Выделение периодической компоненты из временного ряда*

А. А. Токмакова
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

В проекте исследуется временной ряд на наличие периодической компоненты. На основе теории о рядах Фурье строится тригонометрическая интерполяция предложенных временных рядов методом наименьших квадратов. Также производится оценка параметров функции метода наименьших квадратов в зависимости от качества прогнозирования. В вычислительном эксперименте приводятся результаты работы корреляционной функции и метода наименьших квадратов на зашумлённом модельном синусе и реальном временном ряде электрокардиограммы.
Ключевые слова: *корреляционная функция, тригонометрическая интерполяция, метод наименьших квадратов, периодическая компонента.*

Введение

Определение 1. *Временной ряд — последовательно измеренные через некоторые (зачастую равные) промежутки времени данные.*

При прогнозировании некоторых временных рядов, например временных рядов продаж, потребления энергии или электрокардиогрммы, мы сталкиваемся с тем, что данные ряды обладают периодической компонентой. Существует несколько методов выявления периода. В данной работе рассмотрены алгоритмы атвокорреляционной функции и метода наименьших квадратов.

Автокорреляционная функция исследует временной ряд на наличие периодической компоненты, сдвигая ряд на несколько временных отсчетов и сравнивая с самим собой. Более подробно алгоритм автокорреляционной функции представлен в книгах [1, 5, 4].

Метод наименьших квадратов(МНК) вычисляет тригонометрическую аппроксимацию данного на вход ряда. Так как любая последовательность, обладающая периодичностью может быть разложена в ряд Фурье [2], необходимо принять коэффициенты перед синусами и косинусами за коэффициенты регрессии [6] и оценить их величину. Если найденная корреляция (коэффициент при определенном синусе или косинусе) велика, то можно заключить, что существует строгая периодичность на соответствующей частоте в данных.

Далее будет рассмотрена работа алгоритмов на модельных данных, а также на реальном временном ряде электрокардиограммы. Будет исследована зависимость коэффициента корреляции от различных характеристик временного ряда, а также рассмотрена возможность примененения метода наименьших квадратов для прогнозирования данных.

Постановка задачи

Дан временной ряд f_t, где n — длина временного ряда, $t \in \{1, ..., n\}$ — номер отсчета. Предполагаем, что в рассматриваемом временном ряде нет пропущенных значений, и он имеет периодические составляющие с периодом $T = \{\tau_1, ..\tau_p\}$. Работа состоит из трех следующих ступеней.

Научный руководитель В. В. Стрижов

Во-первых, *тестирование на модельной задаче*. Дана зашумлённая синусоида с известным периодом. Необходимо исследовать изменение коэффициента корреляции в следующих ситуациях:

— при увеличении шума;
— при уменьшении числа отсчётов на период;
— при сокращении длины временного ряда.

Во-вторых, *тестирование на реальном временном ряде*. Дан временной ряд электрокардиограммы, включающий периодическую компоненту со сложным строением. Необходимо исследовать его на наличие временных периодичностей, используя алгоритмы автокорреляционной функции и МНК. В-третьих, необходимо выяснить пригодность метода наименьших квадратов для *прогнозирования временных рядов*.

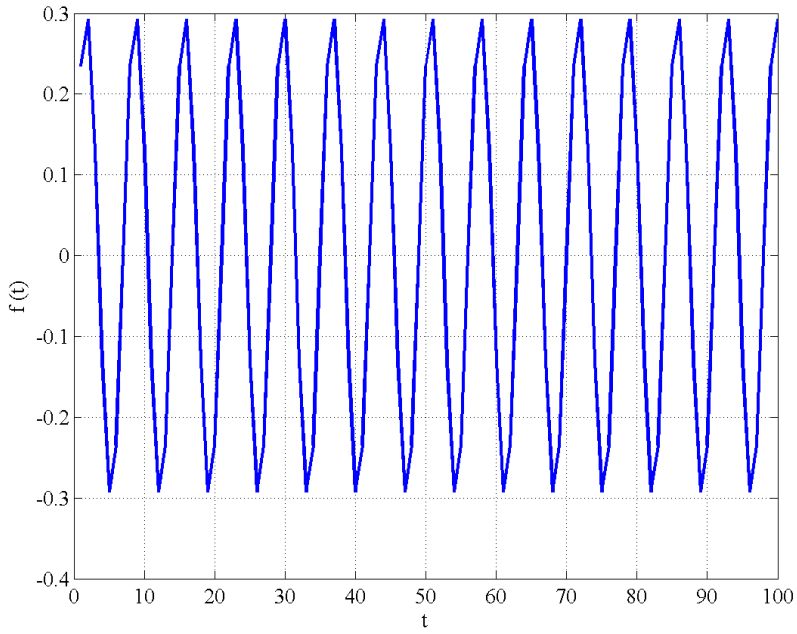

Рис. 1. Модельный временной ряд $f(t) = 0.3\sin(2\pi t/7)$

Для контроля качества алгоритма прогноза будем выделять во временном ряде l последовательных значений (контрольную выборку), которые алгоритм будет прогнозировать по предыдущим значениям. В качестве критерия качества прогноза будем минимизировать следующий функционал:

$$Q = \sum_{t=1}^{l} |\hat{f}_t - f_t|,$$

где $\hat{f}_t$ — прогнозируемое значение в t-ый момент времени, f_t — фактическое значение.

Пути решения задачи

Опишем используемые в работе алгоритмы.

Автокорреляционная функция.

Определение 2. *Автокорреляционная функция — это характеристика временного ряда, которая помогает находить его повторяющиеся участки, скрытые из-за наложений шума или других помех.*

Для дискретного временного ряда $X_1, X_2, ..., X_n$ с известными матожиданием μ и дисперсией σ автокорреляционную функцию можно рассчитать по следующей формуле:

$$R(w) = \frac{1}{(n-w)\sigma^2} \sum_{t=1}^{n-w} [X_t - \mu][X_{t+w} - \mu],$$

где n —длина временного ряда, w —текущая задержка во времени. Таким образом получим функцию $R(w)$, зависящую от лагов (задержек во времени). Исследуя ее на экстремальные значения, получим искомые значения периодов $T = \{\tau_1, ..\tau_p\}$.

Оценка периода осложняется тем, что в некоторой окрестности оцениваемого периода, наблюдаются локальные максимумы коэффициентов корреляции. Следовательно, необходимо усреднение коэффициентов корреляции, а также удаление "близких"и кратных периодов ("близкими"в работе считаются периоды, отличающиеся друг от друга менее, чем на величину δ).

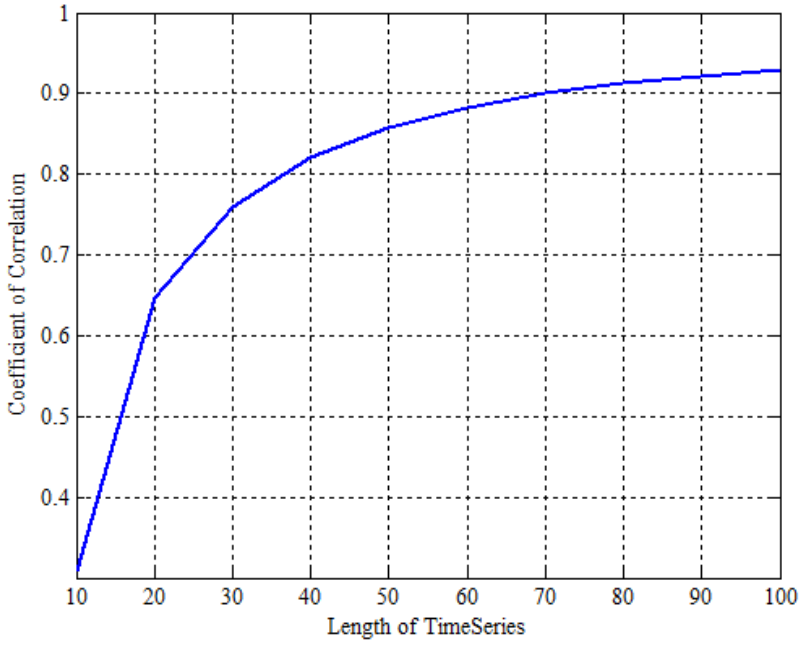

Рис. 2. Зависимость коэффициента корреляции от длины временного ряда

Ряд Фурье. Сделаем предположение о наличии периодики в предлагаемом ряде и обратимся к теореме [2].

Теорема 1. *Если некоторая периодическая функция с периодом $2j$ на интервале $[-j, j]$ удовлетворяет условиям Дирихле (имеет конечное число экстремумов и точек разрыва I рода), то она может быть представлена в виде суммы ряда Фурье (разложена в ряд Фурье).*

Таким образом, рассматриваемые в данной работе временные ряды могут быть представлены в виде бесконечного ряда Фурье. Построим регрессионную модель следующего вида [7].

Можно заметить, что разложение временной последовательности в ряд Фурье позволяет отыскать скрытые периодичности. Одним из возможных способов определения автокорреляционной зависимости является разложение временного ряда в функции синусов и косинусов и нахождение линейной множественной регрессии [6].

$$X_t = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos(\lambda_k t) + b_k \sin(\lambda_k t),$$

где $\lambda_k = 2\pi\eta_k$, $\eta_k = \frac{k}{n}$, $k = 1, 2, .., n$. Коэффициенты a_k и b_k определяются следующими рядами:

$$a_k = \frac{2}{n} \sum_{i=1}^{n} X_i \cos(\lambda_k t), k = 0, 1, 2.., n;$$

$$b_k = \frac{2}{n} \sum_{i=1}^{n} X_i \sin(\lambda_k t), k = 1, 2.., n.$$

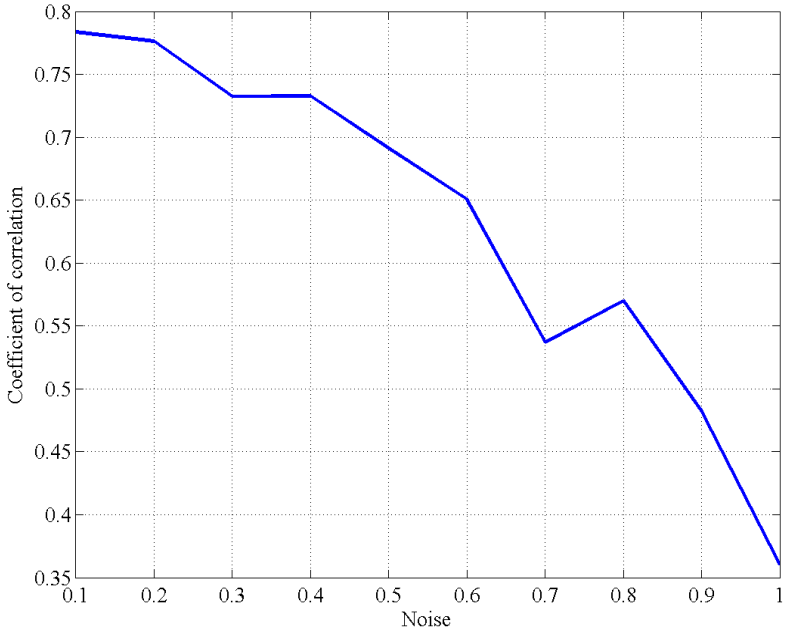

Рис. 3. Зависимость коэффициента корреляции от шума

Коэффициенты при косинусах и синусах — это коэффициенты регрессии. Они показывают степень, с которой соответствующие функции коррелируют с данными. Необходимо заметить, что сами синусы и косинусы на различных частотах ортогональны. Будем рассматривать не более чем n различных синусов и косинусов. В итоге определяется корреляция функций синусов и косинусов различной частоты с наблюдаемыми данными. Если найденная корреляция (коэффициент при определенном синусе или косинусе) велика, то

можно заключить, что существует строгая периодичность на соответствующей частоте в данных.

Данный метод даёт точный результат только тогда, когда длина временного ряда (то есть параметр n) кратен искомому периоду. В противном случае мы получим некую суперпозицию синусов и косинусов, которую достаточно сложно интерпретировать. Поэтому воспользуемся методом тригонометрической интерполяции с помощью метода наименьших квадратов.

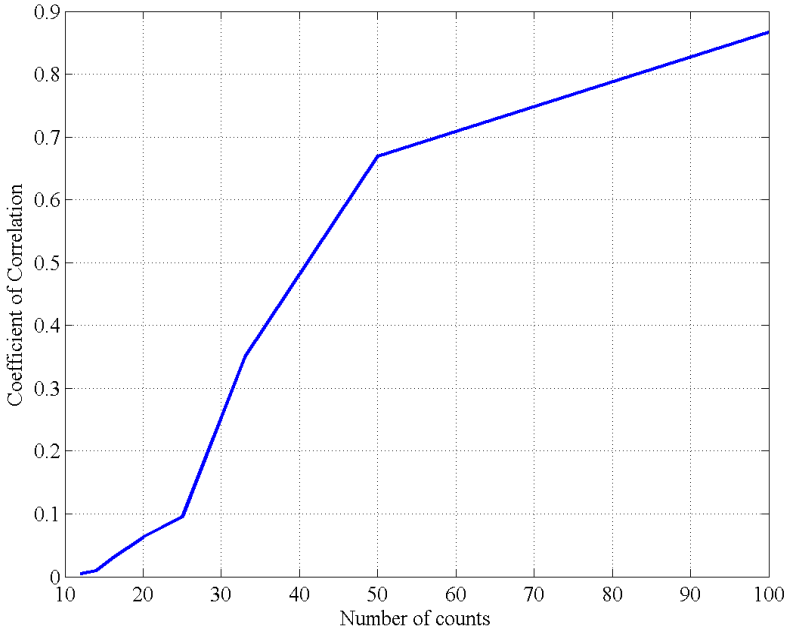

Рис. 4. Зависимость коэффициента корреляции от количества отсчетов за период

Тригонометрическая интерполяция методом наименьших квадратов. Требуется построить кривую, которая воспроизводила бы график исходной экспериментальной закономерности, то есть была бы максимально близка к экспериментальным точкам, но в то же время была бы нечувствительна к случайным отклонениям измеряемой величины.

Введем непрерывную функцию $\varphi(x)$ для аппроксимации дискретной зависимости $g(x_i)$, $i = 1, ..., n$. Будем считать, что $\varphi(x)$ построена при условии наилучшего квадратичного приближения, если:

$$Q = \sum_{i=1}^{n} (\varphi(x_i) - g(x_i))^2 = \min . \tag{1}$$

Рассмотрим случай линейной аппроксимации:

$$\varphi(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + ... + c_m\varphi_m(x),$$

где $\varphi_0, ..., \varphi_m$ — произвольные базисные функции, $c_0, ..., c_m$ — неизвестные коэффициенты. Количество базисных функций должно быть меньше количества заданных точек для того, чтобы их суперпозиция определялась единственным образом.

Для решения задачи линейной аппроксимации в общем случае следует найти условия минимума суммы квадратов отклонений (1). Это можно свести к задаче поиска корня

системы уравнений $\frac{\partial Q}{\partial c_k} = 0,\ k = 1, ..., m$. Вычисление данных производных, при учёте равенства (1) приведёт к следующей системе алгебраических уравнений:

$$\begin{cases} \sum_{i=1}^{n}(c_0\varphi_0(x) + c_1\varphi_1)(x) + \ldots + c_m\varphi_m(x) - f_i)\varphi_0(x) = 0; \\ \sum_{i=1}^{n}(c_0\varphi_0(x) + c_1\varphi_1)(x) + \ldots + c_m\varphi_m(x) - f_i)\varphi_1(x) = 0; \\ \ldots \\ \sum_{i=1}^{n}(c_0\varphi_0(x) + c_1\varphi_1)(x) + \ldots + c_m\varphi_m(x) - f_i)\varphi_m(x) = 0. \end{cases} \tag{2}$$

Далее следует решить полученную СЛАУ относительно коэффициентов $c_0, ..., c_m$. Для решения СЛАУ обычно составляется расширенная матрица коэффициентов, которую называют матрицей Грама, элементами которой являются скалярные произведения базисных функций и столбец свободных коэффициентов:

$$\begin{pmatrix} (\varphi_0, \varphi_0) & (\varphi_0, \varphi_1) & \ldots & (\varphi_0, \varphi_m) & (\varphi_0, f) \\ (\varphi_1, \varphi_0) & (\varphi_1, \varphi_1) & \ldots & (\varphi_1, \varphi_m) & (\varphi_1, f) \\ \ldots & \ldots & \ddots & \ldots & \ldots \\ (\varphi_m, \varphi_0) & (\varphi_m, \varphi_1) & \ldots & (\varphi_m, \varphi_m) & (\varphi_m, f) \end{pmatrix}.$$

$$(\varphi_j, \varphi_k) = \sum_{i=1}^{n} \varphi_j(x_i)\varphi_k(x_i)(\varphi_j f) = \sum_{i=1}^{n} \varphi_j(x_i)f(x_i),$$

где $j = 0, ..., m,\ k = 0, ..., m$.

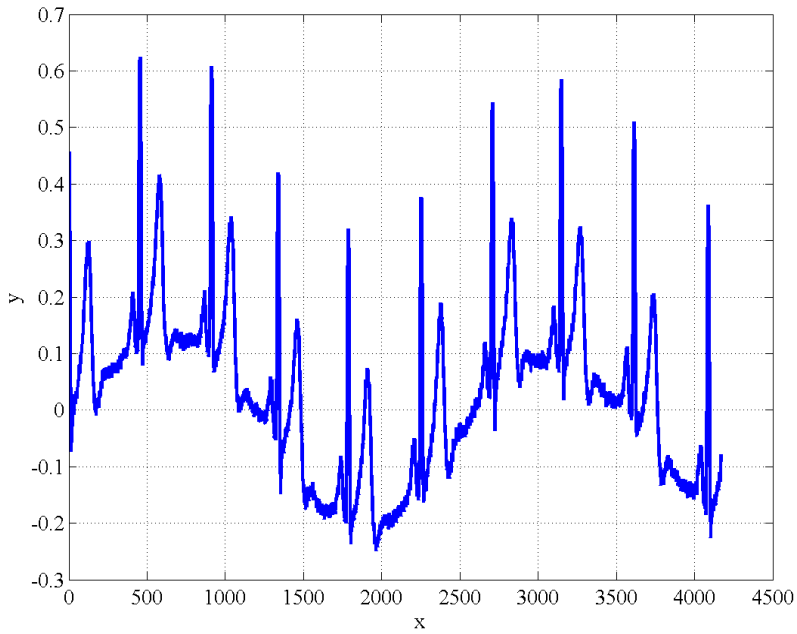

Рис. 5. Временной ряд электрокардиограммы

После того как с помощью, например, метода Гаусса найдены коэффициенты $c_0, ..., c_m$, можно построить аппроксимирующую кривую или вычислить координаты заданной точки. Таким образом, задача аппроксимации решена. Для того чтобы сформировать ортонормированный базис, коэффициенты нормировки будут следующими: для a_0 — $(4/n)^{0.5}$;

для a_k и b_k — $(2/n)^{0.5}$, где a_k — коэффициенты нормировки для косинусов, $k = 0, ..., m$; b_k — коэффициенты нормировки для синусов, $k = 1, ..., m$; n — количество отсчетов.

При работе с временными рядами необходимо учесть возможное наличие тренда или присутствие постоянного слагаемого. Обе эти составляющие исключим из данных, поскольку они могут привести к большим погрешностям при подсчёте функционала $Q = \sum_{t=1}^{l} |\hat{f}_t - f_t|$. Пользуясь методом наименьших квадратов мы учтём и тренд и наличие постоянного слагаемого. Необходимо заметить, что тригонометрическая интерполяция основана на разложении в ряд Фурье, поэтому она также не годится для выявления периодической компоненты ряда. В данной работе она используется для нахождения тренда, содержащего периодическую компоненту. Оставшиеся точки исследуются при помощи автокорреляционной функцией.

Вычислительный эксперимент

Исследование модельного зашумленного синуса. В качестве модельных данных будем использовать функцию $f(t) = 0.3\sin(2\pi t/7)$. Количество отсчетов $n = 100$. При исследовании временных рядов автокорреляционной функцией для поиска "близких"периодов будем считать $\delta = 1\%$.

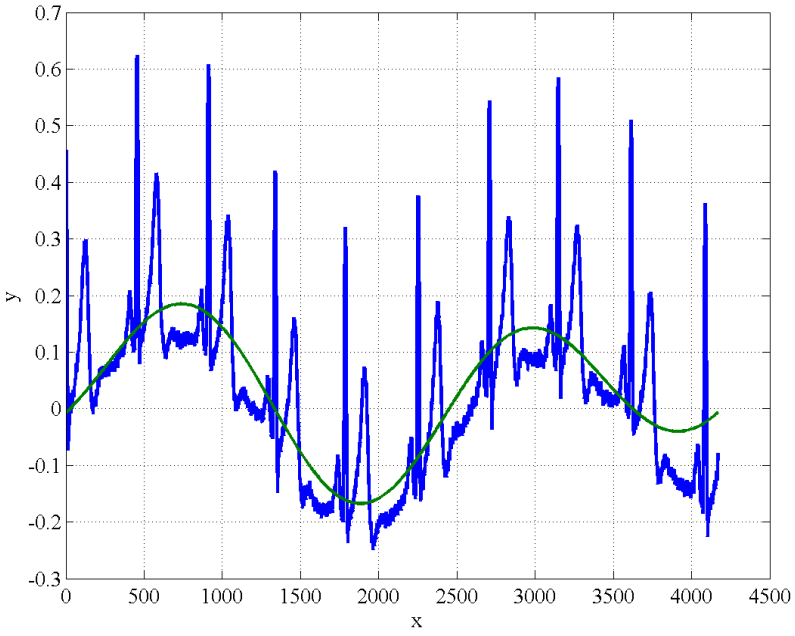

Рис. 6. Выделение тригонометрического тренда

Исследование величины коэффициента корреляции в зависимости от накладываемого шума. Была определена зависимость коэффициента корреляции в зависимости от шума (величина шума определялась в процентах от максимального значения функции). Распределение шума — равномерное. Результаты представлены на рис. 3. Максимальное значение функции $f_{max} = 0.2925$. Максимальный коэффициент корреляции $R_{max} = 0.9234$ соответствует 10%-ому зашумлению. Минимальный коэффициент корреляции $R_{min} = 0.6499$ соответствует 100%-ому зашумлению.

Исследование величины коэффициента корреляции при уменьшении числа отсчетов за период. Были вычислена зависимость коэффициента корреляции от уменьшения количества отсчетов за период. Результаты представлены на рис. 3. Максимальный коэффициент корреляции $R_{max} = 0.9295$ соответствует 100-ти отсчетам за период. Минимальный коэффициент корреляции $R_{min} = 0.4355$ соответствует 7-ми отсчетам за период.

Исследование величины коэффициента корреляции при сокращении временного ряда. Расчет был произведен начиная с длины ряда $n = 100$ до $n = 10$. Величина шага составляла 10 временных отсчетов. Максимальный коэффициент корреляции $R_{max} = 0.9295$ соответствует $n = 100$. Минимальный коэффициент корреляции $R_{min} = 0.3085$ соответствует $n = 10$, где n — длина ряда. Данные выводы напрямую следуют из устройства корреляционной функции:

1. при увеличении шума коэффициент корреляции уменьшается;
2. при уменьшении количества отсчётов за период коэффициент корреляции уменьшается;
3. при уменьшении длины временного ряда коэффициент корреляции уменьшается.

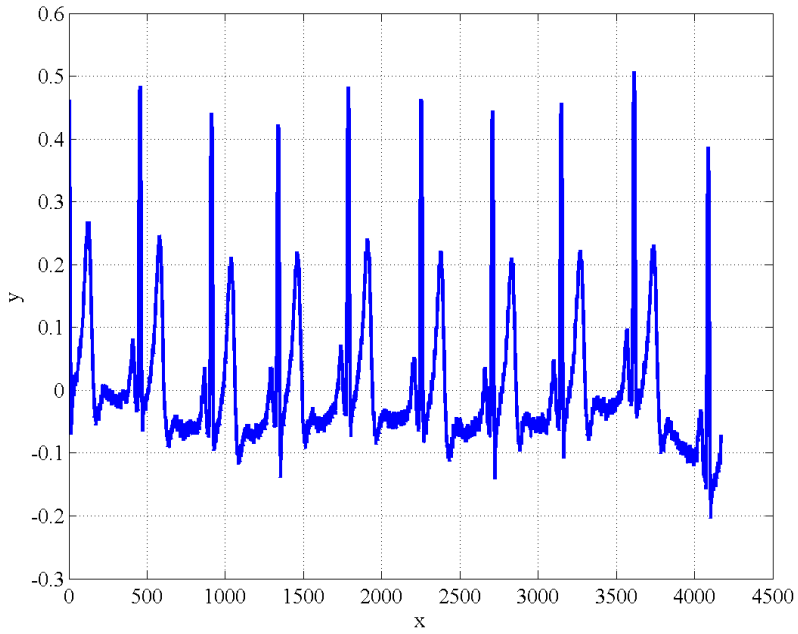

Рис. 7. Ряд электрокардиограммы с исключенным тригонометрическим трендом

Исследование реального временного ряда электрокардиограммы. Рассмотрим реальный временной ряд сложной периодики. Количество отсчетов $n = 4170$. После применения корреляционной функции к первоначальному ряду получим следующие коэффициенты корреляции:

Корреляция	0.6	0.6	0.4	0.4	0.4	0.3	0.2	0.1	0.1
Период	43	121	453	163	205	288	2371	2089	2811

При анализе таблицы и графика получим, что корреляционная функция определила "локальный"период временного ряда, тогда как глобальная периодика осталась не выявленной. Стоит заметить, что локальный период также определён не точно и коэффи-

циенты корреляции очень малы. Применим метод наимеьших квадратов для получения тригонометрического тренда данного временного ряда. Аппроксимация происходит с помощью двух гармоник. Для выявления скрытой периодики необходимо вычесть из исходного ряда его тригонометрическую интерполяцию.

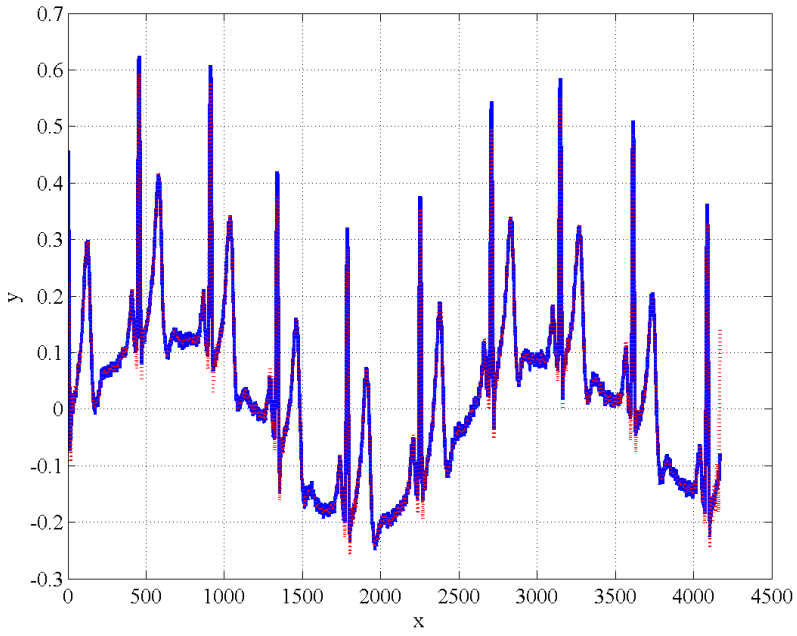

Рис. 8. Тригонометрическая аппроксимация временного ряда

После обработки полученного временного ряда автокорреляционной функцией получим следующие скрытые периодики и коэффициенты корреляции:

Корреляция	0.5	0.2	0.2	0.1	0.1
Период	459	2700	3160	1320	868

Заметим, что хотя коэффициенты корреляции малы, периодика ряда определена верно. Малость коэффициентов объясняется сложным строением ряда.

Исследование применения МНК для прогнозирования реальных временных рядов. Метод наименьших квадратов при тригонометрической интерполяции основывается на добавлении новых гармоник для лучшего совпадения реальной функции и её аппроксимации. Критерием схожести служил функционал:

$$Q = \sum_{t=1}^{l} |\hat{f}_t - f_t|,$$

где $\hat{f}_t$ — прогнозируемое значение в t-ый момент времени, f_t — фактическое значение. В данной работе расчет велся начиная с двух гармоник. Остановка алгоритма происходила в двух случаях:

— номер гармоники больше половины длины ряда;
— относительная ошибка на один отсчёт составляет менее 1%.

Результат работы алгоритма для реального временного ряда приведён на рис. 8. Результат прогнозирования ряда приведён на рис. 9.

Длина входного ряда 3000. Прогноз на 1000 точек. Остановка произошла на 148 гармонике. Функционал качества составил $Q = 163.6817$ на 1000 временных отсчётов, что составляет 16.3%. Плохие результаты объясняются тем, что метод наименьших квадратов берёт за период количество отсчётов равное длине ряда. Следовательно, при прогнозировании происходит копирование первоначальных точек.

Прогнозировать методом наименьших квадратов можно только ряды, которые обладают строгой периодичностью. Однако, метод подходит для интерполяции временного ряда внутри отрезка (создание непрерывных рядов), а также для аналитического описания (представления в виде ряда Фурье) рядов, обладающих периодической компонентой.

Заключение

В работе рассмотрена зависимость коэффициента корреляции от различных входных параметров. Исследованы способы получения периодичности временных рядов, а также метод нахождения скрытых периодичностей. Рассмотрена возможность применения метода наименьших квадратов для прогнозирования временного ряда с периодической компонентой. Анализ проводился на модельных данных и на реальном временном ряде электрокардиограммы.

Необходимый для повторения вычислительного эксперимента код можно найти на сайте:`https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/PeriodicComponents/`

Литература

[1] D. Gujarati. *Basic Econometrics, 4th ed*, The McGraw-Hill Companies, 2004.

[2] А.М. Тер-Крикоров, М.И. Шабунин. *Курс математического анализа*, ФИЗМАТ-ЛИТ, 2001.

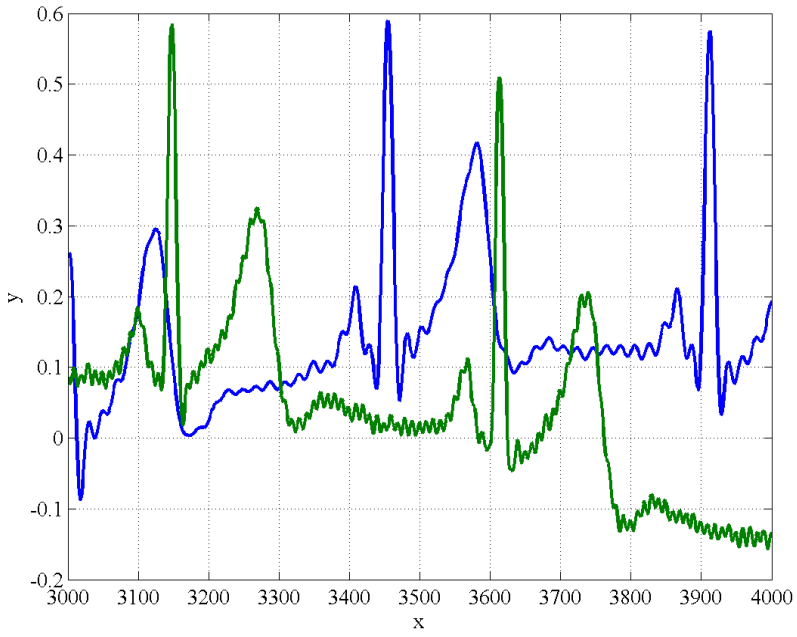

Рис. 9. Реальный и спрогнозированный временной ряд

[3] В.В. Стрижов, Г.О. Пташко. *Алгоритмы поиска суперпозиций при выборе оптимальных регрессинных моделей*, Вычислительный центр им. А.А. Дородницына РАН, 2007.

[4] В.В. Витязев. *Спектрально-корреляционный анализ равномерных временных рядов*, С.-Петербургский государственный университет, 2001.

[5] А.И. Орлов Прикладная статистика, «Экзамен», 2004, №3: 259-279.

[6] Ю.В. Попов *О выделении периодической компоненты из временного ряда показателя количества катастроф*, "Проблемы безопасности полетов 2008.

[7] В.В. Стрижов *Методы индуктивного порождения регрессионных моделей*, Вычислительный центр им. А.А. Дородницына Российской Академии наук, 2008, 37.