Березняк Т.В.

О Методе Наименьших Квадратов

Вступление

Метод МНК в термометрии, как правило, используется для аппроксимации набора градуировочных точек полиномом n-ой степени. Однако МНК работает не только для полиномов, но и для функций вообще любого вида, в том числе для составных функций, имеющих разный вид в различных диапазонах. К сожалению, для составных функций зачастую используется неправильный метод интерполяции, который можно описать так: сначала стандартным (предназначенным для полиномов) методом аппроксимируют похожий на нужную функцию полином, а потом этот полином подгоняют под вид функции. Такой метод не дает минимум погрешности интерполяции. В данной статье показано, что правильное применение МНК позволяет уменьшить погрешность интерполяции при построении индивидуальных градуировочных характеристик рабочих термометров сопротивления и термопар.

Простой пример

Функция Каллендара Ван Дюзена, которая применяется для построения индивидуальной зависимости $R(t)$ для рабочих термометров сопротивления, имеет вид:

$$f(t) = \begin{cases} R_0(1 + At + Bt^2), t \geq 0; \\ R_0(1 + At + Bt^2 + C(t - 100)t^3), t < 0; \end{cases}$$

Рассмотрим искусственно созданный набор данных градуировки, который к платиновой термометрии отношения не имеет, но позволяет наглядно продемонстрировать преимущества правильного применения метода МНК.

Пусть у нас есть такой набор данных:

t, °C	R, ом
0	100
10	100
20	100
30	100
40	100
50	100
60	100
70	100
80	100
90	100
100	100

(во всех точках значение равно 100)

Так как он не содержит температур меньше нуля, то его можно сразу проинтерполировать квадратным полиномом

$$g(t) := R_0(1 + At + Bt^2)$$

Получится, очевидно, так:

R_0	100
A	0
B	0

Добавим к набору одну точку с отрицательной температурой:

t, °C	R, ом
-60	100.073657
0	100
10	100
20	100
30	100
40	100
50	100
60	100
70	100
80	100
90	100
100	100

Метод вычисления коэффициента C достаточно простой - надо посмотреть, какое отклонение дает квадратный полином в этой точке и подобрать коэффициент C так, чтобы нивелировать это отклонение (эта процедура характерна для всех градуировок, которые содержат только одну точку ниже нуля). Полученное таким образом C равно

$$\frac{100.073657 - g(-60)}{R_0 h(-60)} = 2.13128E - 11$$

где $h(t) = (t - 100)t^3$.

Теперь добавим еще точек между -60 и 0:

t, °C	R, ом
-60	100.073657
-50	99.8642355
-40	100.0625892
-30	99.84022808
-20	100.0338502
-10	99.70458042
0	100
10	100
20	100
30	100
40	100
50	100
60	100
70	100
80	100
90	100
100	100

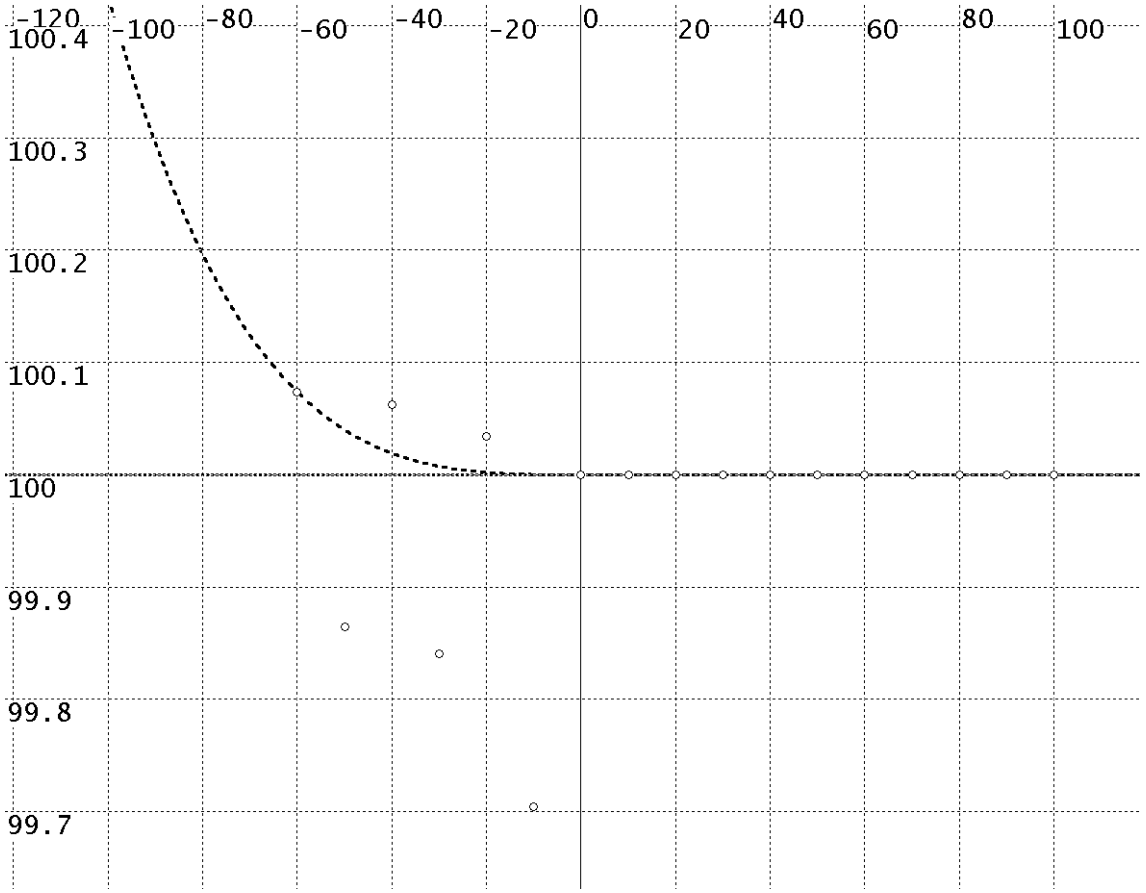
Наличие новых точек с отрицательной температурой делает задачу чуть сложнее. Здесь можно, например, вообще не принимать во внимание их показания, а учитывать только дальнюю (в конце концов, именно на ней сильнее всего сказывается погрешность C). Тогда для

R_0	100
A	0
B	0
C	2.13128E-11

мы получаем погрешность интерполяции $\sum_i (f(t_i) - R_i)^2 = 0.149167536$.

Есть вариант чуть лучше: взять все отклонения $d_i := g(t_i) - R_i$ в отрицательной области и снова применить МНК к функции $R_0 Ch(t)$ для нахождения $R_0 C$, оптимальным образом аппроксимирующим набор $\{d_i\}$. Не будем углубляться в расчеты (все равно этот метод не оптимален), но числа так подобраны, что такое C, минимизирующее погрешность интерполяции при данных A и B, равно нулю. Для данного C погрешность интерполяции равна 0.141720384. Это действительно чуть лучше, чем для C, подобранного исходя из только значения дальней точки.

На графике показано, как расположены точки, и как ведет себя функция, построенная с учетом только последней точки:

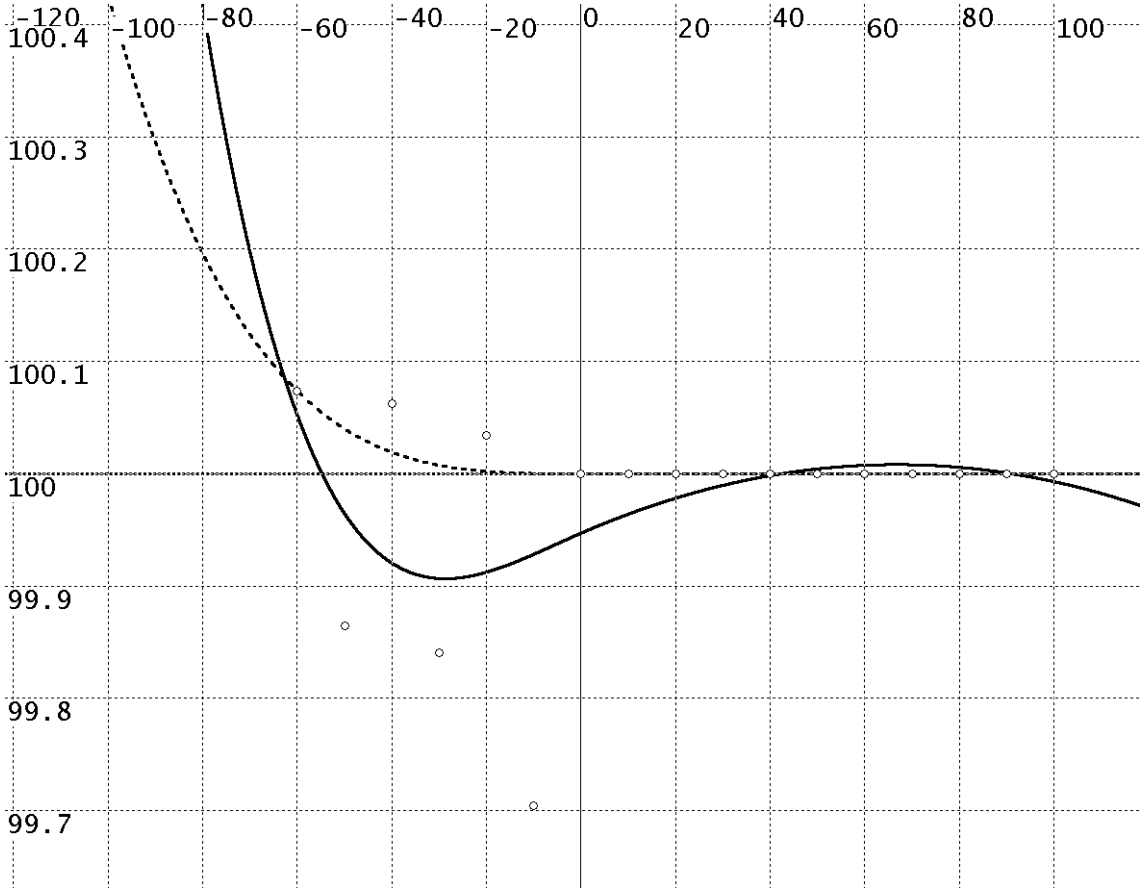


(функция с C=0 тут тоже изображена, она совпадает с линией 100°C).

А теперь самое интересное. Возьмем такой набор $\{R_0, A, B, C\}$:

R_0	99.947
A	0.000018316
B	-1.37215E-07
C	7.65196E-11

Функция КВД с такими коэффициентами выглядит неожиданно, на интервале выше нуля она смотрится, как ошибка (жирная линия на графике):



Но для этой аппроксимации погрешность интерполяции равна 0.104863663, что почти в полтора раза меньше, чем мы получили в предыдущем случае. Как же так получилось, что функция, которая явно не похожа на оптимальную в диапазоне выше нуля (где последний коэффициент ни на что, казалось бы, не влияет), оказалась оптимальнее, чем та, что была описаны выше, и как получились эти числа?

Вот тут настало время перейти к скучной теории.

Теория применения МНК

Пусть есть набор точек $\{x_i, y_i\}$. Есть функция $f(k_1, ..., k_n)(x) := \sum_{j=1}^{n} k_j f_j(x)$, f_j нам даны. k_j нам надо найти так, чтобы минимизировать

$$\Delta(k_1, ..., k_n) := \sum_i (f(k_1, ..., k_n)(x_i) - y_i)^2$$

Чтобы найти $\{k_j\}$, дающие минимум Δ, надо продифференцировать Δ по каждому k_j. В результате получится такая система линейных уравнений:

$$\begin{cases} \sum_i f_1(x_i)f_1(x_i)k_1 + \sum_i f_1(x_i)f_2(x_i)k_2 + ... + \sum_i f_1(x_i)f_n(x_i)k_n = \sum_i f_1(x_i)y_i \\ \sum_i f_2(x_i)f_1(x_i)k_1 + \sum_i f_2(x_i)f_2(x_i)k_2 + ... + \sum_i f_2(x_i)f_n(x_i)k_n = \sum_i f_2(x_i)y_i \\ ... \\ \sum_i f_n(x_i)f_1(x_i)k_1 + \sum_i f_n(x_i)f_2(x_i)k_2 + ... + \sum_i f_n(x_i)f_n(x_i)k_n = \sum_i f_n(x_i)y_i \end{cases}$$

Коэффициенты из левой части образуют квадратную матрицу M, коэффициенты из правой части образуют столбец K, вся система вместе образует матрицу M|K. Числа, образующие M и K, такие:

$$M_{u,v} = \sum_i f_u(x_i) f_v(x_i)$$
$$K_u = \sum_i f_u(x_i) y_i$$

или

$$M_{u,v} = f_u \cdot f_v$$
$$K_u = f_u \cdot y$$

где f_u - вектор значений функции f_u в точках x_j. Это обычная система линейных уравнений, решается стандартными методами. Обратите внимание, что вид функций f_i абсолютно не важен! Важны только их значения в точках x_j. В качестве f_i быть простые степенные функции (самый привычный случай), могут быть полиномы посложнее, могут быть синусы с косинусами. В качестве f_i может быть даже функция Дирихле, и даже, что самое удивительное, такая функция: $\begin{cases} 0, x >= 0 \\ (x-100)x^3, x < 0 \end{cases}$! Да, это та самая функция, которая участвует при коэффициенте R_0C в функции КВД.

Применяя нашу теорию к КВД, получаем такой вид функций $f_1, ..., f_4$ (очевидно, что $n = 4$):

$$\begin{cases} f_1(x) = 1 \\ f_2(x) = x \\ f_3(x) = x^2 \\ f_4(x) = \begin{cases} 0, x >= 0 \\ (x-100)x^3, x < 0 \end{cases} \end{cases}$$

Итак, считаем значения каждой функции в каждой точке (температуре) из набора, также рядом запишем требуемое сопротивление при каждой температуре:

t	f_1	f_2	f_3	f_4	y
-60	1	-60	3600	34560000	100.0737
-50	1	-50	2500	18750000	99.86424
-40	1	-40	1600	8960000	100.0626
-30	1	-30	900	3510000	99.84023
-20	1	-20	400	960000	100.0339
-10	1	-10	100	110000	99.70458
0	1	0	0	0	100
10	1	10	100	0	100
20	1	20	400	0	100
30	1	30	900	0	100
40	1	40	1600	0	100
50	1	50	2500	0	100
60	1	60	3600	0	100
70	1	70	4900	0	100
80	1	80	6400	0	100
90	1	90	8100	0	100
100	1	100	10000	0	100

Теперь составляем систему линейных уравнений. Матрица M|K будет такой:

$$\left(\begin{array}{cccc|c} 17 & 340 & 47600 & 66850000 & 1699.579140 \\ 340 & 47600 & 2584000 & -3495100000 & 34006.935589 \\ 47600 & 2584000 & 276080000 & 189181000000 & 4759866.099924 \\ 66850000 & -3495100000 & 189181000000 & 1639491500000000 & 6684999999.999970 \end{array}\right)$$

Теперь надо найти $Res := M^{-1}K$. Подробно описывать процедуру обращения матрицы смысла нет, есть смысл сразу сказать результат, сразу поделив Res_2, Res_3, Res_4 на Res_1 (поскольку после обращения $Res_1 = R_0$, $Res_2 = AR_0$, $Res_3 = BR_0$, $Res_4 = CR_0$). Получается так:

$$Res = \left(\begin{array}{c} 99.947 \\ 0.000018316 \\ -1.37215E-07 \\ 7.65196E-11 \end{array}\right)$$

То есть получаются те самые коэффициенты R_0, A, B, C, дающие оптимальную кривую аппроксимации, и которые описаны в конце предыдущей секции.

Есть нюанс. Иногда погрешность в положительной области важнее, чем в отрицательной. Не беда, повесим на точки $\{x_i, y_i\}$ массы $\{w_i\}$, то есть получим набор $\{x_i, y_i, w_i\}$. Смысл в том, что там теперь надо минимизировать $\sum_i (w_i(f(k_1, ..., k_n)(x_i) - y_i))^2$, а не просто

$$\sum_i (f(k_1, ..., k_n)(x_i) - y_i)^2$$

Тут все аналогично безмассовому случаю, только теперь M и K выглядят так:

$$M_{u,v} = \sum_i w_i^2 f_u(x_i) f_v(x_i)$$
$$K_u = \sum_i w_i^2 f_u(x_i) y_i$$

или

$$M_{u,v} = f_u \cdot_w f_v$$
$$K_u = f_u \cdot_w y$$

где $\cdot_w$ - это такое специальное скалярное произведение:

$$a \cdot_w b = \sum_i w_i^2 a_i b_i$$

Если взять все w_i равные единице, то получится то же самое, что и в безмассовом случае. Если применить в нахождению функции КВД метод с массами, равными единице в области выше нуля, и равными несоизмеримо малой величине в области ниже нуля, то получится та же функция, что и в случае нахождения функции “кусками”. А если повесить единичные массы на точки, что выше нуля и на самую дальнюю точку ниже нуля, а на остальные точки повесить бесконечно малые массы, то получится та же функция, что и при аппроксимации, учитывающей только самую дальнюю точку из отрицательного диапазона.

То есть “неверный” метод является частным случаем только что описанного (только для пренебрежимо малых масс отрицательных точек).

Возьмем более реалистичные данные, например, такие:

t, °C	R, ом
-60.3	76.1
-40.6356	84
0	100
10.2795	104
92.4758	135.5
132.6969	150.6
161.5362	161.3
207.1301	178
309.6997	214.6

Результат будет аналогичен - функция, построенная “для всех точек сразу” лучше приближает набор точек, чем та, которая сначала строит параболу, а потом ищет коэффициент C.

Это коэффициенты функции, для которой делали сначала параболу, потом искали C:

R_0	100
A	0.0038978
B	-6.37517E-07
C	-4.76236E-11

Для нее погрешность интерполяции будет 0.000126171.

Если же взять функцию, построенную по “всем точкам сразу”:

R_0	99.9977
A	0.00389821
B	-6.38406E-07
C	-4.61369E-11

то для нее погрешность интерполяции будет 0.000087987.

Еще один пример

Еще пример использования МНК - для калибровки термопар типа S по новому алгоритму (утвержденному при аттестации программы для поверочных лабораторий TermoLab), при наличии нескольких температур t_i и измеренных ТЭДС E_i при этих температурах, следует взять набор точек $\{t_i, E_i - E_{ref}(t_i)\}$ и аппроксимировать его функцией такого вида:

$$f(t) = \begin{cases} at + bt^2, t < T \\ ct + d, t \geq T \end{cases} \quad (T = 1064.18)$$

Причем один фрагмент должен непрерывно (и с непрерывной производной) переходить в другой.

Все точки из набора лежат ниже T, и только одна - чуть выше. Типичный пример:

t	E
231.928	1.7121
419.527	3.4466
660.323	5.8653
961.78	9.1658
1064.18	10.3562
1084.62	10.5974

Типичное решение в стиле “сначала аппроксимируем полиномом, потом подгоним”: аппроксимировать набор точек параболой, проходящей через ноль (то есть найти a и b, но так,

будто функция - везде парабола), потом обрубить параболу до T, потом продлить параболу от обрубка по касательной. Для данного набора получаются такие a и b:

a	-1.50472242518E-05
b	3.36389618518E-08

Погрешность интерполяции равна 2.70562346556E-06.

Такой метод действительно почти дает минимум, когда почти все точки ниже T, а если некоторые будут существенно выше, то что делать? А вот тут надо немного внимательнее посмотреть на f. Заметим, что c и d однозначно зависят от a и b (ведь T фиксировано). После небольших вычислений нетрудно прийти к тому, что

$$\begin{cases} c = a + 2bT \\ d = -bT^2 \end{cases}$$

В результате дальнейших преобразований функция f приводится к такому виду:

$$at + b \begin{cases} t^2, t < T \\ T(2t - T), t \geq T \end{cases}$$

Как видим, мы снова свели функцию к виду $\sum_j k_j f_j(t)$. А это значит, что мы можем по аналогии с предыдущими случаями взять три вектора f_1, f_2, y, построить матрицу M|K из их скалярных произведений и найти $M^{-1}K$. Получатся такие a и b:

a	-1.50508442776E-05
b	3.36451742975E-08

Погрешность интерполяции получается 2.70555231073E-06. Это хоть и чуть-чуть, но меньше, чем в предыдущем случае. С практической точки зрения разница несущественна, но она все-таки превышает погрешность вычислений, а значит, с теоретической точки зрения, честное применение МНК “сразу” имеет преимущество перед махинациями с обрубанием парабол. Если же будет много точек выше T, то расхождение в точности аппроксимации будет значительно выше.

Заключение

Правильное применение МНК, основанное только на значениях функций в точках, и никак не затрагивающее вид функций, позволит снизить погрешность интерполяции. Главное - просто применить теорию, не отвлекаясь на ненужные мелочи, например, не надо смотреть, на какой полином похожа функция, потому что эти мелочи отвлекают и подталкивают на применение неправильного метода.

ФГУП ВНИИМ им Д.И. Менделеева
tarasber@mail.ru