Метод наименьших модулей и аппроксимация термодинамических функций

Восков А.Л.

МГУ имени М.В. Ломоносова, Химический факультет

25 февраля 2020 г.

Метод наименьших квадратов (МНК)

Основан на оптимизации параметров модели минимизацией суммы квадратов отклонений:

$$S(\beta) = \sum_{i=1}^{n} e_i^2(\beta)$$

$$e_i(\beta) = (\hat{y}_i(\beta) - y_i)^2$$

e_i — отклонения, $\hat{y}_i$ и y_i — расчётные и экспериментальные значения, β — параметры модели.

Достоинства:

- Аналитическое решение для линейной регрессии
- Широчайшая распространённость реализаций

Недостатки:

- Чувствительность к выбросам (грубым промахам)

Метод наименьших модулей (МНМ)

Минимизация суммы абсолютных отклонений вместо суммы квадратов отклонений:

$$S(\beta) = \sum_{i=1}^{n} |e_i(\beta)|$$

МНК

$$S(a) = \sum_{i=1}^{n} (a - x_i)^2$$

$$\frac{\partial S}{\partial a} = 2(na + n\bar{x}) = 0 \Rightarrow a = \bar{x}$$

a — среднее арифметическое

МНМ

$$S(a) = \sum_{i=1}^{n} |a - x_i|$$

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} \operatorname{sign}(a - x_i) = 0$$

a — медиана

Метод максимального правдоподобия

Основан на максимизации функции правдоподобия (или её логарифма), представляющей из себя произведение плотностей вероятностей.

МНК (нормальное распределение)

$$F = \prod_{k=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_k^2}{2\sigma^2}\right) \Rightarrow \ln F = \frac{n}{\sigma\sqrt{2\pi}} - \sum_{k=1}^{n} \frac{e_k^2}{2\sigma^2}$$

МНМ (распределение Лапласа)

$$F = \prod_{k=1}^{n} \frac{1}{\sigma\sqrt{2}} \exp\left(-\frac{|e_k|\sqrt{2}}{\sigma}\right) \Rightarrow \ln F = \frac{n}{\sigma\sqrt{2}} - \sum_{k=1}^{n} \frac{|e_k|\sqrt{2}}{\sigma}$$

У распределения Лапласа более тяжёлые «хвосты» по сравнению с распределением Гаусса, что даёт меньшую чувствительность к выбросам.

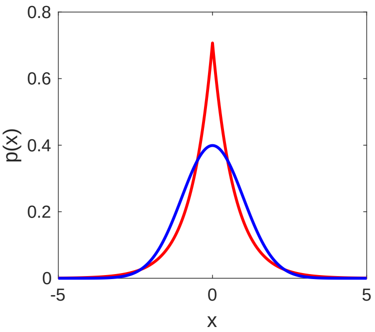
Сравнение распределений Гаусса и Лапласа

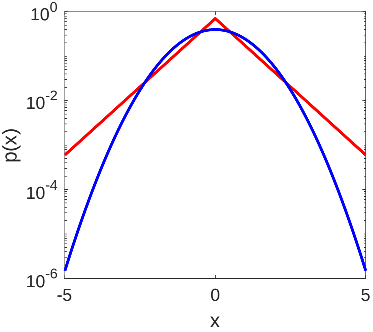
Нормальное распределение

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_k^2}{2\sigma^2}\right)$$

Распределение Лапласа

$$p(x) = \frac{1}{\sigma\sqrt{2}} \exp\left(-\frac{|e_k|\sqrt{2}}{\sigma}\right)$$





Метод наименьших модулей и численные методы

Преимущества

- Наглядность (обобщение медианы)
- Устойчивость к выбросам

Недостатки

- Нет аналитического решения даже для линейной регрессии
- Возможна множественность решений даже для линейной регрессии
- Вычислительная трудоёмкость

Численные методы

- Iterative Reweighted Least Squares (IRLS)
- Симплекс-метод
- Модифицированный метод Гаусса-Ньютона

Iterative Reweighted Least Squares (IRLS)

Постановка задачи

Сводит МНМ к МНК с меняющимися в ходе итераций весами; обычно применяется для линейной регрессии.

$$S = \sum_{k=1}^{n} \omega_k e_k^2; \ \omega_k \approx \frac{1}{|e_k|}$$

Основные шаги

1. $\omega_i = 1$ (т.е. начинаем с невзвешенного МНК)
2. Найти параметры модели, используя взвешенный МНК
3. Пересчитать веса (если $\omega_k \to \infty$ — временно исключаем точку из оптимизации)
4. Если желаемая точность достигнута — выход, иначе перейти к шагу 2

Симплекс-метод: постановка задачи

Минимизацию суммы абсолютных отклонений можно свести к задаче линейного программирования. Минимизируемая функция:

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |u_i - v_i| \leq \sum_{k=1}^{n} u_k + \sum_{k=1}^{n} v_i; \;\; u_i \geq 0; \;\; v_i \geq 0$$

Линейные ограничения ($\beta_i \geq 0$ без потери общности):

$$\begin{cases} x_{11}\beta_1 + \ldots + x_{1m}\beta_m - y_1 - u_1 + v_1 & = 0 \\ x_{21}\beta_1 + \ldots + x_{2m}\beta_m - y_2 - u_2 + v_2 & = 0 \\ \ldots & \ldots \\ x_{n1}\beta_1 + \ldots + x_{nm}\beta_m - y_n - u_n + v_n & = 0 \end{cases}$$

Симплекс-метод: особенности

Особенности применения

- Напрямую применим лишь к линейной регрессии
- Обычно используют специализированные реализации и алгоритмы для МНМ (например, алгоритм Веселовского)

Нелинейная регрессия

Используется линеаризация задачи с итеративным применением симплекс-метода:

$$\sum_{k=1}^{n} |e_k| \approx \sum_{k=1}^{n} \left| e_k\left(\beta^\circ\right) + J_{k*}\left(\beta - \beta^\circ\right) \right|$$

Gonin R., Money A.H. Nonlinear L_p-norm estimation. New York and Basel: Marcel Dekker Inc. 1989. ISBN 978-0-8247-8125-5

Модифицированный метод Гаусса-Ньютона

$$S = \sum_{k=1}^{n} \rho(e_k) \approx \sum_{k=1}^{n} \rho(l_k); \;\; l_k \approx e_k^\circ + J_{k*} p$$

$\rho(t)$ — функция потерь, J — матрица Якоби, $e_k^\circ = e_k(\beta_0)$, $p = (\beta - \beta_0)$; система нормальных уравнений (для поиска минимума) решается методом Ньютона:

$$\frac{\partial S}{\partial \beta_i} = \sum_{k=1}^{n} \dot{\rho}(l_k) J_{ki} = 0 \approx \sum_{k=1}^{n} \left[\dot{\rho}(e_k^\circ) J_{ki} + \sum_{j=1}^{m} \ddot{\rho}(e_k^\circ) J_{ki} J_{kj} p_j \right] = 0$$

В матричном виде (D — диагнональная матрица, $D_{ii} = \ddot{\rho}(e_i^\circ)$):

$$\left(J^\top D J \right) p = -J^\top \dot{\rho}$$

Модифицированный метод Гаусса-Ньютона

Регуляризация (метод Левенберга-Марквардта)

$$\left(J^\top DJ\right)p = -J^\top\dot{\rho} \Rightarrow \left(J^\top DJ + \lambda I\right)p = -J^\top\dot{\rho}$$

$\lambda \geq 0$ — параметр регуляризации, I — единичная матрица

Сглаженные версии функции потерь

Функция $\rho(t) = |t|$ дифференцируема не во всех точках

- Функция Хьюбера

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & \text{если } |a| \leq \delta \\ \delta\left(|t| - \frac{1}{2}\delta\right) & \text{если } |a| > \delta \end{cases}$$

- Модифицированная функция Хьюбера $\rho(t) = \sqrt{t^2+\delta} - \sqrt{\delta}$
- Степенная функция $\rho(t) = |t|^{1+\delta}$

Модифицированный метод Гаусса-Ньютона

Основные шаги

1. Выбрать параметр δ для функции $\rho(t) = \sqrt{t^2 + \delta} - \sqrt{\delta}$ и начальное приближение β°.
2. Решить задачу для заданного δ модифицированным методом Левенберга-Марквардта.
3. Если желаемая точность достигнута — остановиться, если нет — уменьшить δ, т.е. сделать $\rho(t)$ ближе к $|t|$.

Программа lxnormfit

1. Написана на ANSI C и MATLAB (интерфейс)
2. Процедуры для автоматического варьирования δ
3. Расчёт доверительных интервалов

П.2 и 3 нет в MATLAB и Ceres Solver.

Доверительные интервалы

Асимптомтическая формула для доверительных интервалов параметров модели в методе наименьших модулей:

$$\Delta\beta_j = z_{\alpha/2}\lambda\sqrt{C_{jj}}$$

где $C = \left(J^\top J\right)^{-1}$, а λ может быть оценена по следующей асимптотической формуле:

$$\lambda = \frac{1}{2f(m)};\ \left[\hat{f}(m)\right]^{-1} = \frac{\hat{e}_i - \hat{e}_j}{(i-j)/n}$$

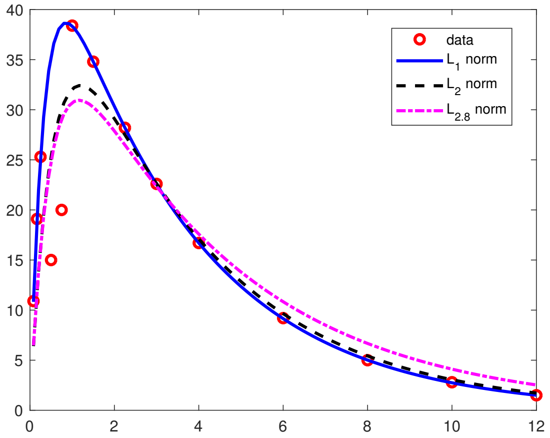
где $f(m)$ — ордината распределения ошибок e в медиане m. i и j выбираются так, чтобы находиться вблизи медианы.

Gonin R., Money A.H. Nonlinear L_p-norm estimation. New York and Basel: Marcel Dekker Inc. 1989. ISBN 978-0-8247-8125-5

Пример: кинетическая кривая

Модель

$$y(x) = \frac{\beta_1\beta_3}{\beta_1 - \beta_2}\left[\exp\left(-\beta_2 t\right) - \exp\left(-\beta_1 t\right)\right]$$



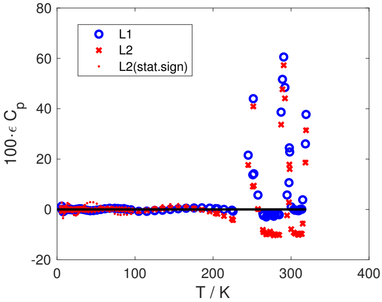
Пример: гексафторид ксенона XeF_6

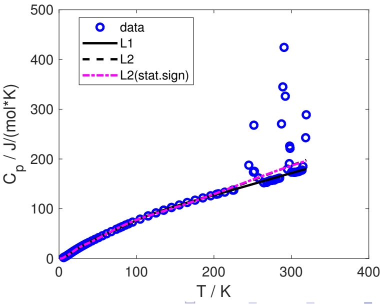
Модель

Взвешенная сумма функций Эйнштейна-Планка

$$C_p(T) = \sum_{i=1}^{m} \alpha_i C_E\left(\frac{\theta_i}{T}\right); \quad \frac{C_E(x)}{R} = \frac{3x^2 e^x}{(e^x - 1)^2}$$

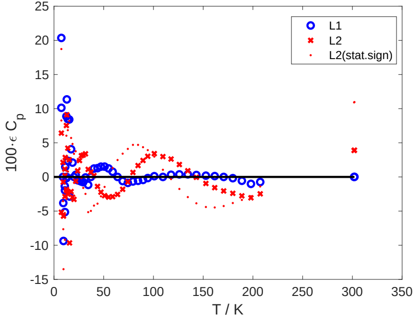
Аппроксимация вместе с аномалией теплоёмкости

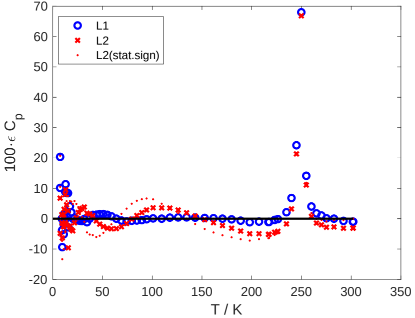




Пример: K-натролит

Два варианта аппроксимации: без аномалии C_p и с ней. При $T < 10$ K у данных большая погрешность, из-за неё наблюдается рост C_p с падением T для двух точек.

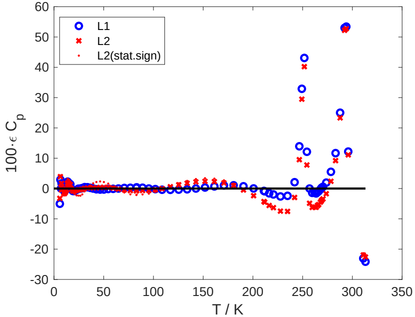


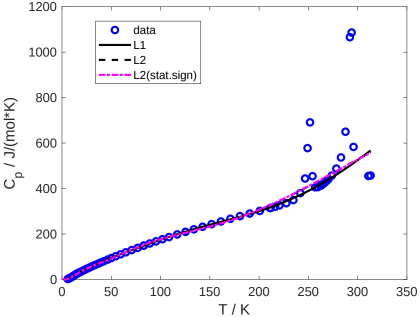


В этом случае МНМ позволяет автоматически обработать как точки при $T < 10$ K, так и аномалию C_p.

Пример: Tl-натролит

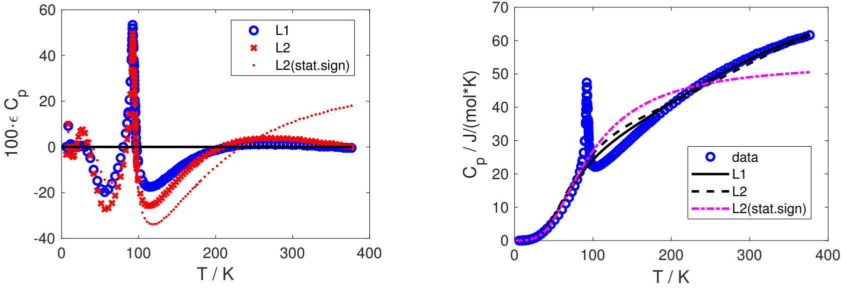
Аппроксимация вместе с аномалией теплоёмкости





Можно видеть, что МНМ лучше МНК исключает пики, но при этом перекрытие двух пиков обрабатывается как базовая линия, а продолжение базовой линии — как выброс.

Пример: диоксид марганца MnO_2



Аномалия влияет на «базовую линию»

Это связано с плавностью начала пика, поэтому МНМ не может автоматически выбрать, какой именно ход кривой $C_p(T)$ правильный.

Рекомендованная литература

- Обзорные монографии

  - Gonin R., Money A.H. Nonlinear L_p-norm estimation. New York and Basel: Marcel Dekker Inc. 1989. ISBN 978-0-8247-8125-5
  - Мудров В.И., Кушко В.Л. Метод наименьших модулей. 2-е изд. — Москва : URSS. 2013. — 64 с. — ISBN 978-5-397-04118-8

- Описание алгоритмов (статьи)

  - Gao Li. Numerical algorithms for nonlinear L_p-norm problem and its extreme case. J. Comput. Appl. Math. 2001. 129(1). P. 139–150. https://doi.org/10.1016S0377-0427(00)00546-X
  - Gao, Li. Using Huber Method to Solve Nonlinear L_1-Norm Problem. Advances in Nonlinear Programming: Proceedings of the 96 International Conference on Nonlinear Programming. P. 263–271. 1998. https://doi.org/10.1007/978-1-4613-3335-7_12
  - Wesolowsky G. O. A new descent algorithm for the least absolute value regression problem // Communications in Statistics — Simulation and Computation. 1981. V.10. N 5. P. 479–491. https://10.1080/03610918108812224

Заключение

- Метод наименьших модулей является устойчивым (робастным) к выбросам.
- Трудоёмкость даже в случае линейной регрессии (в 10–1000 раз медленнее МНК)
- Для нелинейной регрессии целесообразно использовать модифицированный метод Левенберга-Марквардта
- МНМ способен отсеять единичные выбросы, а также узкие аномалии теплоёмкости (лямбда-переходы, предплавление и т.п.), но не более «размытые» пики.
- МНМ может быть неэффективен для отсеивания размытых пиков, а также для различения базовой линии в сложных случаях.