Разделение ошибок в аргументе и значении функции с применением к данным экспрессии генов

Жорникова П.Г., студент кафедры статистического моделирования СПбГУ, polina.zhornikova@mail.ru[1]

Голяндина Н.Э., к.ф.-м.н., доцент кафедры статистического моделирования СПбГУ, n.golyandina@spbu.ru[2]

Аннотация

Реализован алгоритм работы с данными при наличии ошибок в аргументе и при измерениях функции. Рассмотрено два варианта поведения ошибки измерениях функции, аддитивный и мультипликативный. Предложен метод для различения этих двух моделей. Метод проверен на модельных данных, имитирующих поведение реальных данных экспрессии гена, и применен к реальным данным.

Введение

Пусть имеются данные (сигнал), измеренные вдоль пространственной оси X. Пусть в данных содержится ошибка в аргументе измеряемой функции, и есть ошибки в наблюдениях (значениях измеряемой функции). Задача состоит в оценивании дисперсии шума в наблюдениях и устранение шума в аргументе функции. Т.е. предполагается, что шум в наблюдениях имеет, например, биологический смысл, а ошибки в аргументах вызваны процедурой измерения и неинтересны. Примером описанных данных служат данные экспрессии генов, описанные в [1], которые будут рассматриваться в качестве главного примера в данной статье. Заметим, что поставленная задача не является стандартной регрессионной задачей с ошибками в измерениях, когда требуется оценить параметры сигнала. В данном случае сигнал неизвестен, но представляет интерес не он (по крайней мере, не только он), а именно шум.

Рассмотрим две модели наблюдений — аддитивную и мультипликативную. В первом случае ошибки в наблюдениях являются гомоскедастичными,

[1]Работа выполнена при поддержке гранта РФФИ 16-04-00821

[2]Работа выполнена при поддержке гранта РФФИ 16-04-00821

во втором случае гетероскедастичными, т.е. модели наблюдений имеют вид

$$f_i = u(x_i + \varepsilon_i) + \delta_i, \tag{1}$$

$$f_i = u(x_i + \varepsilon_i)(1 + \delta_i), \tag{2}$$

где $i = 0, \ldots, N-1$, ε_i — случайная величина с нулевым математическим ожиданием и дисперсией σ_x^2, δ_i — случайная величина с нулевым математическим ожиданием и дисперсией σ_y^2, ε_i и δ_i независимы.

Будем называть модель (1) аддитивной, а модель (2) мультипликативной. При рассмотрении реальных данных активности генов неизвестно, какая из моделей верна.

Обе модели, используя разложение первого порядка функции $u(x+y) \approx u(x) + u'(x)y$, можем записать в общем виде

$$g_i = u(x_i) + e_i. \tag{3}$$

Будем называть шумом ряд $(e_1, \ldots, e_{N-1})$. Шум имеет разную структуру для разных моделей. Ряд $(u_0, \ldots, u_{N-1})$, где $u_i = u(x_i)$, будем называть трендом.

Таким образом, задача — определить, какая из моделей лучше описывает данные, оценить параметры σ_x^2 и σ_y^2 и дисперсию шума $(e_1, \ldots, e_{N-1})$. В статье будут рассмотрены алгоритмы, решающие эти задачи.

Задача оценивания дисперсии шума с одним параметром рассматривалась и ранее, например, в [2]. В статье [2] приведены ссылки и на другие методы. Рассматриваемые нами модели не удается свести к модели с одним параметром.

Используемые алгоритмы

Рассмотрим первые линейные приближения к моделям (1) и (2) в форме (3):

$$g_i = u(x_i) + \varepsilon_i u'(x_i) + \delta_i, \tag{4}$$

$$g_i = u(x_i) + \delta_i u(x_i) + \varepsilon_i(1 + \delta_i)u'(x_i). \tag{5}$$

Заметим, что члены меньшего порядка можно откинуть только в том случае, когда значение $\sup_{x \in \mathbb{R}} |u''(x)|$ мало в сравнении с g_i. Очевидно, что для обеих моделей верно равенство $\mathbb{E}\, g_i = u(x_i)$. Будем оценивать дисперсии σ_x^2 и σ_y^2 ошибок δ_i и ε_i для моделей (4) и (5).

Для решения рассматриваемой задачи оценки дисперсии были использованы следующие алгоритмы.

1. Алгоритм оценки дисперсии шума $D(u,x) = (\mathbb{D}(e_0), \ldots, \mathbb{D}(e_{N-1}))$ и параметров σ_x^2 и σ_y^2 для аддитивной модели.

   (a) Оцениваем $u(x)$ и $u'(x)$, получаем $\hat{u}(x)$ и $\hat{u}'(x)$. Обозначим за $r_i = f_i - \hat{u}(x_i)$. Рассматриваем ряды $T = (\hat{u}(x_0), \ldots, \hat{u}(x_{N-1}), D = (\hat{u}'(x_0), \ldots, \hat{u}'(x_{N-1})$ и $R = (r_0, \ldots, r_{N-1})$.

   (b) Строим линейную регрессию ряда R^2 на ряд D^2 с неизвестным свободным членом. Получаем оценки $\hat{t}$ (коэффициент перед свободным членом) и $\hat{d}$ (коэффициент перед D^2).

   (c) Получаем искомые оценки для σ_y^2 и σ_x^2:

   $$\hat{\sigma}_y^2 = \hat{t}, \quad \hat{\sigma}_x^2 = \hat{d},$$

   и оценку для дисперсии шума

   $$\hat{D}(u, x_i) = \hat{\sigma}_y^2 + \hat{\sigma}_x^2 \, \hat{d}_i,$$

2. Алгоритм оценки дисперсии шума $D(u,x) = (\mathbb{D}(e_0), \ldots, \mathbb{D}(e_{N-1}))$ и параметров σ_x^2 и σ_y^2 для мультипликативной модели.

   (a) Аналогично, как для аддитивной модели, получаем ряды T, D и R.

   (b) Строим линейную регрессию ряда R^2 на ряды T^2 и D^2 с нулевым свободным членом. Получаем оценки $\hat{t}$ (коэффициент перед T^2) и $\hat{d}$ (коэффициент перед D^2).

   (c) Получаем искомые оценки для σ_y^2 и σ_x^2:

   $$\hat{\sigma}_y^2 = \hat{t}, \quad \hat{\sigma}_x^2 = \frac{\hat{d}}{(1+\hat{t})};$$

   и оценку для дисперсии шума

   $$\hat{D}(u, x_i) = \hat{\sigma}_y^2 \, \hat{u}_i + \hat{\sigma}_x^2 \, (1 + \hat{\sigma}_y^2) \, \hat{d}_i.$$

3. В регрессионных задачах пунктов 1b и 2b дисперсия шума зависит от оцениваемых параметров, поэтому используем итеративный алгоритм оценки (IRLS), алгоритм и его история описаны в [3]. IRLS зависит от двух параметров — невязки τ и максимального числа шагов M, и выглядит следующим образом.

   - Шаг 1. Выбор начального значения.
     Выберем некоторое начальное значение B_0, $i \leftarrow 0$.

- Шаг 2. Вычисление приближенных значений матрицы $\mathbf{W}$ $\mathbf{W}_{(i)} = \mathbf{W}(\widehat{B}_{(i)})$.
- Шаг 3. Нахождение оценок.

$$\widehat{B}_{(i+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_{(\mathbf{i})}{}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_{(\mathbf{i})}{}^{-1}Y.$$

- Шаг 4. Критерий остановки.
  Если $||\widehat{B}_{(i)} - \widehat{B}_{(i+1)}|| < \tau$ или $i+1 = M$, процедура заканчивается и результатом ее работы является $B_{(i+1)}$.
  Иначе $i \leftarrow i + 1$, переход к шагу 2.

4. Для оценивания сигнала $(u(x_0), \ldots, u(x_{N-1}))$ будем использовать метод анализа сингулярного спектра (Singular Spectrum Analysis), подробно описанный в [4]. Важное преимущество метода состоит в том, что он непараметрический и не требует дополнительных знаний о ряде. Метод имеет один параметр L, называемый длиной окна. На одном из этапов метода нужно сгруппировать компоненты разложения ряда в две группы, относящиеся к тренду и шуму. Будем использовать автоматический метод идентификации ([5, раздел 2.4.5],[6]), который имеет несколько параметров: частотный интервал $[\omega_1, \omega_2]$ и порог t. Недостатком метода SSA является то, что он игнорирует значения x_i и рассматривает измерения как равноотстоящие, что, вообще говоря, может быть не так. В частности, поэтому необходимо дополнительное сглаживание. Также, сглаживание должно улучшить точность последующего оценивания производной.

   Будем сглаживать оцененный тренд с помощью алгоритма локальной регрессии (LOESS, [7]) с параметрами p — степень полиномов, которыми аппроксимируем, и α — параметр сглаживания. Заметим, что метод LOESS умеет работать с неравноотстоящими данными.

5. Для оценки производной $w(x) = u'(x)$ в точке x_i будем использовать следующую формулу, считая, что значения x_i упорядочены:

$$\hat{w}(x_i) = \hat{w}_i = \frac{\hat{g}_{i+1} - \hat{g}_{i-1}}{x_{i+1} - x_{i-1}}, i = 1, \ldots, N-2.$$

   Далее сглаживаем производную с помощью алгоритма LOESS с параметрами d в качестве порядка аппроксимирующего полинома и s в качестве меры гладкости.

6. Теперь рассмотрим алгоритм выбора между двумя рассматриваемыми моделями наблюдений, аддитивной и мультипликативной. Сложность

заключается в том, что по средней величине остатков регрессии модели неразличимы. Поэтому алгоритм основан на величине смещения.

Рассмотрим ряд $(n_0, \ldots, n_{N-1})$, где

$$n_i = e_i - \hat{D}(u, x_i),$$

т.е из значения оцененного шума вычли полученную оценку дисперсии (оценку тренда шума). Получившийся ряд будем называть рядом из остатков шума.

К полученному ряду $(n_o, \ldots, n_{N-1})$ из остатков применим скользящую медиану с окном 3, чтобы избавиться от выделяющихся наблюдений. Получим ряд $(\bar{n}_o, \ldots, \bar{n}_{N-3})$. Для получившего ряда построим доверительные интервалы с уровнем доверия γ для скользящих средних $(m_0, \ldots, m_{N-K-2})$ по отрезкам длины K, и будем брать границу доверительного интервала, ближайшую к нулю, если интервал не захватывает ноль, и ноль, если захватывает. Таким образом, получим ряд $(c_0, \ldots, c_{N-K-2})$ длины $N-K-1$, характеризующий смещение, с элементами вида

$$c_i = \begin{cases} \min\left(|m_i \pm z_\gamma\, \hat{s}_i/\sqrt{n}|\right), \text{ если } 0 \notin \left(m_i - z_\gamma\, \hat{s}_i/\sqrt{n},\, m_i + z_\gamma\, \hat{s}_i/\sqrt{n}\right); \\ 0, \text{ иначе,} \end{cases}$$

где z_γ — $(1+\gamma)/2$-квантиль стандартного нормального распределения, $\hat{s}_i$ — выборочная дисперсия выборки $(\bar{n}_i, \ldots, \bar{n}_{i+K-1})$, $i = 0, \ldots, N-K-2$.

В качестве меры, показывающей разницу между моделями, рассмотрим медиану вектора $(c_0, \ldots, c_{N-K-2})$

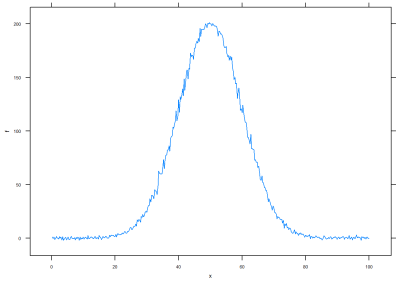
$$\operatorname{med}(c_0, \ldots, c_{N-K-2}). \tag{6}$$

При решении нужно обратить внимание на следующие аспекты.

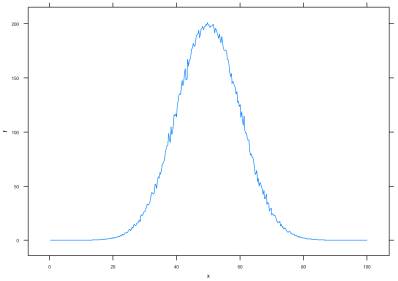
1. Так как рассматриваем аппроксимации (4) и (5), то значений второй производной $|u''(x)|$ должно быть малым в сравнении с аппроксимациями.

2. Нужно уметь хорошо оценивать тренд $u(x)$ и производную $u'(x)$.

3. Чтобы в регрессионной постановке задачи были независимые переменные, в случае аддитивной модели (4) производная тренда должна отличаться от константы, т.е. линейный тренд не подходит. В случае мультипликативной модели (5) производная тренда должна отличаться от самого тренда, т.е., например, экспоненциальный тренд не подходит.



(a) Сигнал $f(x)$ в случае, когда верна аддитивная модель.



(b) Сигнал $f(x)$ в случае, когда верна мультипликативная модель.

Рис. 1: Вид сигнала для модельных данных.

Исследование алгоритмов на модельных данных

Исследуем свойства приведенных алгоритмах на модельных данных, которые повторяют поведение реальных данных активности генов. Возьмем $N = 400$, $u(x) = 200\, e^{-(x-50)^2/200}$, $x = (0.25, 0.5, \ldots, 0.25 \times N)$, $\sigma_x^2 = 0.1$. Для аддитивной модели возьмем $\sigma_y^2 = 1$, для мультипликативной $\sigma_y^2 = 0.0001$. Вид данных для аддитивной и мультипликативной моделей изображен на рис. 1.

В ходе исследования было выяснено, что оценки параметров σ_x^2 и σ_y^2 получаются смещенными, и что основной вклад в смещение оценок дают оценки тренда и производной. Промоделируем модельные данные. Для IRLS алгоритма брались $\tau = 10^{-8}$, $M = 200$, моделирование проводилось $n = 1000$ раз. Когда в алгоритмы оценивания подставлялись истинные тренд производнпя, в обеих моделях истинные значения параметров вошли в 95% доверительные интервалы, хотя небольшое смещение, возможно, есть.

В случае оценки тренда используем следующие параметры: для SSA $L = 30$, $\omega_1 = 0$, $\omega_2 = 0.04$, $t = 0.4$, для LOESS используем $\alpha = 0.2$, $p = 2$ для сглаживания тренда и такие же параметры для сглаживания производной. Результаты показывают, что при оценивании дисперсий шумов возникает смещение (см. табл. 1 для мультипликативной модели? где, в частности, приведены вероятностные уровни проверки гипотез, что значения дисперсий равны тем, которые использовались при моделировании).

Несмотря на смещение в оценках, сравнение моделей по мере (6) неправильно идентифицировало мультипликативную модель всего в 3 случая из 1000. Аддитивная модель была принята за мультипликативную примерно в 50 случаях из 1000.

Таблица 1: Оценки параметров σ_x^2 и σ_y^2 для мультипликативной модели.

	σ_y^2	σ_x^2
Истинное значение	1.00e-04	0.100
Среднее значение оценки	0.96e-04	0.092
p-value для t-test'a	0.0002	0

Применение алгоритма к реальным данным

Применим теперь полученные алгоритмы к реальным данным активности генов, пример профиля которых изображен на рис. 2. Рассматриваем набор эмбрионов одно гена Kruppel.

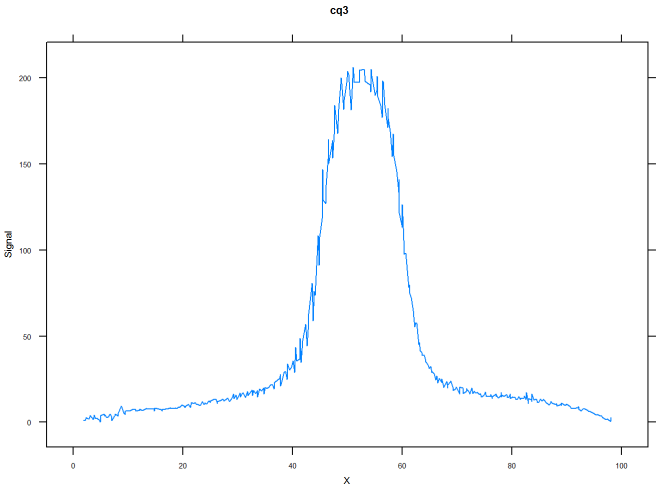


Рис. 2: Профиль данных активности гена Kruppel возраста 4, эмбрион cq3.

Значения всех параметров алгоритмов возьмем такими же, как для модельных данных. Оценим значения параметров σ_x^2 и σ_y^2 для двух моделей и выберем модель с помощью меры (6). Для эмбрионов возраста 4, для 34 эмбрионов из 36 значение меры для мультипликативной модели получилось меньше, чем для аддитивной. Для эмбрионов возраста 8, все 38 эмбрионов были определены алгоритмом как более похожие на мультипликативную модель. Поэтому можно сделать вывод о мультипликативности шума экспрессии гена для рассматриваемых данных.

В мультипликативной модели средние значения оценок параметров σ_x^2 —

0.1, а σ_y^2 — 0.001 для эмбрионов возраста 4. Для возраста 8 соответствующие значения 0.155 и 0.0017.

Заключение

Таким образом, алгоритм оценивания параметров двух моделей был исследован на модельных данных, похожих на реальные данные экспрессии генов для гена Kruppel. Исследование показало, что алгоритм может различать рассматриваемые аддитивную и мультипликативную модель, и поэтому можно сделать вывод, что шум экспрессии генов для анализируемых данных имеет мультипликативную природу.

Список литературы

[1] Measuring Gene Expression Noise in Early Drosophila Embryos: Nucleus-to-nucleus Variability / NE Golyandina, DM Holloway, FJP Lopes et al. // Procedia Computer Science. — 2012. — P. 373–382.

[2] Brown L., Levine M. Variance estimation in nonparametric regression via the difference sequence method // The Annals of Statistics. — 2007. — Vol. 35. — P. 2219–2232.

[3] Daubechies I., Devore R. et al. Iteratively Reweighted Least Squares Minimization for Sparse Recovery // Communications on Pure and Applied Mathematics. — 2010. — Vol. 63. — P. 1–38.

[4] Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques. — Chapman&Hall/CRC, 2001.

[5] Golyandina N., Zhigljavsky A. Singular Spectrum Analysis for time series. Springer Briefs in Statistics. — Springer, 2013.

[6] Alexandrov Th. A method of trend extraction using Singular Spectrum Analysis // RevStat. — 2009. — Vol. 7, no. 1. — P. 1–22.

[7] Cleveland W.S., Grosse E., Shyu W.M. Local regression models // Statistical models in S / Ed. by J.M. Chambers, T.J. Hastie. — Chapman and Hall, London, 1993.