О ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ МЕТОДОВ МОНТЕ-КАРЛО В АНАЛИЗЕ НЕЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Г. И. Рудой

Введение

Символьная регрессия часто используется для построения экспертно интерпретируемых моделей [1–5]. В приложении к естественнонаучным экспериментам речь идет о восстановлении функциональной зависимости между измеряемыми и задаваемыми с некоторой точностью параметрами, как то: зависимость термоэмиссионного тока электронной лампы от температуры катода $I_k(T)$ при неизменных геометрии системы и разности потенциалов, зависимость мощности излучения непрерывного лазера от коэффициента отражения выходного зеркала $W_l(R)$ при постоянных модовой структуре излучения и мощности возбуждения активной среды, зависимость показателя преломления материала от длины волны $n(\lambda)$ при постоянной температуре и т. п., далее мы более подробно рассмотрим именно последний случай.

При регрессионном анализе такого рода экспериментов необходимо учитывать следующие обстоятельства:

1. Все измеряемые (и контролируемые) параметры в каждой экспериментальной точке определяются с некоторой (обычно известной) точностью, причем абсолютная погрешность σ_i соответствующего параметра может существенно изменяться в исследуемом диапазоне. Например, если в качестве спектрального прибора, выделяющего конкретную длину волны λ_i при измерении $n_i(\lambda_i)$, используется дифракционная решетка, то $\frac{\sigma_i}{\lambda_i} \approx \text{const}$, и считать погрешность определения длины волны постоянной некорректно для измерений в достаточно широком спектральном дипазоне.

2. Как правило, эксперимент ставится так, что измеряется функциональная зависимость от одной переменной , то есть, строится зависимость вида $y(x, \boldsymbol{\omega})$, где $\boldsymbol{\omega}$ — набор параметров, которые поддерживаются неизменными. Как отмечалось выше, параметры поддерживаются постоянными с конечной точностью и в ряде случаев при построении модели это обстоятельство необходимо учитывать. Однако обычно эксперт заранее может оценить влияние вариаций условий эксперимента и обеспечить необходимую стабильность проведения измерений. В противном случае необходимо прямо учитывать зависимость измеряемой характеристики от нескольких переменных, что для целей настоящей работы непринципиально.

3. В большинстве случаев эксперт заранее знает вид искомой функциональной зависимости, или же требуется провести выбор между несколькими возможными вариантами, что упрощает задачу регрессии. В то же время для эксперта важнейшее значение имеет не только определение оптимальных численных коэффициентов регрессионной формулы путем минимизации некоторого функционала, но и дисперсия указанных коэффициентов и, что предпочтительнее, связь дисперсии регрессионных коэффициентов с точностью определения измеряемых (контролируемых) в

эксперименте величин. Это особенно существенно в тех случаях, когда коэффициенты регрессионной модели прямо связаны с фундаментальными характеристиками исследуемого процесса и по ним рассчитывается, например эффективная масса электронов в полупроводнике, температура Дебая, резонансная частота и затухание оптического перехода и т. д. — соответственно, точность измерения соответствующих материальных констант определяется точностью вычисления коэффициентов регрессионной модели.

В такой постановке, когда требуется определить не только оптимальные коэффициенты регрессионной модели, но и их погрешность, насколько нам известно, задача нелинейной регрессии не рассматривалась. Известны теоретические результаты для случая линейной регрессии:

$$y = ax + b,$$

в случае, когда дисперсия всех экспериментально измеренных значений y_i зависимой переменной y одна и та же $D(y_i) = \sigma^2$, а значения независимой переменной x_i известны точно: $D(x) = 0$. Тогда при переходе к представлению

$$y_i = a(x_i - \overline{x}) + b + \xi_i \mid i \in \{1, \ldots, n\},$$

где $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, согласно [6], случайные величины a и b независимы и нормально распределены, и, кроме того, их дисперсии выражаются известными соотношениями:

$$D(a) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{1}$$

$$D(b) = \frac{\sigma^2}{n}. \tag{2}$$

В настоящей работе предложен общий метод определения погрешности коэффициентов нелинейной регрессии, и на примере зависимости $n(\lambda)$ для прозрачного полимера определена зависимость погрешности параметров регрессии от точности определения длины волны и показателя преломления прозрачного полимера. Здесь мы ограничиваемся одной независимой переменной λ. Обобщение предлагаемого метода на случай нескольких переменных проводится очевидным образом.

1 Основная гипотеза

Пусть имеется обучающая выборка $(x_i, y_i) \mid i = \{1, \ldots, n\}$, причем для каждого значения x_i, y_i известно распределение вероятности отклонения независимой и зависимой переменных от их среднего значения $P_i^x(x - x_i)$ и $P_i^y(y - y_i)$ соответственно, которые обычно принимаются гауссовыми и для которых считаются известными значения дисперсий σ_i^x, σ_i^y.

Пусть далее с помощью некоторого алгоритма регрессии строится зависимость $y(x, \boldsymbol{\omega})$, минимизирующая некоторый функционал S, например среднеквадратичное отклонение:

$$S = \sum_{i=1}^{n} \big(y(x_i, \boldsymbol{\omega}) - y_i\big)^2 \underset{\boldsymbol{\omega}}{\to} \min.$$

Для таким образом определенного функционала, а также для его модификаций, учитывающих сложность регрессионной модели [5], процедура минимизации эффективно проводится спомощью алгоритма Левенберга-Марквардта (АЛМ) [7, 8].

Далее фиксируем структурный вид полученной зависимости $y(x, \boldsymbol{\omega})$ и многократно повторяем следующую вычислительную процедуру:

1. На k-м шаге генерируется случайная выборка (x_i^k, y_i^k), при этом вероятность появления в выборке значения x_i^k пропорциональна $P_i^x(x_i^k - x_i)$, а вероятность появления y_i^k аналогично пропорциональна $P_i^y(y_i^k - y_i)$.

2. Для таким образом построенной реализации набора экспериментальных данных (x_i^k, y_i^k) (далее — реализация), используя один и тот же алгоритм оптимизации, находим оптимальный (минимизирующий выбранный функционал) набор $\boldsymbol{\omega}_k$ коэффициентов регрессии $y(x, \boldsymbol{\omega})$ для k-й реализации.

   Таким образом, для каждого конкретного коэффициента регрессии $\boldsymbol{\omega}_p$ получаем совокупность его значений в сгенерированных реализациях $\{\boldsymbol{\omega}_p^k\}$.

3. Для достаточно большого числа реализаций M стандартно определим среднее значение и стандартное отклонение соответствующего коэффициента регрессии $\boldsymbol{\omega}_p$:

$$\overline{\boldsymbol{\omega}_p} = \sum_{i=1}^{M} \boldsymbol{\omega}_p^k, \tag{3}$$

$$D(\boldsymbol{\omega}_p) = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (\boldsymbol{\omega}_p^k - \overline{\boldsymbol{\omega}_p})}. \tag{4}$$

Наша гипотеза состоит в том, что полученные согласно (3) и (4) значения соответствуют реальности. Предлагаемый подход к определению погрешности регрессионных коэффициентов, очевидно, представляет собой фактически применение метода типа Монте-Карло к задаче регрессии.

Из предложенной интерпретации также следует очевидный критерий останова вычислительной процедуры, когда с ростом числа реализаций вариация значений $\overline{\boldsymbol{\omega}_p}$ и $D(\boldsymbol{\omega}_p)$ становится меньше экспертно выбранного значения.

Необходимо заметить, что в общем случае пределы выражений типа (3) и (4) при $M \to \infty$ могут и не существовать, что делает предложенную вычислительную схему некорректной. Однако для достаточно гладких функций, которые собственно и представляют практический интерес, корректность предложенной процедуры можно строго доказать, что, однако, выходит за рамки данной работы.

2 Модельный случай

Определим предложенным в предыдущей части методом дисперсию коэффициентов линейной регрессии $y = kx + b \mid k = 3, b = 10$, независимая переменная x определена точно, а погрешность зависимой переменной y постоянна и имеет гауссово распределение. Независимая переменная определена в 10 (30, 50) точках отрезка $[0, 10]$, количество реализаций составляло 10 миллионов, оптимизация регрессии проводилось с помощью АЛМ.

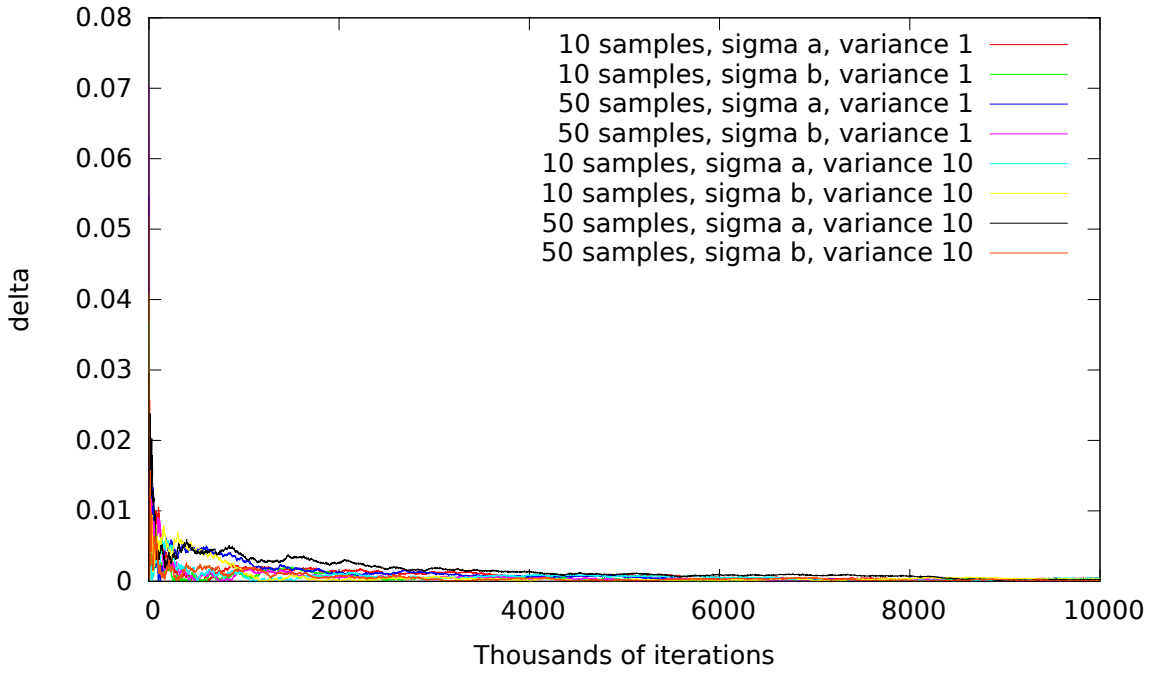

Рис. 1: График зависимости δ от числа итераций (от 0 до 10^7 итераций).

На рис. 1-3 по оси абсцисс указано количество реализаций M, а по оси ординат для каждого из коэффициентов k, b отношение $\frac{\sigma_{c.e.}-\sigma_t}{\sigma_t}$, где $\sigma_{c.e.}$ — значение дисперсии, полученное в вычислительном эксперименте, а σ_t — точное теоретическое значение дисперсии согласно (1) (2).

Как видно, для $M \approx 107$ относительное различие $\frac{\sigma_{c.e.}-\sigma_t}{\sigma_t}$ не превышает 0.05%, что,на наш взгляд, является хорошим результатом и свидетельствует в пользу корректности обсуждаемого подхода и основной гипотезы.

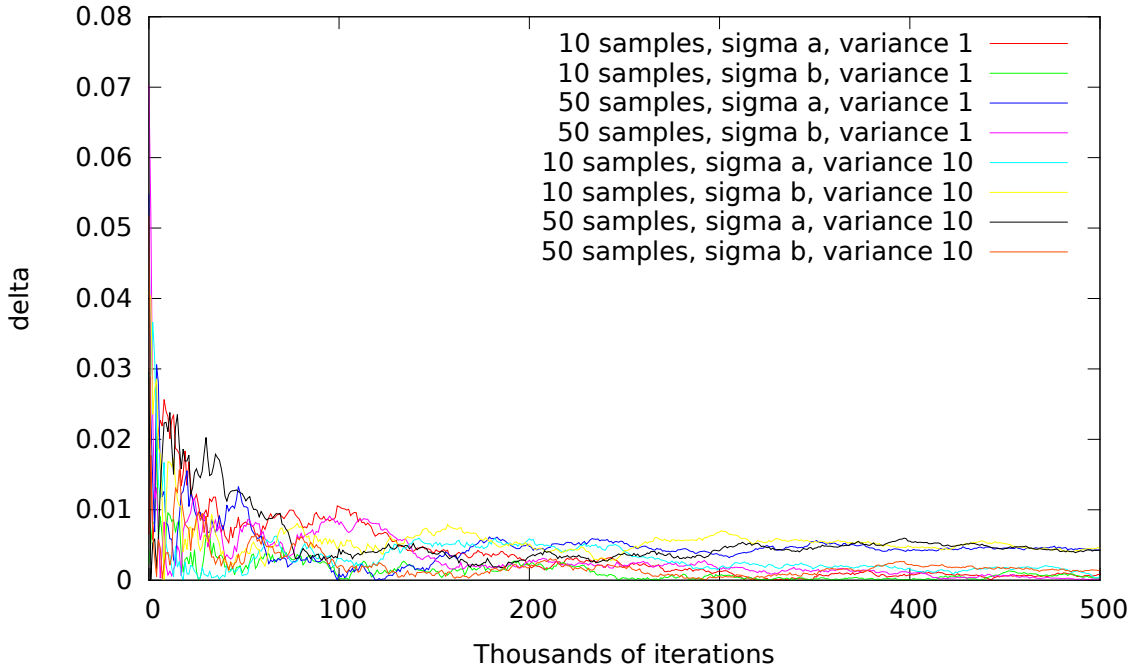

Рис. 2: График зависимости δ от числа итераций (от 0 до $5 \cdot 10^5$ итераций).

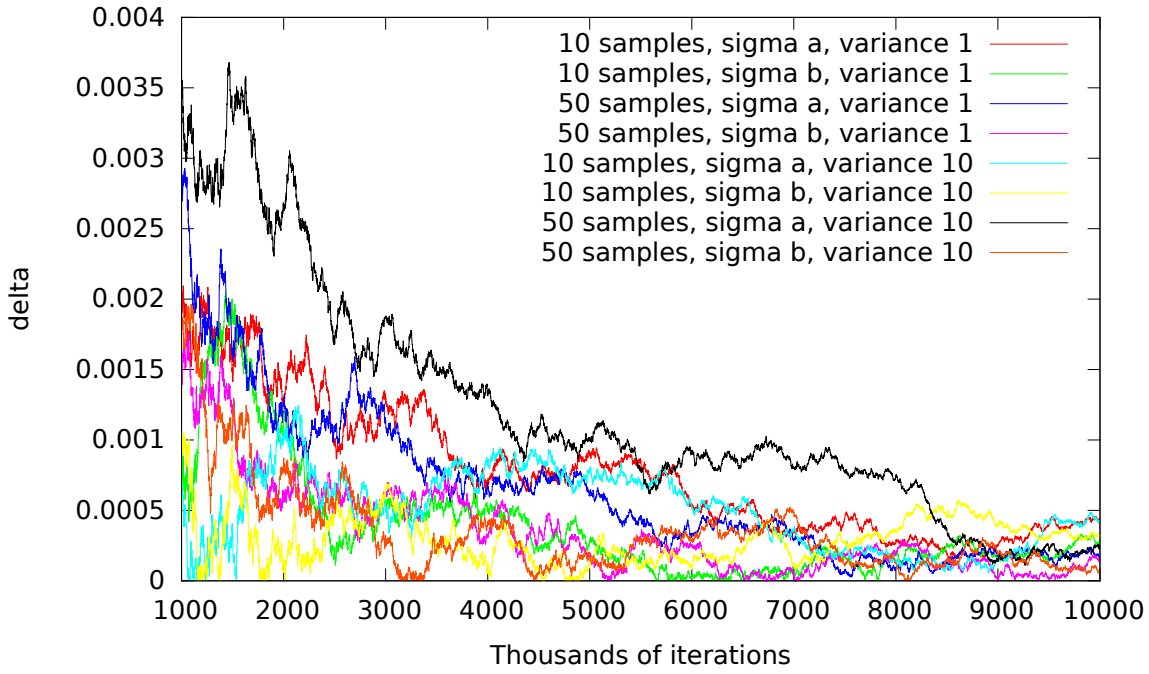

Рис. 3: График зависимости δ от числа итераций (от 10^6 до 10^7 итераций).

Список литературы

[1] Davidson, J. W., Savic, D. A., and Walters, G. A.: *Symbolic and numerical regression: experiments and applications.* In John, Robert and Birkenhead, Ralph (editors): *Developments in Soft Computing*, pages 175–182, De Montfort University, Leicester, UK, 29-30 6 2000. 2001. Physica Verlag, ISBN 3-7908-1361-3.

[2] Sammut, C. and Webb, G. I.: *Symbolic regression.* In Sammut, Claude and Webb, Geoffrey I. (editors): *Encyclopedia of Machine Learning*, page 954. Springer, 2010, ISBN 978-0-387-30768-8. `http://dx.doi.org/10.1007/978-0-387-30164-8`.

[3] Strijov, V. and Weber, G. W.: *Nonlinear regression model generation using hyperparameter optimization.* Computers & Mathematics with Applications, 60(4):981–988, 2010. `http://dx.doi.org/10.1016/j.camwa.2010.03.021`.

[4] Стрижов, В. В.: *Методы индуктивного порождения регрессионных моделей.* Препринт ВЦ РАН им. А. А. Дородницына. — М., 2008.

[5] Рудой, Г. И. и Стрижов, В. В.: *Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных.* Информатика и ее применения, 7(1):44–53, 2013.

[6] Ватутин, В. А., Ивченко, Г. И., Медведев, Ю. И., и Чистяков, В. П.: *Теория вероятностей и математическая статистика в задачах.* Дрофа, 3 редакция, 2005.

[7] Marquardt, D. W.: *An algorithm for least-squares estimation of non-linear parameters.* Journal of the Society of Industrial and Applied Mathematics, 11(2):431–441, 1963.

[8] More, J. J.: *The Levenberg-Marquardt algorithm: Implementation and theory.* In *G.A. Watson*, Lecture Notes in Mathematics 630, pages 105–116. Springer-Verlag, Berlin, 1978. Cited in Åke Björck's bibliography on least squares, which is available by anonymous ftp from `math.liu.se` in `pub/references`.