§1. О постановке задач

Специфика компьютерного анализа данных почти всегда, так или иначе, заключается в присутствии фактора случайности, поскольку любой эксперимент подразумевает наличие погрешностей и шумов. Поэтому соответствующие вычислительные методы будут неразрывно связаны с понятиями теории вероятности и математической статистики.

Начнем разговор о способах обработки экспериментальных данных с классификации наиболее часто встречающихся задач. Для этой цели рассмотрим несколько типичных примеров (главным образом, из области вычислительной экономики и вычислительной геофизики), которыми продолжим пользоваться и в дальнейшем.

Пример: технический анализ рынков

Колебания финансовых показателей финансовых или товарных рынков, например, валютных котировок, – идеальный пример для изучения математической статистики.

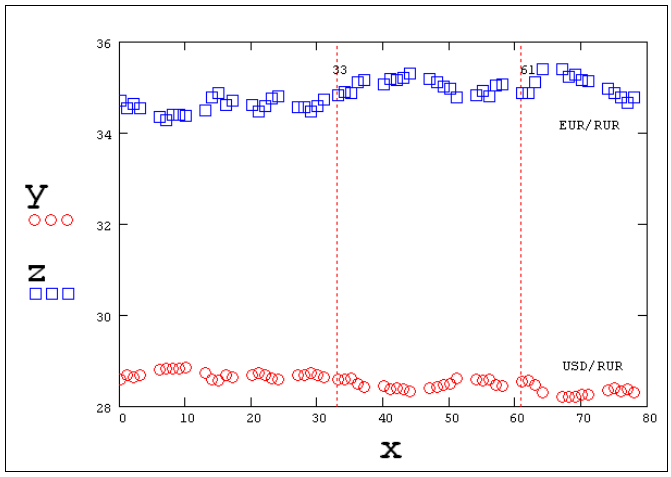

Рис. 1. *Случайный процесс – курсы доллара и евро*

На рис. 1 изображены графики изменения курсов американского доллара и евро по отношению к рублю, устанавливаемые Центробанком России. Каждая точка на графике представляет цену доллара (или евро) в определенный день. За точку отсчета принята некоторая дата, а пунктирные линии выделяют на оси X отрезок, соответствующий одному месяцу (августу 2005 г.).

Типичный объект исследования технического анализа – это выборка некоторого случайного процесса, т.е. зависимость определенной случайной величины от времени y(x). На практике выборка представляет собой массив экспериментальных данных, который состоит из пар чисел (x_i,y_i). В нашем случае, первый вектор **x** – это отсчеты времени, а второй вектор $\mathbf{y}\equiv y(x_i)$ – курс доллара. Другой случайный процесс, показанный на рис. 1, – это динамика курса евро (также по отношению к рублю) z(x).

Первоочередные проблемы, с которыми сталкиваются вычислители при работе с дискретными выборками случайных процессов, связаны с переходом от дискретной зависимости $y(x_i)$ к непрерывной функции f(x). Попросту говоря, встает задача «соединения» экспериментальных точек, чтобы иметь возможность оперировать с непрерывной зависимостью y(x).

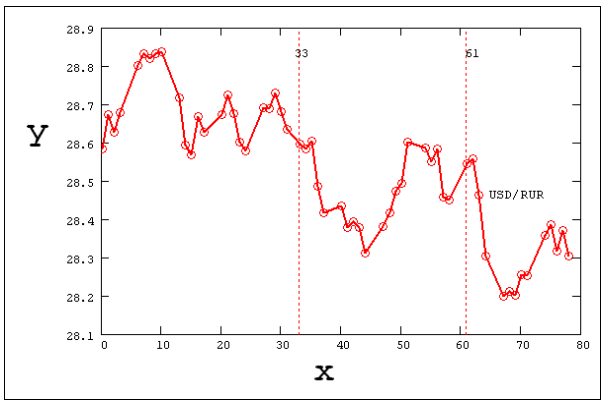

Рис. 2. *Интерполяция*

Заменить дискретную выборку непрерывной функцией f(x) можно по-разному, в зависимости от специфики задачи.

- Во-первых, задать f(x) так, чтобы она проходила через точки (x_i,y_i), т. е. $f(x_i)=y_i$ (рис. 2). В этом случае говорят об *интерполяции* данных функцией f(x) во внутренних точках между x_i или *экстраполяции* (рис. 3, крестики) за пределами интервала, содержащего все x_i;

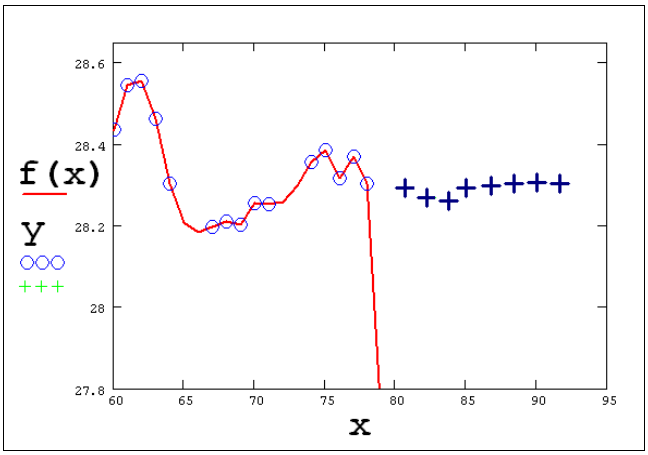

Рис. 3. *Экстраполяция*

- Во-вторых, потребовать, чтобы f(x) определенным образом (например, в виде той или иной аналитической зависимости) приближала $y(x_i)$, не обязательно проходя через точки (x_i,y_i). Тогда говорят о задаче *регрессии* (рис. 4).

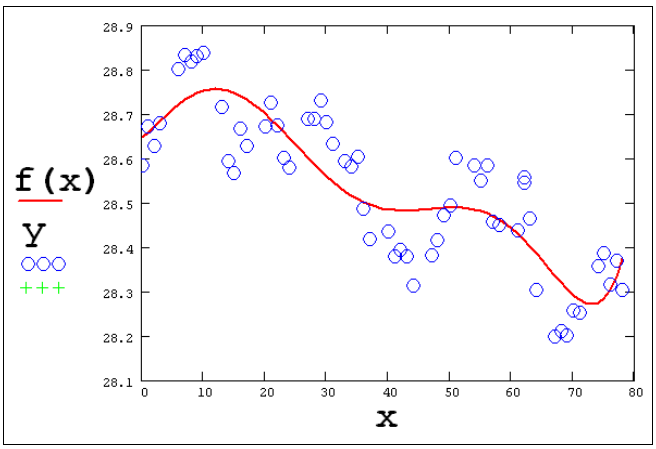

Рис. 4. *Регрессия*

- В-третьих, рассматривать более сложную задачу, связанную с первичной обработкой данных, а именно, выделением из них той или иной компоненты (чаще всего, низко-, средне- или высокочастотной). Такую задачу называют *фильтрацией* данных. Частные случаи фильтрации – это *сглаживание*, служащее для уменьшения (высокочастотной)

шумовой компоненты, и выделение *тренда* (низкочастотной компоненты). Часто выделение тренда (рис. 5) используется в целях его последующего устранения, т.е. вычитания из исходного сигнала (рис. 6). Отметим, что к сглаживанию данных можно отнести и регрессию (рис. 4).

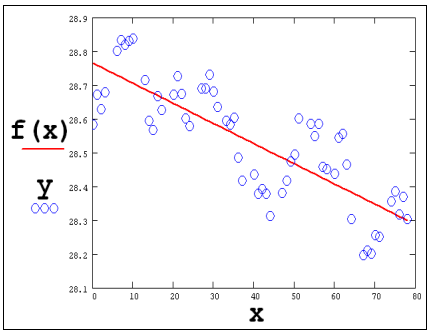

Рис. 5. *Выделение тренда*

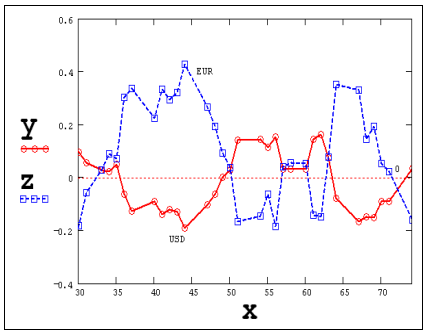

Рис. 6. *Курс минус тренд*

Задачи фильтрации неразрывно связаны с вычислением спектров сигнала. В частности, Фурье-спектр F(ω) (рис. 7 и 8) представляет частотный состав сигнала. Наряду с Фурье-спектром, используются и другие интегральные преобразования сигнала.

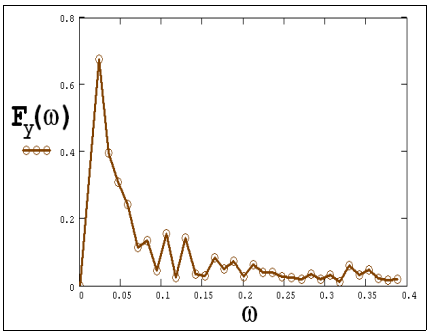

Рис. 7. *Фурье-спектр сигнала y(x)*

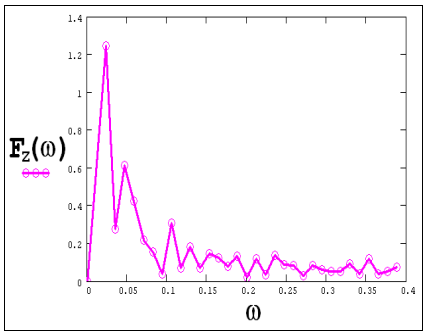

Рис. 8. *Спектр z(x) (курса евро)*

Все сказанное относилось к исследованию случайных процессов. Однако, логичнее нам было бы начать с более простых задач математической статистики, связанных с понятием случайной величины (т.е. со случая, когда зависимость от времени отсутствует). С этой точки зрения, массив **y** (или **z)** можно рассматривать как выборку случайной величины за определенный временной промежуток.

Применительно к данным о курсах валют, такой подход оправдан, к примеру в том случае, когда надо вычислить какие-либо статистические показатели за некоторый период (среднее или дисперсию за неделю, месяц и т.д.). В частности, интерес могут представлять так называемые уровни поддержки и сопротивления (resistance – support), т.е., несколько упрощая, минимальное и максимальное значение курса валют за определенный период. Эти уровни играют важную роль для биржевых брокеров, т.к. задают психологический барьер для движения рынка вверх или вниз. Для их фиксации зависимость векторов **y** и **z** от времени несущественна, т.к. важны только пиковые значения их компонент. Поэтому случайные числа y_i и z_i с этой точки зрения можно считать выборкой случайной величины.

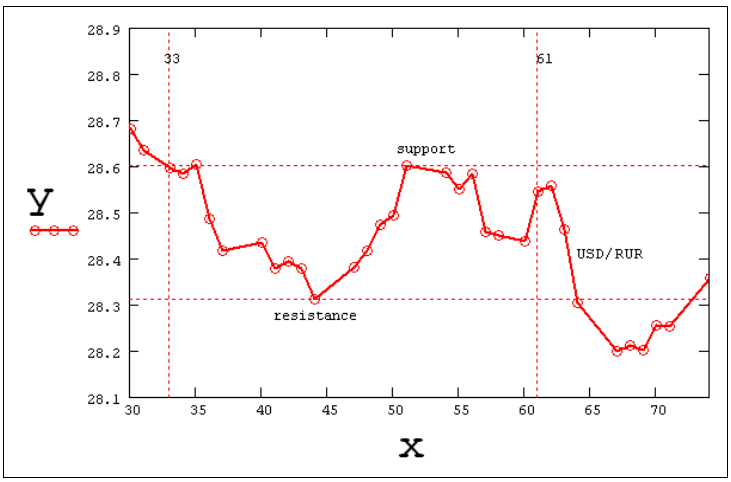

Рис. 9. *Уровни поддержки и сопротивления*

На рис. 9 изображен график колебаний курса доллара с отложенными уровнями поддержки и сопротивления (горизонтальные пунктирные линии) по результатам выделенного месяца наблюдений. Видно, что в первой половине следующего месяца данный уровень поддержки был пробит, и установилось новое значение поддержки.

Важный инструмент анализа выборки случайной величины – это гистограмма, т.е. график p(y), который в виде столбиков представляет частоту попадания p ее выборочных значений y в определенные числовые интервалы (рис. 10).

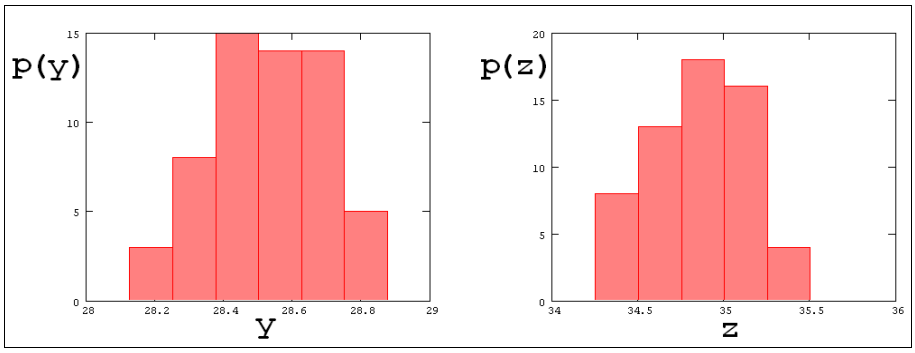

Рис. 10. *Гистограмма распределения курса доллара (слева) и евро (справа)*

Еще один вариант визуализации тех же данных о курсах доллара и евро, когда их можно рассматривать как случайные величины, пренебрегая временной динамикой, приведен на рис. 11. На нем изображен график, состоящий из точек $y_i(z_i)$. Глядя на него, легко убедиться, что случайные величины (рублевые курсы евро и доллара) являются сильно коррелированными, т.е. зависимыми. Более детально корреляция зависимых случайных величин, в том числе, применительно к рассматриваемому примеру динамики валютных курсов доллара и евро, будет разобрана в §4.

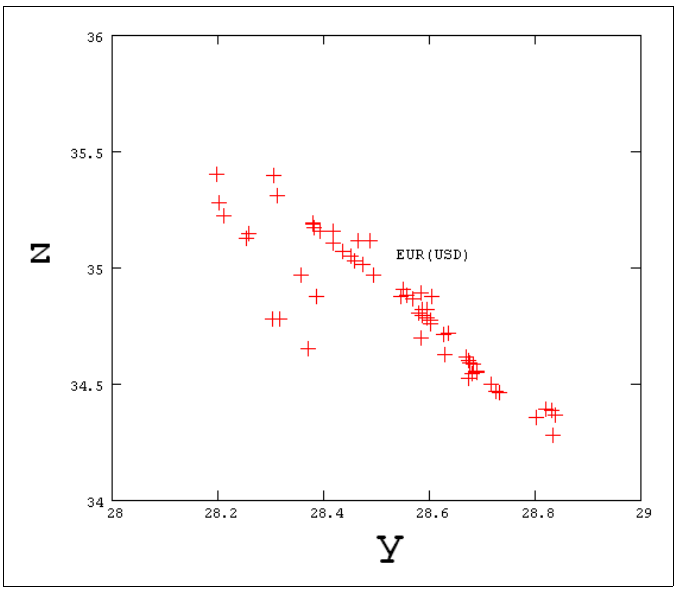

Рис. 11. *Курсы доллара и евро – коррелированные случайные величины*