ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
Научно-учебная лаборатория
Финансовой инженерии и риск-менеджмента

А.В. Косьяненко

ОПЫТ ВОССТАНОВЛЕНИЯ ПРОПУЩЕННОЙ РЫНОЧНОЙ ИНФОРМАЦИИ НА ОСНОВЕ БАЙЕСОВСКОГО ПОДХОДА

Препринт WP16/2007/02
Серия WP16
Финансовая инженерия,
риск-менеджмент и актуарная наука

Москва
ГУ ВШЭ
2007

УДК 336.76
ББК 65.262.2
К72

Редакторы серии WP16
«Финансовая инженерия,
риск-менеджмент и актуарная наука»
С.Н. Смирнов, А.Г. Шоломицкий

К 72 **Косьяненко А.В.** Опыт восстановления пропущенной рыночной информации на основе Байесовского подхода: Препринт WP16/2007/02. — М.: ГУ ВШЭ, 2007. — 32 с.

Работа посвящена актуальной для российского фондового рынка проблеме наличия пропусков в рыночной информации. Приведен краткий обзор методик, наиболее часто используемых для решения данной проблемы. Описан альтернативный метод, позволяющий при определенных предположениях существенно повысить вычислительную эффективность применяемых алгоритмов. Проведен сравнительный анализ результатов заполнения пропусков данных с использованием различных методов.

Работа поддержана Научным фондом ГУ ВШЭ и осуществлена в рамках реализации мероприятия 5.2.5 «Проведение совместных исследований преподавателей, аспирантов и студентов» Инновационного образовательного проекта ГУ ВШЭ по теме № 06-05-0019 «Разработка методологии оценки стоимости портфеля на основе исследования микроструктуры рынков для целей управления рисками».

УДК 336.76
ББК 65.262.2

Препринты ГУ ВШЭ размещаются на сайте:
http://new.hse.ru/C3/C18/preprintsID/default.aspx.

© А.В. Косьяненко, 2007
© Оформление. ГУ ВШЭ, 2007

Содержание

Введение 3

1. Обзор применяемых на практике процедур восстановления пропущенной инфомации 5

2. Численные методы фильтрации и восстановления пропущенной информации 14

3. Анализ результатов численных экспериментов 21

Заключение 27

Литература 29

Введение

В настоящее время в повседневной практике предприятий и организаций, связанных с проведением операций на рынках ценных бумаг и производных финансовых инструментов, широкое применение находят модели, использующие в качестве входной информации результаты рыночных торгов. Цели проведения таких расчетов могут быть различны: прогнозирование стоимости отдельных ценных бумаг и финансовых инструментов, определение стоимости и уровня риска портфеля, построение таких характеристик рынка, как кривые доходности и многие другие. Общей чертой большинства таких моделей является повышенная чувствительность к качеству используемой входной информации. Под качеством в данном случае понимается наличие полной и достоверной информации о рыночной ситуации.

В литературе, посвященной применению такого рода численных методов, проблема некачественной информации, как правило, не обсуждается — подразумевается наличие у исследователя такой информации. Действительно, при анализе результатов дневных торгов на высоколиквидных рынках такой проблемы, как правило, не возникает – большое количество сделок гарантирует наличие информации о ценах, при этом результат каждой отдельной сделки не оказывает существенного влияния на средние показатели.

Ситуация кардинально меняется при анализе низколиквидных рынков или при исследованиях микроструктуры рынков, проводящихся на основе внутридневной информации. В обоих случаях неизбежно возникают периоды, на протяжении которых рыночные сделки отсутствуют. При этом в результате краткосрочных колебаний ликвидности рынков или при совершении нерыночных сделок их цена может существенно отличаться от цен сделок, имевших место ранее, а также выходить за рамки установившихся котировок на покупку и продажу.

Наличие обоих явлений — некачественной информации и пропусков данных — ведет к возникновению неадекватных результатов при применении численных процедур, а иногда и к невозможности их применения. Решение этой проблемы может быть достигнуто двумя способами: либо за счет учета в модели возможности наличия пропущенных данных и сделок, выделяющихся на фоне основного количества нетипичными значениями или объемами, либо за счет проведения процедуры предварительной подготовки данных. Первый путь за редким исключением ведет к потере возможности аналитического решения поставленных задач, применению содержательно и вычислительно сложных численных методов и, как следствие, снижению практической ценности моделей. Второй ведет к необходимости создания специальных процедур фильтрации и восстановления пропущенной информации. При этом сами вычисления могут проводиться без изменений, как если бы на входе подавалась полная и достоверная информация.

Данная работа посвящена разработке и описанию формальных алгоритмов, позволяющих проводить заполнение возникших по тем или иным причинам пропусков. При этом полагается, что вся информация о результатах торгов является достоверной. Таким образом проверка достоверности исходной информации возлагается целиком на пользователя, либо может быть проведена специальными численными методами, описание которых не является целью настоящей работы. Первый раздел работы посвящен обзору описанных в литературе методик заполнения пропущенных данных. Во втором разделе описана разработанная автором методика восстановления информации, а также аналитическое решение задач, возникающих при реализации EM-алгоритма, являющегося стандартом для решения такого рода задач. Наконец, в третьем разделе представлены результаты применения методик к различным наборам рыночной информации.

Работа поддержана Научным фондом ГУ ВШЭ и осуществлена в рамках реализации мероприятия 5.2.5 «Проведение совместных исследований преподавателей, аспирантов и студентов» Инновационного образовательного проекта ГУ ВШЭ по теме № 06-05-0019 «Разработка методологии оценки стоимости портфеля на основе исследования микроструктуры рынков для целей управления рисками»

1. Обзор применяемых на практике процедур восстановления пропущенной информации

Причины возникновения пропусков данных могут быть различны. Во-первых, это отсутствие сделок в результате низкой ликвидности рынка. Во-вторых, возможность временного приостановления торгов как по всем, так и по отдельным видам ценных бумаг. Так, в период, предшествующий выплате купонных доходов по облигациям, торги по ним приостанавливаются на срок, определенный проспектом эмиссии. В-третьих, возможны технологические сбои при накоплении рыночной информации, а также задержки в ее получении. В-четвертых, периоди-

чески на рынке происходят малоправдоподобные или нерыночные сделки, существенно искажающие рыночную ситуацию. Исключение этих сделок, осуществляемое посредством особых процедур фильтрации, ведет к возникновению пропусков. Наконец, возможно применение тех же методов для прогнозирования будущих цен активов, в этом случае будущие значения цен рассматриваются как пропуски.

Наиболее часто используемыми на практике способами восстановления информации о пропущенных по тем или иным причинам ценах являются использование известной цены на предыдущем временном участке и вычисление среднего арифметического между заявками на покупку или продажу на момент пропуска цены. Приписывание цены предыдущего периода может быть отчасти теоретически обосновано. Самуэльсон в своей работе, посвященной поведению фьючерсных цен (Samuelson, 1965), доказывал, что изменение цен представляет собой случайный процесс, обладающий мартингальным свойством:

$$E(x_{t+1} | \mathrm{P}_t) = x_t,$$

где P_t — доступная к моменту t информация. Дальнейшие исследования показали, что в действительности мартингальное свойство присуще в той или иной мере процессу, включающему в себя не только цену, но и дисконтированный поток будущих платежей по активу или финансовому инструменту. Однако на коротких промежутках времени эти процессы не сильно отличаются друг от друга. Исследование данного вопроса не является темой настоящей работы, подробный обзор проведенных в данной области исследований провел, например, Лерой (LeRoy, 1989).

Оба метода имеют недостатки. Во-первых, оба не являются универсальными: имеют место пропуски данных длиной в несколько периодов, при этом отличие между процессами цен и дисконтированных цен с учетом потока платежей по бумаге может стать существенным; возможно отсутствие информации не только о ценах сделок, но и о заявках на покупку или продажу. Использование ценовой информации с предыдущего периода означает, что значения цен активов в разные моменты уже заведомо не будут независимы, что является важным предположением большинства моделей. Наконец, предоставляя исследователю точечную оценку цены, эти методы не дают никакого представления о распределении цены, которое для задач управления рисками имеет наибольшее значение.

Формальные методы, призванные решить задачу отсутствия ценовой информации в отдельные периоды времени (включая задачу прогнозирования), можно разделить на две основные группы. Первые предполагают, что поведение цен описывается случайным процессом с неизвестными, но постоянными параметрами. Соответственно, задача заключается в оценке параметров теми или иными способами и получении выборки из распределения с оцененными параметрами. Данные методы подвержены проблеме большой размерности: при совместном рассмотрении поведения цен даже умеренного количества активов, даже при рассмотрении процессов в широком смысле, т.е. считая, что поведение случайного процесса полностью описывается моментами первых двух порядков, количество параметров очень быстро растет, так что зачастую их количество превосходит количество известных величин. Описания этих методов мы коснемся очень коротко.

Вторая группа методов в основе имеет предположение о случайном характере самих параметров исследуемого случайного процесса цен. Известное априорное распределение этих параметров путем учета известной информации уточняется с целью получения апостериорного распределения неизвестных параметров процесса. При этом в качестве неизвестных параметров могут выступать и пропущенные данные. Этому подходу, носящему название Байесовского, будет уделено основное внимание в данном разделе.

Результатом применения любого из таких методов будет являться уже не набор наблюдений, в котором пропуски данных заполнены их точечными оценками, а большое количество таких наборов. В каждом из них пропуски заполнены результатом выборки из распределения пропущенных данных. Этот способ, носящий название многократного заполнения (Multiple Imputation), был впервые предложен Рубином (Rubin, 1978). При достаточно большом количестве проведенных заполнений таким образом построенный набор данных дает достаточно полное представление о характере совместного распределения пропущенных данных.

Для построения апостериорного распределения параметров широко используются Байесовские методы, в которых вид распределения восстанавливается по определяемым специальными методами модам апостериорного распределения. Последовательность операций при применении таких методов следующая:

1. Фиксируется грубое начальное приближение ненаблюдаемых параметров случайного процесса.

2. Осуществляется поиск мод апостериорного распределения параметров. Для поиска мод совместного распределения параметров используют, например, методы пошагового спуска или метод Ньютона. Другим подходом является поиск мод маргинального апостериорного распре-

деления параметров, в этом случае чаще всего используют EM-алгоритм и его модификации.

3. Построить апостериорное распределение как смесь нормальных или *t*-распределений с центрами, соответствующими найденным модам.

4. Получить выборку из апостериорного распределения.

Такой подход позволяет представить те или иные плотности апостериорного распределения в аналитическом виде. Для применения методов покоординатного спуска и метода Ньютона необходимо знание аналитической формы апостериорного распределения всех параметров. EM-алгоритм применяется в случае, когда маргинальное распределение части параметров имеет слишком сложный аналитический вид или вообще не может быть установлено, при этом условные распределения частей параметров при условии наблюдаемых значений процессов и остальных параметров имеют простой аналитический вид. Разделение параметров может быть проведено различными методами. В данном случае нас больше всего интересует случай, когда в качестве набора неизвестных параметров выступают собственно параметры процесса изменения цен и ненаблюдаемые цены (пропуски данных).

Впервые EM-алгоритм в полном объеме и под этим названием был предложен в работе Демпстера, Лэйрда и Рубина (Dempster, Laird, Rubin, 1977). В неполной форме для решения частных задач алгоритм был предложен Баумом и группой соавторов (Baum et al., 1970), а затем Орчардом и Вудбери (Orchard, Woodbury, 1972). Обзор работ по применению и развитию алгоритма приведен Менгом и Педлоу (Meng, Pedlow, 1992). Для целей настоящей работы мы будем использовать описание EM, приведенное в книге Джелмана и соавторов (Gelman et al., 2004). Пусть y — множество наблюдаемых значений, φ — множество неизвестных параметров процесса, которые подлежат оценке, γ — множество ненаблюдаемых значений (пропусков данных). Задача состоит в построении плотности апостериорного распределения $p(\varphi|y)$. Однако эта плотность имеет очень сложный вид, который помимо прочего зависит от размещения пропущенных данных относительно наблюдавшихся величин. В то же время распределения $p(\gamma|\varphi,y)$ и $p(\varphi|\gamma,y)$ имеют простую структуру.

Название EM происходит от названий двух последовательно выполняемых на этапе переоценки параметров процесса операций: вычисления математических ожиданий функций пропущенных значений (expectation) и максимизации получаемой апостериорной плотности (maximizing).

Алгоритм применяется для поиска локального максимума логарифма плотности апостериорного распределения, что позволяет использовать аналитическое выражение для этой плотности с точностью до множителя. Тогда очевидно:

$$\log(p(\varphi|y)) = \log(p(\gamma,\varphi|y)) - \log(p(\gamma|\varphi,y)).$$

Вычисляя математическое ожидание от обеих частей этого равенства по γ используя вероятностную меру $p(\gamma|\varphi^{old},y)$, где φ^{old} — текущая (предыдущая) оценка значений параметров, и принимая во внимание, что левая часть не зависит от γ, получаем:

$$\log(p(\varphi|y)) = E_{old}(\log(p(\gamma,\varphi|y))) - E_{old}(\log(p(\gamma|\varphi,y))). \quad (1)$$

Важным результатом в основании метода является тот факт, что второе слагаемое в правой части принимает наибольшее значение при $\varphi = \varphi^{old}$. Первое слагаемое выражается в следующем виде:

$$E_{old}(\log(p(\gamma,\varphi|y))) = \int_{\gamma\in\Gamma} \log(p(\gamma,\varphi|y)\,p(\gamma|\varphi^{old},y)d\gamma. \quad (2)$$

Если вычислить это выражение при $\varphi = \varphi^{new}$ таком, что выражение (2) принимает наибольшее значение или просто значение большее, чем на предыдущем шаге, то, принимая во внимание, что второй член в правой части (1) не уменьшится, получаем, что значение $\log(p(\varphi|y))$ увеличится. Максимизация выражения (2) является составной частью EM-алгоритма, простое его увеличение – обобщенного EM-алгоритма.

Последовательность шагов повторяется до момента установления сходимости. В данном случае под сходимостью понимается тот факт, что значения параметров процесса φ и γ стабилизируются.

Другим способом построения апостериорного распределения параметров процесса является метод, называемый в ряде источников пополнения данных (data augmentation) (см., например, Schafer, Olsen, 1998). Наиболее известен этот метод под названием цепи Монте-Карло с Марковским свойством (Markov chain Monte-Carlo) (Jones, 1998, Gelman et al., 2004).

Введем для начала обозначения, которых будем придерживаться до конца работы. В случае отступления от этих обозначений это будет специально указано. Итак, пусть T — набор моментов времени, в которые были произведены изменения, T_k — k-й по счету момент. X_k^o — набор наблюдаемых в момент времени k значений переменных процесса, X^o — совокупность всех наблюдаемых значений процесса. X_k^u и X^u — нена-

блюдаемые значения процесса в момент k и совокупность всех ненаблюдаемых значений процесса. $X = (X^o, X^u)$ — совокупность всех значений процесса. φ — значения параметров процесса, распределение которых Φ. В случае, когда необходимо подчеркнуть, что величины получены на s-м шаге процедуры, будем писать $X_k^{u(s)}$, $X^{u(s)}$, $\varphi^{(s)}$ и $\Phi^{(s)}$ соответственно.

Введя такие обозначения, мы можем кратко описать суть метода цепи Монте-Карло с Марковским свойством:

1. Фиксируется исходное распределение параметров распределения процесса $\Phi^{(0)}$ и значения параметров, полученных выборкой из этого распределения $\varphi^{(0)}$.

2. Производится выборка ненаблюдавшихся значений процесса X^u из распределения с плотностью распределения вероятности $p(X^{u(1)} | \varphi^{(0)}, X^o)$.

3. Производится переоценка параметров распределения $\varphi^{(1)}$ с использованием всего множества значений процесса из распределения $p(\varphi^{(1)} | X^{u(1)}, X^o)$.

Последовательность операций повторяется до установления сходимости. В данном случае сходимость понимается иначе, чем при рассмотрении EM-алгоритма. Если там свидетельством сходимости является незначительное изменение самих параметров наблюдаемого процесса на каждом шаге, то в данном случае под сходимостью следует понимать наличие лишь незначительных изменений распределения этих параметров. Сами же они от итерации к итерации могут существенно изменяться.

В таким образом поставленном процессе наиболее сложными задачами являются построение распределений $p(X^{u(s)} | \varphi^{(s-1)}, X^o)$ и $p(\varphi^{(s)} | X^{u(s-1)}, X^o)$, а также получение выборок из этих распределений. В литературе описывается решение этих задач следующими способами для одного важного частного случая – когда изменение цен описывается моделью Башелье — Самуэльсона, т.е. удовлетворяют следующим свойствам:

1. Изменения цен актива на непересекающихся промежутках времени независимы в совокупности.

2. Распределение изменения цен актива за промежуток времени не зависит от времени начала и конца, а зависит только от длины промежутка.

3. Траектория относительных цен актива является непрерывной.

Хорошо известно, что при сделанных предположениях поведение цен активов описывается стохастическим дифференциальным уравнением

$$dx = \mu x dt + \sigma x dw,$$

а непрерывно начисляемый доход имеет нормальное распределение:

$$\log\left(\frac{x_{t+1}}{x_t}\right) \sim \mathrm{N}\left(\mu - \frac{1}{2}\sigma^2, \sigma^2\right)$$

Сделав замену $\beta = \mu - \frac{1}{2}\sigma^2$, это можно записать следующим образом:

$$\log\left(\frac{x_{t+1}}{x_t}\right) \sim \mathrm{N}\left(\beta, \sigma^2\right)$$

Одной из наиболее часто используемых дискретных аппроксимаций такого стохастического дифференциального уравнения является аппроксимация Эйлера:

$$\hat{X}_{(k+1)h} = \hat{X}_{kh} + h\mu(\hat{X}_{kh}, \phi) + \sqrt{h}\sigma(\hat{X}_{kh}, \phi)\varepsilon_{k+1},$$

где ε_{k+1} — нормальная и одинаково распределенная случайная величина $N(0, I_D)$, I_D — единичная матрица размерности D — количества рассматриваемых активов, h – шаг дискретизации. Для краткости будем писать X_k вместо X_{kh}, учитывая h неявно.

Хакер, Польсен и Росси (Jacquoer, Polson, Rossi, 1994) предложили для рассматриваемого частного случая способ, названный ими циклической цепью Меторполис (cyclic Metropolis chain). Выборка из многомерного распределения с плотностью $p\left(X^u \middle| \phi, X^0\right)$ разбивается на большое число одномерных выборок. Эти одномерные выборки строятся для каждого элемента X^u в зависимости от ϕ, X^0 и всех остальных элементов X^u. Авторы создают цепь Марковских цепей, повторно производя такие условные выборки для каждого элемента X^u и параметров ϕ. Так как одномерные условные распределения неизвестны, то выбрка производится по алгоритму Метрополис — Гастингс (Metropolis — Hasting), описанному, например, Чибом и Гринбергом (Chib and Greenberg, 1995).

Вообще говоря, существует много способов разделения на блоки многомерной выборки с плотностью $p\left(X^u \middle| \phi, X^0\right)$ на одномерные. В данном случае наиболее удобно разбить X^u по компонентам X_k^u.

Пусть X_{-k}^u — это множество всех ненаблюдаемых данных кроме X^u, ненаблюдаемых данных в момент времени k. Если все элементы $X_k = \left(X_k^0, X_k^u\right)$ наблюдаются, то X_k^u равно 0 и, следовательно, нет необходимости производить выборку. Если же X_k полностью не наблюдается, то X_k^u состоит по крайней мере из одного элемента, и нашей целью

является разыграть выборку из условного распределения с плотностью $p\left(X_k^u \middle| X_{-k}^u, X^0, \phi\right)$. Так как рассматривается процесс, обладающий Марковским свойством, значимыми являются только соседние наблюдения:

$$p\left(X_k^u \middle| X_{-k}^u, X^0, \phi\right) = p\left(X_k^u \middle| X_{k-1}, X_k^0, X_{k+1}, \phi\right)$$

Используя формулу условной вероятности можно показать, что эта плотность будет пропорциональна

$$p\left(X_k^u, X_k^0, X_{k+1} \middle| X_{k-1}, \phi\right)$$

или

$$p\left(X_{k+1} \middle| X_k^u, X^0, X_{k-1}, \phi\right) p\left(X_k^u \middle| X_{k-1}, X_k^0, \phi\right)$$

В силу наличия Марковского свойства это произведение эквивалентно следующему выражению:

$$\pi\left(X_k^u\right) = p\left(X_{k+1} \middle| X_k^u, X^0, \phi\right) p\left(X_k^u \middle| X_{k-1}, X_k^0, \phi\right)$$

Это выражение является произведением плотностей нормального распределения, которые мы для краткости обозначим $\pi\left(X_k^u\right)$. Оно легко может быть получено при помощи алгоритма Метрополис-Гастингс. Только в редких случаях это произведение пропорционально известной нормированной плотности, что делает использование этого алгоритма необязательным.

Для построения цепи Метрополис необходимо определить плотность распределения значений-кандидатов $q\left(x, x^*\right)$ для замены имеющихся в цепи значений x вновь полученными x^*. Вероятность принятия кандидата описывается выражением:

$$\alpha\left(x, x^*\right) = \min\left\{\frac{\pi\left(x^*\right) q\left(x^*, x\right)}{\pi\left(x\right) q\left(x, x^*\right)}, 1\right\}$$

Если x^* отвергается, то текущее соответствующее значение переменной в цепи остается равным x. Результатом работы алгоритма Метрополис — Гастингс является Марковская цепь, имеющая инвариантное распределение с плотностью, пропорциональной $\pi\left(x\right)$.

Естественный способ определения плотности $q\left(x, x^*\right)$, который помимо прочего упрощает определение вероятности принятия значения

кандидата $\alpha\left(X_k^u, X_k^{u^*}\right)$, следующий. В имевшем ранее место выражении:

$$p\left(X_k^u \middle| X_{-k}^u, X^0, \phi\right) \infty p\left(X_{k+1} \middle| X_k^u, X_k^0, \phi\right) p\left(X_k^u \middle| X_k^0, X_k\right.$$

Вторая плотность в правой части, взятая отдельно, означает, что X_k^u имеет нормальное распределение. Тогда плотность распределения для разыгрывания кандидатов определяется как

$$q\left(X_k^u, X_k^{u^*}\right) = q\left(X_k^{u^*}\right) = p\left(X_k^{u^*} \middle| X_k^0, X_{k-1}, \phi\right)$$

Заметим, что текущее значение цепи Маркова X_k^u не входит в выражение для плотности кандидатов – q порождает так называемую независимую цепь Метрополис.

Конечно, элементы цепи могут не быть независимыми в действительности, потому что они связаны вероятностью принятия α. Эта вероятность упрощается выбором q:

$$\alpha\left(X_k^u, X_k^{u^*}\right) = \min\left\{\frac{p\left(X_{k+1} \middle| X_k^{u^*}, X_k^0, \phi\right)}{p\left(X_{k+1} \middle| X_k^u, X_k^0, \phi\right)}, 1\right\}$$

По существу моделируем процесс, переходя от шага (k–1) к шагу k для генерации кандидата $X_k^{u^*}$, затем принимаем $X_k^{u^*}$, полученное из предыдущей выборки X_k^u, в зависимости от того, насколько он подходит для значения перед X_{k+1}.

Главным преимуществом для такого последовательного моделирования кандидатов является простота. Для построения цепи Метрополис с использованием последовательного моделирования требуется всего лишь часть вычислений, необходимых для построения более сложных цепей, что значительно снижает вычислительную сложность.

Большинство описанных методов справедливы, если распределение исследуемого процесса отлично от нормального. Однако в этом случае аналитическая форма апостериорного распределения параметров не может быть получена. В этом случае в основном предлагается покомпонентно исследовать параметры распределения процесса. В частности, для получения выборки из совместного распределения параметров применять алгоритм, аналогичный описанной цепи Метрополис.

2. Численные методы восстановления пропущенной информации

В первом разделе настоящей работы приведен обзор описанных в литературе методов, использующихся для восстановления пропущенной информации. Данный раздел посвящен непосредственному описанию методов, использовавшихся для этих целей в настоящей работе. Результаты численных экспериментов вынесены в отдельный раздел.

Прежде чем приступить к описанию использовавшихся методов, рассмотрим в соответствии с идеологией Чатфилда (Chatfield, 1985) состав и качество информации, использованной в данной работе.

Первым набором информации является статистика по результатам торгов 94 наиболее ликвидных государственных, муниципальных и корпоративных облигаций на Московской межбанковской валютной бирже за период с 10 января 2006 г. по 21 февраля 2007 г. Периодичность информации – один торговый день. В ее состав входят:

1. Цены открытия торгов.
2. Цены закрытия торгов.
3. Максимальная цена за день.
4. Минимальная цена за день.
5. Средневзвешенные цены облигации за день.
6. Количество совершенных за день сделок.
7. Суммарный объем совершенных за день сделок.
8. Лучшая котировка Bid на момент закрытия торгов.
9. Лучшая котировка Ask на момент закрытия торгов.

Информация по государственным облигациям является полной, поскольку по данному виду бумаг возможно совершение только биржевых сделок. Информация по муниципальным и корпоративным облигациям содержит также только рыночные сделки, что составляет в среднем около 20% объема дневных торгов. Информация о торгах, проводившихся в режиме переговорных сделок, в данном наборе не представлена.

Набор характеризуется неравномерностью распределения сделок во времени и наличием среднего количества пропущенной информации. Рис. 1, 2 дают представление о плотности распределения пропусков и объемах совершенных сделок за период 1.12.2006—21.02.2007.

В случае наличия информации о сделках возможно отсутствие информации о котировках на момент закрытия, напротив, в случае отсутствия сделок, как правило, имеется информация о котировках. Одновременное отсутствие информации о сделках и котировках является единичным случаем.

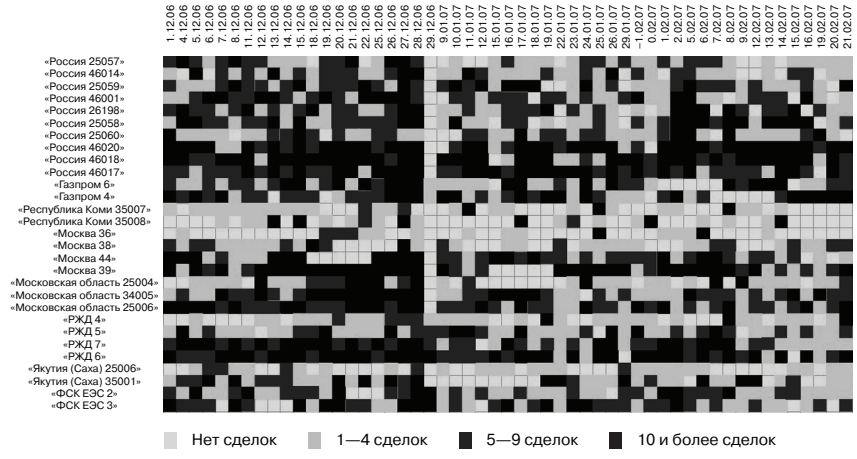

Рис. 1. Распределение количества сделок с наиболее ликвидными облигациями на ММВБ за период 1.12.2006—21.02.2007

Цены открытия и закрытия торгов, а также максимальная, минимальная и средняя цены представлены в данном наборе только при условии наличия сделок. В случае наличия только одной сделки все цены, за исключением цены закрытия, совпадают и равны цене этой сделки.

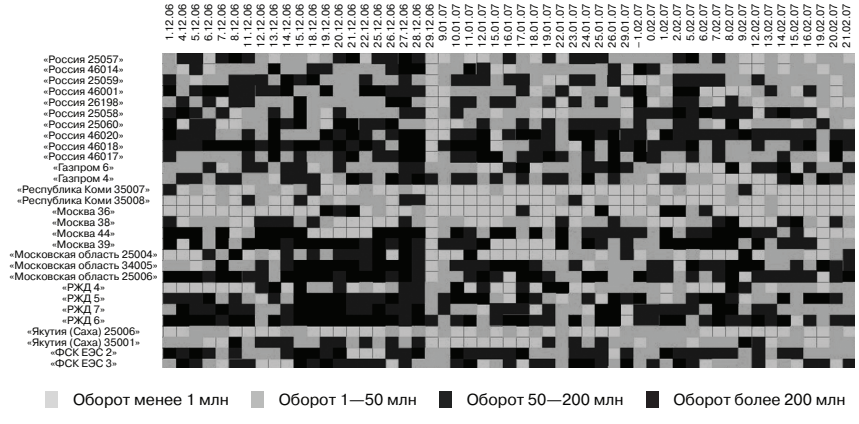

Рис. 2. Распределение объемов сделок с наиболее ликвидными облигациями на ММВБ за период 1.12.2006—21.02.2007 (рублей)

Набор, содержащий рыночную информацию по сделкам с облигациями, используется в данной работе для отработки процедур фильтрации

и восстановления информации. Следующие два набора используются только для работы процедур восстановления информации.

Второй набор содержит информацию о ценах закрытия торгов по наиболее ликвидным акциям на Московской межбанковской валютной бирже. Рассматриваемые акции являются высоколиквидными, число сделок по каждой из них редко падает ниже 10 в день. Количество пропусков не превышает 5%. Данный набор используется для проверки точности получаемых точечных оценок восстановленных данных при произвольном различном количестве случайным образом исключенных значений.

Третий набор используется для оценки точности определения характеристик распределения процесса цен. Он представляет собой несколько реализаций аппроксимации Эйлера многомерного геометрического броуновского движения с известными характеристиками и рядом случайным образом исключенных значений.

Объектом дальнейшего исследования является стохастическое дифференциальное уравнение, описывающее многомерное геометрическое броуновское движение с неизвестными вектором параметров сноса и матрицей ковариации, а, точнее говоря, – дискретная аппроксимация Эйлера такого уравнения. Формальное описание такой аппроксимации приведено в разделе 1. Часть значений переменных не наблюдаются. Задачей является определение неизвестных характеристик процесса в том или ином виде (распределение или точечная оценка) и распределения пропущенных значений переменных.

Как было указано ранее, при сделанных предположениях поведение цен активов описывается стохастическим дифференциальным уравнением

$$dX = \beta X dt + \sigma X dw,$$

а непрерывно начисляемый доход имеет нормальное распределение

$$\log\left(\frac{X_{t+1}}{X_t}\right) \sim \mathrm{N}\left(\beta - \frac{1}{2}\sigma^2, \sigma^2\right)$$

Сделав замену $m = \beta - \frac{1}{2}\sigma^2$, это можно записать следующим образом:

$$\log\left(\frac{X_{t+1}}{X_t}\right) \sim \mathrm{N}\left(m, \sigma^2\right)$$

Одной из наиболее часто используемых дискретных аппроксимаций такого стохастического дифференциального уравнения является аппроксимация Эйлера, которая и используется в настоящей методике:

$$\hat{X}_{(k+1)h} = \hat{X}_{kh} + hm(\hat{X}_{kh}, \phi) + \sqrt{h}\sigma(\hat{X}_{kh}, \phi)\varepsilon_{k+1},$$

где ε_{k+1} — нормальная и одинаково распределенная случайная величина $N(0, I_D)$, I_D — единичная матрица размерности D — количества рассматриваемых активов, h – шаг дискретизации. Для краткости будем писать X_k вместо X_{kh}, учитывая h неявно. Такая дискретизация аналогична описанной Джонсом (Jones, 1998).

Введя такие обозначения, можно кратко описать суть метода цепи Монте-Карло с Марковским свойством:

1. Фиксируется исходное распределение параметров распределения процесса $\Phi^{(0)}$ и значения параметров, полученных выборкой из этого распределения $\varphi^{(0)}$.

2. Производится выборка ненаблюдавшихся значений процесса X^u из распределения с плотностью распределения вероятности $p(X^{u(1)} | \varphi^{(0)}, X^o)$.

3. Производится переоценка параметров распределения $\varphi^{(1)}$ с использованием всего множества значений процесса из распределения $p(\varphi^{(1)} | X^{u(1)}, X^o)$.

4. Шаги 2 и 3 повторяются установленное количество раз. При этом в качестве исходной информации используется результат выборки и переоценки параметров на предыдущем шаге.

На шаге 2 получение выборки пропущенных данных производится в силу марковского свойства процесса цен последовательно для каждого момента времени k. Известными являются параметры распределения m, σ процесса изменения цен, а также значения цен в соседние моменты времени $k-1$, $k+1$. По ним необходимо установить вид распределения пропущенных данных в момент времени k. При сделанных предположениях это распределение является нормальным. Его параметры вычисляются в два этапа. На первом вычисляются праметры распределения всех цен при условии известных значений параметров процесса и цен в соседние моменты времени:

$$\mu = X_{k-1} + m \cdot \Delta t(k) + \\ + \frac{\Delta t(k)}{\Delta t(k+1) + \Delta t(k)}\left(X_{k+1} - X_{k-1} - m \cdot (\Delta t(k+1) + \Delta t(k))\right), \qquad (3)$$

$$\Sigma^2 = \frac{\Delta t(k)\Delta t(k+1)}{\Delta t(k+1)+\Delta t(k)}\sigma^2 \qquad (4)$$

где $\Delta t(k) = t(k) - t(k-1)$ — разница между соседними моментами времени.

На втором этапе вычисляются параметры распределения ненаблюдаемых цен в момент времени k при условии фиксированных значений параметров μ, Σ, вычисленных ранее, и наблюдаемых значениях цен в момент времени k:

$$\tilde{\mu} = E\left(X^u \middle| X_{k-1}, X_k^0, X_{k+1}, \phi^{(0)}\right) = \mu^u + \Sigma_{ou}^2 \cdot \Sigma_{oo}^{2\ -1} \cdot (X^o - \mu^o), \qquad (5)$$

$$\tilde{\Sigma}^2 = \left\{\mathrm{cov}\left(X_{ki}^u, X_{kj}^u \middle| X_{k-1}, X_k^0, X_{k+1}, \phi^{(0)}\right)\right\}_{i,j} = \Sigma_{uu}^2 + \Sigma_{ou}^2 \cdot \Sigma_{oo}^{2\ -1} \cdot \Sigma_{uo}^2 \qquad (6)$$

Здесь μ, Σ — вектор математических ожиданий и ковариационная матрица апостериорного распределения процесса. Индексы *u* и *o* означают, как и ранее, что величины соответствуют ненаблюдаемым и наблюдаемым изменениям цен. У ковариационных матриц указываются два индекса, например, Σ^{ou} — матрица ковариаций ненаблюдаемых и наблюдаемых изменений цен.

Таким образом, распределение ненаблюдаемых значений цен в момент времени является в общем случае многомерным нормальным, а его параметры выражены через параметры распределения процесса цен и значения цен в соседние моменты времени. Для получения выборки из такого распределения его ковариационная матрица раскладывается по базису из ее собственных векторов:

$$\Sigma^2 = S \cdot \Lambda \cdot S'. \qquad (7)$$

После чего искомая многомерная случайная величина получается следующим образом:

$$\xi = \mu + S \cdot \varepsilon, \qquad (8)$$

где $\varepsilon \in N(\vec{0}, \Lambda)$.

Такой способ получения выборки многомерного нормального распределения в отличие от традиционно используемого преобразования Халецкого является устойчивым и не приводит к накоплению вычислительных ошибок при многократном применении.

На третьем этапе построения цепи Монте-Карло с Марковским свойством необходимо произвести переоценку распределения параметров случайного процесса и выборку значений параметров из апостериорного распределения, использующего в качестве входной информации как

наблюдаемые значения, так и результат выборки из распределения ненаблюдаемых значений на предыдущем шаге. В данном случае нормальность закона распределения процесса позволяет применить метод сопряженных семейств распределений. Решение задачи определения характеристик распределения ϕ при известных значениях X^0, X^u наблюдаемых и ненаблюдаемых параметров получается на основе следующего известного факта из теории случайных процессов (см., например, де Гроот, 1974).

Рассмотрим задачу выбора из многомерного нормального распределения, для которого неизвестны вектор средних и матрица точности (матрица, обратная к ковариационной матрице). Пусть $x_1, ..., x_n$ — выборочные значения наблюдений $X_1, ..., X_n$. Обозначим через s симметрическую неотрицательно определенную $k \times k$-матрицу, определяемую равенством

$$s = \sum_{i=1}^{n} (x_i - \bar{x})((x_i - \bar{x})'. \qquad (9)$$

Пусть $X_1, ..., X_n$ — повторная выборка из многомерного нормального распределения с неизвестным вектором средних M и неизвестной матрицей точности B. Пусть, далее, априорное совместное распределение M и R таково, что условное распределение M при $R = r$ является многомерным нормальным распределением с вектором средних μ и матрицей точности vr, $\mu \in \Re^k, \nu > 0$, а маргинальное распределение R есть распределение Уишарта с α степенями свободы и матрицей точности τ, причем $\alpha > k - 1$ и τ — симметрическая положительно определенная матрица[1]. Тогда апостериорное совместное распределение M и R при $X_i = x_i$ $(i = 1, ..., n)$ описывается следующим образом. Условное распределение M при $R = r$ есть многомерное нормальное распределение с вектором средних μ^* и матрицей точности $(v + n)r$, где

$$\mu^* = \frac{v\mu + n\bar{x}}{v + n}, \qquad (10)$$

а маргинальное распределение R есть распределение Уишарта с $\alpha + n$ степенями свободы и матрицей точности

[1] Пусть $X_1, ..., X_n$ – случайная выборка, состоящая из k-мерных векторов, распределенных по многомерному нормальному закону с вектором средних $\vec{0}$ и ковариационной матрицей Σ размера $k \times k$. Пусть V — случайная симметрическая матрица размера $k \times k$, определяемая равенством: $V = \sum_{i=1}^{n} X_i \cdot X_i'$. Эта случайная матрица имеет распределение Уишарта с n степенями свободы и параметрической матрицей Σ.

$$\tau^* = \tau + s + \frac{vn}{v+n}(\mu - \overline{x})(\mu - \overline{x})'. \qquad (11)$$

Следует, однако, подчеркнуть, что следующий результат получен для результатов повторной выборки, т.е. для случая, когда величины $X_1,...,X_n$ являются независимыми одинаково распределенными. Это верно по отношению к значениям приращений процессов цен, но не для самих значений цен. Соответственно, предварительным этапом при перерасчете параметров распределения процесса цен является переход к приращениям.

Таким образом, описаны основные этапы построения цепи Монте-Карло с Марковским свойством. Далее приведено описание EM-алгоритма, реализованного для проведения сопоставления с ним результатов работы предложенной процедуры. Джелманом и соавторами (Gelman et al., 2004) приводится результат аналитического решения на обоих шагах EM-алгоритма. Это решение будет использовано без изменений, поэтому имеет смысл привести его здесь полностью.

Пусть y — повторная выборка из n элементов d-мерного нормального распределения с неизвестными вектором средних и матрицей ковариаций. Будем обозначать y_{ij} j-ю компоненту i-го элемента выборки. В остальном обозначения аналогичные использованным в разделе 1. Достаточными статистиками нормального семейства распределений являются

$$\sum_{i=1}^{n} y_{ij}, \quad j = 1,...,d, \qquad (12)$$

$$\sum_{i=1}^{n} y_{ij} y_{ik}, \quad j,k = 1,...,d \qquad (13)$$

Обозначая $\theta^{old} = (\mu^{old}, \Sigma^{old})$ — значения параметров нормального распределения на предыдущем шаге алгоритма, выпишем в аналитическом виде результаты обоих шагов алгоритма.

На шаге E, вычисления математических ожиданий,

$$E\left(\sum_{i=1}^{n} y_{ij} \middle| y^o, \theta^{old}\right) = \frac{1}{n}\sum_{i=1}^{n} y_{ij}^{old}, \quad j = 1,...,d, \qquad (14)$$

$$E\left(\sum_{i=1}^{n} y_{ij} y_{ik} \middle| y^o, \theta^{old}\right) = \frac{1}{n}\left(\sum_{i=1}^{n} (y_{ij}^{old} \cdot y_{ik}^{old} + c_{ijk}^{old})\right), \quad j,k = 1,...,d, \qquad (15)$$

где использованы следующие обозначения:

$$y_{ij}^{old} = \begin{cases} y_{ij}, \ y_{ij} \in y_i^o, \\ E\left(\sum_{i=1}^{n} y_{ij} \middle| y^o, \theta^{old}\right), \ y_{ij} \in y_i^u; \end{cases}$$

$$c_{ijk}^{old} = \begin{cases} 0, \ y_{ij} \in y_i^o \ \textit{и} \ y_{ik} \in y_k^o, \\ \mathrm{cov}\left(y_{ij}, y_{ik} \middle| y^o, \theta^{old}\right), \ \textit{в противном случае}. \end{cases}$$

Вычисление условных математических ожиданий и ковариаций ненаблюдаемых компонент при известных значениях наблюдаемых компонент и значениях параметров является стандартной задачей регрессии и описывается выражениями (3) и (4).

На шаге M, максимизации функции правдоподобия, новые значения параметров распределения процесса вычисляются следующим образом:

$$\mu_j^{new} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}^{old}, \quad j = 1,...,d, \qquad (16)$$

$$\sigma_{jk}^{new} = \frac{1}{n}\left(\sum_{i=1}^{n} y_{ij}^{old} \cdot y_{ik}^{old} + c_{ijk}^{old}\right) - \mu_j^{new}\mu_k^{new}, \quad j,k = 1,...,d \qquad (17)$$

Таким образом, все шаги EM-алгоритма имеют аналитическое представление. Последовательно повторяя шаги алгоритма, с каждой итерацией увеличивается значение функции правдоподобия. Тем самым в пределе можно оказаться в сколь угодно малой окрестности моды апостериорного распределения параметров. Недостатком алгоритма следует считать низкую скорость сходимости. Это особенно важно, поскольку каждый раз при вычислении математических ожиданий на пункте E необходимо проводить обращение плохо обусловленных матриц, в связи с чем возможна неустойчивая работа алгоритма и отсутствие его сходимости.

3. Анализ результатов численных экспериментов

Данный раздел посвящен интерпретации результатов, полученных при проведении численных экспериментов по восстановлению данных на примерах, описанных в разделе 2. Основное внимание будет уделено

наиболее полному набору информации, описывающему результаты торгов облигациями на Московской межбанковской валютной бирже.

Результаты восстановления информации и способы их применения могут быть различны. Зачастую, как при решении задачи о построении кривой доходности облигаций, достаточно получения точечной оценки пропущенных значений и характеристики точности.

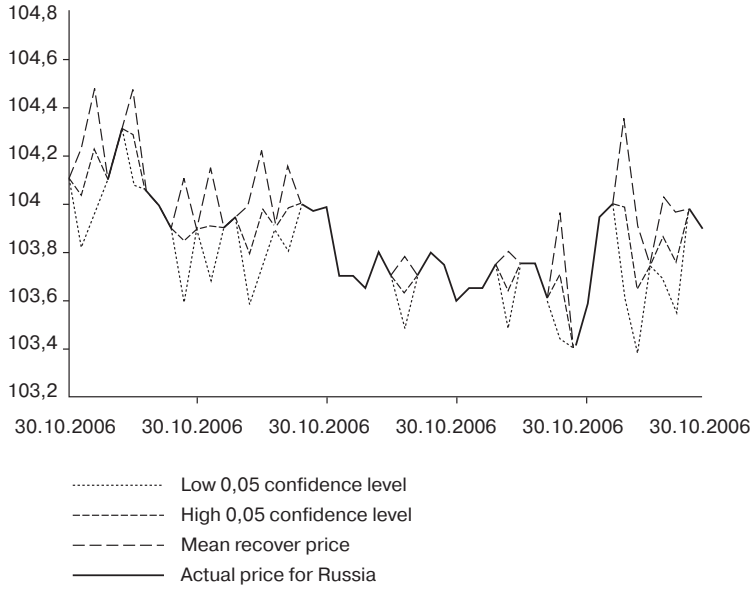

Рис. 3. Исходные значения цен и результаты восстановления пропущенной информации по торгам облигации «Россия 25057» за период 30.10.2006—17.01.2007

Сплошная линия показывает фактические значения цен, жирная пунктирная – оценку среднего значения цен при помощи цепи Монте-Карло, тонкие пунктирные линии показывают границы 10% доверительного интервала.

Рис. 3 иллюстрирует типичное поведение процесса цен отдельной облигации и результаты восстановления данных. Среднее значение цены получено как среднее арифметическое результатов итераций. Верхняя и нижняя границы доверительного интервала представляют собой выборочные квантили с уровнями значимости 0,05 и 0,95 соответственно. Количество шагов в цепи Монте-Карло с Марковским свойством 2200, из них первые 200 отброшены, чтобы исключить эффекты, связанные

со сходимостью процесса. Пример распределения результатов для одного отдельно взятого пропущенного значения приведен на рис. 4.

Определить качество оценки с точки зрения способности алгоритма воспроизвести неизвестные значения параметров можно только при работе с распределением с заранее известными характеристиками. С такой целью был проведен ряд численных экспериментов. В качестве исходного процесса использовалось многомерное геометрическое броуновское движение с параметрами, более или менее правдоподобно воспроизводящими динамику рыночных процессов. Случайным образом проводилось исключение элементов указанного процесса с последующим восстановлением совместного распределения параметров и пропущенных данных. Рис. 5, 6 иллюстрирует качество полученных оценок с точки зрения точности восстановления параметров процесса.

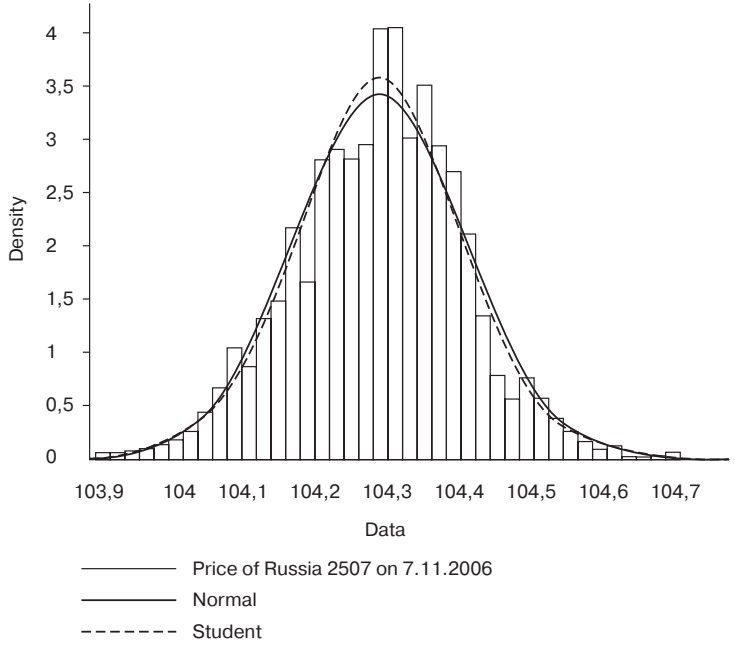

Рис. 4. Эмпирическое распределение восстановленной цены облигации «Россия 25057» за 7.11.2006

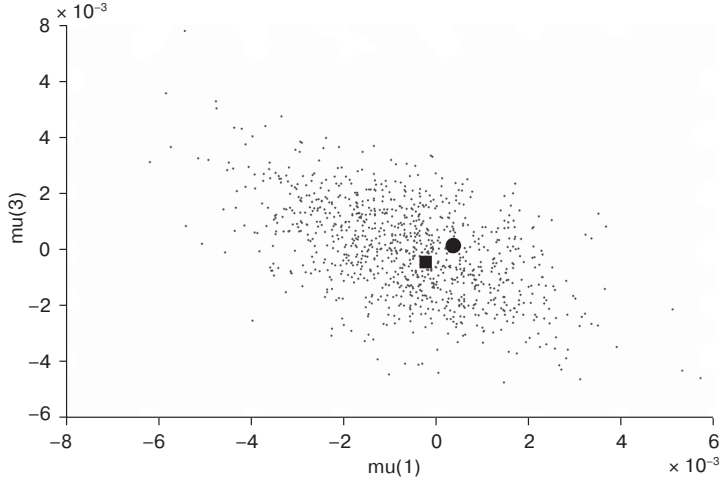

Рис. 5. Совместное эмпирическое распределение значений среднего двух компонент

Жирная круглая точка показывает реальное значение параметров. Жирная квадратная точка – оценку тех же параметров при помощи EM-алгоритма

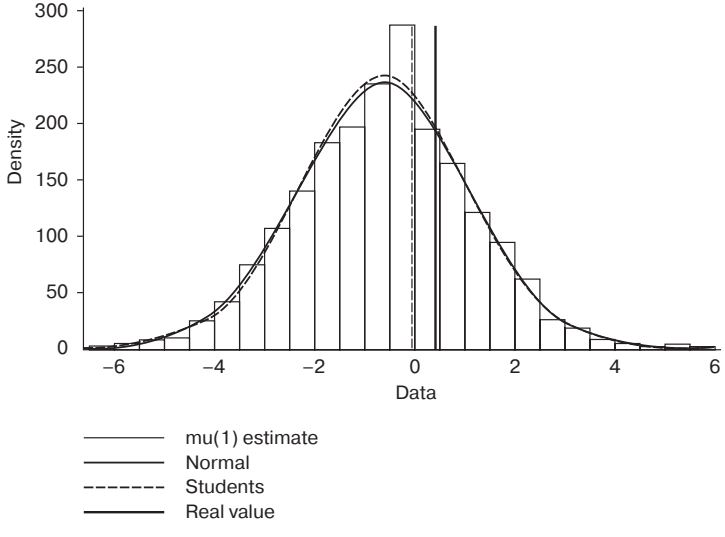

Рис. 6. Гистограмма эмпирического распределения значения среднего

Сплошная вертикальная линия показывает реальное значение параметра. Пунктирная вертикальная линия показывает положение оценки параметра при помощи EM-аглоритма.

Рис. 6 дает представление о точности восстановления пропущенных данных. Следует отметить, что получаемые оценки в большинстве случаев лежат в зоне наибольшей концентрации полученных эмпирических распределений. По результатам реальные значения параметров никогда не выходили за пределы 95% доверительного интервала получаемых эмпирических распределений.

Одновременно была проведена оценка параметров распределения цен и восстановление пропущенных данных при помощи EM-алгоритма. Результаты сравнения использованных методов можно трактовать двояко.

С одной стороны, EM-алгоритм дает оценки параметров распределения процесса, совпадающие с модами эмпирического распределения параметров, полученного при построении цепи Монте-Карло с Марковским свойством. С другой стороны, оценки значений получаемых величин значительно отличаются как от их реальных значений, так и от оценок, получаемых при помощи восстановления распределения (см. рис. 6, 7). При этом наблюдается очень медленная динамика изменений оценок значений восстанавливаемых пропущенных данных (см. рис. 8), в результате чего восстановленный временной ряд получается переглаженным, а оценки волатильности цен, полученные на основании таких данных, являются заниженными. Оценки, получаемые при помощи цепи Монте-Карло с Марковским свойством, лишены указанных недостатков.

Следует отметить еще одно достоинство метода восстановления данных при помощи Марковской цепи. Полученные результаты могут быть непосредственно использованы для получения распределения стоимости портфеля активов в будущем. Для этого к интервалу, содержащему доступную информацию, присоединяется интервал прогнозирования необходимой длины. После восстановления данных пользователь получает совместное распределение пропусков и будущих значений цен в предположении, что характер процесса цен сохранится. По указанному распределению могут быть получены любые характеристики стоимости портфеля – точечные и интервальные оценки стоимости, разброса. Оценка показателя Value-at-Risk при этом является тривиальной и заключается в вычислении выборочного квантиля распределения относительных потерь портфеля. Применение такого подхода иллюстрирует рис. 9.

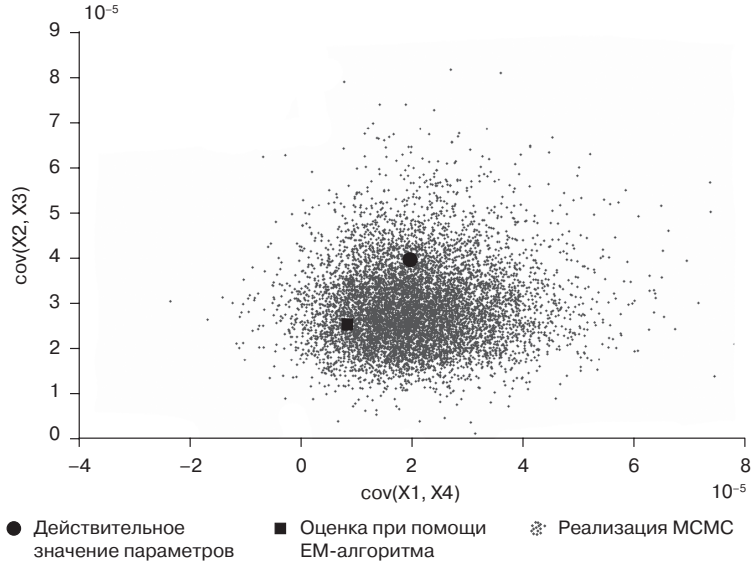

Рис. 7. Эмпирическое распределение значений коэффициентов ковариации

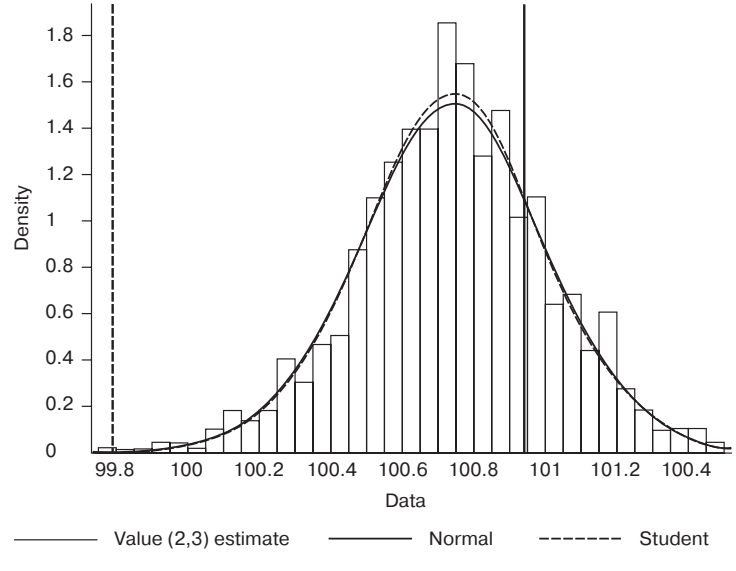

Рис. 8. Гистограмма эмпирического распределения пропущенного значения

Сплошная вертикальная линия показывает реальное значение величины, пунктирная вертикальная – оценку при помощи EM-алгоритма

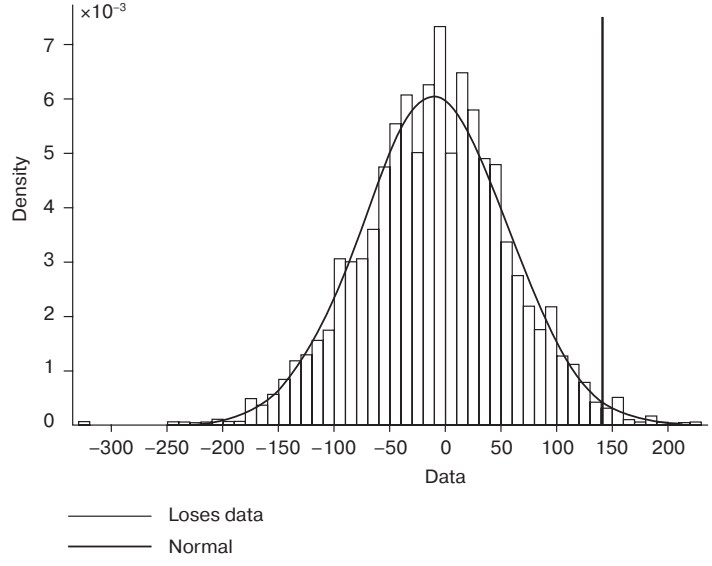

Рис. 9. Распределение относительных потерь условного портфеля

Вертикальная линия показывает значение 1% показателя VaR

Заключение

В современном обществе, основанном на использовании информации, ее своевременность, полнота и достоверность играют большую роль. Особое значение эти характеристики информации приобретают, когда она в больших объемах используется автоматическими системами. Ярким примером является область управления рисками, где широко используются автоматические системы, регулярно обрабатывающие огромные массивы информации о результатах торгов различных активов. Природа процесса торгов такова, что наличие полной и достоверной информации является скорее исключением чем правилом.

Фактически существуют два способа борьбы с проблемой некачественных данных – модификация численных методов с целью адаптации их к некачественной входной информации или ее предварительная подготовка. Однако зачастую даже незначительная модификация численных методов ведет к многократному возрастанию их сложности. Как

правило, это означает невозможность использования аналитических выражений, а последовательное применение различных численных методов означает потерю точности вычислений. На первое место, однако, выходят проблемы быстродействия – коммерческие предприятия не могут позволить себе использовать медленные численные процедуры в повседневной практике.

Настоящая работа призвана снять остроту проблемы некачественной информации путем ее предварительной подготовки – восстановления пропущенных данных.

В области восстановления информации основным направлением развития является использование все более сложных видов вероятностных распределений рассматриваемых процессов, в частности, типична ориентация на распределения с так называемыми тяжелыми хвостами. Однако, как показывают результаты проведения фильтрации, подавляющее большинство значений цен, составляющих тяжелые хвосты, имеют крайне низкую правдоподобность, что определяется в первую очередь вновь включенной информацией. Фактически это позволяет использовать для описания рыночных процессов геометрическое броуновское движение, что в свою очередь позволяет широко использовать при восстановлении информации аналитические выражения. В частности, это позволило применить метод сопряженных семейств распределений, который по причине очень малого числа известных семейств крайне редко применяется на практике.

Результаты работы предложенных методов могут быть использованы как в виде входной информации для других вычислительных процедур (например, по построению безрисковой кривой доходности), так и непосредственно, например, для решения задачи оценки показателя Value-at-Risk портфеля. Полученные результаты могут служить подтверждением практической значимости предложенного подхода.

Литература

Гроот М. де. Оптимальные статистические решения. М.: Мир, 1974.

Baum L.E., Petrie T., Soules G., Weiss N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains // Annals of Mathematical Statistics. Vol. 41. P. 164—171.

Chatfield C. (1985) The Initial Examination of Data // Journal of the Royal Statistical Society. Series A (General). Vol. 148. No. 3. P. 214—253.

Chib S., Greenberg E. (1995) Understanding the Metropolis-Hastings algorithm // American Statistician. Vol. 49. P. 327—335.

Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion) // Journal of the Royal Statistical Society. Vol. 39. P. 1—38.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (2004) Bayesian Data Analysis // 2nd ed. A CRC Press Company, Boca Raton, Florida.

Jacquier E., Polson N.G, Rossi P.E. (1994) Bayesian Analysis Models // Journal of Business and Economic Statistics. Vol. 12. P. 371—389.

Jones C.S. (1998) Bayesian Estimation of Continuous-Time Finance Models // http://www.ssb.rochester.edu/fac/c_s_jones.

Leropy S.F. (1989) Efficient Capital Markets and Martingales // Journal of Economic Literature. 1989 (Dec.). Vol. 27. No. 4. P. 1583—1621.

Meng X.L., Pedlow S. (1992) EM: a bibliographic review with missing articles // Proceedings of the Statistical Computing Section, American Statistical Association, 24—27.

Rubin D.B. (1978b) Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse (with discussion). Proceedings of the American Statistical Association, Survey Research Methods Section, 20—34.

Samuelson P.A. Proof That Properly Anticipated Prices Fluctuate Randomly // Industrial Management Review. 1965 (Spring): 6:2. P. 41.

Schafer J.L., Olsen M.K. (1998) Multiple imputation for multivariate missing-data problems: a data analyst's perspective // The Pennsylvania State University.

Препринт WP16/2007/02
Серия WP16
Финансовая инженерия,
риск-менеджмент и актуарная наука

А.В. Косьяненко

Опыт восстановления пропущенной рыночной информации на основе Байесовского подхода

Публикуется в авторской редакции

Зав. редакцией *А.В. Заиченко*
Технический редактор *Ю.Н. Петрина*

ЛР № 020832 от 15 октября 1993 г.
Отпечатано в типографии ГУ ВШЭ с представленного оригинал-макета.
Формат 60×84 $^{1}/_{16}$. Бумага офсетная. Тираж 150 экз. Уч.-изд. л. 1,9.
Усл. печ. л. 1,9. Заказ № . Изд. № 818.

ГУ ВШЭ. 125319, Москва, Кочновский проезд, 3
Типография ГУ ВШЭ. 125319, Москва, Кочновский проезд, 3
Тел.: (495) 772-95-71; 772-95-73

Для заметок

Для заметок