Оптимизация смещений и дисперсий оценок параметров математических моделей при обработке сглаженных экспериментальных данных

Д. А. Потапов

Аннотация. В работе исследуется влияние степени сглаживающего полинома на смещение и дисперсию получаемых оценок параметров на примере кластерной модели растворов. Разработан алгоритм оптимизации смещений и дисперсий оценок параметров при обработке сглаженных экспериментальных данных.
Ключевые слова: *моделирование свойств растворов, проверка адекватности модели, математическая модель, метод наименьших квадратов, смещенность оценки.*

Введение

При исследовании растворов важным этапом в определении их термодинамических характеристик является оптимизация функции невязки математической модели. При этом в большинстве случаев в литературе недоступны данные, полученные непосредственно из эксперимента, а приводится результат их сглаживания полиномами некоторой степени n. Коэффициенты полинома определяются методом наименьших квадратов [1]. При адекватно выбранной степени полинома n сглаживание позволяет уменьшить дисперсию оценок параметров [2]. В качестве результата предоставляются значения найденного полинома в некоторых точках, отличных от точек, в которых измерялись экспериментальные значения. Таким образом, часть информации об эксперименте утрачивается.

Полученные данные используются другими исследователями для проверки адекватности математических моделей. На данном этапе изменить степень описывающего полинома и получить информацию о значениях аргумента, в которых проводились непосредственные измерения, обычно не представляется возможным. Из-за погрешностей измерения экспериментальные данные представляют собой случайные величины, поэтому коэффициенты полинома также являются случайными величинами, вследствие чего параметры модели также случайны. Идентификацию параметров осуществляют с помощью метода наименьших квадратов, который приводит к смещенности оценок в случае нелинейности модели [3]. В данной работе показано, что от степени сглаживающего полинома зависят математическое ожидание и дисперсия искомых параметров модели, однако эта зависимость индивидуальна для различных функций. В результате эксперт, осуществляющий сглаживание, не может выбрать степень полинома n с учетом специфики модели, по которой другие исследователи будут проводить расчет. Из вышесказанного следует, что лучшим решением является публикация непосредственно измеренных данных, что предоставит исследователям, работающим с этими данными, возможность самостоятельного выбора степени полинома.

В настоящей работе исследуется влияние степени сглаживающего полинома на статистические характеристики искомых параметров нелинейных моделей, также предлагается методика обработки сглаженных данных, обеспечивающая более точные и устойчивые значения параметров модели по сравнению с обычным применением метода наименьших квадратов.

1. Влияние степени сглаживающего полинома на математическое ожидание и дисперсию параметров модели

Исследование проводилось на примере кластерной модели растворов [4], связывающей осмотический коэффициент раствора φ с моляльностью раствора m. Ниже приведены уравнения модели:

$$n_1 = 55,508; \tag{1}$$

$$A_1 = 0,5115 \ln 10; \tag{2}$$

$$A = A_1 z_1 z_2; \quad (3)$$

$$q = q_1 + q_2; \quad (4)$$

$$I(m) = m\frac{q_1 {z_1}^2 + q_2 {z_2}^2}{2}; \quad (5)$$

$$X(m) = \frac{qm}{qm + n_1}; \quad (6)$$

$$fe_D(B,m) = \frac{-(A\sqrt{I(m)})(1 + B\sqrt{I(m)})}{(B\sqrt{I(m)})^3} - \\ - 2\ln(1 + B\sqrt{I(m)}) - \frac{1}{1 + B\sqrt{I(m)}}; \quad (7)$$

$$fe_A(A_s,m) = \frac{-A_s X(m)}{1 + A_s X(m)(1 - X(m))}; \quad (8)$$

$$fe_h(q,h,r,m) = \frac{qhX(m)(1 - X(m))^{r-1}}{(1 - hX(m)(1 - X(m))^r)}; \quad (9)$$

$$\varphi\ (m,h,r,A_s,B) = 1 + \frac{fe_h(q,h,r,m)}{q} + \\ + fe_A(A_s,m) + fe_D(B,m), \quad (10)$$

где z_1, z_2 — заряд соответственно катиона и аниона растворенного вещества, q_1, q_2 — количество катионов и анионов в молекуле, m — моляльность растворенного вещества, B — коэффициент Дебая, A_s — коэффициент, характеризующий степень ассоциации в растворе, φ — осмотический коэффициент, h — средняя степень гидратации иона, r — коэффициент, характеризующий дисперсию распределения ионов в растворе по степеням гидратации.

В целях уменьшения влияния алгоритма оптимизации на нахождение минимума функции невязки, а также возможности отображения результатов на графике, было положено $A_s = 0{,}0001$, $B = 2{,}1497$, таким образом четырехпараметрическая модель была преобразована в двухпараметрическую с параметрами h и r.

Для определения точности решения необходимо знать истинные значения измеряемой величины. Пусть для целей моделирования

$$\varphi_{ист}(m) = 1 + \frac{fe_h(q,h,r,m)}{q} + \\ + fe_A(A_s,m) + fe_D(B,m), \quad (11)$$

где

$$q = 4,\ h = 6{,}6918,\ r = 2{,}7068,\\ A_s = 0{,}0001,\ B = 2{,}1497.$$

График $\varphi_{ист}(m)$ приведен на рис. 1.

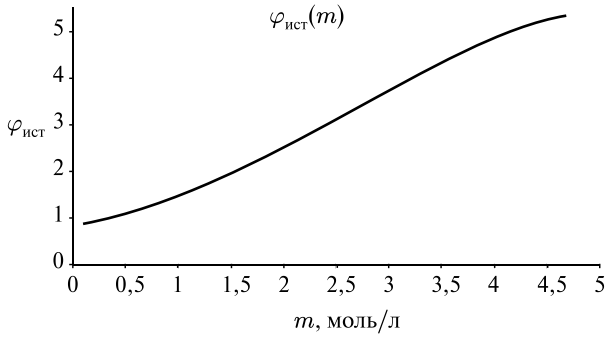

Рис. 1. Зависимость значений осмотического коэффициента, принятых за истинные для целей моделирования, от моляльности раствора

Экспериментальные данные могут быть получены из истинных с помощью выражения (12):

$$\varphi_{эксп}\ (m) = \varphi_{ист}\ (m) + \delta(m), \quad (12)$$

где $\delta(m)$ — погрешность измерений. Для моделирования будем считать $\delta(m)$ случайной величиной, имеющей равномерное распределение в диапазоне [-5 %*$\varphi_{ист}(m)$, 5 %*$\varphi_{ист}(m)$].

В качестве аргументов сглаженных значений возьмем вектор $M_{сгл}$:

$M_{сгл}$ = [0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1; 1,2; 1,4; 1,6; 1,8; 2; 2,2; 2,4; 2,6; 2,8; 3; 3,2; 3,4; 3,6; 3,8; 4; 4,2; 4,4; 4,6; 4,7]

В общем случае вектор $M_{сгл}$ не совпадает с вектором значений $M_{изм}$, в которых производились непосредственные измерения. Для рассматриваемой системы вектор $M_{изм}$ содержал 79 значений, тогда как размерность вектора $M_{сгл}$ равна 29.

Возьмем полиномы степени $n = 2{,}3, \ldots, 9$ с неизвестными коэффициентами. Для каждой степени вычислим коэффициенты полинома с помощью метода наименьших квадратов, рассчитаем значения этого полинома в точках $M_{сгл}$, после чего с помощью того же метода рассчитаем параметры модели h и r. Данный расчет был осуществлен $N = 2000$ раз для каждой степени n, причем на каждом шаге генерировались новые значения $\varphi_{эксп}(m)$. Аналогичная процедура была проделана для сглаженных данных с добавлением случайного шума, после чего были рассчитаны дисперсии найденных параметров модели и их отклонения от истинных значений. На рис. 2 показан пример распределения оценок параметров модели для $n = 4$.

Значения дисперсий D и отклонений математических ожиданий от истинных значений Δ для параметров, полученным по сглаженным и сглаженным с добавлением шума экспериментальным данным, приведены в табл. 1.

Таблица 1

Смещения и дисперсии оценок параметров модели при различных степенях сглаживающего полинома

n	2	3	4	5	6	7	8	9
$\Delta_{сгл}$	0,2162	0,0168	0,0026	0,0019	0,0023	0,0037	0,0032	0,0028
$\Delta_{шум}$	0,1961	0,0317	0,0084	0,0007	0,0017	0,0021	0,0094	0,0143
$D_{сгл}$	0,0031	0,0048	0,0053	0,0054	0,0054	0,0057	0,0060	0,0061
$D_{шум}$	0,0136	0,0183	0,0137	0,0197	0,0172	0,0185	0,0155	0,0144

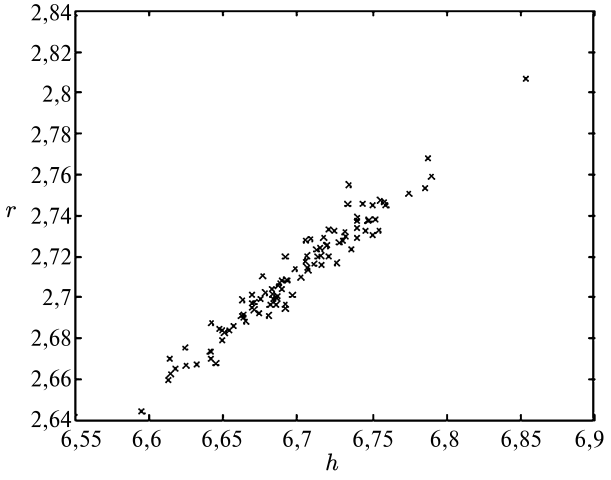

Рис. 2. Распределение оценок параметров модели при аппроксимации по сглаженным данным для n = 4

При моделировании непосредственно по экспериментальным данным, без использования сглаживания, результаты получаются следующими:

$$D = 0{,}0040;$$
$$\Delta = 0{,}0057.$$

На рис. 3 приведены графические зависимости полученных дисперсий от степени n аппроксимирующего полинома.

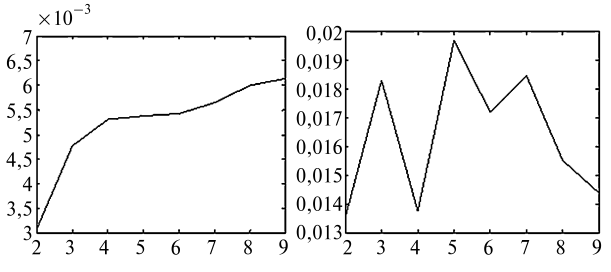

Рис. 3. Зависимости полученных дисперсий от степени n аппроксимирующего полинома для параметров, полученных по сглаженным и сглаженным с добавлением шума экспериментальным данным

2. Корректировка полученных значений параметров при работе со сглаженными данными

Из табл. 1 и рис. 2 видно, что явной закономерности между дисперсией D и n в случае зашумленных данных не наблюдается, в случае обработки сглаженных данных дисперсия растет с ростом степени сглаживающего полинома n. Минимум дисперсии для сглаженных данных соответствует $n = 2$, однако при данном значении n возникает существенная смещенность оценок. При этом при $n \geq 4$ отклонения от истинного значения малы (h и r обычно рассчитываются с точностью до второго знака после запятой). Дисперсия результатов, полученных по зашумленным данным, существенно больше полученной по сглаженным данным при всех значениях n.

Если смещение по каждому из параметров известно, то при его вычитании из найденных значений возможно получить параметры с минимальной дисперсией, являющиеся несмещенными. Однако прямое вычисление значения смещений невозможно ввиду отсутствия у исследователя информации об истинных значениях моделируемой величины.

Приведенные ниже результаты численного моделирования показывают, что в случае сглаживания экспериментальных данных полиномом, обеспечивающим относительно малые дисперсии оценок и малые смещения (для рассматриваемой модели это $n \geq 4$), в целях определения значений смещений в качестве истинных значений параметров можно взять найденные по сглаженным данным методом наименьших квадратов. Численное моделирование проводилось на основании следующих соображений. При аппроксимации сглаженных с помощью полинома степени n данных уравнениями математической модели, значения найденных параметров приобретают некоторое смещение и разброс относительно истинных значений. И смещение, и дисперсия могут быть определены с помощью исследования уравнения модели. Для рассматриваемой модели эти величины приведены в табл. 1. Возьмем несколько точек из области, в которую могут попадать найденные параметры, и пересчитаем табл. 1, принимая в качестве истинных найденные значения. Пересчет будем проводить только для n, соответствующего минимальной

Таблица 2

Смещения и дисперсии оценок параметров, полученные в случае принятия рассчитанных параметров модели в качестве истинных при степени сглаживающего полинома $n = 2$

Номер эксперимента	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$\Delta_{сгл}$	0,2168	0,2165	0,2161	0,2159	0,2163	0,2166	0,2165	0,2161
$D_{сгл}$	0,0033	0,0032	0,0035	0,0031	0,0033	0,0036	0,0034	0,0031

дисперсии. Для рассматриваемой модели $n = 2$. Результаты пересчета приведены в табл. 2.

Из таблицы видно, что для $n = 2$ смещение и дисперсия, получаемые по истинным значениям ($\Delta_{сгл} = 0{,}2168$, $D_{сгл} = 0{,}0031$), близки к значениям смещения и дисперсии, полученным в случае принятия рассчитанных параметров модели в качестве истинных. Таким образом, в качестве параметров распределения оценок для истинных значений могут быть взяты данные табл. 2, которые могут быть рассчитаны исследователем, не имеющим доступ к истинным или непосредственно измеренным значениям.

Ранее упоминалось, что для применения описанной процедуры необходимо, чтобы для сглаживания данных использовался полином степени, обеспечивающей малость смещения и дисперсии оценок параметров модели. Однако в большинстве случаев непосредственная информация о степени сглаживающего полинома недоступна исследователю. В этом случае ее можно определить, последовательно аппроксимируя сглаженные данные полиномами различной степени. При соответствии степеней сумма квадратов невязок резко устремится к нулю, что свидетельствует о нахождении искомого n.

3. Алгоритм оптимизации дисперсии и смещения оценок параметров

В настоящем разделе на основании изложенных ранее результатов исследования приведено обобщенное описание алгоритма оптимизации дисперсии и смещения оценок параметров. Блок-схема алгоритма приведена на рис. 4.

Исходными данными для его применения являются векторы $M_{сгл}$ и $\varphi_{сгл}$, известна погрешность измерений δ.

Также имеется уравнение модели с неизвестными значениями параметров, которые требуется идентифицировать.

Алгоритм получения оптимизированных по дисперсии и смещению оценок параметров состоит из следующих этапов:

1) Получение оценок параметров обычным методом наименьших квадратов.
2) Определение степени сглаживающего полинома n последовательной аппроксимацией сглаженных данных полиномами различной степени. При нахождении искомого n функция невязки резко устремится к нулю.
3) Принятие найденных оценок в качестве истинных и численное моделирование на основе них параметров распределения оценок, получаемых методом наименьших квадратов.
4) В случае малости дисперсий и смещений при использованной для сглаживания степени полинома n выбор степени полинома n_2 и значения смещения Δ, соответствующих минимальной дисперсии.
5) Вычитание из значений параметров, найденных на этапе 1, величин Δ, определенных на этапе 4.

Критерием применимости описанного алгоритма является малость дисперсий и смещений, соответствующих степени сглаживающего полинома n, т. к. в противном случае дисперсия и смещение оценок параметров для истинных значений, измеренных с некоторой погрешностью, и найденных значений могут существенно различаться и в этом случае последние не позволят сделать вывод о значениях первых.

Построение таблицы зависимости дисперсий и смещений оценок параметров на этапе 3 алгоритма, осуществляется с помощью зашумления теоретической зависимости с найденными оценками параметров с последующей их аппроксимацией полиномом степени n_2 и нахождением новых оценок. Процедура повторяется многократно (в данной работе $N = 2000$ раз) для получения репрезентативных данных о распределении оценок параметров. Вносимый шум должен соответствовать погрешности измерений при проведении эксперимента δ.

Заключение

При моделировании свойств растворов исследователи обычно имеют дело с нелинейными моделями. Кроме того, данные для моделирования представляют собой измеренные данные, сглаженные полиномом неизвестной степени n. При этом нарушаются условия теоремы Гаусса—Маркова [3] и применяемый большинством исследователей метод наименьших квадратов не обеспечивает несмещенности и эффективности найденных оценок параметров.

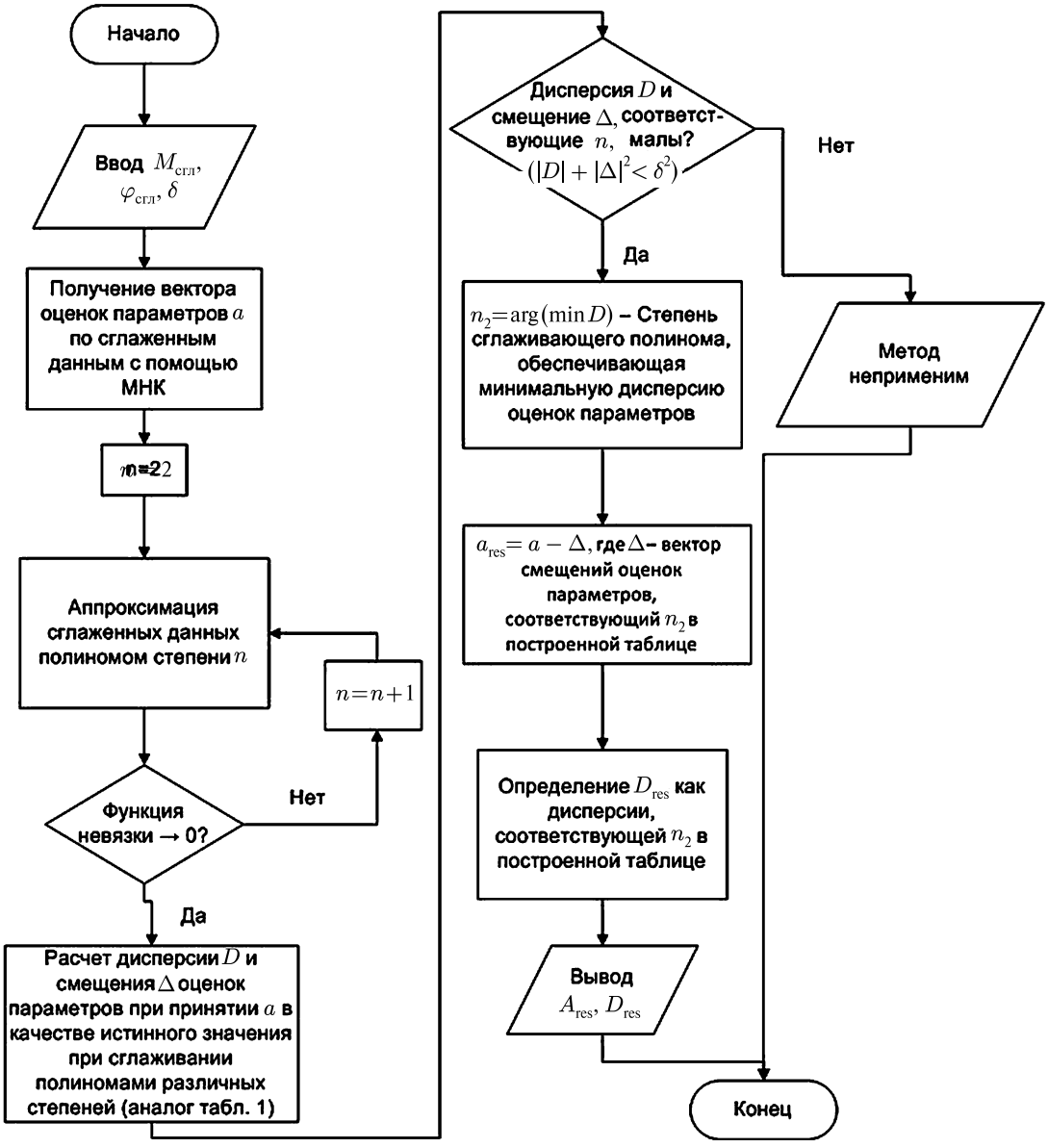

Рис. 4. Блок-схема алгоритма оптимизации смещений и дисперсий оценок параметров математических моделей

К основным термодинамическим характеристикам растворов относятся осмотический коэффициент φ и коэффициент активности $\gamma_{\pm}$. Обычно идентификацию параметров исследуемых моделей растворов осуществляют с использованием концентрационной зависимости осмотического коэффициента от моляльности раствора m, а совпадение экспериментальной зависимости коэффициента активности с полученной теоретической служит дополнительным подтверждением адекватности исследуемой модели. Часто возникает ситуация, когда при точном описании экспериментальной зависимости $\varphi(m)$, в описании $\gamma_{\pm}(m)$ возникает систематическая ошибка. Одна из причин такого поведения заключается в смещенности оценок параметров, получаемых с помощью метода наименьших квадратов в случае нелинейности модели. В результате исследователем может быть сделан ошибочный вывод о неадекватности исследуемой им математической модели. Описанный в данной работе алгоритм оптимизации смещений и дисперсий оценок параметров модели позволяет существенно уменьшить возникающую в подобных ситуациях систематическую ошибку, позволяя делать более объективные выводы об адекватности исследуемых моделей.

Литература

1. *Wolberg J. R.* Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments. Springer, 2005.
2. *Седелев Б. В.* Регрессионные модели и методы оценки параметров и структуры экономических процессов. М.: Московский инженерно-физический институт (государственный университет), 2009.
3. *Plackett R. L.* Some Theorems in Least Squares // Biometrika, 1950. Vol. 37. № 1–2. P. 149–157.
4. *Рудаков А. М., Майкова Н. С., Сергиевский В. В.* Исследование сольватации и ассоциации в бинарных растворах на основе кластерной модели // Проблемы сольватации и комплексообразования в растворах, Иваново, 2011, 14 с.

Потапов Дмитрий Александрович. Аспирант НИЯУ «МИФИ», кафедра «Информатика и процессы управления». Окончил МИФИ в 2009 г. Количество печатных работ: 1. Область научных интересов: компьютерное моделирование физико-химических процессов. E-mail: div-x15@yandex.ru