АСТРОФИЗИКА

ТОМ 57 АВГУСТ, 2014 ВЫПУСК 3

ОБЗОРЫ

О ПОИСКЕ ПЕРИОДИЧЕСКИХ КОМПОНЕНТ В НАБЛЮДАТЕЛЬНЫХ ДАННЫХ

Р.В.БАЛУЕВ
Поступила 18 марта 2014
Принята к печати 30 апреля 2014

Обзор посвящен задаче поиска периодичностей в наблюдательных данных при помощи периодограмм, основанных на общем статистическом критерии отношения правдоподобия, и его частных вариантах, к которым относится и классическая периодограмма Ломба-Скаргла. Основное внимание уделяется задаче оценки статистической значимости выявляемых при этом периодичностей. Мы предлагаем универсальное решение этой задачи с помощью эффективной методики, в которой периодограмма рассматривается как случайный процесс (или случайное поле), а приближение для требуемой "вероятности ложной тревоги" строится при помощи обобщенного метода Райса. Помимо единой методики определения ожидаемых шумовых уровней (или уровней значимости) таких периодограмм, мы также рассматриваем некоторые важные частные случаи с различными моделями периодического сигнала (как линейными, так и нелинейными). Вероятность ложной тревоги, связанная с наблюдаемым сигналом, в большинстве случаев аппроксимируется формулами вида $e^{-z}P(\sqrt{z})$, где z - наблюдаемый максимальный отсчет периодограммы, а P - некоторый алгебраический многочлен, коэффициенты которого зависят от условий задачи. Кроме того, мы рассматриваем задачу выделения из шума составных сигналов, содержащих несколько частот. В этом случае для корректного анализа данных оказывается необходимым применять так называемые многочастотные периодограммы, основанные на моделях сигнала, содержащих несколько периодических компонент. Мы показываем, что для полного решения данной задачи необходимо построить 2^n-1 таких периодограмм, где n - полное количество возможных частот. Наконец, в работе описываются программные пакеты, разработанные нами для облегчения практических задач частотного анализа временных рядов с применением данной теории.

Ключевые слова: *периодические компоненты: поиск*

1. *Введение*. Одной из важнейших практических задач современной наблюдательной астрономии является задача выделения в наблюдательных данных (точнее, в одномерных временных рядах) периодичностей различной природы. Эта задача возникает в весьма разных областях астрономии и астрофизики. Сюда входит поиск экзопланет разными наблюдательными методами (доплеровский метод, фотометрический метод - по прохождениям, метод тайминга различных периодических явлений), исследование переменных звезд (включая и Солнце) и более экзотических астрофизических объектов, демонстрирующих периодическую переменность любой природы. Задача частотного анализа наблюдательных или экспериментальных данных часто

возникает и за пределами астрономии и астрофизики, например в таких науках, как климатология и геофизика.

Задача поиска периодичностей обычно решается астрономами при помощи разного рода периодограмм - функций частоты предполагаемого сигнала, которые так или иначе представляют собой оценки спектра мощности наблюдаемой переменности. Простейшими такими оценками являются периодограмма Ломба-Скаргла [1,2] и Шустера [3]. Большое значение периодограммы на какой-то частоте свидетельствует о вероятном наличии на этой или близкой частоте периодического сигнала.

На практике поиск периодичностей в зашумленном временном ряду осложнен рядом обстоятельств. В первую очередь сюда относится неравномерность (или нерегулярность) моментов наблюдений, которая может создавать значительные практические трудности, связанные в первую очередь с неоднозначностью выделяемых периодов, или эффектом элайзинга [4,5]. Заметим, что неравномерные временные ряды для астрономии довольно характерны, являясь скорее правилом, чем исключением. Впрочем, в данном обзоре проблема элайзинга и связанных с ним явлений не является основной проблемой, и будет находиться несколько в стороне.

Шум в исходных данных создает, очевидно, и шум на периодограммах. Поэтому на практике возникает задача определения ожидаемого на периодограмме уровня шума или, говоря более строгим языком, определение уровней статистической значимости периодограмм. Эти уровни значимости оказываются необходимы, когда нам нужно определить, является ли конкретный найденный нами пик периодограммы шумовым, или же он скорее всего отражает реальную периодичность. В конечном счете нам нужно иметь некоторый метод оценки "вероятности ложной тревоги" - вероятности того, что наблюдаемый нами наивысший пик периодограммы мог быть произведен одним лишь шумом. Если эта вероятность (ниже обозначаемая как FAP - от "false alarm probability") достаточно мала, то мы можем считать соответствующую периодичность реальной, а не шумовой.

До недавнего времени удовлетворительного решения данной задачи известно не было. Применяемые подходы требовали, так или иначе, весьма затратного численного моделирования по методу Монте-Карло, либо не обладали достаточной математической строгостью в тех практических ситуациях, в которых реально применялись. Подробное обсуждение этой проблемы и связанных с ней исследований можно найти в работах [4,6-16].

В данном обзоре мы описываем новую универсальную методику оценки статистической значимости пиков периодограмм, разработанную нами за последние несколько лет. Эта методика основана на одном методе из теории экстремальных значений случайных процессов и полей - обобщенном методе Райса. Основной результат применения данного подхода - оценка вероятности ложной тревоги в форме

$$\mathrm{FAP}(z_{max}) \le M(z_{max}), \tag{1}$$

где z_{max} есть максимальный отсчет периодограммы, а функция $M(z_{max})$ зависит от условий задачи. Знак "$\lesssim$" в этой формуле означает, что $M(z_{max})$ представляет собой *одновременно* и верхнюю границу для FAP, и некоторую ее аппроксимацию[1]. Такая оценка обладает рядом практически важных свойств:

1. Функцию M как правило удается выразить в явной аналитической форме, поэтому у нас не возникает нужды в проведении, например, какого-либо дополнительного численного моделирования. Во многих случаях эта функция оказывается весьма проста, так что формула (1) работает по принципу "взял и подставил".

2. Благодаря свойству верхней оценки, применение формулы (1) не увеличивает количество ложных тревог свыше заданного предела. Если полученное значение M показалось нам достаточно малым, чтобы выделить какой-то пик периодограммы, то реальное значение FAP устроит нас и подавно.

3. Даваемая методом Райса функция $M(z)$ является не просто какой-то оценкой сверху; она одновременно представляет собой асимптотическую аппроксимацию для оцениваемой вероятности ложной тревоги при больших z. Большие z соответствуют наиболее полезным *малым* уровням FAP, для которых нам как раз и нужна хорошая точность.

4. Точность функции $M(z)$ как аппроксимации на практике оказывается обычно весьма неплохой, а в худших случаях все еще удовлетворительной.

В статье мы сначала даем максимально обобщенное определение периодограммы на основе статистических критериев наименьших квадратов и отношения правдоподобия. Это определение содержит произвольные модели данных, включая модель базовой переменности (например, тренд), модель ожидаемой периодичности, и даже модель шума. Далее мы приводим целостное описание общей процедуры применения нашей методики к периодограммам как с линейными, так и нелинейными моделями данных, а также кратко приводим основные результаты ее применения к некоторым конкретным периодограммам. Кроме того, мы рассмотрим задачу многочастотного анализа временных рядов, когда сигнал моделируется суммой нескольких периодических компонент с независимыми друг от друга частотами (в частности, синусоид). В рамках решения данной задачи мы применяем т.н. "многочастотные" периодограммы, для которых также

---

[1] *Ниже, для упрощения обозначений мы будем обычно опускать индекс у* z_{max}, *но читателю не следует забывать, что вероятность ложной тревоги есть функция именно максимального отсчета периодограммы, а не произвольного.*

приводим законченную теорию уровней статистической значимости в виде функции $M(z)$ из (1). Наконец, мы описываем несколько вычислительных программных пакетов, в которых реализована разработанная теория частотного анализа временных рядов.

2. *Обобщенное определение периодограмм наименьших квадратов и периодограмм отношения правдоподобия.*

2.1. *Вводные определения.* Введем ряд предварительных определений. Пусть мы имеем временной ряд из N скалярных измерений x_i, сделанных в моменты времени t_i. Ошибки ε_i различных измерений x_i мы будем обычно считать статистически независимыми, а также нормально распределенными со стандартными отклонениями σ_i. Для x_i у нас как правило будет какая-то параметрическая модель вида $\mu(\theta,t)$, так что $x_i=\mu(\theta,t_i)+\varepsilon_i$. В зависимости от условий задачи, величины σ_i сами могут быть как известны заранее, так и выражаться какой-то параметрической моделью $\sigma_i(\mathbf{p})$ с заранее неизвестным $\mathbf{p}$. Например, распространена модель с масштабируемым шумом $\sigma_i^2=\kappa/w_i$, где w_i есть заданные веса наблюдений, а κ - неизвестный параметр.

Кроме того, определим следующую операцию взвешенного суммирования по временному ряду:

$$\langle\phi\rangle=\sum_{i=1}^{N}\frac{\phi(t_i)}{\sigma_i^2}. \tag{2}$$

Немного необычное обозначение $\langle 1\rangle$ (единица в треугольных скобках) представляет, таким образом, собой сумму всех величин $1/\sigma_i^2$.

2.2. *Применяемые методы точечного оценивания.* Хорошо известно, что если известны сами погрешности σ_i или хотя бы веса w_i, то во многих смыслах наилучшую оценку параметров модели по имеющемуся временному ряду можно получить при помощи метода наименьших квадратов (МНК). Введем функцию

$$\chi^2(\theta)=\left\langle (x-\mu(\theta,t))^2\right\rangle. \tag{3}$$

Требуемая МНК-оценка θ^* определяется путем минимизации данной функции в пределах некоторой заданной области Θ:

$$\theta^*=\arg\min_{\theta\in\Theta}\chi^2(\theta). \tag{4}$$

Если же модель для σ_i более сложна, то можно применить метод максимального правдоподобия, являющийся обобщением метода наименьших квадратов. Соответствующие формулы выглядят как:

$$\{\theta^*, \mathbf{p}^*\} = \arg\max_{\theta, p} \ln \mathcal{L}(\theta, \mathbf{p}),$$

$$\ln \mathcal{L}(\theta, \mathbf{p}) = -\sum_{i=1}^{N} \left(\ln \sigma_i(\mathbf{p}) + \frac{(x_i - \mu(\theta, t_i))^2}{2\sigma_i^2(\mathbf{p})} \right) + \text{const}, \tag{5}$$

где $\mathcal{L}$ есть функция правдоподобия задачи (для независимых и нормально распределенных ошибок ε_i). К примеру, в работе [17] такой подход применяется для модели шума с аддитивным "дрожанием": $\sigma_i^2(p) = \sigma_{meas,i}^2 + p$, которая лучше подходит для задач поиска экзопланет по доплеровским временным рядам. Заметим также, что в данной работе предложено ввести некоторую поправку в функцию правдоподобия $\mathcal{L}$, которая позволяет значительно сократить смещение оценок $\mathbf{p}^*$:

$$\ln \tilde{\mathcal{L}}(\theta, \mathbf{p}) = -\sum_{i=1}^{N} \left(\ln \sigma_i(\mathbf{p}) + \frac{(x_i - \mu(\theta, t_i))^2}{2\gamma\sigma_i^2(\mathbf{p})} \right) + \text{const}, \tag{6}$$

где корректирующий делитель γ равен $1 - \dim\theta/N$. Эта поправка выполняет редукцию систематического эффекта, возникающего из-за того, что невязки "O-C" (или $x-\mu$) в точке наилучшего приближения θ^* всегда оказываются систематически ниже реальных ошибок ε_i, что приводит к занижению получающихся значений $\sigma_i(\mathbf{p}^*)$ в (5) и к смещению в $\mathbf{p}^*$ порядка $\dim\theta/N$. Это смещение может оказаться значимым уже когда $\dim\theta$ имеет порядок $\sqrt{N}$, а не N, так как это смещение нужно сравнивать со случайной ошибкой в $\mathbf{p}^*$, которая падает как $1/\sqrt{N}$.

В нашей статье нам не понадобится определять погрешности и корреляции получаемых оценок $(\theta^*, \mathbf{p}^*)$, поэтому данный вопрос мы оставим за кадром.

2.3. *Определение периодограммы*. Сформулируем теперь саму задачу обнаружения периодического сигнала в наблюдательных данных в общей форме. Пусть у нас есть некоторая базовая модель данных вида $\mu_{\mathcal{H}}(\theta_{\mathcal{H}}, t)$, которая описывает ожидаемое поведение величины x в отстутствии как шума, так и потенциального периодического сигнала, существование которого мы хотим проверить. Эта модель зависит от $d_{\mathcal{H}}$ свободных параметров в векторе $\theta_{\mathcal{H}}$, подлежащих оценке из наблюдений. К примеру, $\mu_{\mathcal{H}}$ может быть полиномиальным трендом с неизвестными коэффициентами или просто неизвестной константой, отвечающей за центрирование временного ряда. Для большей общности предположим, что $\theta_{\mathcal{H}}$ лежит внутри некоторой заданной $d_{\mathcal{H}}$-мерной области $\Theta_{\mathcal{H}}$.

В качестве альтернативной модели данных примем модель $\mu_{\mathcal{K}}(\theta_{\mathcal{K}}, f, t) = \mu_{\mathcal{H}}(\theta_{\mathcal{H}}, t) + \mu(\theta, f, t)$, которая аддитивно включает в себя периодический сигнал заданного функционального вида $\mu(\theta, f, t)$. При этом полный вектор параметров $\theta_{\mathcal{K}}$ есть объединение $\{\theta_{\mathcal{H}}, \theta\}$, а f - пробная частота сигнала. В принципе, мы могли бы включить f в вектор θ, но предпочли его оттуда выделить ввиду особого статуса частоты в данной задаче. Можно также ввести область определения параметров θ как Θ, и тогда аналогичная область для $\theta_{\mathcal{K}}$ будет выражаться как Декартово произведение $\Theta_{\mathcal{K}} = \Theta_{\mathcal{H}} \times \Theta$. Кроме того, введем обозначения для размерностей: $d = \dim\theta$, $d_{\mathcal{K}} = \dim\theta_{\mathcal{K}}$ (очевидно, $d_{\mathcal{K}} = d_{\mathcal{H}} + d$).

Рассмотрим сначала более простой случай, когда все σ_i известны. Тогда первую периодограмму $z(f)$ можно определить на основе χ^2-критерия:

$$z(f) = \left(g_{\mathcal{H}} - g_{\mathcal{K}}(f)\right)/2, \tag{7}$$

где

$$g_{\mathcal{H}} = \min_{\theta_{\mathcal{H}} \in \Theta_{\mathcal{H}}} \chi^2_{\mathcal{H}}(\theta_{\mathcal{H}}), \quad g_{\mathcal{K}}(f) = \min_{\theta_{\mathcal{H}} \in \Theta_{\mathcal{H}}, \theta \in \Theta} \chi^2_{\mathcal{K}}(\theta_{\mathcal{H}}, \theta, f); \tag{8}$$

а $\chi^2_{\mathcal{H},\mathcal{K}} = \left\langle \left(x - \mu_{\mathcal{H},\mathcal{K}}\right)^2 \right\rangle$. Большие значения функции $z(f)$ на какой-то частоте f говорят о том, что на этой частоте вероятно наличие периодического колебания. При этом модель μ определяет, под какие именно сигналы наша периодограмма будет "заточена" лучше всего.

Задача усложняется, когда σ_i заранее неизвестны. Например, для модели с масштабируемым шумом $\sigma_i^2 = \kappa/w_i$ мы уже не можем определить функции χ^2 в (8), так как они пропорциональны неизвестному κ. Однако положения их минимумов все еще можно определять как раньше, а от множителя κ можно избавиться, рассматривая отношение величин g, а не их разность. Можно ввести, например, такие периодограммы:

$$z_1 = \frac{N_{\mathcal{H}}}{2g_{\mathcal{H}}}\left(g_{\mathcal{H}} - g_{\mathcal{K}}\right), \quad z_2 = \frac{N_{\mathcal{K}}}{2g_{\mathcal{K}}}\left(g_{\mathcal{H}} - g_{\mathcal{K}}\right), \quad z_3 = \frac{N_{\mathcal{K}}}{2} \ln \frac{g_{\mathcal{H}}}{g_{\mathcal{K}}}, \tag{9}$$

где $N_{\mathcal{H}} = N - d_{\mathcal{H}}$ и $N_{\mathcal{K}} = N - d_{\mathcal{K}}$. Как известно, при фиксированной f величины z_1 и z_2 можно трактовать как статистики бета- и F-критериев, а z_3 связана со статистикой отношения правдоподобия [11,18]. Заметим, что три введенные периодограммы можно легко выразить друг через друга, поэтому на практике они взаимно эквивалентны и приводят к совершенно идентичным результатам (конечно, если при этом используются правильные функции распределения для оценки значимости).

Для более сложной модели шума можно использовать общий критерий отношения правдоподобия. Помимо уже упомянутой модели шума с аддитивным "дрожанием", можно отметить и другие: в работах [19,20] рассматривался коррелированный ("красный") Гауссов шум, а авторы работы [21] применяли периодограмму отношения правдоподобия для Пуассоновского шума (дробовой шум наблюдений солнечных нейтрино). Правда, в данной статье мы не будем рассматривать столь уж далекие обобщения, предполагая шум белым и Гауссовым.

Если действовать строго канонически, то периодограмму отношения правдоподобия следовало бы определить так:

$$Z(f) = \max_{\mathbf{p}, \theta_{\mathcal{H}}, \theta} \ln \mathcal{L}_{\mathcal{K}}(\mathbf{p}, \theta_{\mathcal{H}}, \theta, f) - \max_{\mathbf{p}, \theta_{\mathcal{H}}} \ln \mathcal{L}_{\mathcal{H}}(\mathbf{p}, \theta_{\mathcal{H}}), \tag{10}$$

где величины $\mathcal{L}_{\mathcal{H}}$ и $\mathcal{L}_{\mathcal{K}}$ есть функции правдоподобия (5) для соответствующей модели данных. Мы все же предпочитаем, где это возможно, немного подправленное определение из [17], учитывающее описанную выше поправку (6). Это модифицированное определение выглядит так:

$$\tilde{Z}(f) = \frac{N_{\mathcal{K}}}{N}\left(\max_{\mathbf{p}, \theta_{\mathcal{H}}, \theta} \ln \mathcal{L}_{\mathcal{K}}(\mathbf{p}, \theta_{\mathcal{H}}, \theta, f) - \max_{\mathbf{p}, \theta_{\mathcal{H}}} \ln \mathcal{L}_{\mathcal{H}}(\mathbf{p}, \theta_{\mathcal{H}})\right) + \frac{N_{\mathcal{K}}}{2} \ln \frac{N_{\mathcal{H}}}{N_{\mathcal{K}}}. \tag{11}$$

Модификация (6) и (11) вносит поправки относительной величины порядка $1/N$, поэтому она не нарушает асимптотические свойства (при больших N) как получаемых оценок параметров, так и статистики отношения правдоподобия (и самой периодограммы). Однако она улучшает их поведение при умеренных N. Нормализация и смещение в (11) выбраны так, чтобы $\tilde{Z}(f)$ не принимала отрицательные значения (в приближении первого порядка по $1/N$), а для модели масштабируемого шума $\tilde{Z}(f)$ совпадала с $z_3(f)$.

3. ***Уровни значимости периодограмм и обобщенный метод Райса***. Периодограммы со сложными моделями данных на практике используются реже, что обусловлено объективными факторами. Первый из них очевиден: сложные модели требуют трудоемких и медленных вычислений. В общем нелинейном случае вычисления по формулам (7) и (8) можно выполнить лишь путем численной минимизации. Впрочем, для современных компьютеров эта задача не так сложна, как например всего десятилетие назад. Но следом возникает второе препятствие, преодолеть которое значительно труднее. Помимо того, чтобы уметь вычислять саму периодограмму $z(f)$, нам нужно уметь находить ее уровни статистической значимости, по которым можно отделить реальные периодичности от шумовых.

Мы обычно ищем максимальное значение периодограммы в некотором априори заданном интервале частот $f \in [f_{min}, f_{max}]$. Полученный наивысший пик - наш главный кандидат на возможную периодичность. Если его высота z_{max} окажется достаточной, то мы можем считать соответствующую периодичность реальной. Вопрос о том, достаточна ли эта высота в конкретном случае, решается в вероятностном смысле. Нам нужно вычислить вероятность "ложной тревоги", связанную с наблюдаемой величиной z_{max}:

$$\mathrm{FAP}(z_{max}) = \Pr\left\{\max_{[f_{min}, f_{max}]} z(f) > z_{max}\right\}, \tag{12}$$

где вероятность берется при базовой гипотезе $\mathcal{H}$, то есть для периодограмм $z(f)$, построенным по данным без реального сигнала. Если полученная вероятность оказалась ниже некоторого критического малого значения (обычно в интервале 0.1% - 10%), то искомый периодический сигнал вероятно существует, потому как иначе столь большая величина z_{max} была бы труднообъяснима.

Из формулы (12) следует, что вероятность ложной тревоги есть дополнительная вероятность к функции распределения максимального отсчета периодограммы шума. Для ее оценки мы применяем метод Райса, уже кратко описанный во *Введении*. Теперь мы лишь немного конкретизируем формулу (1):

$$\mathrm{FAP}(z_{max}) \le M(z_{max}) = \mathrm{FAP}_{\mathrm{single}}(z_{max}) + \tau(z_{max}). \qquad (13)$$

Здесь член $\mathrm{FAP}_{\mathrm{single}}$ выражает собой вероятность ложной тревоги для *одиночного* значения периодограммы (т.е. фиксируя f); в нашей задаче им как правило можно пренебречь. Основной член - функция τ - представляет собой математическое ожидание числа выбросов периодограммы за уровень z_{max}. Для последнего математического ожидания имеется явная формула (формула Райса), выражающая его через некий интеграл от плотности совместного распределения значений $z(f)$ и $z'(f)$. Метод назван по имени пионера по теории передачи информации С.О.Райса, который вывел эту формулу для стационарного Гауссового случайного процесса [22]. Детали теории по экстремальным значениям и выбросам случайных процессов, и по обобщенному методу Райса в частности, можно найти в обзорах [23,24].

Несколько слов о точности метода Райса. Главная его слабость - чувствительность к корреляциям между удаленными друг от друга значениями случайного процесса. Для периодограмм такая коррелированность может возникать при наличии эффекта *элайзинга*, порождаемого скважностью временного ряда (периодическими пропусками наблюдений). Элайзинг создает на периодограммах системы ложных пиков-"двойников", что в конечном счете снижает точность метода Райса. При этом оценка вероятности ложной тревоги (13) может оказываться завышенной (но не заниженной). Впрочем, это завышение не имеет таких уж серьезных последствий в пересчете на более практически важные уровни значимости и в интересной для нас области $\mathrm{FAP} \ge 10\%$. Исходя из результатов численного моделирования [18,25,26], нам еще ни разу не удалось наблюдать, чтобы такие уровни оказывались завышены (по z) более чем на ~30%.

Заметим, что нарушение неравенства в (13), то есть занижение вероятности ложной тревоги, все же возможно. Такое явление может возникнуть для нелинейных моделей, когда эта нелинейность не учитывается при построении оценки Райса (см. раздел 5). Это еще один потенциальный недостаток метода Райса, который все же далеко не всегда дает на практике серьезные негативные последствия, и вероятно сможет быть ликвидирован в будущем путем совершенствования теории.

4. *Периодограммы с линейными моделями данных.*

4.1. *Общие замечания.* Предположим, что наши модели $\mu_{\mathcal{H}}$, μ, $\mu_{\mathcal{K}}$ линейны по всем параметрам, кроме частоты f. Это допускает запись

$$\mu_{\mathcal{H}}(\theta_{\mathcal{H}},t)=\theta_{\mathcal{H}}\cdot\varphi_{\mathcal{H}}(t),\quad \mu(\theta,f,t)=\theta\cdot\varphi(f,t),\quad \mu_{\mathcal{K}}(\theta_{\mathcal{H}},f,t)=\theta_{\mathcal{K}}\cdot\varphi_{\mathcal{K}}(f,t),\quad(14)$$

где вектора $\varphi_{\mathcal{H}}$ и φ содержат некоторые заданные базисные функции времени, а $\varphi_{\mathcal{K}}=\{\varphi_{\mathcal{H}},\varphi\}$. Если зафиксировать частоту f, то мы имеем простую задачу линейной регрессии. Минимизацию (7) и (8) можно в этой задаче выполнить элементарно, поскольку функция χ^2 во всех случаях оказывается квадратичной формой. В конечном счете, мы получим

$$g_{\mathcal{H}}=\langle x^2\rangle-\mathbf{v}_{\mathcal{H}}^T\mathbf{Q}_{\mathcal{H}}^{-1}v_{\mathcal{H}},\quad g_{\mathcal{K}}=\langle x^2\rangle-\mathbf{v}_{\mathcal{K}}^T\mathbf{Q}_{\mathcal{K}}^{-1}v_{\mathcal{K}},\quad z=\frac{1}{2}\left(\mathbf{v}_{\mathcal{K}}^T\mathbf{Q}_{\mathcal{KK}}^{-1}v_{\mathcal{K}}-\mathbf{v}_{\mathcal{H}}^T\mathbf{Q}_{\mathcal{HH}}^{-1}v_{\mathcal{H}}\right),\quad(15)$$

где вектора $v_{\mathcal{H},\mathcal{K}}=\langle x\varphi_{\mathcal{H},\mathcal{K}}\rangle$ содержат корреляции исходных данных с компонентами функционального базиса соответствующих моделей, а в матрицах

$$\mathbf{Q}_{\mathcal{HH}}=\langle\varphi_{\mathcal{H}}\otimes\varphi_{\mathcal{H}}\rangle,\quad \mathbf{Q}_{\mathcal{KK}}=\langle\varphi_{\mathcal{K}}\otimes\varphi_{\mathcal{K}}\rangle\qquad(16)$$

мы можем узнать информационные матрицы Фишера для этих моделей[2]. Очевидно, что периодограммы (9) также можно легко выразить через введенные вектора и матрицы.

Алгоритм вычисления такой обобщенной периодограммы можно значительно упростить, ортогонализовав базис $\varphi_{\mathcal{K}}(f,t)$ в смысле скалярного произведения $(a(t),b(t))=\langle a(t)b(t)\rangle$ [11]. После этого матрицы Фишера становятся диагональными или даже единичными, и тогда мы имеем $z=|\mathbf{v}|^2/2$, где $\mathbf{v}=\langle x\varphi\rangle$. Такую ортогонализацию базиса необходимо заново выполнять для каждого значения f.

4.2. *Теория уровней значимости: метод Райса для случайных процессов.* Известно [7,10], что при фиксированной f все свободные параметры наших моделей линейны, и поэтому распределение величины $2z(f)$ совпадает с χ^2 распределением с d степенями свободы, а распределения величин $2z_1(f)$ и $2z_2(f)$ - с B- (бета) или F- распределением с $(d,N_{\mathcal{H}})$ или $(d,N_{\mathcal{K}})$ степенями свободы, соответственно. Указанные распределения не зависят при этом от конкретного значения f.

В случае неизвестной частоты хорошо работает метод Райса. Явные формулы для величины τ из формулы (13) в линейном случае даны в [18]. Они были получены из работ [27-29], где рассматривалась общая задача линейной регрессии с одним нелинейным параметром, вырождающимся при $\mathcal{H}$ (у нас таким параметром является частота). Для

[2] *Операция $x\otimes y$ обозначает диадное произведение векторов, т.е. матрицу с элементами $x_i y_j$.*

периодограммы z, например, функцию τ можно представить как

$$\tau=\frac{A(f_{max})-A(f_{min})}{2\pi}\left(\frac{z}{\pi}\right)^{(d-1)/2} e^{-z}, \tag{17}$$

где функция $A(f)$ зависит от базисных функций $\varphi_{\mathcal{K}}$, а также от моментов времени t_i и весов w_r. Для периодограмм $z_{1,2,3}$ аналогичные формулы сложнее, хотя по-прежнему элементарны:

$$\tau=\frac{A(f_{max})-A(f_{min})}{2\pi}\frac{\Gamma\left(\frac{N_{\mathcal{H}}}{2}\right)}{\Gamma\left(\frac{N_{\mathcal{K}}+1}{2}\right)}\times\begin{cases}\left(\frac{2z_1}{\pi N_{\mathcal{H}}}\right)^{(d-1)/2}\cdot\left(1-\frac{2z_1}{N_{\mathcal{H}}}\right)^{(N_{\mathcal{K}}-1)/2}, \\ \left(\frac{2z_2}{\pi N_{\mathcal{K}}}\right)^{(d-1)/2}\left(1+\frac{2z_2}{N_{\mathcal{K}}}\right)^{(-N_{\mathcal{H}}/2)+1}, \\ \left(\frac{2}{\pi}\operatorname{sh}\frac{z_3}{N_{\mathcal{K}}}\right)^{(d-1)/2} e^{-z_3(1+(d-3)/2N_{\mathcal{K}})}.\end{cases} \tag{18}$$

Коэффициент $A(f)$, который мы не будем здесь выписывать в общей форме, имеет вид интеграла по частоте от некоторой функции, обычно содержащей величины типа $\left\langle t^n \begin{smallmatrix}\cos\\ \sin\end{smallmatrix}\omega t\right\rangle$ (с небольшими $n\geq 0$). В такой ситуации оказывается весьма эффективным приближение, которое мы будем называть "приближением равномерного покрытия фазы". Оно состоит в том, чтобы заменить операцию осреднения по временному ряду $\langle * \rangle/\langle 1 \rangle$ на операцию осреднения при помощи непрерывного интеграла по времени. При этом получается, что суммами $\left\langle t^n \begin{smallmatrix}\cos\\ \sin\end{smallmatrix}\omega t\right\rangle$ мы можем просто пренебречь в сравнении с $\left\langle |t|^n \right\rangle$ и иными похожими величинами. Этот прием позволяет добиться значительного упрощения конечных формул. Такое приближение хорошо работает для частот, на которых наблюдения покрывают почти равномерно все фазы колебания, соответствующего частоте ω. Следует признать, что это условие часто нарушается, например когда на какой-то частоте моменты t_i показывают периодические пропуски (явление элайзинга). К счастью, на практике подобный эффект может держаться лишь в узких частотных интервалах шириной $\Delta f\sim 1/T$, где T - длина временного ряда (обычно большая). А поскольку $A(f)$ включает в себя интегрирование по f в значительно большем частотном диапазоне, эти аномальные интервалы не оказывают на конечный результат почти никакого влияния. Для практической применимости такого приближения нужно, чтобы интервал интегрирования по частоте $f_{max}-f_{min}$ был велик в сравнении с $1/T$.

В приближении равномерного покрытия фазы функция $A(f)$ обычно оказывается просто пропорциональна f, что мы увидим на примере частных случаев ниже.

4.3. ***Периодограмма Ломба-Скаргла и мультигармонические периодограммы.*** В этом случае мы имеем $d_{\mathcal{H}}=0$, $d=2n$, а базис модели сигнала зададим в форме

$$\varphi=\{\cos(2\pi kft),\sin(2\pi kft),k=1,2,...,n\},\quad d=2n. \tag{19}$$

При этом модель $\mu_{\mathcal{H}}$ предполагается пустой, что неявно предполагает предварительное центрирование данных. Эти периодограммы были предложены для произвольного n в работе [8]. Сигнал здесь моделируется отрезком ряда Фурье заданной степени n. При $n=1$ мы имеем простую периодограмму Ломба-Скаргла с синусоидальной моделью сигнала, а при $n>1$ такие периодограммы применяются, когда потенциальный периодический сигнал может иметь несинусоидальную форму [30].

Вычисление мультигармонической периодограммы можно заметно ускорить в сравнении с общим алгоритмом линейной регрессии, пользуясь некоторыми особенностями данного функционального базиса [8,31].

Как показано в [18,25,32], в этом случае приближение равномерного покрытия фазы дает $A(f)\approx 2\pi^{n+1/2}\alpha_n T_{\mathit{eff}} f$, где

$$\alpha_n=\frac{2^n}{(2n-1)!!}\sum_{k=1}^{n}\frac{(-1)^{n-k}k^{2n+1}}{(n+k)!(n-k)!},\quad T_{\mathit{eff}}=\sqrt{4\pi\left(\frac{\langle t^2\rangle}{\langle 1\rangle}-\left(\frac{\langle t\rangle}{\langle 1\rangle}\right)^2\right)}. \tag{20}$$

Здесь T_{eff} - некоторая эффективная длина временного ряда, которая на практике оказывается близка к его обычной длине T. Вводя безразмерный параметр $W=(f_{max}-f_{min})T_{\mathit{eff}}$, для общей мультигармонической периодограммы z получаем $\tau\approx\alpha_n We^{-z}z^{n-1/2}$, а для собственно периодограммы Ломба-Скаргла $\tau\approx We^{-z}\sqrt{z}$ и значит $\mathrm{FAP}(z)\le e^{-z}\left(W\sqrt{z}+1\right)$. Аналогичные формулы можно применить и к нормализованным периодограммам $z_{1,2,3}$, исходя из (18).

Для применимости приближения равномерного покрытия фазы нужно, чтобы рассматриваемый частотный интервал был велик в сравнении с $1/T_{\mathit{eff}}$, то есть чтобы велика была W. На практике описанного приближения для $A(f)$ бывает совершенно достаточно уже для $W>10$, причем даже для рядов с предельно мощным элайзингом [18,25].

Эта теория остается в равной степени справедливой и тогда, когда модель $\mu_{\mathcal{H}}$ содержит полиномиальный тренд низкой степени. Сюда относится, например, случай $\mu_{\mathcal{H}}(c,t)=c$ (со свободным параметром c), соответствующий известной "периодограмме с плавающим средним" [12,33,34]. Заметим, что подобные периодограммы с "непустой" $\mu_{\mathcal{H}}$ не эквивалентны периодограммам с $\mu_{\mathcal{H}}\equiv 0$, примененным к предварительно центрированному временному ряду или к временному ряду с предварительно удаленным трендом. Строгий подход со встроенной нетривиальной $\mu_{\mathcal{H}}$ более эффективен в смысле чувствительности к слабым сигналам [12,35].

4.4. ***Периодограмма Шустера.*** Как и периодограмма Ломба-Скаргла, эта периодограмма определена для синусоидальной модели сигнала и пустой (нулевой) базовой модели. В наших обозначениях (принимая немного нетрадиционную нормировку) ее можно выразить формулой

$$S(x)=\frac{\left|\left\langle xe^{2\pi ift}\right\rangle\right|^2}{\langle 1\rangle}=\frac{1}{\langle 1\rangle}\left(\langle x\cos 2\pi ft\rangle^2+\langle x\sin 2\pi ft\rangle^2\right). \qquad (21)$$

Интересно, что данную периодограмму можно получить из периодограммы Ломба-Скаргла путем приближения равномерного покрытия фазы, пренебрегая величинами $\langle\sin 4\pi ft\rangle$ и $\langle\cos 4\pi ft\rangle$ в сравнении с $\langle 1\rangle$.

Строго говоря, периодограмма Шустера вовсе не относится к введенным нами периодограммам линейной регрессии. Ее достоинство состоит в том, что для ее вычисления можно применять алгоритм быстрого преобразования Фурье, поэтому она до сих пор часто применяется на практике. Наиболее важный недостаток - ее статистические распределения куда сложнее распределений периодограммы Ломба-Скаргла. Соответствующее распределение одиночного отсчета S неэлементарно; оно дано в [4,36]. Распределение максимального отсчета S неизвестно.

Впрочем, на практике разница между этими периодограммами обычно невелика. Обозначив периодограмму Ломба-Скаргла как $L(f)$, мы можем написать неравенство

$$1-\sqrt{\Omega(2f)}\le\frac{S(f)}{L(f)}\le 1+\sqrt{\Omega(2f)}, \qquad (22)$$

где $\Omega(f)=\left(\langle\cos 2\pi ft\rangle^2+\langle\sin 2\pi ft\rangle^2\right)/\langle 1\rangle^2$ есть спектральное окно временного ряда, нормированное так, что $\Omega(0)=1$. Неравенство (22) следует из известного в матричной алгебре факта, что $\lambda_{min}|\mathbf{x}|^2\le\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\le\lambda_{max}|\mathbf{x}|^2$, где λ_{max}, λ_{min} есть максимальное и минимальное собственные числа $\mathbf{A}$ (в нашем случае $\mathbf{A}=\mathbf{Q}_{\chi\chi}^{-1}$, а $\mathbf{x}=\mathbf{v}_{\chi}$).

Из (22) очевидно, что $L(f)$ и $S(f)$ могут сильно различаться лишь в пиках спектрального окна $\Omega(2f)$, которые соответствуют частотам периодических пропусков в данных, т.е. частотам элайзинга. Для большинства же частот $L(f)$ мало отличается от $S(f)$. Влияние эффекта элайзинга на общий уровень шума в *широком* частотном интервале невелико: даже если на некоторых частотах шумовые пики оказываются немного занижены или завышены, максимум по всему интервалу частот будет почти наверняка определяться не этими отдельными пиками, а какими-то другими.

Таким образом, у нас есть основания подозревать, что когда дело касается широкого интервала частот, уровни значимости периодограмм L и S должны быть близки, по крайней мере с точки зрения приближения (13). Для проверки этой догадки мы предприняли попытку применить метод Райса к $S(f)$. В общем случае формулу Райса для $S(f)$ не удается, к сожалению, довести до элементарных функций. При этом "портится" не только множитель A в формулах (17), но и сама зависимость от z. Однако в условиях приближения "равномерного покрытия фазы" результат получается совершенно идентичным результату для L.

Исходя из сказанного, мы полагаем, что хотя бы при каких-то условиях для уровней значимости периодограммы Шустера может быть допустимо применять теорию периодограммы Ломба-Скаргла, то есть приближение $\mathrm{FAP}(z)\le e^{-z}\left(W\sqrt{z}+1\right)$. Аналогично (9), можно ввести нормализованные периодограммы для S, например $S_1(f)=NS(f)/\langle x^2\rangle$, распространив на нее теорию уровней значимости $z_1(f)$. Впрочем, на данный момент это пока лишь некоторая гипотеза, требующая проверки и уточнения, а не рекомендация для практического частотного анализа. Статистические уровни значимости периодограммы Шустера все еще требуют дальнейшего изучения; эту задачу пока нельзя считать решенной.

5. *Нелинейные периодограммы.*

5.1. ***О роли нелинейности***. Что будет, если хотя бы базовая модель $\mu_{\mathcal{H}}$ нелинейна? Такое может быть, например, когда мы уже надежно выделили в данных одну или несколько синусоидальных или других периодичностей, и хотим знать, нет ли там еще одной. При этом нелинейными параметрами будут как минимум частоты уже выделенных периодичностей. Такие периодограммы в работах [17,19,20] назывались "остаточными периодограммами" ("residual periodograms"), а в [35] "рекурсивными". Сюда можно также отнести и периодограммы отношения правдоподобия (10) с параметризованным шумом, так как модель $\varpi_i(\mathbf{p})$ также, в сущности, есть некоторая базовая модель, одинаковая для гипотез $\mathcal{H}$ и $\mathcal{K}$, и генерирующая эффекты нелинейности.

Нам хотелось бы уяснить, каких изменений можно ожидать в теории уровней значимости нелинейных периодограмм. Ввиду того, что в вектор $\theta_{\mathcal{H}}$ теперь входят нелинейные параметры, мы уже не можем записать $\mu_{\mathcal{H}}=\theta_{\mathcal{H}}\cdot\varphi_{\mathcal{H}}$. Однако в большинстве практических задач мы можем линеаризовать $\mu_{\mathcal{H}}$ в некоторой малой окрестности параметров (например, вокруг оценок $\theta_{\mathcal{H}}$ при базовой гипотезе). При этом, чем больше число наблюдений N, тем с меньшими уклонениями по $\theta_{\mathcal{H}}$ мы имеем дело, и тем точнее оказывается такая линеаризация. Таким образом, мы можем ожидать что теория уровней значимости линейных периодограмм будет в целом выполняться и для слабонелинейных моделей, правда лишь в асимптотическом смысле ($N\to\infty$).

Довольно трудно дать строгую формулировку условий "слабой" нелинейности. Мы этого здесь и не будем делать. И все же заметим, что одним из ключевых моментов здесь является то, допустима ли вообще линеаризация $\mu_{\mathcal{H}}$ в требуемой окрестности. Если при этом $\mu_{\mathcal{H}}$ не слишком уж сложна, то, как показывает опыт, теория линейных периодограмм для оценки соответствующей вероятности ложной тревоги может работать неплохо [17-20,37].

Однако если нелинейность заключена в самом сигнале, то все обстоит в точности наоборот. Дело в том, что модель μ надо линеаризовать сразу по всем параметрам θ в окрестности точки, принадлежащей базовой гипотезе $\mathcal{H}$ (поскольку нам нужно получить распределение периодограммы именно в отсутствии сигнала). Оказывается довольно трудным представить себе периодическую модель μ, за исключением синусоиды, которую можно было бы адекватно линеаризовать в окрестности нулевой амплитуды сигнала. Как минимум фаза сигнала не будет поддаваться такой линеаризации. Причина такого поведения - в неопределенности большинства параметров сигнала при гипотезе $\mathcal{H}$ [26,38]. Синусоида представляет собой счастливое исключение: мы можем переписать ее в форме $A\cos(x+\lambda) = a\cos x + b\sin x$, где оба параметра a и b при $A=0$ вполне осмысленны (в отличие от λ).

Таким образом, выводов в данном разделе у нас два. Во-первых, нелинейность в базовой модели (включая сюда и нетривиальные модели шума) часто несущественна; при достаточно больших N она обычно допускает оценку значимости пиков периодограммы по формулам линейной теории. Во-вторых, нелинейность в модели сигнала как правило принципиальна и неустранима; значимость пиков таких периодограмм недопустимо оценивать при помощи линейной теории. Именно такие сильно нелинейные сигналы мы и будем рассматривать в этой главе. Здесь мы приведем лишь основные полученные результаты; подробности представлены в работе [26].

5.2. *Теория уровней значимости: метод Райса для случайных полей*. В совсем общем случае о форме модели сигнала μ мало что можно сказать: просто некая нелинейная функция. Однако практически во всех приложениях из набора параметров θ можно выделить как минимум один линейный - амплитуду K. При этом мы можем положить $K \equiv \theta_1$, а оставшиеся $d-1$ параметров θ обозначить вектором ξ. Тогда имеем

$$\mu(\theta, f, t) = Kh(\xi, f, t), \tag{23}$$

где h есть модель *формы* сигнала. Тогда в определениях (7) и (8) функции χ^2 можно минимизировать по K, сохранив зависимость от ξ и f. Несложно показать, что если $\mu_{\mathcal{H}}$ все еще линейна, то в результате этой процедуры получается

$$z(f) = \frac{1}{2}\max_{\xi\in\Xi} \eta^2(\xi, f), \quad \eta(\xi, f) = \langle x\,\psi(\xi, f, t)\rangle, \tag{24}$$

где функция ψ получается в результате некоторого смещения и нормированния фунции h, так что $\langle\mu_{\mathcal{H}}\psi\rangle = 0$ и $\langle\psi^2\rangle = 1$. Например, при $\mu_{\mathcal{H}} = 0$ имеем $\psi = h\Big/\sqrt{\langle h^2\rangle}$.

Таким образом максимизация z сводится к максимизации η, так что $\max|\eta| = \sqrt{2\max z}$. Если x_i содержат Гауссов шум, то $\eta(\xi, f)$ - Гауссово случайное поле. При этом в отсутствии сигнала (то есть при гипотезе $\mathcal{H}$) мы имеем $\mathbb{E}\eta = 0$, $\mathbb{D}\eta = 1$, то есть это поле центрировано и нормировано.

Для таких полей разработано многомерное обобщение метода Райса [24,39]. Поскольку понятие "выброса" для случайного поля не определено, то вместо выбросов за заданный уровень подсчитываются локальные максимумы η, расположенные выше того же уровня. Вероятность ложной тревоги оценивается сверху математическим ожиданием числа таких максимумов, которое выражается аналогом формулы Райса.

Итак, приведем основные результаты, которые удалось получить, применяя данный метод к нелинейным периодограммам. Введем пока небольшое техническое ограничение на K вида $K \geq 0$. Это значит, что в ходе нашего анализа мы не допускаем "переворота" формы сигнала ($h \mapsto -h$). Для синусоиды такое ограничение не имеет никакого значения ввиду ее симметрии, но для несимметричных моделей оно может быть значимо и часто даже практически полезно. Для нас оно удобно, потому что позволяет отбросить значения $\eta < 0$, так что максимум z всегда совпадает с максимумом η (в противном случае, нам пришлось бы рассматривать еще и минимумы η). Тогда мы можем записать аналог формулы (13) в виде

$$\mathrm{FAP}(z_{max}) \leq M(z_{max}) \approx M_{boundary}(z_{max}) + \tau(z_{max}), \tag{25}$$

где τ есть математическое ожидание числа локальных максимумов выше z_{max}, а член $M_{boundary}$ отвечает за максимумы η, расположенные на границе области параметров (ξ, f). Этот член аналогичен члену FAP_{single} в (13); он обычно имеет относительную величину (в сравнении с τ) порядка $1/\sqrt{z}$ и сильно зависит от геометрии границы. На практике им можно или нельзя пренебрегать в зависимости от условий задачи. Этот член обычно вычисляется, с некоторыми оговорками, аналогично самому M, рассматривая ограничение задачи на границу области параметров.

Основным членом все же остается τ; он представим в форме

$$\tau(z) \approx \sum_{k=0}^{[n/2]} c_k P_{n+1-2k}(z), \quad P_k(z) = \frac{1}{\Gamma(k/2)}\int_z^{\infty} x^{k/2-1} e^{-x} dx, \tag{26}$$

где n есть число нелинейных параметров (в нашем случае $n = d$). Данная аппроксимация имеет асимтотический характер при $z \to \infty$; ее относительная ошибка стремится к нулю быстрее, чем $\exp(-z\delta)/z^{(n-1)/2}$ с некоторым $\delta > 0$. На практике мы обычно как раз и заинтересованы в области больших z, где находятся малые вероятности.

К сожалению, из коэффициентов c_k легко вычислить только старший c_0; остальные даже в простых приложениях оказываются слишком сложными. К счастью, члены с $k \geq 1$ вносят в сумму (26) небольшой относительный вклад $O(1/z^k)$, поэтому на практике ими часто можно пренебречь. Оставляя в сумме (26) только величины высшей степени по z, имеем

$$\tau(z) \approx A e^{-z} z^{(n-1)/2}, \tag{27}$$

где множитель A (заметим, это немного не то A, что возникало для линейных периодограмм) выражается как

$$A = \frac{1}{2\pi^{(n+1)/2}} \int_{f_{min}}^{f_{max}} df \int_{\Xi} \sqrt{\det \mathbf{G}}\, d\xi, \quad \mathbf{G} = \mathrm{Var}(\eta') = \langle \psi' \otimes \psi' \rangle. \tag{28}$$

В последней формуле η' и ψ' есть вектора градиентов соответствующих функций по всем нелинейным параметрам (ξ и f).

5.3. *Нелинейный сигнал + нелинейная базовая модель.* Описанная выше теория шумовых уровней нелинейных периодограмм справедлива лишь для ненормированной периодограммы z, основанной на χ^2-критерии, а также предполагает линейность $\mu_{\mathcal{H}}$. Для нормализованных периодограмм $z_{1,2,3}$ эта теория напрямую не работает. Для построения аналогичной теории нелинейных периодограмм $z_{1,2,3}$, необходим метод Райса для негауссовых случайных полей, который пока не разработан. Тем не менее, как мы обсуждали выше, теория χ^2-периодограмм (как линейных, так и не линейных) должна работать в *асимптотическом* смысле для соответствующих периодограмм отношения правдоподобия и, как следствие, для периодограммы z_3. При этом также допускается нелинейность базовой модели $\mu_{\mathcal{H}}$.

Фактически, это утверждение справедливо и для линейных периодограмм; выражения τ для z и z_3 при $N \gtrsim z$ асимптотически эквивалентны (ср. (17) и (18)). Последнее условие выражает тот факт, что особенно большие z соответствуют слишком большим отклонениям параметра κ от "номинала", что ухудшает точность его линеаризации. Заметим, что уровни значимости периодограммы $z_{1,2}$ демонстрируют аналогичную эквивалентность с z, но лишь при $N \gtrsim z^2$, чего для практических нужд уже недостаточно.

Это свойство представляет собой большую ценность, так как χ^2-периодограмму мы обычно просто не можем использовать из-за плохого знания σ_i. Вместо этого мы можем построить периодограмму z_3, применив к ней теорию уровней значимости z (с той же моделью μ) в асимптотическом смысле. Аналогичный подход как правило работает и для периодограмм отношения правдоподобия (10) и (11). При этом условие аппроксимации имеет вид $N \gtrsim Z$, $\widetilde{Z}$ [17]. Следует иметь в виду, что в конкретных случаях эта аппроксимация может оказаться все же плохой из-за недостаточной величины N (критическое N зависит от степени нелинейности конкретной задачи). Заметим, что эффекты нелинейности базовой модели могут в принципе нарушать неравенство в оценке (13), т.е. если нелинейность в $\mu_{\mathcal{H}}$ сильна, то оценка вероятности ложной тревоги может все же оказаться заниженной, а статистическая значимость выделенного колебания - напротив, завышенной. Для полного устранения данной проблемы необходимо еще обобщить теорию метода Райса, внеся в нее дополнительные расширения.

5.4. ***Периодограмма с заданной несинусоидальной формой сигнала***. В работе [26] рассмотрены две нелинейных модели сигнала. Первая имеет вид

$$h(\lambda, f, t) = u(2\pi ft + \lambda), \tag{29}$$

где $u(x)$ задает форму периодического сигнала. При этом вектор ξ состоит из одного параметра λ. Для периодограммы Ломба-Скаргла $u(x) = \cos x$, а для произвольного u, пользуясь описанной выше методикой и приближением равномерного покрытия фазы, можно получить

$$\mathrm{FAP}_{single}(z) \le M_{single}(z) \approx \sqrt{q}\, e^{-z}, \quad \mathrm{FAP}(z) \le M(z) \approx e^{-z}\left(qW\sqrt{z} + \sqrt{q}\right), \tag{30}$$

где коэффициент q зависит только от формы сигнала:

$$q = \int_0^{2\pi} (u'(x))^2\, dx \Bigg/ \int_0^{2\pi} (u(x) - u_0)^2\, dx, \quad u_0 = \frac{1}{2\pi}\int_0^{2\pi} u(x)dx. \tag{31}$$

Для синусоиды $q = 1$, в общем же случае $q \ge 1$.

5.5. ***Нелинейная "периодограмма фон Мизеса" для переменных звезд и экзопланетных прохождений***. Вторая периодическая модель из [26] имеет вид

$$h(\xi, f, t) = \exp(\nu \cos(2\pi ft + \lambda)), \quad \xi = \{\lambda, \nu\}. \tag{32}$$

В теории вероятностей нормированный вариант функции (32) вида $e^{\nu\cos(x+\lambda)}/I_0(\nu)$, где I_0 - модифицированная функция Бесселя, описывает т.н. плотность распределения фон Мизеса, которое представляет собой один из аналогов распределения Гаусса для угловых переменных. Ввиду этого мы называем данную периодограмму периодограммой фон Мизеса.

Благодаря лишь одному дополнительному параметру ν, модель (32) охватывает весьма широкий класс периодичностей: от синусоиды (предел при $\nu \to 0$) до периодичностей с острыми пиками или провалами, которые характерны для затменных двойных звезд или для экзопланетных прохождений. Благодаря такой универсальности, данная периодограмма может оказаться довольно полезной для указанных астрономических объектов, даже несмотря на ее повышенную вычислительную трудоемкость.

Подробности применения метода Райса к этой периодограмме описаны в [26], где в качестве онлайн-приложения опубликован и вспомогательный пакет вычислительных програм на C++. Если мы сканируем периодограмму в прямоугольнике $f_{min} \le f \le f_{max}$, $\nu_{min} \le \nu \le \nu_{max}$ (и, понятно, $0 \le \lambda \le 2\pi$), то

$$\mathrm{FAP}(z) \le M(z) \approx \left(X(\nu_{max}) - X(\nu_{min})\right) W e^{-z} z + \left(Y(\nu_{max}) + Y(\nu_{min})\right) \frac{W}{2} e^{-z} \sqrt{z}, \quad (33)$$

где функции X и Y, помимо аргументов ν, также слабо зависят от f_{min}, f_{max}, и сильнее - от моментов времени и весов временного ряда. За подробностями и дополнительными разъяснениями по периодограмме фон Мизеса см. работу [26].

5.6. *Кеплерова периодограмма*. Еще один пример периодограммы с нелинейной моделью сигнала - Кеплерова периодограмма, которая используется для поиска экзопланет по колебаниям лучевой скорости звезды [13,40]. Сигнал моделируется выражением кеплеровской лучевой скорости компонент двойной системы:

$$h = \cos\left(\omega + \upsilon(e, 2\pi f t + \lambda)\right) + e\cos\omega, \quad (34)$$

где ω есть аргумент перицентра орбиты невидимого спутника (экзопланеты), e - ее эксцентриситет, а $\upsilon(e,M)$ - истинная аномалия, зависящая от эксцентриситета e и соответствующей средней аномалии M. Последняя представляет собой фазовый угол наблюдаемого доплеровского сигнала.

В сравнении с синусоидальной моделью, Кеплерова модель содержит уже два дополнительных нелинейных параметра, e и ω. Эта периодограмма разрабатывалась для нужд доплеровского поиска экзопланет.

Хотя эта периодограмма уже около 10 лет применяется на практике разными авторами, для ее уровней значимости до сих пор не было известно аналитического выражения. При помощи методики раздела 5.2 вероятность ложной тревоги удается оценить в виде выражения, похожего на формулу (33):

$$\mathrm{FAP}(z) \le M(z) \approx 2X(e_{max}) W e^{-z} z^{3/2} + Y(e_{max}) W e^{-z} z \sqrt{\pi}, \quad (35)$$

где функции X и Y теперь зависят от максимального допустимого эксцентриситета (это конечно не те же самые функции, что в формуле (33), а некоторые другие). Вывод данного выражения с подробным обсуждением и детализацией будет дан в отдельной работе, которая находится в подготовке.

6. *Многочастотный анализ и составные сигналы*. Строго говоря, все описанные выше периодограммы основаны на предположении, что во временном ряду присутствует лишь одна периодичность (синусоидальная или нет), чья статистическая значимость находится под вопросом. Если сигнал состоит из априори неопределенного числа периодических компонент с неизвестными частотами, напрашивается механизм последовательного извлечения компонент сигнала:

1. Строим периодограмму с предельно простой базовой моделью $\mu_{\mathcal{H}}$ (скажем, включающей только неопределенную константу). Проверяем статистическую значимость максимального пика.

2. Если найденная периодичность значима, строим периодограмму с базовой моделью $\mu_{\mathcal{H}}$, включающей уже выделенную периодичность. Опять ищем максимальный пик периодограммы и проверяем его значимость.

3. Если вторая периодичность значима, включаем и ее в модель $\mu_{\mathcal{H}}$, строим с этой $\mu_{\mathcal{H}}$ новую периодограмму, снова проверяем ее максимальный пик.

4. Повторяем такие шаги, пока с остаточной периодограммы не исчезнут все значимые пики.

Однако при ближайшем рассмотрении оказывается, что такой алгоритм содержит недостатки. Во-первых, бывают случаи, когда максимальный пик на периодограмме является ложным, представляя собой результат интерференции частот регулярных пропусков в наблюдениях с частотами реальных сигналов. При этом истинные периодичности соответствуют каким-то побочным пикам. В этом случае было бы полезно "заглянуть" в ходе анализа на ступеньку вперед, проверив, не получим ли мы в итоге лучшей модели, приняв за сигнал не главный пик периодограммы, а несколько побочных. Последовательный алгоритм такой возможности не дает. Кроме того, такой алгоритм содержит концептуально ошибочную интерпретацию статистической значимости выделяемых компонент. Положим, мы таким образом выделили из шума 10 компонент сигнала, и каждая имеет приемлемо низкую вероятность ложной тревоги, скажем 1%. Это значит, что при совершении *каждого* из 10 актов выделения очередной компоненты, мы рисковали ошибиться (принять шум за сигнал) с вероятностью 1%. Однако какова будет итоговая вероятность того, что мы ошиблись хотя бы *в одной* компоненте из выделенных десяти? Очевидно, эта вероятность будет значительно выше, чем 1%. К примеру, если считать события выделения компонент статистически независимыми, то вероятность ошибиться хотя бы один раз из десяти будет $1-(1-0.01)^{10} \approx 9.5\%$. Если вероятность ошибки в 1% была вполне приемлема, то вероятность в 9.5% уже слишком велика. В реальности предположение о статистической независимости компонент не выполняется, поэтому эти 9.5% лишь весьма грубая оценка, однако в любом случае получается, что мы все еще не можем быть уверены в статистической значимости *всех* найденных частотных компонент. Даже если каждая из этих компонент проходит тест значимости по отдельности, итоговая вероятность ошибки по всему их ансамблю будет значительно выше, чем индивидуальные вероятности. Необходимо модифицировать описанный последовательный алгоритм, чтобы учесть указанный эффект структуры сигнала.

Этот вопрос рассматривался в работе [41], где он был решен с применением так называемых многочастотных периодограмм. В этих периодограммах сигнал моделируется суммой нескольких однотипных периодических компонент с разными частотами. Каждая такая периодограмма является уже функцией *нескольких* частот, которые считаются независимыми друг от друга. В простейшем случае сигнал представляет собой сумму синусоид:

$$\mu(\theta, \mathbf{f}, t) = \sum_{i=1}^{n} A_i \cos(2\pi f_i t) + B_i \sin(2\pi f_i t). \quad (36)$$

Здесь параметры A_i, B_i формируют вектор θ, а вместо единственной частоты имеем теперь частотный вектор $\mathbf{f}$.

Итак, допустим с помощью описанного выше алгоритма последовательного извлечения компонент или какого-то сходного алгоритма мы выделили n вероятных компонент сигнала. Как нам удостовериться, что *все* эти компоненты обладают достаточной статистической значимостью, как индивидуально, так и рассматриваемые в ансамбле? Как показано в [41], для этого нам нужно построить полный набор всевозможных многочастотных периодограмм, содержащих указанные n частотных компонент. При этом в модель сигнала может вводиться лишь некоторое подмножество размера $m \le n$, а оставшиеся $n-m$ компонент следует при этом поместить в базовую модель $\mu_{\mathcal{H}}$. Для каждого m мы таким образом имеем всего C_n^m значений различных периодограмм, а в сумме - $\sum_{m=1}^{n} C_n^m = 2^n - 1$ значений. Каждое из полученных значений представляет собой тестовую статистику в виде максимума случайного поля (многочастотной периодограммы), для которой вероятность ложной тревоги может быть вычислена с помощью многомерного метода Райса из раздела 5.2 (итоговая формула дается ниже). Итоговая вероятность ложной тревоги по всему ансамблю представляет собой наихудшее, то есть максимальное значение из всех полученных. Если эта максимальная вероятность ниже заданного малого уровня, то мы можем с уверенностью утверждать, что скорее всего все выделенные компоненты сигнала реальны. Если же эта максимальная вероятность выше заданного порога, то таких гарантий дать нельзя, и остается констатировать, что не все компоненты выделены надежно. Заметим, что в последнем случае не всегда можно сказать, какие именно компоненты сомнительны, так как индивидуально они могут оказаться значимы все (незначимыми могут оказаться лишь какие-то многочастотные комбинации).

Заметим также, что идея многочастотных периодограмм сама по себе не нова - для решения проблем частотного анализа временных рядов похожие статистики применялись еще Фостером [42-44] в его алгоритме CLEANest. Однако задача определения статистической значимости таких периодограмм не нашла в соответствующей литературе должного внимания, вероятно из-за отсутствия простого способа решения. Метод Райса как раз и позволяет эту задачу легко решить. Согласно [41], вероятность ложной тревоги, связанная с наивысшим пиком многочастотной периодограммы порядка n (т.е. с n гармониками в модели сигнала), аппроксимируется выражением

$$\mathrm{FAP}(z) \leq M(z) \simeq \tilde{A}_n e^{-z} z^{3n/2-1}, \quad \tilde{A}_n = \frac{W^n}{n!} I_1 I_2 \ldots I_{n-1}, \quad I_k = B\left(\frac{3k}{2}, \frac{3}{2}\right), \qquad (37)$$

где $B(p, q)$ есть бета-функция, z - наблюдаемый максимальный отсчет многочастотной периодограммы, а W - нормализованная ширина поискового частотного диапазона, такая же как для периодограммы Ломба-Скаргла.

7. ***Новые программные пакеты для частотного анализа временных рядов.*** Очевидно, что многие аспекты описанных статистических методов выделения периодичностей из шума довольно сложны, и их может быть трудно применить в практических вычислениях без глубокого понимания соответствующей теории. По этой причине мы предлагаем ряд готовых программных пакетов, значительно облегчающих такие практические вычисления.

В работе [26] помимо самой математической теории уровней значимости периодограммы фон Мизеса был опубликован программный пакет для вычисления как самой этой периодограммы, так и ее уровней значимости, а также для проведения связанного с ней численного моделирования Монте Карло и для решения некоторых вспомогательных задач. Этот пакет может оказаться полезен при анализе фотометрических временных рядов с целью поиска в них сильно несинусоидальных периодичностей, характерных для экзопланетных прохождений и многих типов звездной переменности. Пакет первоначально был опубликован на сервере издательства в качестве скачиваемого онлайн-дополнения к статье. Позднее его стало возможным скачать в свободном доступе по URL http://sourceforge.net/projects/vonmises, где в будущем могут появиться также и обновленные версии.

В работе [45] представлено подробное описание другого программного инструмента для проведения частотного анализа временных рядов. Данный пакет, FREDEC - "Frequency Decomposer", реализует теорию многочастотного анализа, основные принципы которой были описаны выше в разделе 6. Алгоритм анализа реализован в виде "трубопровода", то есть выполняет самостоятельно всю работу от начала до конца, не требуя вмешательства пользователя. "Трубопровод" или "конвейер" алгоритма содержит следующие основные стадии:

1. Нахождение предварительного набора кандидатов в возможные периодичности при помощи гибридного сочетания методов QUICK и SLICK из [42].

2. Перебор всевозможных комбинаций найденных компонент (в синусоидальной модели) и фильтрация комбинаций, прошедших полный тест значимости (см. раздел 6).

3. Отбрасывание из получившегося набора альтернативных моделей сигнала таких, которые в сравнении с другими слишком плохо согласуются с исходным временным рядом. Эта фильтрация альтернатив выполняется при помощи критерия Вуонга, применимого для статистического сравнения параметрически невложенных моделей [46,47].

На выходе такого алгоритма мы получаем набор альтернативных многочастотных комбинаций, которые лучше всего описывают исходный временной ряд. При этом в рамках каждой такой комбинации все выделенные компоненты гарантированно значимы (если считать, что верна именно данная модель, а не какая-то из прочих альтернатив). FREDEC является пакетом общего назначения, то есть не содержит какой-либо специфической адаптации под конкретную узкую научную задачу или под какой-то заданный тип исходных данных. Пакет можно скачать по URL http://sourceforge.net/projects/fredec. Этот алгоритм способен переносить вычисления на доступную в системе видеокарту (графический ускоритель), где они ведутся в параллельном режиме. Это позволяет значительно - в десятки раз - увеличить производительность вычислений.

Наконец, отметим пакет PlanetPack, созданный для обработки высокоточных доплеровских временных рядов с целью поиска экзопланет и исследования экзопланетных систем [48]. PlanetPack включает весьма обширный набор методов статистического анализа, объединенных в целостную систему при помощи собственной командной оболочки. В части задачи поиска периодичностей данный пакет позволяет работать с расширениями периодограммы Ломба-Скаргла и мультигармонических периодограмм, учитывающих аддитвную и иные неклассические модели доплеровского шума, включая коррелированный шум. Начиная с версии PlanetPack 1.6 имеется возможность работы с Кеплеровой периодограммой (пока еще без аналитической оценки уровней значимости). Готовящаяся версия PlanetPack 2.0 включает возможность ускорения вычислений путем их параллелизации на современных многоядерных процессорах. Пакет можно скачать по адресу http://sourceforge.net/projects/planetpack, вместе с техническим руководством.

8. *Заключение*. Описанную в данном обзоре теорию уровней значимости периодограмм мы в целом оцениваем как серьезный прорыв. На протяжении примерно трех десятилетий были доступны, в сущности, лишь два основных метода оценки значимости пиков периодограммы Ломба-Скаргла. Это "чистый" метод Монте-Карло и метод Монте-Карло с экстраполяцией "формулы независимых частот" [6]. И тот и другой

требуют затратного численного моделирования, а надежность получаемого во втором методе результата еше и весьма сомнительна [14,18].

Представленный в данной статье единый подход имеет строгую математическую основу метода Райса, обладает высокой практической эффективностью, а также легко обобшается. С его помощью удается решать весьма сложные и даже на первый взгляд неприступные задачи, как, например, определение распределений периодограмм с сильно нелинейной моделью сигнала. Применение такой методики в практических задачах анализа астрономических и иных временных рядов по меньшей мере чрезвычано полезно.

Данная работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект 12-02-31119 мол_а), гранта Президента РФ для молодых ученых - кандидатов наук МК-733.2014.2, а также программы Президиума РАН "Нестационарные явления в объектах Вселенной".

Главная (Пулковская) астрономическая обсерватория РАН, Санкт-Петербургский государственный университет, НИАИ им. В.В.Соболева, Россия, e-mail: r.baluev@spbu.ru

REVEIWS

ON THE DETECTION OF PERIODIC COMPONENTS IN OBSERVATIONAL DATA

R.V.BALUEV

The review is devoted to the task of period search in observational data by means of periodograms based on the general statistical likelihood-ratio test and its special cases, including the classic Lomb-Scargle periodogram. The main attention is paid to the task of assessing the statistical significance of detected periodicities. We suggest the universal solution to this task using the efficient technique, in which the periodogram is considered as a random process (or random field), and the approximation to the desired false alarm probability is obtained by means of the generalized Rice method. Additionally to the general technique of determining the expected periodogram noise (or significance) levels, we consider several important special cases involving different models of the periodic signal (linear as well as non-linear). In the most cases, the false alarm probability, associated to an observed putative signal, is expressed as $e^{-z}P\left(\sqrt{z}\right)$, where z is the observed periodogram maximum value, and P is an algebraic polynomial with coefficients depending on the particular task. We also cosider the task of detecting the multicomponent signals, containing several frequencies. In this case we must apply so-called multifrequency periodograms to achieve the self-consistency of the analysis. These periodograms represent the signal as a sum of several periodic terms. We show that a complete analysis should deal with 2^n-1 such periodograms, where n is the cumulative number of possible frequencies. Finally, the paper describes software packages that we developed to facilitate the handling of practical frequency-analysis tasks with the use this theory.

Key words: *periodic components: detection*

ЛИТЕРАТУРА

1. *N.R.Lomb*, Astrophys. J. Suppl. Ser., **39**, 447, 1976.
2. *J.D.Scargle*, Astrophys. J., **263**, 835, 1982.
3. *A.Schuster*, Terrestial Magnetism and Atmospheric Electricity, **3**, 13, 1898.
4. *В.В.Витязев*, Анализ неравномерных временных рядов, Изд. СПбГУ, СПб., 2001.
5. *R.Vio*, *M.Diaz-Trigo*, *P.Andreani*, Astronomy & Computing, **1**, 5, 2013.
6. *J.H.Horne*, *S.L.Baliunas*, Astrophys. J., **302**, 757, 1986.
7. *C.Koen*, Astrophys. J., **348**, 700, 1990.
8. *A.Schwarzenberg-Czerny*, Astrophys. J., **460**, L107, 1996.
9. *A.Schwarzenberg-Czerny*, Astrophys. J., **489**, 941, 1997.
10. *A.Schwarzenberg-Czerny*, Mon. Notic. Roy. Astron. Soc., **301**, 831, 1998.
11. *A.Schwarzenberg-Czerny*, Baltic Astron., **7**, 43, 1998.
12. *A.Cumming*, *G.W.Marcy*, *R.P.Butler*, Astrophys. J., **526**, 890, 1999.
13. *A.Cumming*, Mon. Notic. Roy. Astron. Soc., **354**, 1165, 2004.
14. *F.A.M.Frescura*, *C.A.Engelbrecht*, *B.S.Frank*, Mon. Notic. Roy. Astron. Soc., **388**, 1693, 2008.
15. *M.J.Graham*, *A.J.Drake*, *S.G.Djorgovski et al.*, Mon. Notic. Roy. Astron. Soc., **434**, 3423, 2013.
16. *В.В.Витязев*, Спектрально-корреляционный анализ равномерных временных рядов, Изд. СПбГУ, СПб., 2001.
17. *R.V.Baluev*, Mon. Notic. Roy. Astron. Soc., **393**, 969, 2009.
18. *R.V.Baluev*, Mon. Notic. Roy. Astron. Soc., **385**, 1279, 2008.
19. *R.V.Baluev*, Celest. Mech. Dyn. Astron., **111**, 235, 2011.
20. *R.V.Baluev*, Mon. Notic. Roy. Astron. Soc., **429**, 2052, 2013.
21. *G.Ranucci*, *M.Rovere*, Phys. Rev. D., **75**, 013010, 2007.
22. *S.O.Rice*, Bell System Tech. J., **23**, 282, 1944.

23. *M.F.Kratz*, Probability Surveys, **3**, 230, 2006.
24. *J.-M.Azaïs*, *M.Wschebor*, Level Sets and Extrema of Random Processes and Fields, Wiley, 2009.
25. *R.V.Baluev*, Mon. Notic. Roy. Astron. Soc., **395**, 1541, 2009.
26. *R.V.Baluev*, Mon. Notic. Roy. Astron. Soc., **431**, 1167, 2013.
27. *R.B.Davies*, Biometrika, **64**, 247, 1977.
28. *R.B.Davies*, Biometrika, **74**, 33, 1987.
29. *R.B.Davies*, Biometrika, **89**, 484, 2002.
30. *S.S.Vogt*, *R.P.Butler*, *N.Haghighipour*, Astron. Nachr., **333**, 561, 2012.
31. *D.M.Palmer*, Astrophys. J., **695**, 496, 2009.
32. *Р.В.Балуев*, Вестн. СПбГУ (Сер.1), **2**, 129, 2009.
33. *S.Ferraz-Mello*, Astron. J., **86**, 619, 1981.
34. *M.Zechmeister*, *M.Kürster*, Astron. Astrophys., **496**, 577, 2009.
35. *G.Anglada-Escudé*, *M.Tuomi*, Astron. Astrophys., **548**, A58, 2012.
36. *V.V.Vityazev*, Astron. Astrophys. Trans., **11**, 159, 1996.
37. *R.V.Baluev*, Celest. Mech. Dyn. Astron., **102**, 297, 2008.
38. *D.Dacunha-Castelle*, *E.Gassiat*, Ann. Stat., **27**, 1178, 1999.
39. *J.-M.Azaïs*, *C.Delmas*, Extremes, **5**, 181, 2002.
40. *S.J.O'Toole*, *H.R.A.Jones*, *C.G.Tinney et al.*, Astrophys. J., **701**, 1732, 2009.
41. *R.V.Baluev*, Mon. Notic. Roy. Astron. Soc., **436**, 807, 2013.
42. *G.Foster*, Astron. J., **109**, 1889, 1995.
43. *G.Foster*, Astron. J., **111**, 541, 1996.
44. *G.Foster*, Astron. J., **111**, 555, 1996.
45. *R.V.Baluev*, Astronomy & Computing, **3-4**, 50, 2013.
46. *Q.H.Vuong*, Econometrica, **57**, 307, 1989.
47. *R.V.Baluev*, Mon. Notic. Roy. Astron. Soc., **422**, 2372, 2012.
48. *R.V.Baluev*, Astronomy & Computing, **2**, 18, 2013.