (Neal and Hinton, 1998) отмечено соответствующее ключевое свойство ЕМ-алгоритма, но при этом ЕМ-алгоритм формально не идентифицирован как PP-алгоритм). Это замечание существенно упрощает исследование свойства монотонности ЕМ-алгоритма. При этом главную роль играет следующее представление функции .

Из соотношения вытекает, что

По этому

Согласно принятой выше терминологии, обычная статистическая процедура поиска оценок максимального правдоподобия направлена на максимизацию по логарифма неполной функции правдоподобия, который при известном значении равен первому слагаемому в правой части (3.2.1)

### ЕМ алгоритм как проксимальный

Рассмотрим расстояние Кульбака–Лейблера (Kullback and Leibler,1951) (см. также, например, (Кульбак, 1967), (Cover and Thomas, 1991)) между условными плотностями и , определяемое как условное математическое ожидание логарифма отношения правдоподобия при известном значении наблюдаемой компоненты относительно меры :

Расстояние Кульбака–Лейблера можно вычислить по формуле

При этом, применяя неравенство Иенсена, в силу вогнутости логарифмической функции легко убедиться, что

Таким образом, расстояние Кульбака–Лейблера удовлетворяет условиям:

1.   
2.

Положив

мы таким образом замечаем, что свойства 1 и 2 расстояния Кульбака–Лейблера соответствуют условиям, которые определяют штрафную функцию в определении проксимального алгоритма.

Осталось убедиться, что с и расстоянием Кульбака–Лейблера в качестве штрафной функции соотношение (5.3.11), определяющее проксимальный алгоритм, трансформируется в соотношение, определяющее М-этап ЕМ-алгоритма. Действительно, при таких и имеем

Заметим, что второе слагаемое в правой части последнего соотношения (равное дифференциальной энтропии условного распределения ) не зависит от . Поэтому

Но выражение в фигурных скобках в правой части оказывается в точности равным функции , откуда вытекает требуемое соотношение

Другими словами, рекуррентные соотношения  
и

задают одну и ту же последовательность, то есть ЕМ-алгоритм является специальным проксимальным алгоритмом. А так как любой проксимальный алгоритм обладает свойством монотонности, то свойство монотонности оказывается присущим и ЕМ-алгоритму в том смысле, что, если последовательность вычисляется в соответствии с правилами, определяющими ЕМ-алгоритм, то  
то есть

[1]

### Скорость сходимости

Скорость сходимости ЕМ-алгоритма обычно меньше, чем квадратичная, доступная с Ньютоновскими методами. В работе (Dempster et al., 1977) показано, что скорость сходимости ЕМ алгоритма линейна и скорость зависит от доли информации в наблюдаемых данных. Следовательно, в сравнении с задачей с полными данными, если большая часть данных отсутствует, сходимость может быть достаточно медленной.

Определим отображение из параметрического пространства ***Ω*** в само себя, таким, что . Функция ***М*** называется ЕМ отображением.

Если сходится к какой-то точке и непрерывна, тогда стационарная точка алгоритма; такая что, должна удовлетворять условиям . Разложив в ряд Тейлора в точке , в окрестности получим что

Где есть матрица Якоби для , имеющая (i,j)-й элемент равный

Где и *d* размерность **.** Таким образом, в окрестности , ЕМ алгоритм представляет собой, по существу, линейную итерацию с матрицей отношений поскольку обычно не равен нулю. По этой причине часто называют матрицей сходимости. Для вектора , мера фактически наблюдаемой скорости сходимости есть глобальная скорость сходимости определенная как

где это любая норма на *d*-мерном Евклидовом пространстве . Отмечено, что наблюдаемая скорость сходимости равна наибольшему собственному значению при определённых условиях регулярности (Meng and van Dyk 1997). Поскольку большое значение r подразумевает медленную сходимость, глобальная скорость сходимости определена как (Meng, 1994); см. так же McLachlan and Krishnan (2008, Sect. 3.9).

### Выбор начального приближения

ЕМ алгоритм будет сходиться медленно если выбрано плохое начальное значение . Действительно в некоторых случаях, когда функция правдоподобия стремится в бесконечность, на краю параметрического пространства, последовательность оценок генерированная ЕМ алгоритмом может расходиться если выбран слишком близко к границе. Так же с приложением, где уравнение правдоподобия имеет несколько корней, соответствующих локальным максимумам, ЕМ алгоритм следует применять из широкого набора начальных значений в поиске всех локальных максимумов. Вариант ЕМ алгоритма (Wright and Kennedy, 2000) использует метод интервального анализа, для отыскания стационарных точек логарифма функции правдоподобия внутри любого обозначенного региона параметрического пространства; см. также McLachlan and Krishnan (2008, Sect. 7.9).

Различные способы определения начального значения были рассмотрены специально в рамках модели смесей. С EMMIX программой (McLachlan and Peel 2000, pp. 343–344), значение начального параметра может быть получено, автоматически используя случайные части данных, или алгоритм кластеризации k-средних, или иерархический метод кластеризации. С случайными начальными значениями, эффект центральной предельной теоремы делает параметры изначально похожими, по крайней мере в больших выборках. С программой EMMIX, есть дополнительная опция для случайного старта, чтобы уменьшить этот эффект путем случайного выбора подвыборки из данных, которая потом случайно приписывается *g* компонентам. Как описано в McLachlan and Peel (2000, Sect. 2.12), подвыборка должна быть достаточно большой, чтобы убедиться, что первый M-шаг может произвести недегенеративную оценку вектора параметров .

Ueda and Nakano (1998) предложили алгоритм детерминированного отжига EM (DAEM) для того, чтобы итерационный процесс EM мог восстановиться после плохого выбора начальных значений. Они предложили использовать принцип максимальной энтропии и аналог статистической механики, когда, параметр, скажем θ, вводится с соответствующей «температуре» в смысле отжига. С их алгоритмом DAEM. E-шаг осуществляется путем усреднения по распределению взятому пропорционально к текущей оценке условной плотности полных данных (имея наблюдаемые данные) в степени θ; см пример McLachlan and Peel (2000, pp. 58–60). В 2005 Pernkopf and Bouchaffra (2005) соединили генетический алгоритм (GA) и EM алгоритм для подгонки смесей нормальных распределений, где предложенный алгоритм менее чувствителен к начальному приближению и позволяет выбираться из локально оптимальных решений.

### Правила остановки

Необходимо иметь разумные критерии для остановки итерационного ЕМ-алгоритма. Существует несколько подходов:

Выберем малое число .

1. Условимся останавливать ЕМ-алгоритм, если расстояние между эмпирической функцией распределения и теоретической смесью функций распределения, в которую подставлены параметры, вычисленные с помощью ЕМ-алгоритма, становится меньше .

Где

- теоретическая смесь нормальных функций распределения, в которую подставлены параметры, вычисленные с помощью ЕМ-алгоритма на m-итерации,

-эмпирическая функция распределения (, если , и в противном случае).

1. ЕМ-алгоритм можно останавливать, когда разность между значениями функции правдоподобия, вычисленными на соседних итерациях, мала, то есть дальнейшая работа ЕМ-алгоритма практически не увеличит текущее правдоподобие:
2. ЕМ-алгоритм можно останавливать, когда расстояние между оценками параметров, вычисленными на соседних итерациях мало, то есть дальнейшая работа ЕМ-алгоритма практически не изменит уже полученные оценки:

В соотношениях (3.3) и (3.5) нормы можно определять по-разному.

У всех перечисленных критериев остановки есть и достоинства, и недостатки.

Критерий (3.3) основанный на близости эмпирической функции распределения и теоретической смеси функции распределения, в которую подставлены параметры, вычисленные с помощью ЕМ-алгоритма, дает хорошую возможность судить о том насколько хороша полученная аппроксимация. Однако ЕМ-алгоритм может никогда не достичь таких значений параметров теоретической функции распределения, которые обеспечивают хорошее согласие с имеющимися наблюдениями.

Критерий (3.4), ориентирующийся на функцию правдоподобия естественно согласуется с методом, используемым в ЕМ-алгоритме, но как бы сводит все имеющиеся возможности, так сказать, степени свободы алгоритма, к единственному значению. Согласно этому критерию, алгоритм может быть остановлен если последовательно вычисленные значения *L* близки, но параметры далеки от точки максимума (функция правдоподобия может принимать близкие значения на 2х достаточно отдаленных друг от друга множествах значений параметров).

Критерий (3.5) использующий расстояние между последовательно полученными оценками вектора параметров, конечно же, остановит ЕМ-алгоритм, если последний достигнет точки максимума функции правдоподобия. К сожалению, этот критерий так же остановит алгоритм в случае, когда алгоритм достигнет области, далекой от точки максимума, но в которой скорость сходимости очень низка. [1]

## Применение ЕМ алгоритма

### Разделение конечных смесей нормальных распределений

Определим функцию плотности вероятности смеси нормальных распределений как

Где это p-мерная нормальная функция плотности вероятности с математическим ожиданием и ковариационной матрицей . Вектор неизвестных параметров состоит из пропорций смеси , элементов мат ожидания и отдельных компонент ковариационной матрицы **.** Задача оценки это пример проблемы разделения смесей или в языке распознавания образов «задача обучения без учителя».

Рассмотрим соответствующую задачу «обучения с учителем», где наблюдения над случайным вектором **,** такие, что . Здесь является ненаблюдаемым компонентами-индикатора вектором, где *i*-ый элемент берется нулем или единицей, в соответствии с тем, является ли j-м наблюдение приходит из i-го компонента .Задача нахождения Оценок Максимального Правдоподобия гораздо проще в данном случае. Векторы классификаторы можно называть отсутствующими данными. Задача обучения без учителя так же может называться задачей с неполными данными, а задача обучения с учителем, задачей с полными данными. Относительно простой итерационный метод для вычисления Оценок Максимального Правдоподобия для задачи обучения без учителя может быть дан, используя простоту метода Оценок Максимального Правдоподобия для обучения с учителем. Это суть ЕМ алгоритма.

Логарифмическая функция правдоподобия на полных данных для θ определяется как

ЕМ алгоритм для этой задачи начинается с некоторого начального значения . Так как линейная функция ненаблюдаемых данных *z* для задачи, вычисление на Е шаге осуществляется просто заменой его текущим условным математическим Ожиданием с учетом наблюдаемых данных y, которые являются обычно апостериорной вероятностью *j*-го наблюдения из *i*-го компонента.

Из (5.2) следует что

Для смесей с нормальными плотностями, численно выгоднее работать в терминах достаточных статистик (Ng and McLachlan 2003) данных

Дифференцируя (5.4) по θ на базисе достаточных статистик в (4.4.5), М шаг представляется как

Е и М шаги повторяют до выполнения условия сходимости. В отличии от метода Максимального Правдоподобия для задачи с учителем, на М шаге задачи без учителя, используются апостериорные вероятности . Вектор математических ожиданий и ковариационная матрица вычисляются, используя как веса во взвешенных средних.

В случае неограниченной компонент-ковариационной матрицы , не ограничена, так как каждая точка данных дает прирост к сингулярности на границе параметрического пространства. (McLachlan and Peel 2000, Sect. 3.8) На практике, компонент-ковариационные матрицы может быть наложено ограничение, , где не определена. В этом случае гомоскедастических нормальных компонент, обновленная оценка общей компонентно-ковариационной матрицы определяется как  
где определяется (5.6), а обновление и также как выше в гетероскедастичном случае.

## Модификации ЕМ алгоритма

### Медианные модификации ЕМ-алгоритма

Как было экспериментально установлено, ЕМ-алгоритм обладает сильной неустойчивостью по начальным данным. Например, в случае четырехкомпонентной смеси нормальных законов при объеме выборки 200–300 наблюдений замена лишь одного наблюдения другим может кардинально изменить итоговые оценки, полученные с помощью ЕМ-алгоритма. Возможно, эта неустойчивость обусловлена тем, что стандартные (наиболее правдоподобные для случая нормального распределения) оценки математического ожидания и дисперсии (среднее арифметическое и выборочная дисперсия) при “засорении” (контаминации) выборки “посторонними” или “паразитными” наблюдениями становятся заметно менее эффективными по сравнению со, скажем, выборочной медианой. Этот эффект обнаружен Дж. Тьюки (Tukey, 1960) и описан, например, в (Айвазян и др., 1983) и (Королев, 2006). Формально модель контаминации Тьюки сводится к тому, что вместо “чистого” модельного распределения, интерпретируемого как однородная модель, в качестве модельного распределения рассматривается неоднородная модель, имеющая вид смеси исходного “чистого” распределения и некоторого другого закона, описывающего “засоряющие” наблюдения. В задаче разделения смесей по самой сути модели, когда оцениваются параметры одной компоненты смеси, наблюдения с распределениями, соответствующими другим компонентам, являются «загрязняющими». Это обстоятельство может сыграть особенно важную роль при реализации SEM-алгоритма, описываемого ниже.

Для противодействия указанной неустойчивости ЕМ-алгоритма можно использовать медианные модификации ЕМ-алгоритма. Смысл этих модификаций в том, что наиболее “неустойчивые” этапы выполнения ЕМ-алгоритма заменяются более устойчивыми. В частности, на М-этапе неустойчивые моментные оценки наибольшего правдоподобия (которые для нормальных компонент минимизируют квадратичный риск) заменяются более устойчивыми (робастными) оценками медианного типа, оптимальными в смысле среднего абсолютного отклонения.

Далее будут приведены конечные результаты для смеси нормальных распределений. Математическое обоснование данного метода можно посмотреть в [1].

### Первая медианная модификация

Введем следующие величины:

имеющие смысл некой вероятности. Введем также «фиктивные» случайные величины принимающие значения с вероятностями . Переупорядочим значения случайной величины по неубыванию, одновременно переставляя соответствующие данным значениям вероятности. Пусть — вероятность, соответствующая значению. Положим:

Тогда

Пусть тогда

### Вторая медианная модификация

Матожидание оценивается так же, как и в первой медианной модификации. Рассмотрим дисперсию .

Обозначим

Тогда

### Stochastic EM(SEM) алгоритм

Классический ЕМ-алгоритм относится к категории так называемых «жадных» алгоритмов (greedy algorithms) в том смысле, что он «бросается» на первый попавшийся локальный максимум. Другими словами, будучи методом *локальной оптимизации*, он приводит не к глобальному максимуму функции правдоподобия, а к тому локальному максимуму, который, грубо говоря, является ближайшим к начальному приближению.

Самый простой способ противодействия этому свойству заключается в том, чтобы, не ограничиваясь единственным начальным приближением и, соответственно единственной траекторией ЕМ-алгоритма, реализовать несколько траекторий, задавая (например, случайно) нескольких различных начальных приближений, а затем выбрать тот из результатов, для которого правдоподобие является наибольшим среди всех реализованных траекторий ЕМ-алгоритма. Однако при таком подходе остаётся неясным ответ о том, каким механизмом разумнее всего пользоваться при переходе от одного начального приближения к другому. В частности, когда начальное приближение задается случайно, без дополнительной информации нельзя исчерпывающим образом определить распределение вероятностей, в соответствии с которым следует генерировать очередное начальное приближение.

Другой, оказавшийся весьма эффективным, способ заключается как бы в случайном, но целенаправленном «встряхивании» наблюдений (выборки) на каждой итерации. Этот способ лежит в основе SEM-алгоритма, название которого является аббревиатурой термина Stochastic EM-algorithm (Стохастический (или случайный) ЕМ-алгоритм). SEM-алгоритм, предложенный в работах (Broniatowski, Celeux and Diebolt, 1983), (Celeux and Diebolt, 1984), (Celeux and Diebolt, 1985), весьма прост.

* SEM работает относительно быстро, и его результаты практически не зависят от начального приближения;
* как правило, SEM находит экстремум , близкий к глобальному.

Пусть вся выборка разбита на кластеры : каждый элемент относится к единственному(!) кластеру , то есть утверждается, что данный элемент взят из компоненты смеси.

S-шаг

На первом этапе SEM-алгоритма производится так называемое стохастическое моделирование. Для каждого , m генерируется вектор как реализация случайного вектора из полиномиального распределения с параметрами 1 и , где — это вероятность того, что величина равна 1. По векторам определяется разбиение выборки на кластеры и соответствующие численности кластеров .

Е-шаг

Остается без изменений.

М-шаг

Пересчитываются веса:

и вместо максимизации взвешенного правдоподобия:

решается задача обычного невзвешенного правдоподобия:

Для SEM-алгоритма также существуют медианные модификации. Ознакомиться с ними можно в [1].

/\* дописать про median SEM\*/

# Скользящее (динамическое) разделение смесей вероятностных распределений. СРС-метод

## Общая идея метода, скользящего разделения смесей

Общая идея метода скользящего (динамического) разделения смесей (СРС-метода) такова. Как уже отмечалось, есть все основания считать, что не существует универсальной смеси нормальных законов, которая может адекватно описать состояние рынка. По всей видимости, все параметры, включая – количество компонент смеси, то есть число параметров смеси само по себе, – изменяются во времени. Следовательно, предметом исследования должно стать описание характера изменения параметров смеси во времени. При этом следует опираться на принцип “локальной эргодичности”, согласно которому усреднение вероятностных характеристик наблюдаемого процесса “по ансамблю” (то есть по пространству элементарных исходов) можно заменить усреднением этих характеристик по времени. К такому принципу приходится вынужденно прибегать при применении практически любого метода анализа временных рядов, ведь в наличии имеется только одна реализация случайного процесса – временной ряд. С этой целью параметры смеси будут оцениваться по выборке, представляющей собой некоторый отрезок временного ряда, содержащий наблюдения, накопленные в течение некоторого интервала времени фиксированной длины. Затем интервал будет сдвигаться (скользить) в направлении астрономического времени, и процедура оценивания параметров смеси будет повторяться всякий раз при очередном сдвиге интервала. Таким образом, используемый метод будет называться методом скользящего (динамического) разделения смесей.

При этом для интерпретации получаемых результатов очень важно, что, коль скоро был принят принцип “локальной эргодичности”, то статистические характеристики изучаемого процесса (вынужденно) считаются неизменными внутри каждого интервала времени – отрезка временного ряда, используемого для текущего оценивания параметров смеси. Практическая адекватность этого принципа определяет длину скользящего интервала, также называемого окном (при этом говорят не о длине интервала, но о ширине окна). Очевидно, что, чем ширина окна меньше, тем больше оснований считать принцип “локальной эргодичности” выполненным. К сожалению, также довольно очевидно, что при уменьшении ширины окна одновременно уменьшается точность получаемых оценок параметров смеси, так как уменьшается объем выборки. Отыскание требуемого баланса – это прерогатива исследователя в зависимости от целей анализа.

Основным объектом проводимого анализа является временной ряд (реализация процесса), состоящий из приращений индекса, наблюдаемого в равноотстоящие моменты времени

Здесь M – длина исходного временного ряда (объем выборки), далее будем иметь дело лишь с временными рядами указанной структуры.

## Оценивание предполагаемого диффузионного спектра с помощью CРС-метода

Поскольку заранее нет информации о том, как эволюционируют параметры рассматриваемых процессов, до поры до времени придется руководствоваться некоторыми упрощающими предположениями. Несмотря на то, что уже имелась возможность (например, на рис. 1) убедиться в том, что смешиваемые распределения (компоненты смеси) не всегда имеют нулевые или почти нулевые средние, сначала предположим, что смешивающие распределения имеют нулевые математические ожидания. В обоснование этого предположения можно заметить, что, как правило, в реальных процессах, описывающих более или менее стабильное развитие рассматриваемой системы, компоненты с ненулевыми средними играют второстепенные роли (их веса относительно невелики). Другими словами, сейчас внимание будет сосредоточено на изучении предполагаемых диффузионных спектров рассматриваемых процессов и будем оценивать лишь диффузионные компоненты волатильности в предположении отсутствия (равенства нулю) динамических компонент. Это существенно упростит визуализацию полученных результатов. Несмотря на то, что такое предположение неизбежно ведет к некоторой потере общности, полученных в рамках такого ограничения результатов будет все же достаточно для отладки СРС-метода и иллюстрации его возможностей. С другой стороны, в разделе 4.3.6 уже было показано, что иногда такое упрощение может оказаться весьма полезным с точки зрения интерпретации результатов. Позднее вернемся к общей ситуации.

Сделанное предположение позволяет в некоторой степени упростить вычисления в рамках ЕМ-алгоритма и не проводить оценивание средних значений компонент. Упрощенная таким образом итерационная процедура имеет вид

где

В общей ситуации, конечно же, такое упрощение довольно опасно, поскольку, поступая так, приходится считать, что вся волатильность имеет лишь диффузионную природу.

Тем не менее качество получаемых результатов будет все же достаточно высоким, так как упомянутое ограничение оказывает не очень существенное влияние на реализацию СРС-метода, поскольку довольно маловероятно, что ЕМ-алгоритм в итоге припишет существенные веса тем значениям , которые существенно отличны от нуля. Здесь под “существенностью” подразумевается то, что соответствующие веса превышают некоторый (довольно малый) порог, например, .

В качестве результатов получим процессы (последовательности) диффузионных компонент волатильности и весов вида

для процесса (5.1.1) Здесь L – длина скользящего интервала (ширина окна), k – максимальное число компонент смеси.

## СРС-метод с оцениванием значений компонент волатильности

Существует несколько модификаций СРС-метода среди которых можно выделить два основных класса. В рамках первого предусматривается возможность оценивания значений параметров масштаба (диффузионных компонент) и их весов при каждом фиксированном окне. Второй класс содержит модификации СРС-метода, в которых оцениванию подлежат только веса диффузионных компонент, а множество (достаточно богатое) возможных значений самих параметров диффузии заранее фиксировано. При этом внутри обоих классов возможны модификации метода за счет изменения правил выбора начальных приближений и остановки.

В данной работе будем исследовать первый класс. Как было показано, в работе (Королёв, 2011) основную роль влияют величины диффузных компонент, а веса быстро настраиваются алгоритмом по ходу его работы.

### Критерий остановки

В качестве критерия остановки будем использовать (4.3.4) так как он обеспечивает достаточно уверенное определение приближения решения к стационарной точке и в то же время не требует большой вычислительной нагрузки.

### Зависимость СРС-метода от выбора начальных приближений

Так как мы ничего не знаем о решении начальное приближение придется выбирать с помощью псевдослучайной генерации. Однако все же мы установим некоторые разумные рамки основываясь на выборочной дисперсии при выборе среднеквадратичного отклонения и здравого смысла, что все компоненты в среднем должны иметь одинаковую вероятность, при выборе весов компонент. Математическое ожидание мы положим равным нулю.

В качестве начальных приближений будут использоваться следующие псевдослучайные векторы:

*(символом обозначено, что случайная величина ξ имеет равномерное распределение на отрезке ).*

Случайный выбор начальных значений можно осуществлять как сразу для всего алгоритма (то есть одни и те же начальные значения, сгенерированные в самом начале, используются затем при каждом сдвиге скользящего интервала), так и для каждого нового положения окна (то есть при каждом сдвиге скользящего интервала начальные приближения генерируются заново). Мы же будем пользоваться первым вариантом, поскольку это дает более гладкие оценки.

### О влиянии ширины окна на результат применения СРС-метода

Рассмотрим вопрос выбора длинны скользящего интервала (ширины окна) и и ее взаимосвязь с интервалом времени между наблюдениями (шагом временного ряда). Основной интерес представляют интервалы кратные торговой неделе, это позволяет избежать искажений, вызванных систематическими колебаниями торговой активности в течении недели.

Когда выбирается длина скользящего интервала, приходится находить компромисс между гладкостью получаемого результата и его актуальностью. Если выбрать малую длину интервала, то для получения оценок будут использованы данные, лежащие не так далеко в прошлом, но при этом оценки могут быть подвержены сильному влиянию лишь нескольких изменяющихся значений, то есть при сдвиге только на один шаг можно получить существенно отличную картину. Напротив, если использовать интервал большой длины, то резких изменений оценок при сдвигах на малое число шагов как правило не будет, но при этом оценки будут зависеть от большого числа наблюдений, далеко отстоящих от текущего момента в прошлое и потому, возможно, утративших информативность.

При выборе ширины окна и частоты, будем руководствоваться следующими правилами: (i) число наблюдений, попадающих внутрь скользящего интервала, должно быть достаточным для приемлемого оценивания параметров смеси (обычно более ста) и (ii) интервал между каждыми двумя соседними наблюдениями достаточно велик, чтобы можно было считать, что в течение такого интервала заключается большое число контрактов.

## Применение СРС метода к анализу финансовых индексов

Рассмотрим применение метода скользящего разделения смесей на примере индекса МосБиржи. Индекс МосБирэжи это ценовой, взвешенный по рыночной капитализации (free-float) композитный индекс российского фондового рынка, включающий наиболее ликвидные акции крупнейших и динамично развивающихся российских эмитентов, виды экономической деятельности которых относятся к основным секторам экономики. В индекс входит 41 акция (на 23.04.2019), основными являются Сбербанк, Лукойл, Газпром, ГМК Норникель, Новатек, ТатНефть, Рос нефть и др. (с весом <5%).

Для исследования взят интервал 23.03.18 по 07.06.18. На рисунке 1 представлены значения индекса с шагом 10 минут.

Выбор интервала для исследования объясняется наличием на нем как моментов высокой торговой активности с высокой волатильностью, так и «боковых» интервалов с низкой волатильностью индекса. В этот период произошло несколько событий, повлиявших на индекс. Первое и самое крупное событие — это «черный понедельник» 9 апреля, США ввели антироссийские санкции, что повлекло за собой резкое падение индекса с дальнейшем периодом нестабильности и высокой волатильности. Вторым событием можно отметить 26 апреля рост вслед за мировыми фондовыми площадками и нефтью на фоне коррекции вниз доходностей американских гособлигаций индекс МосБиржи на фоне ослабления рубля превысил 2285 пунктов, отыграв все потери с "черного понедельника" 9 апреля.

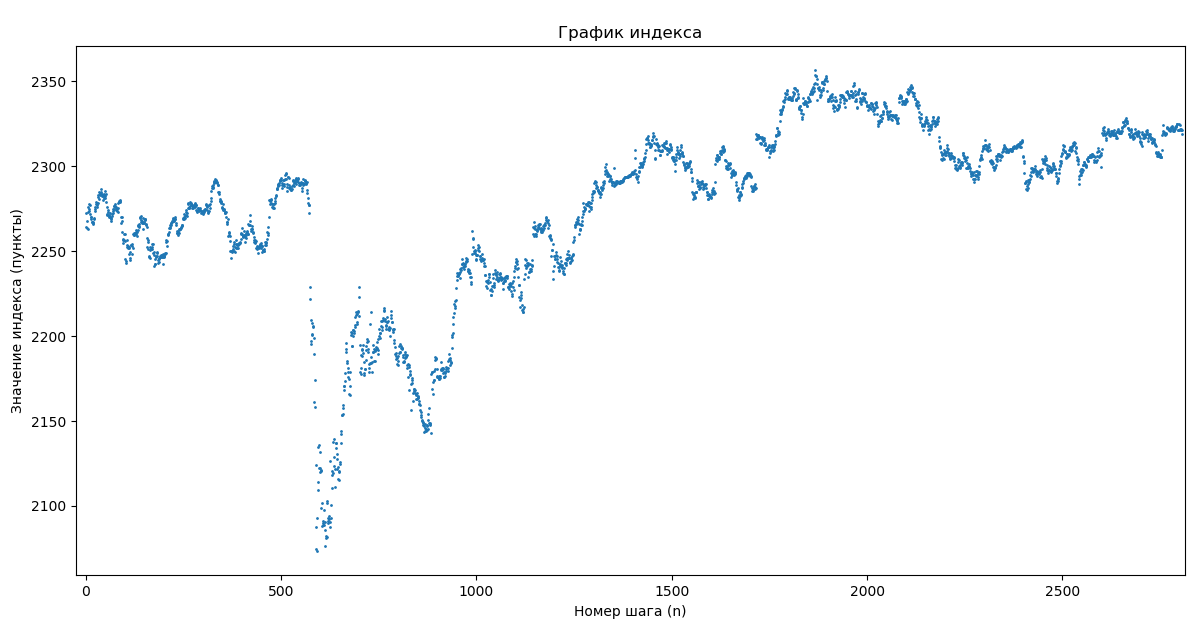


Рисунок 1

Далее представлены результаты декомпозиции волатильности данного индекса в исследуемом периоде с различными параметрами метода.

Пример 1

Минутные приращения, шаг окна 5 минут, окно 520 значений (1 торговый день).

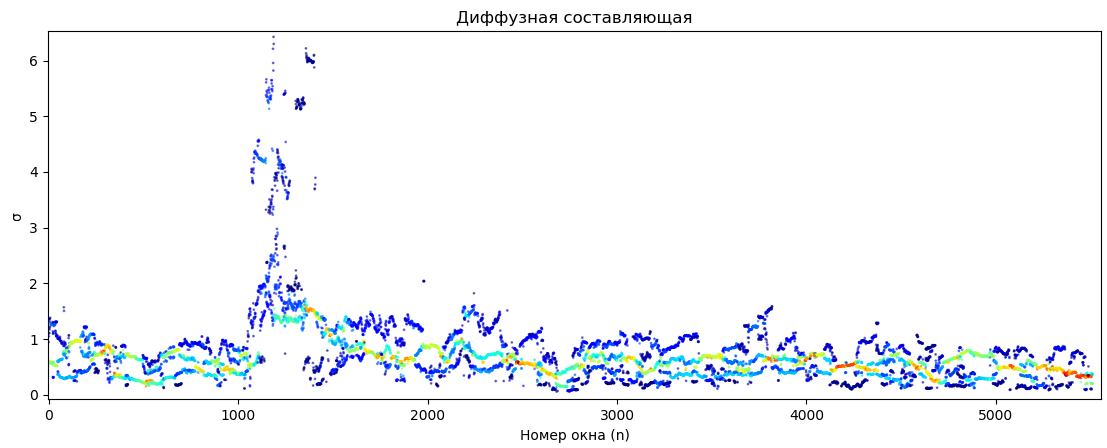


Рисунок 2

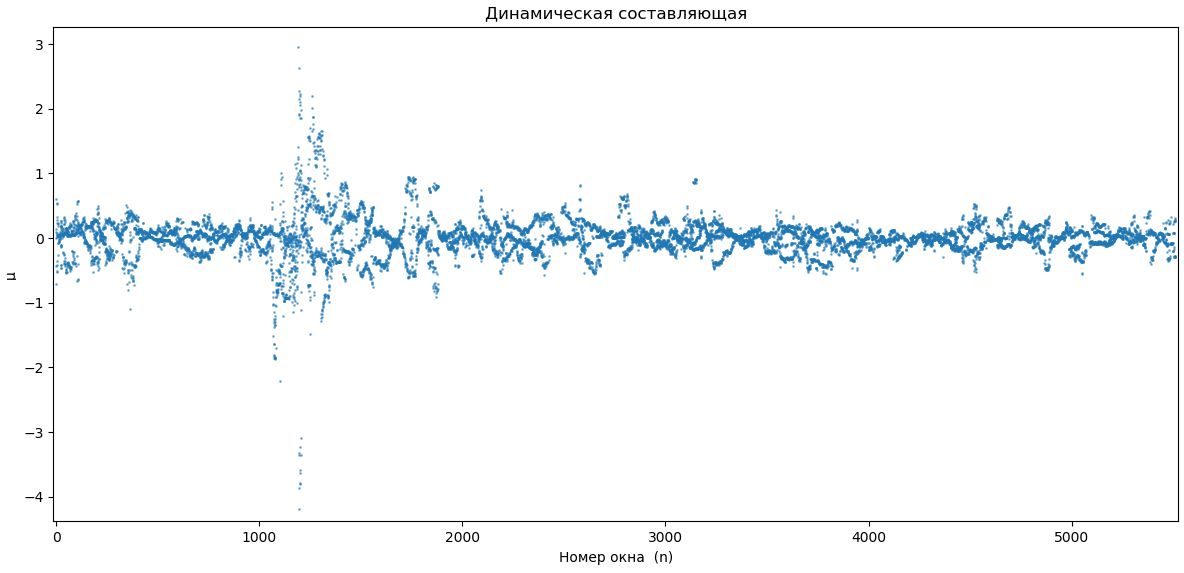


Рисунок 3

Пример 2

Минутные приращения, шаг окна 5 минут, окно 1560 значений (3 торговых дня). Здесь и далее значения компонент с весом менее не показаны.

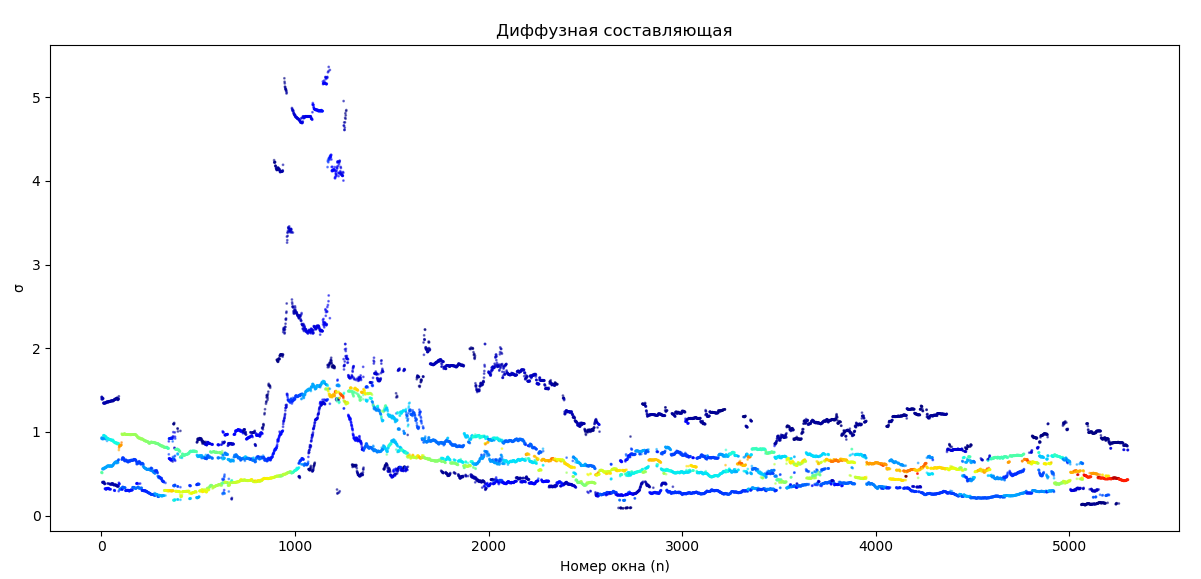


Рисунок 4

На Рисунке 2 мы видим 4 ярко выраженные компоненты волатильности. При таком выборе окна для данного шага как видно график получается достаточно гладкий и в то же время отображает актуальную информацию без излишнего сглаживания. На шаге n=832 (9 апреля) отчетливо виден момент всплеска компоненты высокой волатильности, а также через некоторое время, синхронный всплеск компонент низкой волатильности, вызванный наличием ярко выраженной трендовой составляющей, что так же видно на Рис 3.

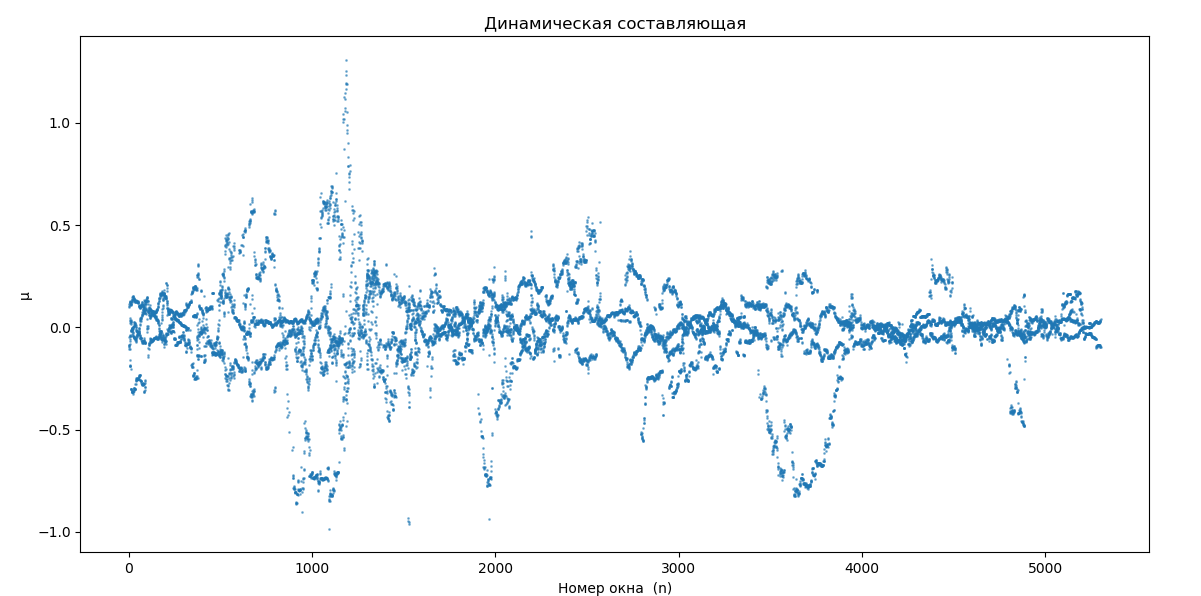


Рисунок 5

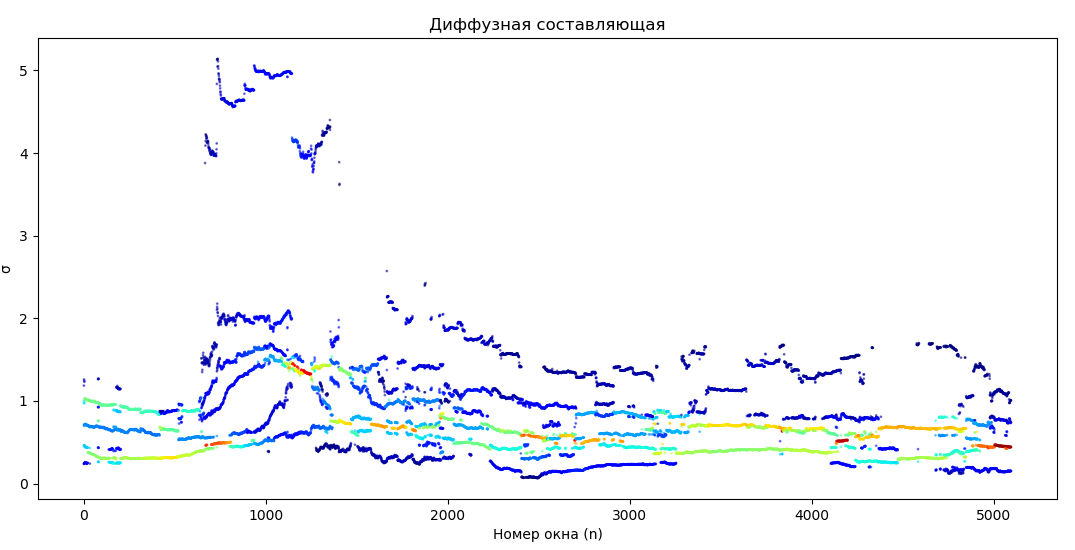
На Рисунке 3 показана динамическая составляющая. Отчетливо прослеживаются те же 4 компоненты. Анализируя рисунок можно заметить, что даже в моменты резкого падения/роста индекса, все равно остаются положительные динамические компоненты, что можно интерпретировать как наличие игроков на рынке, которые, несмотря на падение индекса, все равно «играют в плюс».

Проход окном 9 апреля.

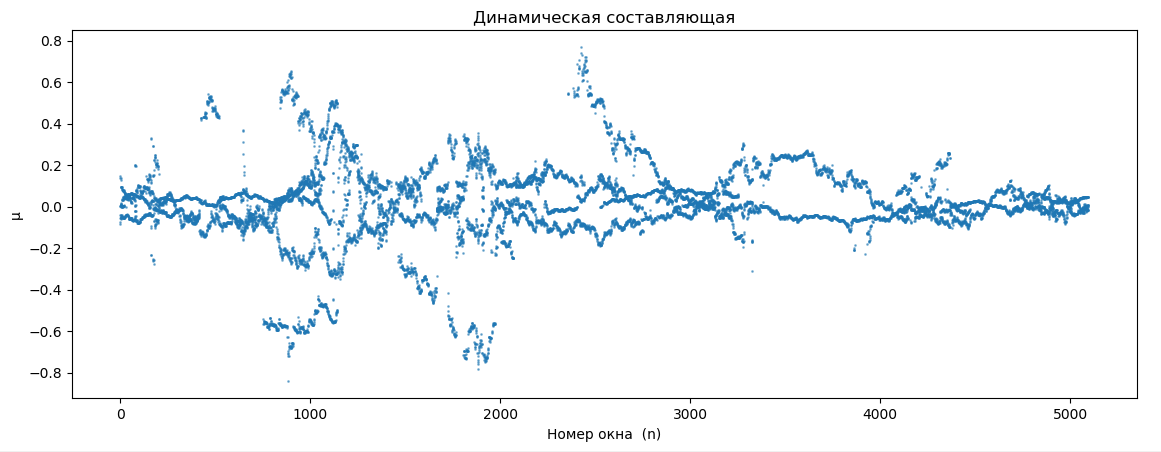
Проход окном 26 апреля

Пример 3

Минутные приращения, шаг окна 5 минут, окно 2600 значений (1 торговая неделя).  
значения компонент с весом менее не показаны

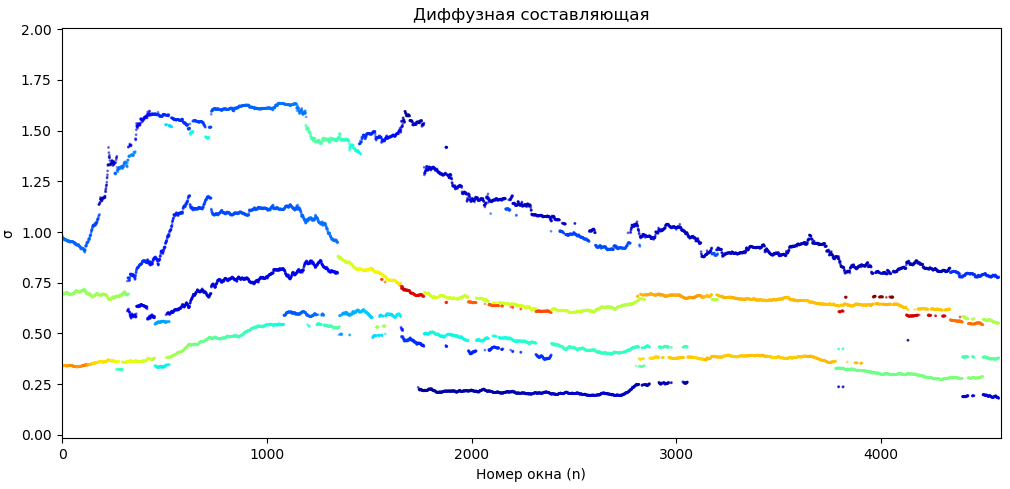
Рисунок 6

На рис 8 видна более полная картина относительно 1 примера. Удалось выделить до 6 основных компонент волатильности

  
Рисунок 7

Пример 4

Минутные приращения, шаг окна 5 минут, окно 5200 значений (2 торговых недели).

  
Рисунок 8

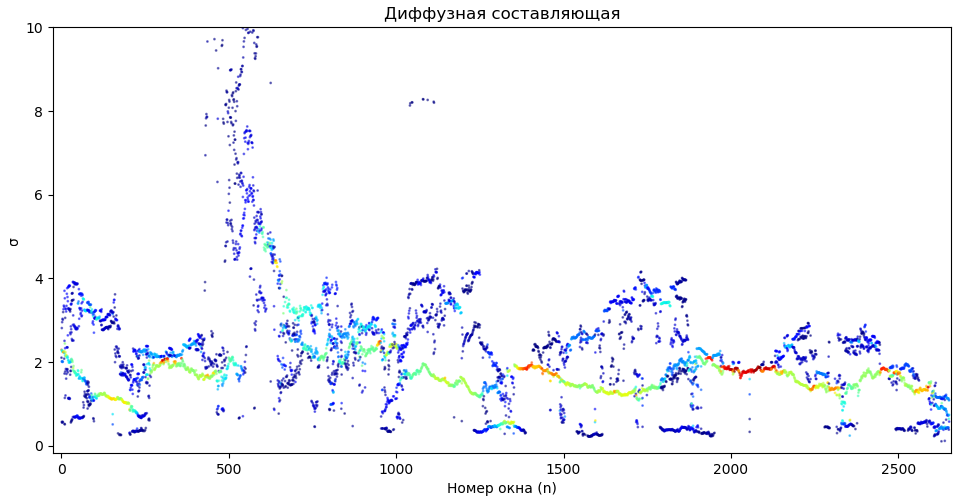
Явно выделены 4 компоненты волатильности

  
Рисунок 9

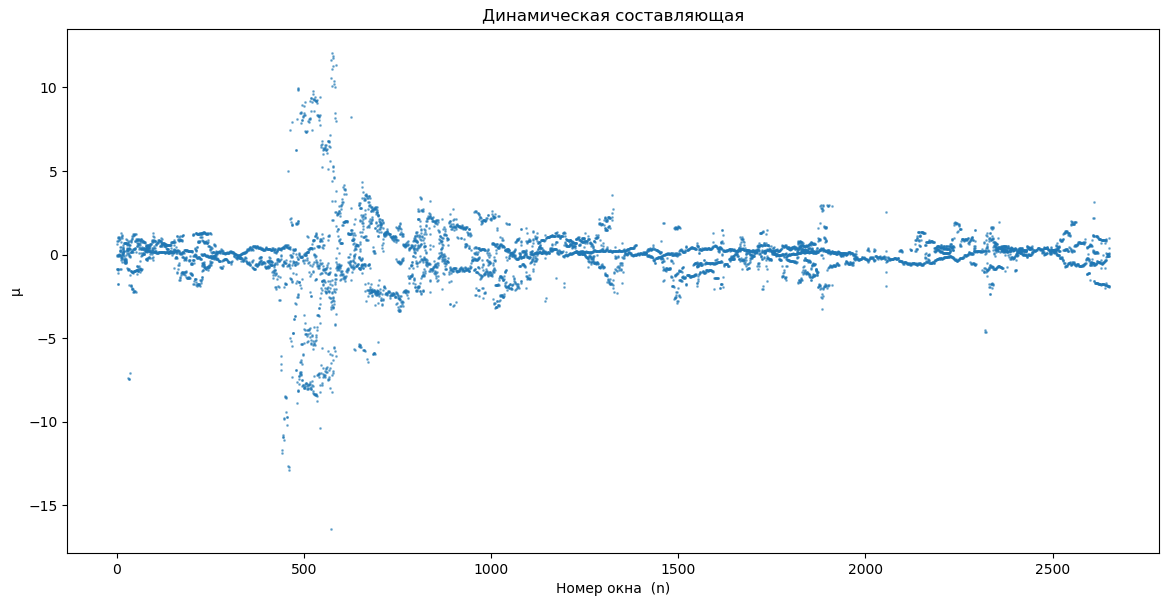
щшщо

Пример 5

10 минутные приращения, шаг окна 10 минут, окно 156 значений (3 торговых дня).

Рисунок 10

Как видно из рисунка, хотя и удается выделить основные 2-3 компоненты волатильности, выборка размером 156 значений не дает приемлемого результата. Компоненты высокой волатильности получаются сильно зашумленными. Хотя, анализируя компоненту высокой волатильности по-прежнему можно различить основные события на рынке в данный период. Компонента низкой волатильности так же имеет зашумлённый вид, и ее график является разрывным в моменты высокой волатильности рынка.

  
Рисунок 11

В динамической картине аналогично видна высокая зашумленность, что затрудняет ее анализ.

Пример 6

10 минутные приращения, шаг окна 1 элемент, окно 260 значений (1 торговая неделя).  
значения компонент с весом менее не показаны

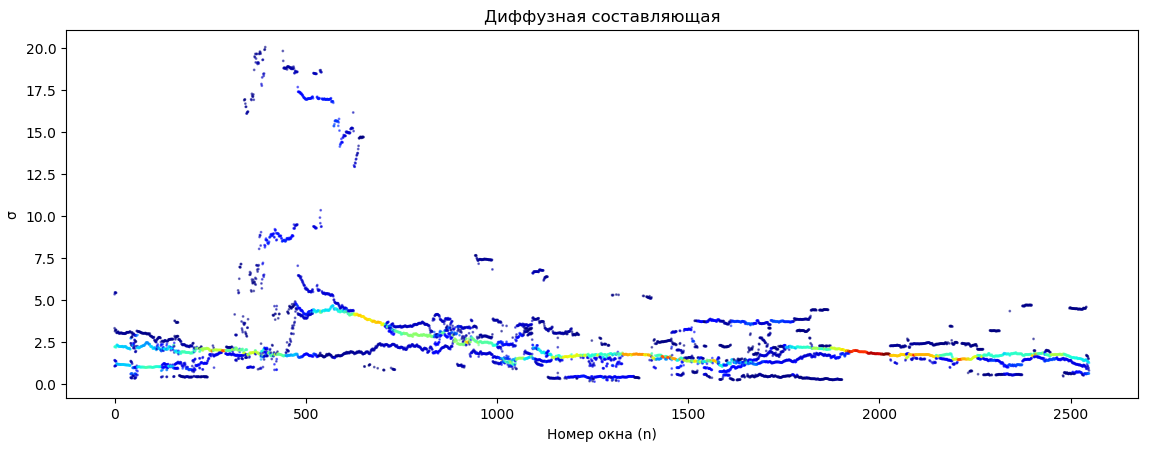
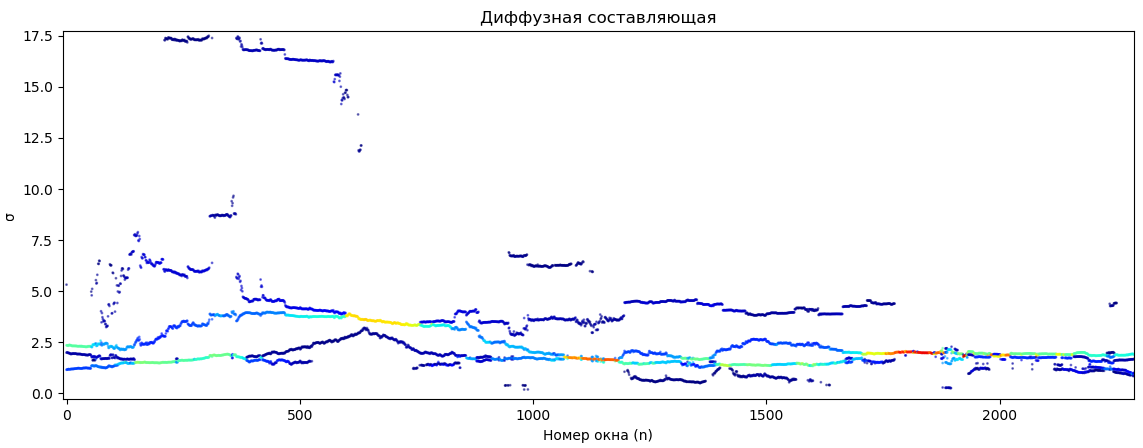
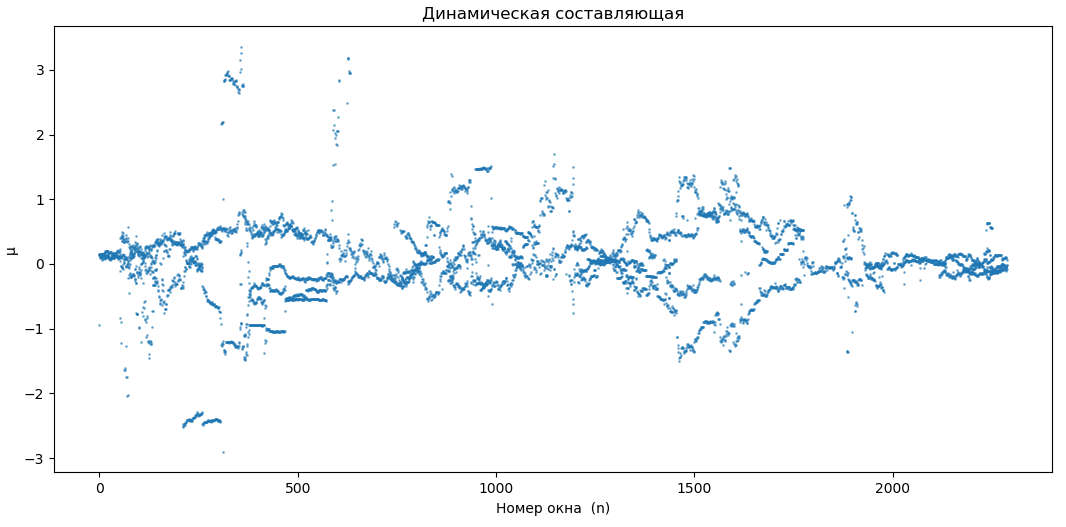
  
Рисунок 12

  
Рисунок 13

Пример 7

10 минутные приращения, шаг окна 10 минут, окно 520 значений (2 торговых недели).

  
Рисунок 14

додло  
  
Рисунок 15

Пример 8

30 минутные приращения, шаг окна 30 минут, окно 54 значения (3 торговых дня).

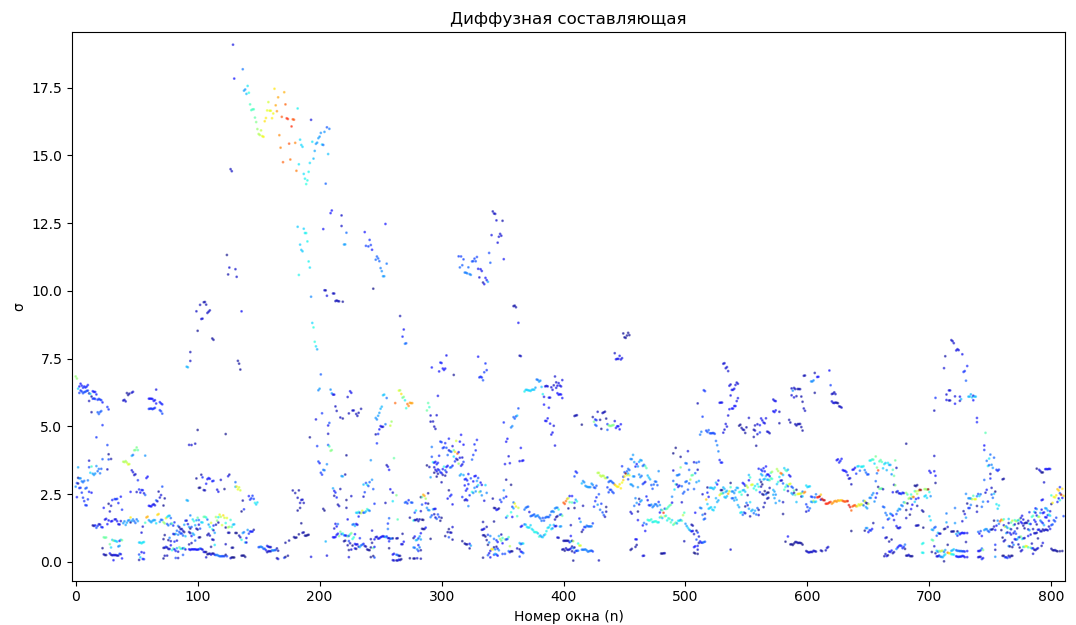


Рисунок 16

В качестве крайнего примера, на Рисунке 6 видно, что никаких существенных данных получить при таком размере выборки нельзя, выделяются только короткие отрезки непрерывных компонент, в остальном они являются разрывными и не представляют никакого интереса для анализа.

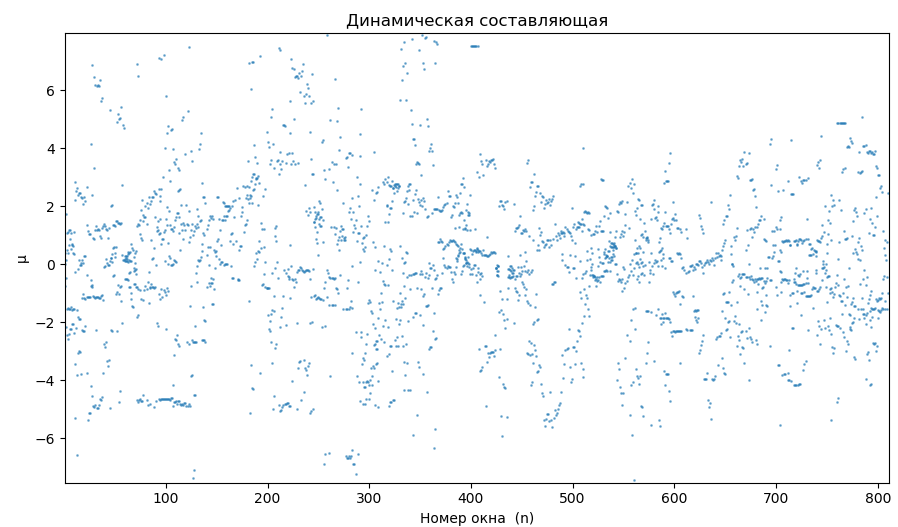
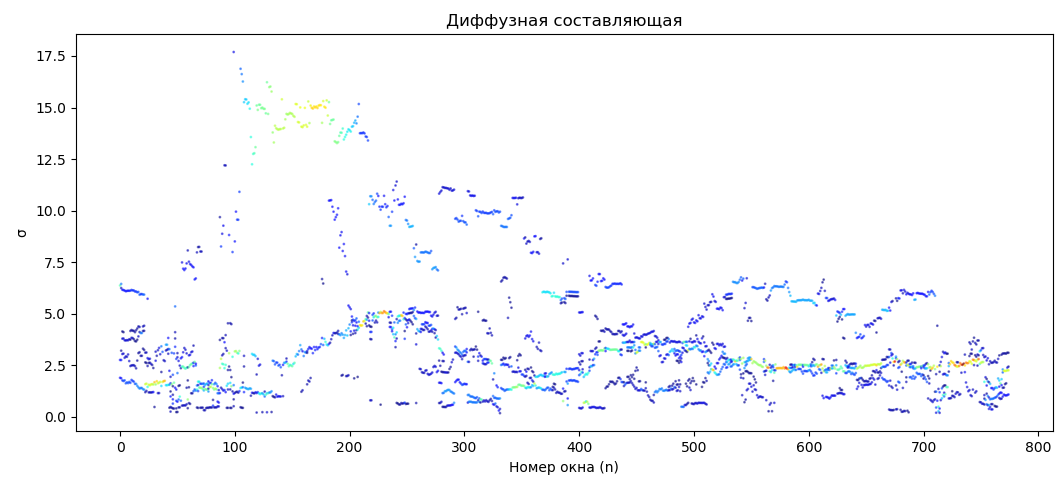


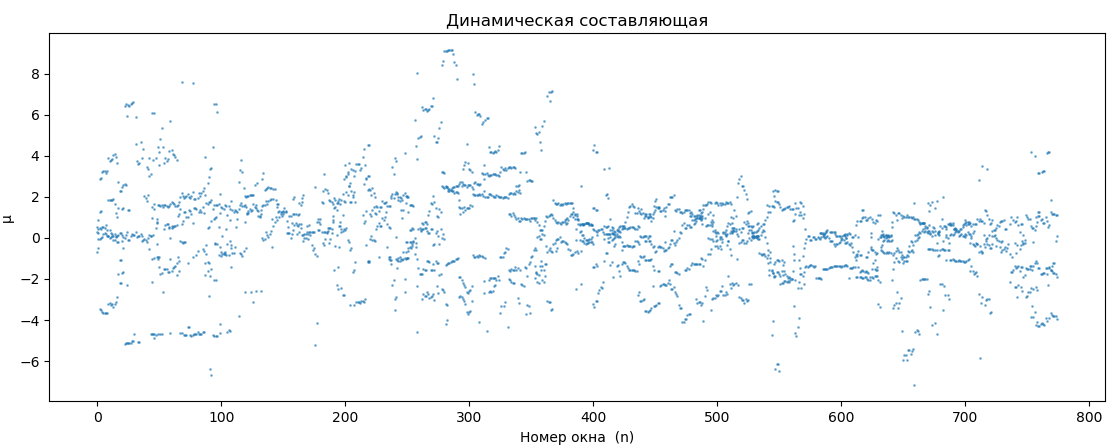
Рисунок 17

рисунок 7 подтверждает данную картину, представляя практически хаотичный набор точек.

Пример 9

30 минутные приращения, шаг окна 1 элемент, окно 90 значений (1 торговая неделя).

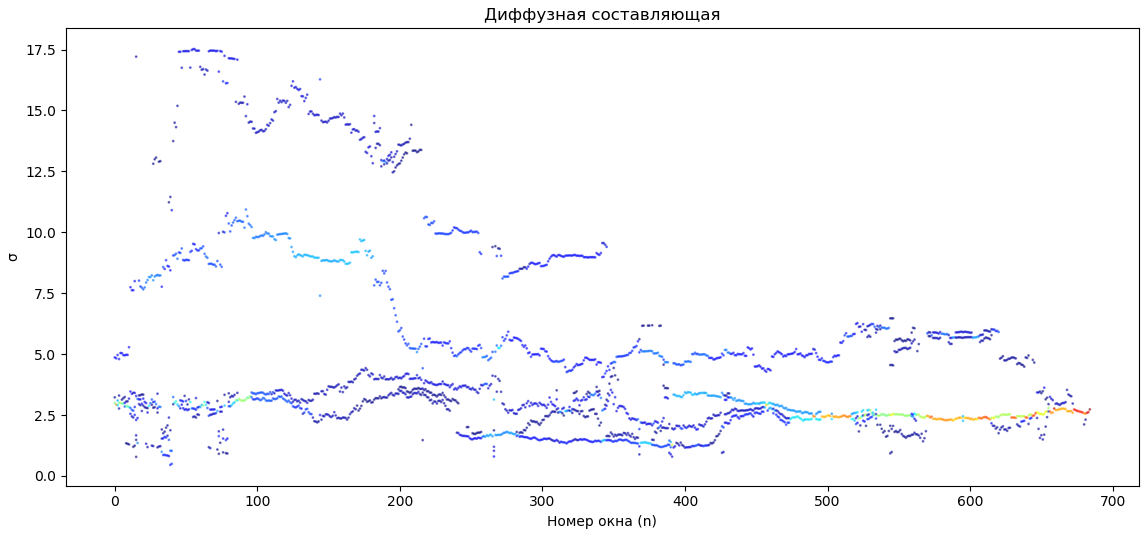
  
Рисунок 18

  
Рисунок 19

Hg

Пример 10

30 минутные приращения, шаг окна 30 минут, окно 180 значений (2 торговых недели).

  
Рисунок 20

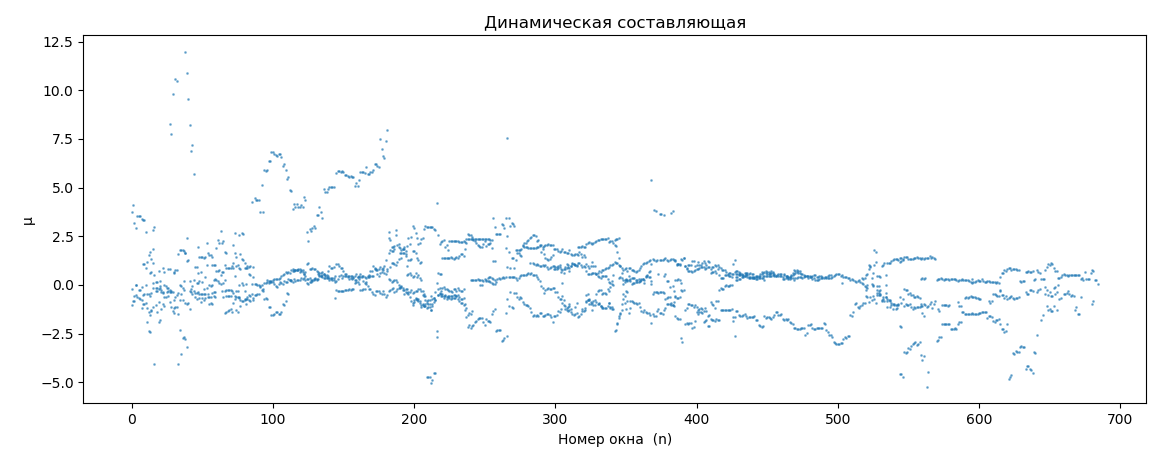
вапвап  


Рисунок 21

### О влиянии ширины окна на результат применения СРС метода

На примерах показано различие в результатах, полученных при трех различных значениях ширины окна, равных трем, пяти и десяти торговым дням и трех различных значениях шага 30, 10 и 1 минута. Сравнение графиков говорит о том, что при выборе ширины окна следует обратить внимание на две характеристики: гладкость доминирующего процесса низкой волатильности (самые темные линии) и степень детализации остальных компонент волатильности.

При сравнении первых и вторых графиков видно, что процесс, описывающий компоненту низкой волатильности, на вторых графиках более гладкий, нежели на первом, поведение же компонент высокой волатильности в основном одинаково, за исключением первого графика, где менее гладкий характер вызван, скорее всего малым количеством элементов попавшим в выборку. Поэтому первый график не позволяет нам получить надежные данные оценки параметров смеси, поэтому второй и третий случай более предпочтительны. Однако хотя и дает более гладкую картину, все же менее предпочтителен, так как идет усреднение по данным прошедших двух недель, что не позволяет получать картину актуальную на данный момент.

Таким образом, главный вопрос заключается в том, какой из двух интервалов – недельный или двухнедельный – следует предпочесть. Ответ зависит от цели исследования, если требуется проводить исследование свойств ЕМ алгоритма или долгосрочный анализ индекса, то следует предпочесть более длинный интервал, так как он дает более гладкие, ярко выраженные, оценки компонент и в тоже время данные остаются все еще актуальными.

Более короткий интервал пригоден для прогнозирования будущего значения индекса, так как в нем более выражены локальные тенденции.

Так же, окно исследования можно двигать как в направлении астрономического времени (с целью предсказания будущих результатов) так и в обратном направлении при анализе прошлых значений. Так же при анализе прошлых значений полезно «прогнать» алгоритм в обоих направлениях, что позволит убрать эффект инерции, присущий СРС-методу с большой шириной рабочего окна.

### О влиянии масштаба (длины интервала времени между наблюдениями) на результат применения СРС-метода

Рассмотрим теперь некоторые вопросы, связанные с влиянием длины интервала времени между наблюдениями на получаемый результат. Эти вопросы совсем не просты. Периодичность измерения значений индекса играет чрезвычайно важную роль для интерпретации результатов. Ее суть заключается в том, что коэффициент диффузии характеризует стандартное отклонение приращения индекса за вполне определенный промежуток времени – интервал времени между наблюдениями, то есть стандартное отклонение разности значений индекса в равноотстоящие моменты времени. При этом характер эволюции процесса между последовательными наблюдениями (измерениями, отсчетами) остается неизвестным.

Более того, при неадекватном выборе временного масштаба – длины интервала времени между наблюдениями – может быть получена картина, совершенно не соответствующая действительности. В этом легко убедиться с помощью следующего тривиального примера. Рассмотрим функцию . Сначала предположим, что она “наблюдается” в точках, кратных периоду, скажем, точках вида   
 Очевидно, что тогда “временной ряд” абсолютно не содержит никакой информации об изменении “наблюдаемой” функции. Но если та же функция “наблюдается” в точках вида , где – малое число, то “временной ряд” в зависимости от величины может вместо гладкой плавной кривой выглядеть как пила с неровными и довольно острыми зубцами, причем в точках оси времени, соответствующих зубцам, запросто можно предположить (а с помощью соответствующих статистических процедур и “научно обосновать”) отсутствие производной у исходной функции, что, естественно, не имеет никакого отношения к реальности. К сожалению, выбор интервала времени между наблюдениями (измерениями) часто не зависит от исследователя-статистика.

Как уже говорилось, от выбора интервала времени между наблюдениями сильно зависит интерпретация результатов. Предположим, что этот интервал достаточно мал для того, чтобы быть уверенным, что процесс не слишком сильно “виляет” между соседними точками отсчетов. Забегая вперед, на время отойдем от принятого ранее упрощения, заключающегося в том, что средние значения компонент смеси равны нулю, и будем считать, что ЕМ-алгоритм на каждом шаге оценивает все параметры смеси. Так как изучаемый временной ряд представляет собой приращения индекса, то оцененные средние значения компонент смеси (параметры положения, математические ожидания нормальных законов, составляющих смесь) являются средними (по “ансамблю”, то есть по элементарным исходам) перемещениями индекса. Но так как интервалы времени между всеми отсчетами одинаковы, то, приняв длину этих интервалов за единицу времени, можно заключить, что среднее значение каждой компоненты, будучи средним перемещением за единицу времени, есть не что иное как усредненная по “ансамблю” средняя (по времени – за принятую единицу времени) скорость изменения соответствующей составляющей исходного процесса. Таким образом, при оценивании параметров смеси на каждом шаге автоматически получается (статистически оценивается) распределение скоростей изменения компонент той системы, которую характеризует временной ряд значений рассматриваемого индекса. Так, наблюдая финансовый индекс, с помощью СРС-метода реализуется возможность оценивать изменение во времени скоростей изменения компонент финансового рынка. Ясно, что понятие средней скорости сильно зависит от длины интервала времени, усреднение по которому проводится. Таким образом, залогом успеха каждого применения СРС-метода служит адекватность длины интервала времени между наблюдениями характерному масштабу времени, соответствующему интенсивности изучаемых изменений рассматриваемой системы.

Сравнивая графики на рисунках 1-9, можно заметить разницу в результатах анализа компонент волатильности по данным, полученным при разной длине интервала времени между наблюдениями. Хотя оба графика фактически соответствуют одному и тому же процессу, можно заметить, что тот график, где используются минутные данные, имеет большее разрешение, особенно это заметно на компонентах большей волатильности.

Также можно заметить, что на больших интервалах между отсчетами компоненты с высокой волатильностью более значимы (имеют большие веса. Отметим, однако, что выбор большой ширины рабочего окна для ЕМ- алгоритма может оказаться опасным вследствие возможности получения неадекватных результатов из-за возрастающей инерции СРС-метода. Соотношение между распределениями вероятностей, соответствующими данным разных типов (то есть с разной длиной интервала времени между последовательными отсчетами).

## Применение СРС метода к анализу метеорологических данных

Попробуем применить метод к метеорологической информации. Поскольку погода тоже формируется с участием множества факторов.

В качестве источника данных возьмём метеорологическую сводку FM-15 за период с 01.01.2010 по 02.05.2019 со станции №72392523190 NOAA США, находящейся в региональном аэропорту города Санта Барбара, Калифорния, США. Будем использовать данные ежечасного приращения температуры сухого термометра. Выбор данного источника обусловлен мягким климатом, что уменьшает наличие сильных трендовых составляющих и наличием ежечасной информации о температуре.

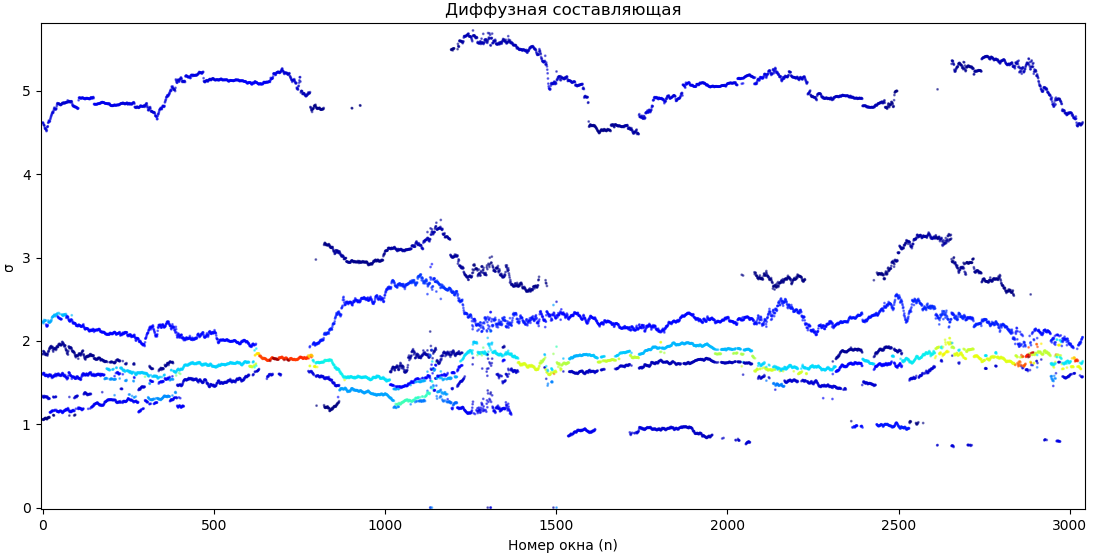
.

Зима 2013/2014 года

Февраль 2015.

Пример 1

Часовые приращения, шаг 24 часа, окно 8760 часов (365 дней)

  
Рисунок 22

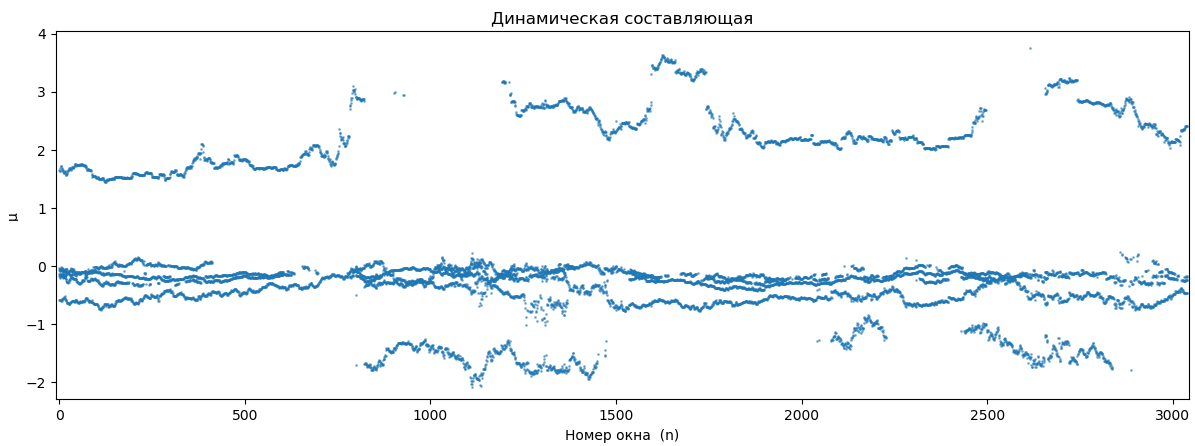


Рисунок 23

Такой выбор окна позволил сгладить сезонные изменения, но не потерять информацию об аномальной погоде.

Ярко выделяются 2 события

Зима 2013/2014

Лето 2015

Пример 2

Часовые приращения, шаг 24 часа, окно 2232 часов (93 дня)

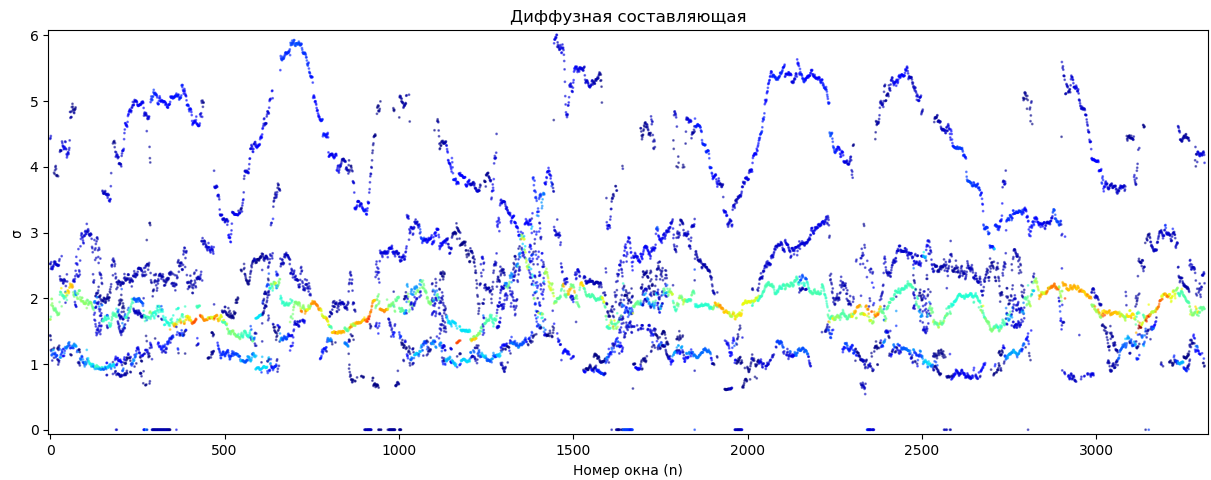


Рисунок 24

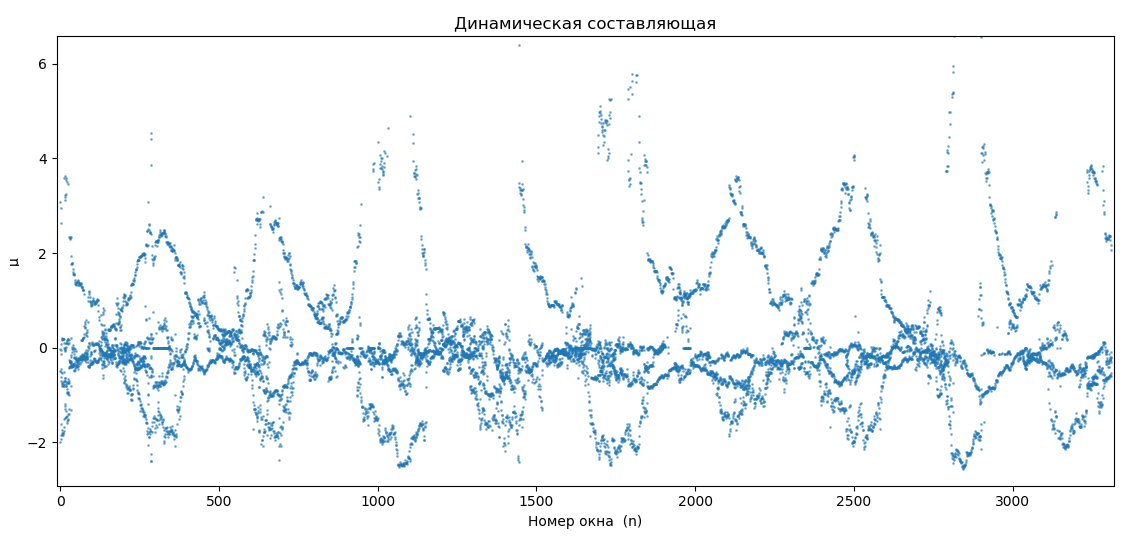


Рисунок 25

# Реализация метода СРС с применением CUDA.

Для получения экспериментальных результатов, в рамках этой работы был реализован метод скользящего разделения смесей на языке Python с использованием библиотек NumPy, SciPy, Matplotlib. Однако скорость вычислений на CPU в данной реализации оказалась слишком мала. Для вычисления отрезка в 1300 шагов, с размером окна порядка 1000 элементов потребовалось чуть более 2х суток с применением процессора Intel core i7-3770(В режиме Boost до 4,5 ГГц). Поэтому была написана реализация алгоритма с применением параллельных вычислений с использованием графического адаптера, используя библиотеку CUDA, компании Nvidia. Вычисления с применением графического адаптера NVIDIA Tesla K40, позволяют произвести тот же объем работы за ~2 минуты, что дает прирост производительности расчетов на 3-4 порядка.

CUDA это платформа для параллельных вычислений с применением графических карт (graphical process units, GPU) созданная компанией Nvidia. Оптимизация платформы позволяет запустить вычисления параллельно. Последовательную часть программы, оптимизированную для выполнения в один поток, на CPU, а основную часть, с «тяжелыми» вычислениями, на ядрах GPU (которых более тысячи) параллельно. При разработке программы для CUDA может использоваться язык C, C++, Fortran, MATLAB.

Большая разница в вычислительных мощностях при вычислении с плавающей запятой у CPU и GPU вызваны тем, что GPU специально предназначены для выполнения «тяжелых», хорошо распараллеливаемых вычислений, а именно рендеринг графики, по этой причине при разработке GPU большая часть процессора отдается под обработку данных нежели под их кеширование или управление. Точнее сказать GPU особенно хорошо приспособлены для решения задач, которые могут быть представлены в виде задач для параллельного вычисления. Одинаковая программа выполняется на множестве вычислительных блоков с большой интенсивностью вычислений (отношении операций вычисления к операциям обращения к памяти). Поскольку одна и та же программа выполняется для каждого элемента из набора данных не требуются сложные системы управления потоком команд, и поэтому, поскольку программа выполняется на многих элементах и имеет высокую интенсивность вычислений, задержки доступа к памяти могут быть спрятаны за вычислениями вместо использования больших блоков кеширования.

Благодаря тому что в EM алгоритме основные вычислительные затраты приходятся на вычисление матрицы размером весов на Е- шаге, где каждый элемент матрицы вычисляется по одной и той же формуле, что позволяет отлично распараллелить эти вычисления на множество ядер. Так же и вычисления на М- шаге, хотя и являются менее затратными, тоже хорошо поддаются распараллеливанию. Все вместе позволяет практически полностью осуществить выполнение алгоритма на GPU, оставив CPU лишь контроль последовательности выполнения, что позволяет производить вычисления используя лишь GPU, его графический процессор и встроенную графическую память.

# Заключение

В заключении проведенного нами исследования можно сделать следующие основные выводы по теме.

Применение СРС-метода в анализе реальных данных показывает, что распределения приращений исследуемых финансовых индексов отличны от нормальных. Это подтверждает эмпирическое наблюдение, что в реальности такие распределения не являются нормальными, хотя и близки к ним, но имеют «тяжелые хвосты».

В случае если бы мы наблюдали одну нормальную диффузную компоненту, это могло бы говорить о том, что мы имеем дело с стохастической волатильностью (коэффициентом диффузии) в модели броуновского движения.

В нашем же случае подобный эффект наблюдается лишь при очень малой ширине окна, и выборе в качестве начального приближения результата предыдущей итерации. В других же случаях мы наблюдаем набор из нескольких нормальных компонент.

В результатах разложения СРС-методом, при адекватном выборе параметров, явно наблюдается несколько гладких компонент волатильности. Т.е. дискретное смешивающее распределение с конечным носителем демонстрирует высокую адекватность. Можно было ожидать, что смешивающее распределение будет иметь не столь явно конечный носитель и окажется более гладким, но полученные результаты свидетельствуют об обратном.

Это показывает, что состояние рынка определяется все же конечным числом фактором, которые СРС-метод, выделяет.

# Список источников

1. Королёв В.Ю. Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. Москва: Издательство Московского университета, 2011. 512 с
2. Dempster A.P., Laird N.M., Rubin D.B. Maximum Likelihood from Incomplete Data Via the EM Algorithm // J. R. Stat. Soc. Ser. B. 1977. Т. 39. № 1. С. 1–38.
3. Nvidia. CUDA C PROGRAMMING GUIDE v10.1 // 2019. С. 301.