«Оптика атмосферы и океана», **17**, № 11 (2004)

УДК 536.55

С.А. Ташкун

Использование робастных методов оценивания спектроскопических параметров колебательных полос линейных молекул из экспериментальных данных

Институт оптики атмосферы СО РАН, г. Томск

Поступила в редакцию 1.09.2004 г.

Традиционный линейный метод наименьших квадратов имеет нулевую устойчивость по отношению к выбросам в экспериментальных данных, что ведет к искаженным оценкам параметров и построению моделей, неадекватных полученным данным. Поэтому для получения подогнанных спектроскопических параметров колебательных полос линейных молекул предлагается использовать вместо указанного метода метод наименьшей медианы, который является устойчивым к наличию выбросов в данных и позволяет получать статистически обоснованные значения параметров. Важность и полезность использования метода наименьшей медианы продемонстрированы на примере получения спектроскопических параметров полосы 40002–01101 молекулы $^{12}C^{16}O_2$.

Введение

В подавляющем большинстве работ, посвященных измерениям и моделированию частот колебательно-вращательных переходов линейных молекул, используется полиномиальное представление вращательной зависимости уровней энергии $E(J)$ колебательного состояния v:

$$E_v(J) = G_v + B_v[J(J+1)] - \\ - D_v[J(J+1)]^2 + H_v[J(J+1)]^3, \qquad (1)$$

где G_v, B_v, D_v, H_v – спектроскопические параметры состояния, определяемые из подгонки к экспериментальным значениям E_v. В качестве минимизируемой величины используется сумма квадратов отклонений между экспериментальными и вычисленными значениями энергий. Методом минимизации является линейный метод наименьших квадратов (ЛМНК), поскольку параметры G_v, B_v, D_v, H_v входят в выражение для $E_v(J)$ линейно. Имея подогнанные значения параметров, можно проводить интерполяционные и экстраполяционные расчеты по вращательному квантовому числу J. Недостатком такого подхода является невозможность корректного описания состояний, вовлеченных в локальные по J резонансы, поскольку формула (1) справедлива лишь для случая отсутствия резонансных взаимодействий. При использовании ЛМНК неявно предполагается, что в подгоняемых данных нет выбросов, т.е. модель (1) в состоянии описать все данные с приемлемыми невязками. К сожалению, малость невязок еще не гарантирует отсутствие выбросов.

ЛМНК и робастные методы оценивания для линейных моделей

Теория ЛМНК хорошо разработана. Полезный обзор использования ЛМНК в приложении к задачам молекулярной спектроскопии приведен в работе [1]. Оценки параметров $\hat{G}_v$, $\hat{B}_v$, $\hat{D}_v$ и $\hat{H}_v$ модели (1) являются несмещенными и имеют наименьшую дисперсию в классе линейных оценок (теорема Гаусса–Маркова). Статистические характеристики оценок (стандартные ошибки, матрица корреляций, инфляционные коэффициенты дисперсии) дают исследователю теоретически обоснованную информацию о свойствах модели, руководствуясь которой можно пытаться построить модель, оптимальную в некотором смысле. Кроме того, имеется ряд высококачественных программ, реализующих этот метод.

Краеугольным камнем всей теории ЛМНК является предположение о достаточном объеме выборки данных и нормальности распределения их ошибок. При его нарушении большинство выводов теории теряют силу. К сожалению, большие выборки с нормальным распределением ошибок измерения являются скорее исключением, чем правилом. Для работы с короткими выборками, отклоняющимися от нормального закона распределения случайных ошибок, следует использовать робастные методы оценивания параметров моделей [2, 3]. Было сделано несколько попыток (см., например, [4]) использования таких методов в молекулярной спектроскопии. Однако они являются более ресурсоемкими по сравнению с ЛМНК, а также имеют более сложную алгоритмическую реализацию. Возможно

поэтому эти методы все еще не получили широкого распространения в спектроскопических задачах.

После проведения подгонки данные из выборки, имеющие большие невязки, требуют дальнейшего изучения. Причина наличия этих выбросов двоякая. С одной стороны, они могут быть дефектами наблюдений (ошибками измерений, зашумленностью, неверной квантовой идентификацией и т.д.) и не иметь значения с точки зрения физических явлений, описываемых моделью. С другой стороны, именно они могут свидетельствовать о неадекватности используемой модели. Ясно, что выбор между этими вариантами может сделать лишь исследователь на основе анализа сложившейся ситуации.

Известно, что ЛМНК имеет нулевую устойчивость по отношению к выбросам [2]: достаточно лишь одного выброса, чтобы кардинальным образом изменить оценки параметров и, следовательно, структуру модели. Причем сам этот выброс может и не иметь большую невязку. Для устранения этого недостатка и используются робастные методы. Разницу между ЛМНК и робастными методами можно проиллюстрировать на рис. 1 и 2, заимствованных из книги [3]. Пусть имеются данные, состоящие из трех групп точек A, B, C, которые мы хотим подогнать с использованием линейной модели. Если воспользоваться ЛМНК, то подгоночная линия пройдет примерно так, как показано на рис. 1.

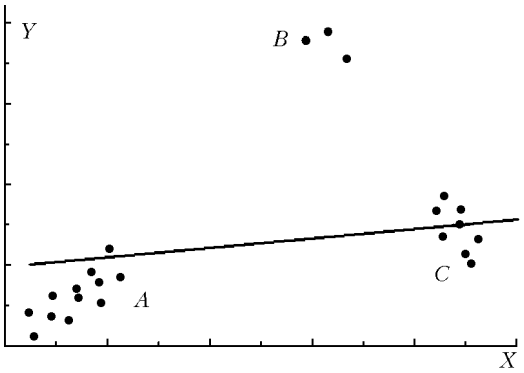

Рис. 1. Подгонка линейным методом наименьших квадратов

При этом точки группы B, имеющие наибольшие невязки, классифицируются как выбросы и не описываются моделью. Такой вид прямой объясняется тем, что ЛМНК избегает появления больших невязок и в нашем случае дает усредненное мнение обо всех точках. С другой стороны, используя высокоробастную подгонку, мы получим картину, приведенную на рис. 2. Видно, что этот метод дает явное мнение о большинстве точек. Точки группы B оказываются при этом хорошо описанными, зато точки группы C являются явными выбросами. Очевидно, необходим дальнейший анализ групп данных B и C с целью выявления источника их анормальности. Являются ли эти точки шумами или же несут важную информацию о моделируемом явлении, которая не описывается линейной моделью, — все эти вопросы для исследователя.

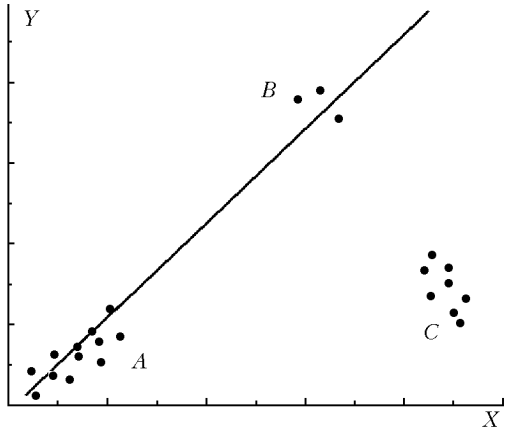

Рис. 2. Подгонка методом наименьшей медианы

Наиболее устойчивым методом оценивания является метод наименьшей медианы (МНМ) [2]. Разработанный для линейных моделей он имеет устойчивость по отношению к выбросам 50%. Это значит, что метод в состоянии указать до 49 выбросов для выборки из 100 данных. Однако он требует чрезвычайно большого времени счета, поскольку реализует комбинаторный перебор всевозможных подвыборок определенной длины из всей выборки.

Применение МНМ к анализу полос молекулы CO_2

В качестве примера, показывающего полезность применения МНМ в спектроскопии, рассмотрим задачу определения параметров G_v, B_v, D_v, верхнего состояния полосы 40002—01101 молекулы $^{12}C^{16}O_2$. Экспериментальные частоты ν_{obs} были взяты из работы [5]. Рассчитанные частоты моделировались выражением

$$\nu_{calc} = E_{v'}(m) - E_v(m-1), \qquad (2)$$

где $v' = 40002$, $v = 01101$; $m = -J$ для P-ветви и $m = J + 1$ для R-ветви. Значения параметров состояния 01101 были взяты из работы [6].

Подгонка с использованием ЛМНК дала величину среднеквадратического отклонения (СКО) 0,0053 $см^{-1}$ и невязки, приведенные в третьей колонке таблицы. В колонке 1 приведена идентификация линий, в колонке 2 — экспериментальные значения частоты перехода. Все данные имели вес, равный единице. Наибольшая невязка не превосходит 0,012 $см^{-1}$, и, по-видимому, модель (2) адекватна данным.

Рассмотрим те же данные с использованием МНМ. Текст программы, реализующий этот метод, был взят из [2]. МНМ обнаружил 9 выбросов в 34 данных, которым был приписан нулевой вес, что отражено в колонке 5 таблицы. После этого с оставшимися 25 данными была проведена повторная подгонка с помощью ЛМНК. СКО подгонки стало 0,0014 $см^{-1}$, а невязки приведены в четвертой колонке таблицы. Рассчитанные невязки выброшенных данных подчеркнуты. Наибольшая невязка этих

данных равна 0,0027 см$^{-1}$. Из выброшенных данных наибольший интерес представляют линии *R*39-*R*43, для которых невязка достигает 0,21 см$^{-1}$. Остальные линии, очевидно, являются зашумленными.

Подгонка спектроскопических постоянных верхнего состояния полосы 40002–01101 с использованием ЛМНК и МНМ

Линия	ν_{obs}, см$^{-1}$	ЛМНК, см$^{-1}$	МНМ, см$^{-1}$	Вес
P5	4804,2693	–0,01151	–0,01481	0,0
P7	4802,7082	–0,00452	–0,00816	0,0
P9	4801,1437	0,00140	–0,00269	1,0
P11	4799,5747	0,00537	0,00083	1,0
P13	4798,0009	0,00737	0,00249	1,0
P15	4796,4086	–0,00600	–0,01097	0,0
P17	4794,8386	0,00637	0,00167	1,0
P19	4793,2507	0,00457	0,00057	1,0
P21	4791,6578	0,00176	–0,00107	1,0
P23	4790,0614	–0,00036	–0,00162	1,0
P25	4788,4613	–0,00191	–0,00135	1,0
P27	4786,8574	–0,00303	–0,00072	1,0
P29	4785,2504	–0,00323	0,00029	1,0
P31	4783,6415	–0,00172	0,00176	1,0
P33	4782,0302	0,00030	0,00155	1,0
P35	4780,4175	0,00287	–0,00158	1,0
R3	4811,3016	–0,00998	–0,01433	0,0
R5	4812,8643	–0,00389	–0,00890	0,0
R7	4814,4257	0,00333	–0,00231	1,0
R9	4815,9807	0,00684	0,00073	1,0
R13	4819,0752	0,00771	0,00173	1,0
R15	4820,6162	0,00728	0,00211	1,0
R17	4822,1483	0,00201	–0,00177	1,0
R19	4823,6814	0,00213	0,00027	1,0
R21	4825,2068	–0,00079	–0,00035	1,0
R23	4826,7270	–0,00409	–0,00123	1,0
R25	4828,2445	–0,00519	–0,00027	1,0
R27	4829,7579	–0,00561	0,00036	1,0
R29	4831,2690	–0,00385	0,00122	1,0
R31	4832,7766	–0,00165	–0,00064	1,0
R33	4834,2823	0,00177	–0,00602	0,0
R39	4838,7915	0,00911	–0,07542	0,0
R41	4840,2943	0,00622	–0,13047	0,0
R43	4841,7919	–0,00908	–0,21753	0,0
СКО, см$^{-1}$		**0,0053**	**0,0014**	

Что является причиной неадекватности модели для случая $J \sim 40$? Ответ был найден при глобальной подгонке центров линий $^{12}C^{16}O_2$ с моделью эффективного гамильтониана, приведенной [7]. Оказалось, что состояние 40002 находится в сильном резонансе с состоянием 21113, причем максимум взаимодействия приходится как раз на $J \sim 40$ [8]. Ясно, что простая полиномиальная модель (2) низкого порядка не в состоянии корректно описать локальное по J взаимодействие. Таким образом, метод робастной подгонки оказался в состоянии выделить область данных, которые неадекватны модели, в то время как с помощью обычного ЛМНК это сделать невозможно.

Заключение

Целью настоящей статьи является привлечение внимания исследователей, занимающихся интерпретациями спектров линейных молекул, к возможности использования МНМ, имеющего большую устойчивость по отношению к выбросам в экспериментальных данных по сравнению с широкоиспользуемым ЛМНК. МНМ осуществляет комбинаторный перебор подвыборок определенной длины из выборки подгоняемых данных и является поэтому достаточно ресурсоемким с точки зрения времени счета.

Кроме моделей, рассмотреных в этой статье, он может быть также использован для извлечения экспериментальных уровней энергии из наблюдаемых частот переходов с использованием фундаментального квантово-механического принципа Рица [см., например, 9].

1. *Albritton D.L., Schmeltekopf A.L., Zare R.H.* Molecular spectroscopy: modern research / Ed. K.N. Rao. New York: Academic Press, 1976. V. 2. P. 1–67.
2. *Rousseeuw P.J., Leroy A.M.* Robust regression and outlier detection. NY: Wiley, 1987. 328 p.
3. *Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В.* Робастность в статистике. М.: Мир, 1989. 512 с.
4. *Ruckstuhl A.F., Stahel W.A., Dressler K.* Robust estimation of term values in high-resolution spectroscopy: application to the $e^3\Sigma_u^+ \rightarrow a^3\Sigma_g^+$ spectrum of T_2 // J. Mol. Spectrosc. 1993. V. 160. N 2. P. 434–445.
5. *Arcas Ph., Arie E., Cuisenier M., Maillard J.P.* The infrared spectrum and molecular constants of CO_2 in the 2 μm region // Can. J. Phys. 1983. V. 61. P. 857–866.
6. *Rothman L.S., Hawkins R.L., Wattson R.B., Gamache R.R.* Energy levels, intensities, and linewidths of atmospheric carbon dioxide bands // J. Quant. Spectrosc. and Radiat. Transfer. 1992. V. 48. N 5/6. P. 537–566.
7. *Tashkun S.A., Perevalov V.I., Teffo J.-L., Rothman L.S., Tyuterev Vl.G.* Global fitting of $^{12}C^{16}O_2$ vibrational-rotational line positions using the effective Hamiltonian approach // J. Quant. Spectrosc. and Radiat. Transfer. 1998. V. 60. N 5. P. 785–801.
8. *Bailly D., Tashkun S.A., Perevalov V.I., Teffo J.-L., Arcas Ph.* CO_2 emission in the 4 μm region: the 21113 → 21103 transition revisited // J. Mol. Spectrosc. 1998. V. 190. N 1. P. 1–6.
9. *Tashkun S.A., Perevalov V.I., Teffo J.-L., Bykov A.D., Lavrentieva N.N.* CDSD-1000, the high-temperature carbon dioxide spectroscopic databank // J. Quant. Spectrosc. and Radiat. Transfer. 2003. N 6/8. V. 82. P. 165–196.

S.A. Tashkun. **Usage of robust methods to estimate spectroscopic parameters of vibrational bands of linear molecules from experimental data.**

In order to obtain fitted spectroscopic parameters of vibrational bands of linear molecules it is suggested to use the least median method instead of the widely used least squares method. The former is known to be robust with respect to presence of outliers in data and provides statistically justified estimates. On the contrary, the latter has zero resistance against outliers that may lead to distorted estimates and deficient models. Importance and usefulness of the least median method is illustrated with an example of deriving fitted spectroscopic parameters of the 40002–01101 band of the $^{12}C^{16}O_2$ molecule.