

# 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  $\sigma_i$  соответствующего параметра может существенно изменяться в исследуемом диапазоне. Например, если в качестве спектрального прибора, выделяющего конкретную длину волны  $\lambda_i$  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 диапазоне.
2. Как правило, эксперимент ставится так, что измеряется функциональная зависимость от одной переменной, то есть, строится зависимость вида  $y(x, \omega)$ , где  $\omega$  — 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 не принципиально.
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

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

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

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

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

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  $\lambda$ . Обобщение предлагаемого метода на случай нескольких переменных проводится очевидным образом.

## 1 Основная гипотеза

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

## 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.