

АНАЛИЗ УСТОЙЧИВОСТИ СУЩЕСТВЕННО НЕЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ К ПОГРЕШНОСТЯМ В ИЗМЕРЯЕМЫХ ДАННЫХ

Г. И. Рудой

Аннотация

Для восстановления нелинейной зависимости показателя преломления среды от длины волны рассматривается набор индуктивно порожденных моделей с целью выбора оптимальной. Применяется алгоритм индуктивного порождения допустимых существенно нелинейных моделей. Предлагается критерий определения погрешности коэффициентов порожденных суперпозиций, называемый устойчивостью, а также метод оценки устойчивости полученного решения. Приводятся результаты численного моделирования на данных, полученных в ходе эксперимента по определению состава смеси по суммарной дисперсии.

Ключевые слова: *символьная регрессия, нелинейные модели, индуктивное порождение, устойчивость решений, дисперсия прозрачной среды.*

Введение

Для анализа результатов физического эксперимента, как правило, требуется восстановить функциональную зависимость, описывающую соотношение измеряемых величин. При этом необходимо, чтобы эксперт имел возможность интерпретировать полученную зависимость, исходя из соответствующих теоретических моделей. Во многих случаях вид функциональной зависимости заранее известен, либо необходимо сделать выбор между несколькими (также заранее известными) вариантами моделей.

Одним из методов, позволяющих строить интерпретируемые модели, является символьная регрессия [1–5], порождающая, в том числе, и структурно сложные нелинейные модели. Различные приближения сравниваются согласно ошибке на измеряемых данных, при этом оптимизация параметров модели проводится, например, с помощью алгоритма Левенберга-Марквардта [6, 7].

Однако при анализе физического эксперимента важны не только значения самих параметров искомой функциональной зависимости, но и погрешности их определения, обусловленные погрешностями измеряемых в эксперименте величин. Для задачи линейной регрессии соответствующая задача решена в частном случае, когда погрешность определения регрессора пренебрежимо мала, а погрешность определения зависимой переменной во всех экспериментальных точках одинакова [8]. Для более сложного случая нелинейной регрессии и ситуации, когда необходимо учитывать погрешности как регрессора, так и зависимой переменной (которые при этом могут быть разными в различных экспериментальных точках), подобная задача, насколько нам известно, не ставилась.

В настоящей работе метод нелинейной регрессии применяется для восстановления зависимости показателя преломления n от длины волны λ в полосе прозрачности полимера, включающей видимую и ближнюю инфракрасную области спектра. Цель экспериментаторов состояла в том, чтобы по известной дисперсии для каждого полимера с учетом того, что показатель преломления смеси химически инертных полимеров равен взвешенной сумме (с соответствующими весами) показателей преломления компонентов, определить состав смеси по экспериментально определенной зависимости $n(\lambda)$. Другими словами, для

случая двух полимеров, заранее измерив и вычислив зависимости $n_1(\lambda)$ и $n_2(\lambda)$, необходимо экспериментально определить суммарную зависимость $n(\lambda) = \alpha n_1(\lambda) + (1 - \alpha)n_2(\lambda)$ и по ней вычислить коэффициент α , имеющий смысл концентрации первого полимера в смеси.

Поскольку показатели преломления для прозрачных полимеров близкого химического состава различаются незначительно, учет погрешности определения коэффициентов функциональной зависимости $n(\lambda)$ и их связи с погрешностями экспериментального определения длины волны λ и показателя преломления n имеет принципиальное значение. Указанная связь важна еще и потому, что именно она определяет требования к точности и чувствительности измерительной аппаратуры и, следовательно, влияет на стоимость и продолжительность эксперимента.

Обычно в рефрактометрах используются источники широкополосного (непрерывного) спектра, а погрешность выделения конкретной длины волны определяется аппаратной функцией используемого монохроматора (прибора, выделяющего узкий спектральный диапазон) и подробно рассматривается, например, в [9, 10]. В большинстве случаев погрешность λ может быть рассчитана, а также определена экспериментально с использованием узкополосных источников света (лазеров, известных атомных переходов вроде триплета ртути или дублета натрия, и т. д.). Характерная относительная погрешность определения длины волны в рассматриваемой задаче обычно составляет $0.03 \div 0.5\%$, а абсолютная погрешность определения длины волны, как правило, меняется с изменением самой длины волны. Экспериментальная погрешность показателя преломления n зависит от выбранного способа его измерения и, например, при определении n по углу полного внутреннего отражения обусловлена непараллельностью используемых световых пучков, погрешностями в измерении углов и т. д. и составляет от $(1 \div 2) \cdot 10^{-5}$ для приборов высокого класса точности до $(1 \div 10) \cdot 10^{-4}$. Для рассматриваемых в настоящей работе задач существенно то, что величины погрешностей могут считаться известными и, возможно, различными для каждой экспериментальной точки.

В настоящей работе на примере дисперсионной зависимости показателя преломления исследуется возможность применения предложенного в [5] алгоритма восстановления нелинейной регрессии. Результаты его работы сравниваются с результатами применения SVM-регрессии. Кроме того, рассмотрено влияние штрафа за сложность на качество и структурную сложность порождаемых суперпозиций. Формально поставлена задача определения устойчивости регрессионной зависимости от произвольного набора независимых переменных в общем виде, предложен метод оценки устойчивости решения к погрешностям измерений, и изучена зависимость этих характеристик от параметров модели для конкретного случая определения дисперсии полимеров.

В первой части работы поставлена задача восстановления дисперсии полимера и предложено понятие устойчивости суперпозиции, позволяющее учитывать и анализировать зависимость погрешностей различных коэффициентов порожденной суперпозиции от погрешности измеряемых данных. Во второй части вкратце описывается алгоритм [5], используемый для порождения аналитической функции-суперпозиции, аппроксимирующей данные. В третьей части предложен численный метод нахождения устойчивости суперпозиции. В четвертой части приводятся результаты вычислительного эксперимента на реальных данных. Рассматривается два полимера в области прозрачности, для каждого из которых имеется 17 экспериментальных точек, соответствующих величине коэффициента преломления при различных значениях длины волны.

1 Постановка задачи

Задача регрессии. Дана выборка D из ℓ результатов измерений коэффициента преломления для некоторого полимера: $D = \{\lambda_i, n_i \mid i \in \{1, \dots, \ell\}\}$, где λ_i — длина волны, а n_i — измеренный коэффициент преломления в i -ом измерении.

Требуется найти функцию $\hat{f} = \hat{f}(\lambda)$, минимизирующую стандартный функционал потерь в предположении о нормальности случайной ошибки эксперимента:

$$S(f, D) = \sum_{i=1}^{\ell} (f(\lambda_i) - n_i)^2 \rightarrow \min_{f \in \mathcal{F}}, \quad (1)$$

где $D = \{\lambda_i, n_i\}$, а \mathcal{F} — некоторое множество суперпозиций, из которого выбирается оптимальная.

Иными словами,

$$\hat{f}(\lambda) = \hat{f}_D(\lambda) = \arg \min_{f \in \mathcal{F}} S(f, D). \quad (2)$$

Задача оценки устойчивости. Введем в общем виде понятие устойчивости суперпозиции f , характеризующей поведение коэффициентов суперпозиции \hat{f} при небольшом случайном изменении исходной обучающей выборки $D = \{\mathbf{x}_i, y_i\}$, где \mathbf{x}_i — исходное (полученное в ходе эксперимента) признаковое описание i -го объекта, а y_i — соответствующее экспериментально измеренное значение функции, которую требуется восстановить.

Функционал потерь (1) в этом случае выглядит следующим образом:

$$S(f, D) = \sum_{i=1}^{\ell} (f(\mathbf{x}_i) - y_i)^2 \rightarrow \min_{f \in \mathcal{F}}. \quad (3)$$

Условимся также обозначать матрицу плана $X = \|x_{ij}\|$, строками которой являются признаковые описания объектов выборки D . Иными словами, x_{ij} обозначает j -ую компоненту признакового описания i -го объекта.

Рассмотрим вектор параметров $\boldsymbol{\omega}_f = \{\omega_i^f \mid i \in \{1, \dots, l_f\}\}$ некоторой суперпозиции f : $f(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\omega}_f)$. Пусть для некоторой выборки $D = \{\mathbf{x}_i, y_i\}$ и функции f вектор параметров $\hat{\boldsymbol{\omega}}_f(D)$ минимизирует функционал (3) с суперпозицией f , имеющей фиксированную структуру:

$$\hat{\boldsymbol{\omega}}_f(D) = \arg \min_{\boldsymbol{\omega}_f} S(f, D).$$

Пусть также дана матрица стандартных отклонений независимых переменных $\Sigma^{\mathbf{x}} = \|\sigma_{ij}^{\mathbf{x}}\|$, где $\sigma_{ij}^{\mathbf{x}}$ характеризует стандартное отклонение j -ой компоненты признакового описания \mathbf{x}_i i -го объекта обучающей выборки, и вектор стандартных отклонений $\boldsymbol{\sigma}^y$, где σ_i^y характеризует стандартное отклонение зависимой переменной, соответствующей i -му объекту. Рассмотрим выборку \hat{D} , полученную из исходной выборки D добавлением к каждой компоненте реализаций нормально распределенных случайных величин с нулевым матожиданием и соответствующей $\Sigma^{\mathbf{x}}$ и $\boldsymbol{\sigma}^y$ дисперсией:

$$\hat{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}^y) = \{\mathbf{x}_i + \boldsymbol{\xi}_i^{\mathbf{x}}, y_i + \xi_i^y \mid i \in 1, \dots, \ell; \boldsymbol{\xi}_i^{\mathbf{x}} \sim \mathcal{N}(0; \Sigma^{\mathbf{x}}); \xi_i^y \sim \mathcal{N}(0; \sigma_i^y)\}. \quad (4)$$

Для этой выборки \dot{D} найдем оптимальный вектор $\hat{\omega}_f(\dot{D}(\Sigma^{\mathbf{x}}, \sigma_y))$ параметров суперпозиции f , минимизирующий функционал (1):

$$\hat{\omega}_f(\dot{D}(\Sigma^{\mathbf{x}}, \sigma_y)) = \arg \min_{\omega_{f_D} \in R^{|\hat{\omega}_f|}} S(f_D(\cdot, \omega_{f_D}), \dot{D}(\Sigma^{\mathbf{x}}, \sigma_y)). \quad (5)$$

Понятно, что $\hat{\omega}_f(\dot{D}(\Sigma^{\mathbf{x}}, \sigma_y))$ — векторная случайная величина, и, следовательно,

$$\Delta \hat{\omega}_f(\dot{D}(\Sigma^{\mathbf{x}}, \sigma_y)) = \hat{\omega}_f(D) - \hat{\omega}_f(\dot{D}(\Sigma^{\mathbf{x}}, \sigma_y))$$

также векторная случайная величина.

Пусть дано множество \dot{D}_N из N таких выборок, где каждая выборка соответствует отдельным реализациям случайных величин из (4):

$$\dot{D}_N(\Sigma^{\mathbf{x}}, \sigma_y) = \{\dot{D}_1(\Sigma^{\mathbf{x}}, \sigma_y), \dots, \dot{D}_N(\Sigma^{\mathbf{x}}, \sigma_y)\},$$

и пусть $\bar{\sigma}_i$ — эмпирическое стандартное отклонение i -ой компоненты векторной случайной величины $\Delta \hat{\omega}_f(\dot{D}(\Sigma^{\mathbf{x}}, \sigma_y))$ на множестве $\dot{D}_N(\Sigma^{\mathbf{x}}, \sigma_y)$.

Определение 1. *Относительной устойчивостью* (или просто *устойчивостью*) параметра ω_i относительно $\dot{D}_N(\Sigma^{\mathbf{x}}, \sigma_y)$ при исходной обучающей выборке D будем называть следующий вектор из $|\mathbf{x}| + 1$ компонент:

$$\mathbf{T}_f^N(i) = \left\{ \frac{\bar{\sigma}_i}{\hat{\omega}_i}, \dots, \frac{\bar{\sigma}_i}{\hat{\omega}_i}, \frac{\bar{\sigma}_i}{\hat{\omega}_i} \right\}, \quad (6)$$

где $r(\boldsymbol{\alpha}, \mathbf{a}) = r(\frac{\alpha_1}{a_1}, \dots, \frac{\alpha_{|\mathbf{a}|}}{a_{|\mathbf{a}|}})$ — функция, переводящая вектор, составленный из отношений соответствующих компонент векторов $\boldsymbol{\alpha}$ и \mathbf{a} , в скаляр.

Функция r позволяет сопоставить, вообще говоря, различным отношениям стандартного отклонения зависимой переменной y_i i -го объекта обучающей выборки и самого значения y_i единственное число (аналогично и для x_i -й независимой переменной и ее стандартного отклонения). Примерами такой функции могут являться среднее арифметическое всех отношений или максимальное значение отношения, а конкретная функция выбирается экспертом в зависимости от физического смысла задачи. В настоящей работе предлагается выбирать значения стандартных отклонений так, чтобы все значения соответствующих переменных были равны, поэтому функция r может выбирать любой из своих аргументов.

Каждый компонент вектора $\mathbf{T}_f^N(i)$ показывает, как относится стандартное отклонение параметра $\hat{\omega}_i$, нормированное на значение этого параметра, к характерному стандартному отклонению соответствующего элемента признакового описания, нормированного на значение этого элемента. Например, если это отношение больше единицы, то погрешности определения коэффициента растут быстрее погрешностей измерения параметра.

В частности, в искомой задаче восстановления дисперсионной зависимости с учетом неизменной относительной ошибки эксперимента:

$$\mathbf{T}_f(i) = \left\{ \frac{\bar{\sigma}_i}{\hat{\omega}_i}, \frac{\bar{\sigma}_i}{\hat{\omega}_i} \right\}.$$

Матрицу, столбцами которой являются векторы $\mathbf{T}_f(i) \mid i \in \{1, \dots, l_f\}$, будем называть устойчивостью функции f и обозначать \mathbb{T}_f .

Требуется исследовать зависимость устойчивости $\mathbb{T}_{\hat{f}}$ относительно σ_n и σ_λ .

2 Алгоритм индуктивного порождения суперпозиций

Опишем предложенный в [5] алгоритм.

Пусть задано некоторое множество $G = \{g_1, \dots, g_k\}$ порождающих функций. Набор суперпозиций $\mathcal{F} = \{f\}$ инициализируется случайными суперпозициями функций $g \in G$. Суперпозиции из \mathcal{F} содержат как свободные переменные, соответствующие компонентам вектора-описания объектов из генеральной совокупности, так и константы, которые оптимизируются на каждом шаге алгоритмом Левенберга-Марквардта согласно введенному функционалу потерь (1). Также на каждой итерации над суперпозициями выполняется набор модифицирующих операций с целью улучшения качества Q_f суперпозиций.

Качество Q_f суперпозиции f вычисляется по совокупности точности приближения экспериментальных данных и структурной сложности суперпозиции по следующей формуле:

$$Q_f = \frac{1}{1 + S(f)} \left(\alpha + \frac{1 - \alpha}{1 + \exp(C_f - \tau)} \right), \quad (7)$$

где:

$S(f)$ — значение функционала потерь (1) на данной выборке D ;

C_f — сложность суперпозиции, соответствующая количеству элементарных функций, свободных переменных и констант;

$\alpha - 0 \ll \alpha < 1$, характеризует влияние штрафа за сложность на качество суперпозиции (большие значения α отдают предпочтение более точным моделям, а меньшие — более простым);

τ — коэффициент, характеризующий желаемую сложность модели.

Второй множитель в (7) выполняет роль штрафа за слишком большую сложность суперпозиции, что подавляет эффект переобучения и позволяет получать более простые суперпозиции ценой большей ошибки на обучающих данных при большей обобщающей способности.

Отметим, что параметры α и τ выбираются экспертом.

Таким образом, исходная задача минимизации функционала (1) заменяется на задачу минимизации функционала (7):

$$Q_f = \frac{1}{1 + S(f)} \left(\alpha + \frac{1 - \alpha}{1 + \exp(C_f - \tau)} \right) \rightarrow \min_{f \in \mathcal{F}}. \quad (8)$$

3 Метод исследования стабильности решения

Для оценки устойчивости $T_{\hat{f}}$ решения \hat{f} задачи (7), как предложено выше, фиксируется структурный вид суперпозиции \hat{f} и исследуется зависимость стандартного отклонения ее коэффициентов как функция стандартного отклонения нормально распределенной случайной добавки в исходных данных.

Иными словами, выбираются значения σ_λ и σ_n , затем для этих значений генерируется выборка $\hat{D}(\sigma_n, \sigma_\lambda)$ согласно (4). Для этой выборки вычисляются значения коэффициентов суперпозиции \hat{f} , минимизирующие функционал (1) согласно (5), методом Левенберга-Марквардта.

Данная процедура для фиксированной пары σ_λ и σ_n повторяется до достижения некоторого критерия останова (например, по количеству итераций), после которого и рассчитывается $\mathbb{T}_{\hat{f}}$.

Повторяя описанные выше шаги для различных σ_λ и σ_n , можно оценить зависимость стандартного отклонения коэффициентов суперпозиции от стандартного отклонения шума.

Из физических соображений ясно, что гладкая зависимость означает устойчивое в физическом смысле решение, тогда как отклонения от гладкости означают ту или иную ошибку в суперпозиции и могут являться свидетельством переобучения: чем меньше коэффициенты зависят от случайных шумов в данных, тем больше обобщающая способность.

Кроме того, сравнение различных суперпозиций может также производиться по критерию устойчивости в дополнение к сравнению по сложности и по значению функционала (1). В ряде практических приложений критерий устойчивости может иметь приоритетное значение.

4 Вычислительный эксперимент

В вычислительном эксперименте используются данные, полученные в ходе изучения возможности определения состава смеси прозрачных полимеров по суммарной дисперсионной зависимости, если известна экспериментальная зависимость дисперсии для каждого конкретного полимера. Рассматривается два полимера, для каждого из которых имеется 17 экспериментальных точек, соответствующих коэффициенту преломления при разных значениях длины волны. Значения приведены в таблице 1.

| λ , нм | Первый полимер | Второй полимер |
|----------------|----------------|----------------|
| 435.8 | 1.36852 | 1.35715 |
| 447.1 | 1.36745 | 1.35625 |
| 471.3 | 1.36543 | 1.35449 |
| 486.1 | 1.36446 | 1.35349 |
| 501.6 | 1.36347 | 1.35275 |
| 546.1 | 1.36126 | 1.35083 |
| 577.0 | 1.3599 | 1.34968 |
| 587.6 | 1.3597 | 1.34946 |
| 589.3 | 1.35952 | 1.34938 |
| 656.3 | 1.35767 | 1.34768 |
| 667.8 | 1.35743 | 1.34740 |
| 706.5 | 1.35652 | 1.34664 |
| 750 | 1.35587 | 1.34607 |
| 800 | 1.35504 | 1.34544 |
| 850 | 1.3544 | 1.34487 |
| 900 | 1.35403 | 1.34437 |
| 950 | 1.35364 | 1.34407 |

Таблица 1: Экспериментальные значения коэффициентов преломления.

Предполагается, что дисперсионные свойства полимеров описываются одной и той же функциональной зависимостью, так как подчиняются одним и тем же физическим закономерностям. Поэтому сначала получена суперпозиция \hat{f} , минимизирующая (7) для

первого полимера, а затем для каждого из полимеров находятся соответствующие векторы параметров $\hat{\omega}_f$ и оценивается устойчивость полученного решения.

Разделение на обучающую и контрольную выборку не производилось, однако переобучения удастся избежать и без такого разделения, опираясь целиком на штраф за сложность.

Из физических соображений следует [11], что зависимость коэффициента преломления n от длины волны λ должна выражаться суммой четных степеней длины волны, поэтому множество элементарных функций состоит из стандартных операций сложения и умножения:

$$g_1(x_1, x_2) = x_1 + x_2,$$

$$g_2(x_1, x_2) = x_1 x_2,$$

а также из функции

$$g_3(\lambda, p) = \frac{1}{\lambda^{2p}}.$$

В ходе вычислительного эксперимента константы, меньшие 10^{-7} , заменялись на 0.

В результате применения описанного выше алгоритма со значениями $\alpha = 0.05$, $\tau = 10$ получена следующая суперпозиция (константы округлены до пятой значащей цифры):

$$f(\lambda) = 1.3495 + \frac{3.5465 \cdot 10^3}{\lambda^2} + \frac{2.023 \cdot 10^3}{\lambda^4}, \quad (9)$$

со сложностью 13, среднеквадратичной ошибкой $2.4 \cdot 10^{-8}$ и значением $Q_f \approx 0.095$. Длины волн выражаются в нанометрах.

Отметим, что обычно в приложениях учитывают только квадратичный член, а более высокими степенями пренебрегают. Величина поправки, вносимой в результирующее значение суперпозиции последним слагаемым, указывает на полное согласие полученных результатов с принятой практикой.

Влияние штрафа за сложность. Исследуем, как влияет добавление нечетных степеней на результат решения задачи (8), заменив функцию g_3 в порождающем наборе на

$$g_3(\lambda, p) = \frac{1}{\lambda^p}.$$

Следует отметить, что при тех же $\alpha = 0.05$ и $\tau = 10$ результирующей функцией остается (9).

Увеличим τ до 30. Получим следующую формулу (константы округлены до третьей значащей цифры):

$$n(\lambda) = 1.34 + \frac{11.6}{\lambda} + \frac{17.37}{\lambda^2} + \frac{0.0866}{\lambda^3} + \frac{2.95 \cdot 10^{-4}}{\lambda^4} + \frac{8.54 \cdot 10^{-7}}{\lambda^5}, \quad (10)$$

сложность которой составляет 31, и для которой среднеквадратичная ошибка на выборке составляет $\approx 3.9 \cdot 10^{-9}$, а значение $Q_f \approx 0.31$.

Иными словами, при большей желаемой сложности, регулируемой параметром τ , выигрывает более сложная (а в данном случае и физически некорректная) модель, которая лучше описывает экспериментальные данные.

Как и следовало ожидать, чрезмерное увеличение τ ведет к переобучению.

SVM. В качестве базового алгоритма используется SVM-регрессия с RBF-ядром [12]. Параметр γ ядра подбирался по методу скользящего контроля, наилучшим результатом является комбинация из 15 опорных векторов с $\gamma \approx 2 \cdot 10^{-6}$, при этом среднеквадратичная ошибка при кросс-валидации с тестовой выборкой, содержащей по 2 объекта, составляет $8.96 \cdot 10^{-8}$. Однако, проинтерпретировать полученную решающую функцию не представляется возможным.

Исследование стабильности решения. Для оценки стабильности решения фиксировалась формула (9) и исследовалась зависимость стандартного отклонения ее коэффициентов от стандартного отклонения нормально распределенного случайного шума в исходных данных описанным выше методом. Критерием останова в нем являлось достижение 10000 итераций для каждой пары $(\sigma_\lambda, \sigma_n)$.

В таблице 2 представлены поверхности уровня дисперсии для первого, второго и третьего коэффициентов каждого из полимеров соответственно.

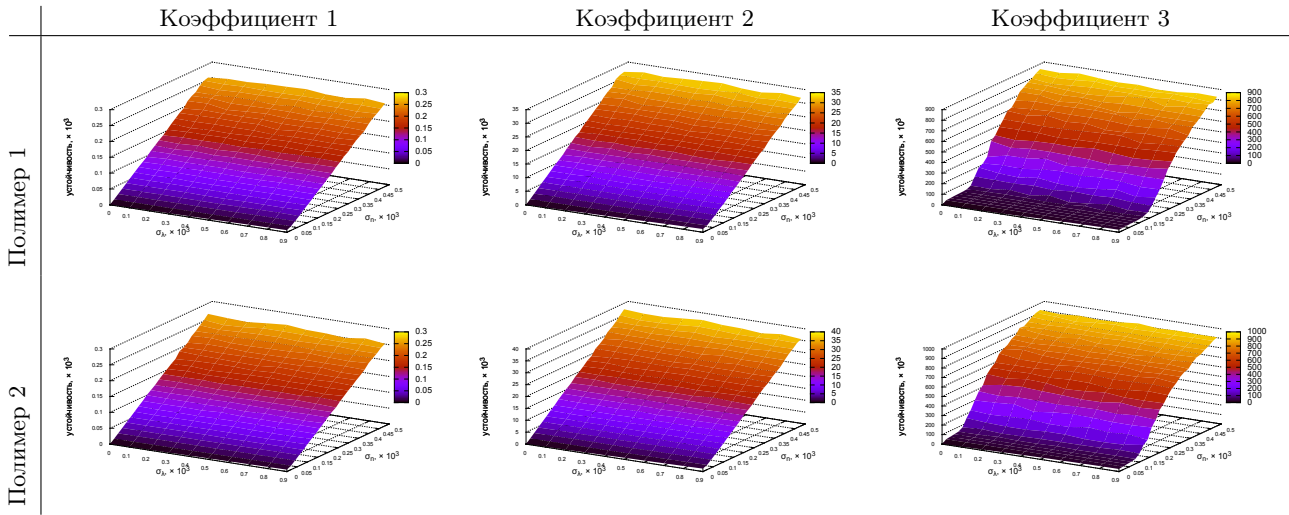


Таблица 2: Поверхности дисперсии для формулы (9).

| Полимер | Коэффициент 1 | Коэффициент 2 | Коэффициент 3 | MSE |
|----------|----------------------|----------------------|---------------------|---------------------|
| 1 | 1.34946 | 3558.95 | 1924.33 | $2.2 \cdot 10^{-8}$ |
| 2 | 1.34047 | 3118.84 | 1578.59 | $1.4 \cdot 10^{-8}$ |
| Разность | $6.71 \cdot 10^{-3}$ | $1.41 \cdot 10^{-1}$ | $2.2 \cdot 10^{-1}$ | |

Таблица 3: Значения коэффициентов для формулы (9) и их относительная разность.

| Коэфф. | $\frac{\sigma_\lambda}{\lambda} = 2 \cdot 10^{-4}; \frac{\sigma_n}{n} = 2 \cdot 10^{-5}$ | $\frac{\sigma_\lambda}{\lambda} = 6 \cdot 10^{-4}; \frac{\sigma_n}{n} = 6 \cdot 10^{-5}$ | $\frac{\sigma_\lambda}{\lambda} = 9 \cdot 10^{-4}; \frac{\sigma_n}{n} = 2 \cdot 10^{-4}$ |
|--------|--|--|--|
| 1 | $1.22 \cdot 10^{-5}$ | $3.59 \cdot 10^{-5}$ | $1.19 \cdot 10^{-4}$ |
| 2 | $1.48 \cdot 10^{-3}$ | $4.38 \cdot 10^{-3}$ | $1.44 \cdot 10^{-2}$ |

Таблица 4: Значения стандартного отклонения для коэффициентов формулы (9) для первого полимера в зависимости от относительных дисперсий.

Из графиков видно, что от шума, накладываемого на значения длины волны, дисперсия значений первого и второго коэффициентов практически не зависит в достаточно

широком диапазоне точности определения длины волны, представляющем практический интерес. В то же время дисперсия значений первого коэффициента зависит от дисперсии шума коэффициента преломления практически линейно, тогда как для второго коэффициента после некоторого характерного значения зависимость становится слабой.

Физическая интерпретация этих результатов — при построении прибора для измерения дисперсии такого типа полимеров в их полосе прозрачности значительное внимание следует уделять точности измерения коэффициента преломления, тогда как измерения длины волны могут быть неточны вплоть до нескольких нанометров. Кроме того, предложенный метод прямо указывает, на каких интервалах шума каким будет выигрыш в точности определения параметров регрессионной модели в зависимости от увеличения точности измерений.

Принципиально важно, что значения стандартного отклонения параметров регрессионной модели существенно меньше разности между самими значениями этих параметров по порядку величины (см. таблицы 3 и 4), что означает, в частности, что полимеры могут быть различены даже не очень точным рефрактометром.

Стабильность некорректного решения. Аналогично исследуем стабильность решения (10). Приведем только графики зависимости первых трех коэффициентов, см. таблицу 5.

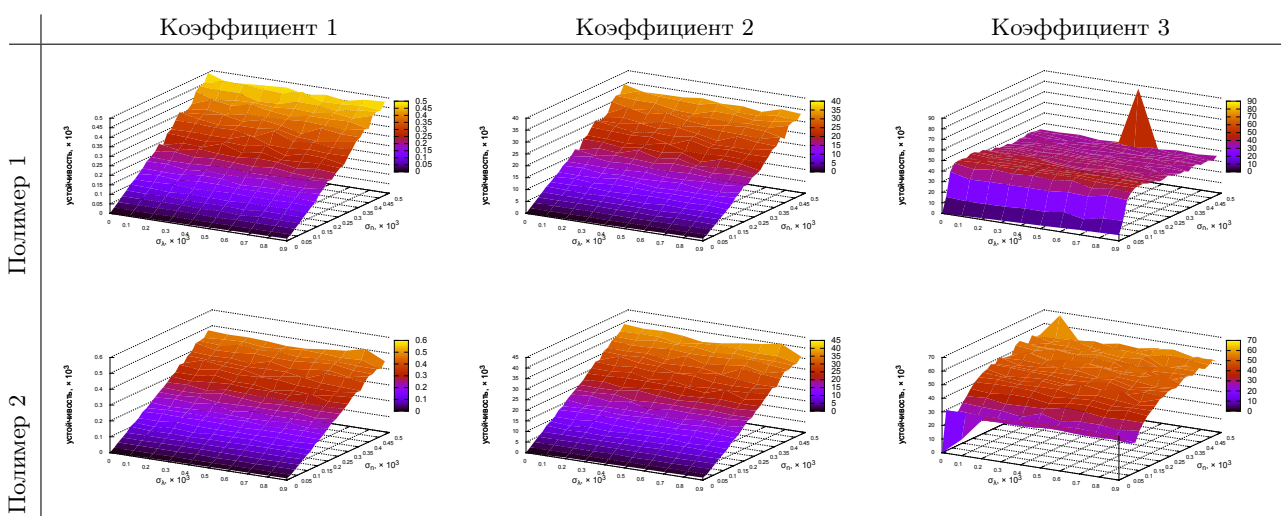


Таблица 5: Поверхности дисперсии для формулы (10).

Из графиков видно, что в случае формулы (10) дисперсия соответствующих параметров существенно превышает таковую для (9). В частности, второй, третий и четвертый коэффициенты имеют дисперсию, на порядки превышающую характерные значения самих коэффициентов.

Данные результаты свидетельствуют о переобучении, и что полученная модель не может быть использована для надежного приближения экспериментальных данных ввиду большой чувствительности к шумам.

5 Сходимость к классическому случаю

Рассмотрим случай, когда зависимость линейна:

$$y = ax + b,$$

и с учетом ошибок измерений представима в виде

$$y_i = ax_i + b + \xi_i \mid i \in \{1, \dots, n\},$$

где ошибки ξ_i независимы, $E(\xi_i) = 0$; $D(\xi_i) = \sigma^2$ [8]. То есть, рассматривается случай независимости ошибки измерения от точки измерения, при этом независимая переменная измеряется точно.

Перейдем к представлению

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

для которого, согласно [8], случайные величины a и b независимы и нормально распределены, и, кроме того, их дисперсии выражаются известными соотношениями:

$$D(a) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (11)$$

$$D(b) = \frac{\sigma^2}{n}. \quad (12)$$

Рассмотрим, насколько результаты предложенного метода отличаются от значений, полученных согласно (11) и (12). Для этого исследуем зависимость относительной разности между этими значениями и эмпирическими значениями устойчивости от числа итераций N :

$$\delta_1 = \frac{|\mathbf{T}_y^N(1) - D(a)|}{D(a)},$$
$$\delta_2 = \frac{|\mathbf{T}_y^N(2) - D(b)|}{D(b)}.$$

Соответствующие графики для функции $y = 2x + 1 + \xi_i$ на интервале $x \in [0; 10]$ при различном количестве порожденных точек (10 и 50) приведены на рис. 1 для значений $D(\xi_i)$, равных 1 и 10. По оси x отложено число итераций в тысячах (до 10^7), по y — значения δ_1 и δ_2 .

На рис. 2 и 3 приведены отдельные графики, соответствующие интервалам от 0 до $5 \cdot 10^5$ итераций и от 10^6 до 10^7 итераций соответственно. Рис. 2 демонстрирует сходимость в начале работы алгоритма, в то время как рис. 3 показывает характер сходимости при достаточно больших N .

Из графиков видно, что значения разности стабилизируются в районе $(1.5 \div 3) \cdot 10^6$ итераций и не демонстрируют явной зависимости от числа точек или дисперсии погрешности.

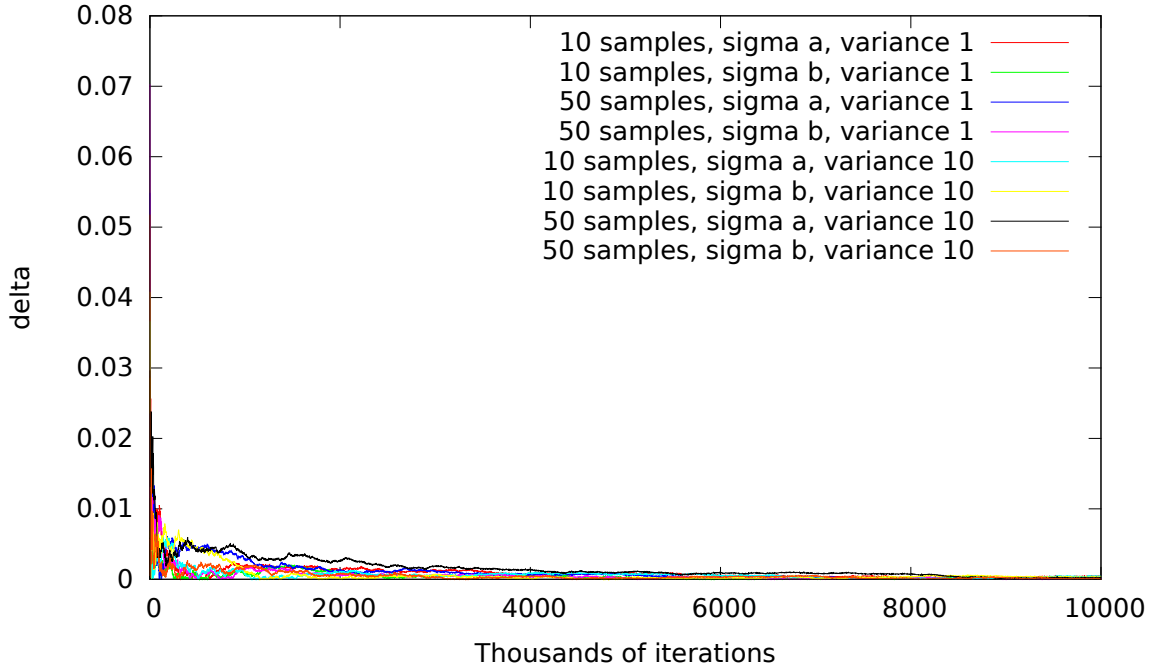


Рис. 1: График зависимости δ от числа итераций (от 0 до 10^7 итераций).

6 Заключение

Предложенный в [5] алгоритм позволяет получить интерпретируемую аналитическую формулу, описывающую зависимость коэффициента преломления среды от длины волны. Введенный штраф за сложность позволяет избежать переобучения без использования методов вроде скользящего контроля, и, таким образом, отпадает необходимость в контрольной выборке.

Хотя другие алгоритмы, такие как SVM-регрессия, могут демонстрировать более высокое качество приближения данных, их результаты неинтерпретируемы и не защищены от переобучения «по построению», поэтому требуют разделения выборки на обучающую и контрольную. Кроме того, их структурные параметры так же требуют оценки по методам вроде кросс-валидации.

Предложенный в настоящей работе метод оценки стабильности решения позволяет исследовать вклад различных членов результирующей суперпозиции и зависимость изменения этих членов от случайных шумов во входных данных. В частности, в прикладных областях данный метод позволяет выявить, какие именно элементы признакового описания объектов в генеральной совокупности наиболее чувствительны к шуму. Кроме того, для корректных с экспертной точки зрения решений оказывается, что они стабильны, в то время как некорректные результаты нестабильны.

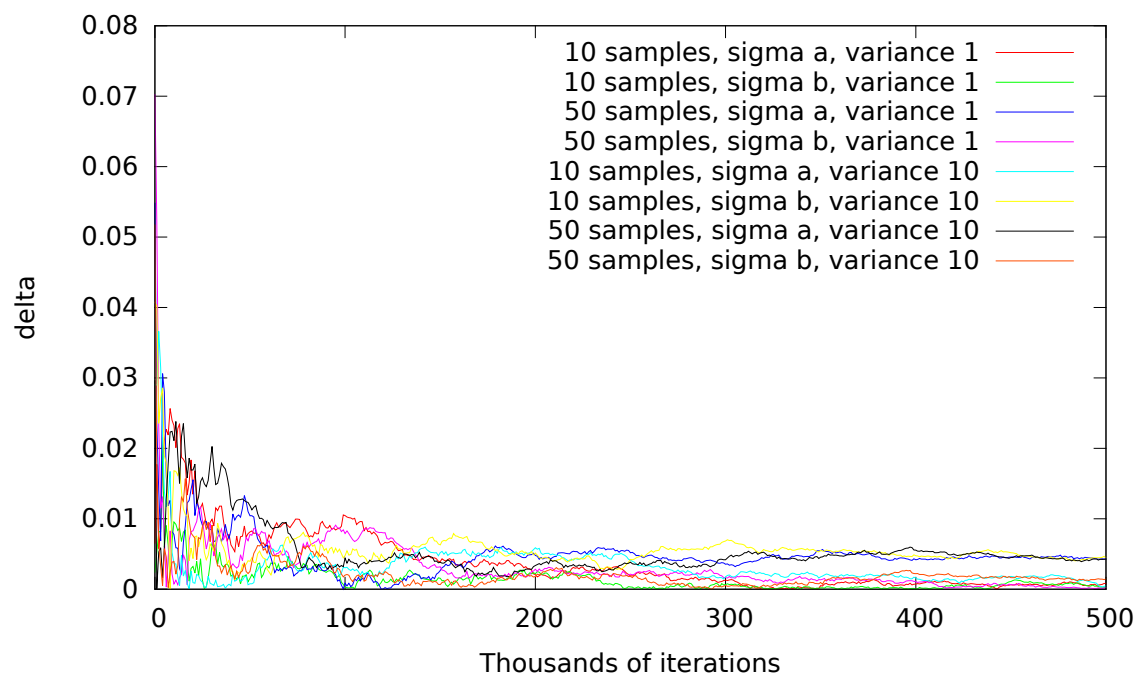


Рис. 2: График зависимости δ от числа итераций (от 0 до $5 \cdot 10^5$ итераций).

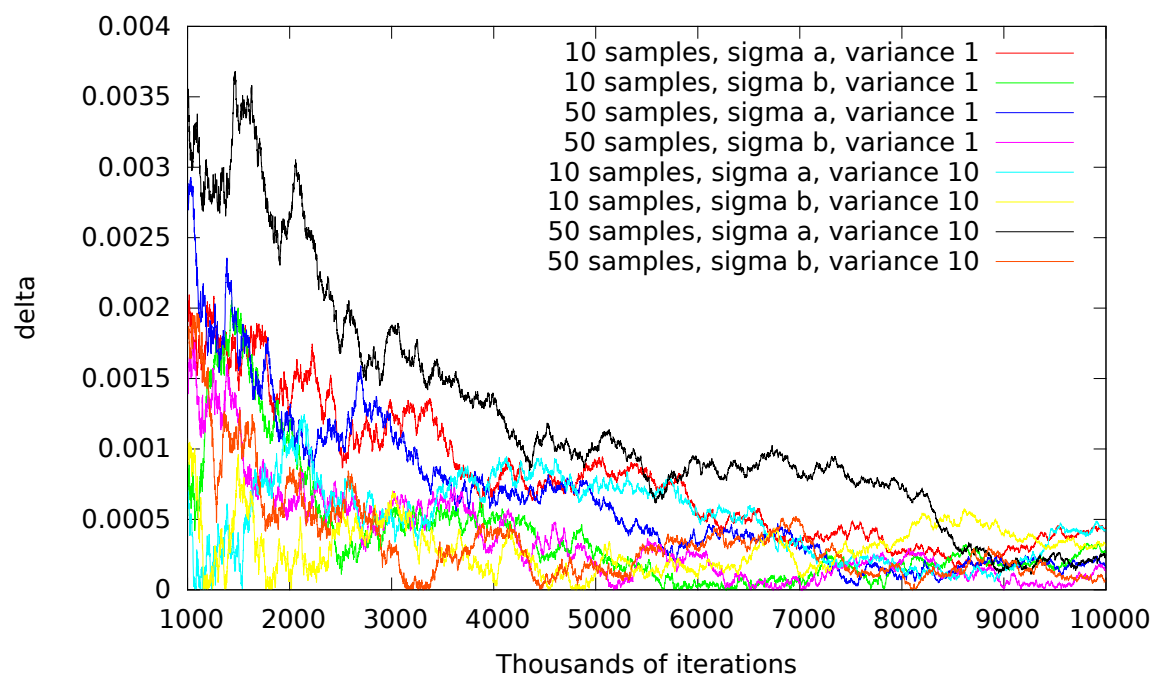


Рис. 3: График зависимости δ от числа итераций (от 10^6 до 10^7 итераций).

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] Marquardt, D. W.: *An algorithm for least-squares estimation of non-linear parameters*. Journal of the Society of Industrial and Applied Mathematics, 11(2):431–441, 1963.
- [7] More, J. J.: *The Levenberg-Marquardt algorithm: Implementation and theory*. In G.A. Watson, Lecture Notes in Mathematics 630, pages 105–116. Springer-Verlag, Berlin, 1978. Cited in Åke Björck’s bibliography on least squares, which is available by anonymous ftp from math.liu.se in `pub/references`.
- [8] Ватутин, В. А., Ивченко, Г. И., Медведев, Ю. И., и Чистяков, В. П.: *Теория вероятностей и математическая статистика в задачах*. Дрофа, 3 редакция, 2005.
- [9] Малышев, В. И.: *Введение в экспериментальную спектроскопию*. Наука, 1979.
- [10] Зайдель, И. Н.: *Техника и практика спектроскопии*. Наука, 1972.
- [11] Серова, Н. В.: *Полимерные оптические материалы*. Научные основы и технологии, 2011.
- [12] Вапник, В. Н.: *Восстановление зависимостей по эмпирическим данным*. М.: Наука, 1979.