ОПТИМИЗАЦИЯ ПАРАМЕТРОВ СУЩЕСТВЕННО НЕЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ С УЧЕТОМ ПОГРЕШНОСТИ КАК ЗАВИСИМЫХ, ТАК И НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ В ОБУЧАЮЩЕЙ ВЫБОРКЕ

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 и модифицированный алгоритм Левенберга-Марквардта. 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.

Стандартные методы нахождения оптимальных параметров существенно нелинейных моделей (например, алгоритм Левенберга-Марквардта), основываются на предположении о точно измеряемых величинах в обучающей выборке. Однако, в настоящей работе данное предположение не рассматривается, поэтому использование подобных методов представляется неточным. Задача оценки параметров линейных функций рассматривается, например, в [11] для случая неточно измеренных зависимых переменных в линейной регрессии. В частности, показывается, что для случая гетероскедастичности ошибок при отсутствующей корреляции предложенный в [11] метод вырождается во взвешенную сумму квадратов, где веса обратно пропорциональны дисперсиям ошибок. В связи с этим представляет интерес дальнейшая проработка методов оценки параметров существенно нелинейных моделей при наличии ошибок в измерениях.

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. Предложена модификация алгоритма Левенберга-Марквардта, основывающаяся на предположении об отсутствии автокорреляции ошибок, а также о наличии ошибок измерения независимых переменных. 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: $D = \{\lambda_i, n_i \mid i \in \{1, \ldots, \ell\}\}$, 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:

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

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

Иными словами,

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

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:

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

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 $\boldsymbol{\omega}_f = \{\omega_i^f \mid i \in \{1, \ldots, l_f\}\}$ 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 $\acute{D}$, 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:

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

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

$$\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)) = \arg\min_{\boldsymbol{\omega}_{f_D} \in R^{|\hat{\boldsymbol{\omega}}_f|}} S(f_D(\cdot, \boldsymbol{\omega}_{f_D}), \acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)). \tag{5}$$

Понятно, что $\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y))$ — векторная случайная величина, и, следовательно,

$$\Delta\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)) = \hat{\boldsymbol{\omega}}_f(D) - \hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y))$$

также векторная случайная величина.

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

$$\acute{\mathcal{D}}_N(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y) = \{\acute{D}_1(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y), \ldots, \acute{D}_N(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)\},$$

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

Определение 1. *Относительной устойчивостью* (или просто *устойчивостью*) параметра ω_i относительно j-ой компоненты векторного описания при исходной обучающей выборке D и параметрах $\acute{\mathcal{D}}_N(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)$ будем называть следующую величину:

$$T_{ij}(f) = \begin{cases} \dfrac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{r\left(\left\{\frac{\sigma_{kj}^{\mathbf{x}}}{x_{kj}}\right\}_{k\in\{1,\ldots,\ell\}}\right)} & j \le |\mathbf{x}| \\ \dfrac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{r\left(\left\{\frac{\sigma_k^y}{y_k}\right\}_{k\in\{1,\ldots,\ell\}}\right)} & j = |\mathbf{x}| + 1 \end{cases} \tag{6}$$

где r — некоторая функция, переводящая вектор, составленный из соответствующих дробей, в единственный 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.

Величина $T_{ij}(f)$ показывает, как относится стандартное отклонение параметра $\hat{\omega}_i$, нормированное на значение этого параметра, к характерному стандартному отклонению j-го элемента признакового описания, нормированного на значение этого элемента. Например, если это отношение больше единицы, то погрешности определения коэффициента $\hat{\omega}_i$ растут быстрее погрешностей измерения параметра.

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

$$T_{i0}(f) = \frac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{\frac{\sigma_n}{n}},$$

$$T_{i1}(f) = \frac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{\frac{\sigma_\lambda}{\lambda}}.$$

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

Отметим, что для простых случаев одномерной регрессии (каковым является исследование дисперсионной зависимости полимеров), возможно, более наглядным является исследование *абсолютной стабильности* — зависимости величины $\frac{\overline{\sigma}_i}{\hat{\omega}_i}$ от нормализованных вариаций в обучающей выборке, $\frac{\sigma_n}{n}$ и $\frac{\sigma_\lambda}{\lambda}$ соответственно. В этом случае возможно проиллюстрировать искомую зависимость интерпретируемым графиком.

2 МЕТОД ОПТИМИЗАЦИИ ПАРАМЕТРОВ СУЩЕСТВЕННО НЕЛИНЕЙНЫХ СУПЕРПОЗИЦИЙ

Требуется минимизировать функционал (1) в предположении об измерении как зависимых, так и независимых переменных с погрешностями. Кроме того, учитываются следующие требования:

- При определении минимального расстояния от каждой точки обучающей выборки до регрессионной кривой учитывается наличие погрешностей как зависимых, так и независимых переменных.
- Метод оптимизации функционала является развитием имеющихся хорошо развитых итеративных методов оптимизации, вроде алгоритма Левенберга-Марквардта.

Рассматривается случай зависимости $y = f(x, \boldsymbol{\omega})$, где x — скалярная переменная, $\boldsymbol{\omega}$ — вектор параметров модели f.

Пусть стандартное отклонение измерения y_i — σ_{y_i}, стандартное отклонение измерения x_i — σ_{x_i}.

Определим расстояние от точки (x_i, y_i) до кривой $y = f(x, \boldsymbol{\omega})$ следующим образом:

$$\rho^2(x_i, y_i, f|\boldsymbol{\omega}) = \arg\min_x \left(\frac{x_i - x}{\sigma_{x_i}}\right)^2 + \left(\frac{y_i - f(x, \boldsymbol{\omega})}{\sigma_{y_i}}\right)^2.$$

К такому определению приводят следующие экспертные соображения:

- Члены суммы, имеющие большее стандартное отклонение, должны иметь меньший вес при оптимизации. Иными словами, к неточно определенным данным подстраиваться необходимо не настолько строго, как к более точным.
- Числители имеют ту же размерность, что и соответствующие знаменатели, поэтому получается сумма двух безразмерных чисел, что корректно с точки зрения анализа размерностей.

Линеаризуем f в окрестности x_i:

$$f(x, \boldsymbol{\omega}) = f(x_i, \boldsymbol{\omega}) + \frac{\partial f}{\partial x}(x_i, \boldsymbol{\omega})(x - x_i) + \mathrm{o}(x - x_i).$$

Получаем следующее выражение для расстояния:

$$\rho^2(x_i, y_i, f|\boldsymbol{\omega}) = \arg\min_x \Big(\frac{x_i - x}{\sigma_{x_i}}\Big)^2 + \Big(\frac{y_i - f(x_i, \boldsymbol{\omega}) - \frac{\partial f}{\partial x}(x_i, \boldsymbol{\omega})(x - x_i)}{\sigma_{y_i}}\Big)^2.$$

Минимизируя, получим выражение для минимального расстояния:

$$\rho^2(x_i, y_i, f|\boldsymbol{\omega}) = \frac{(y_i - f(x_i, \boldsymbol{\omega}))^2}{\sigma_{y_i}^2 + (\frac{\partial f}{\partial x}(x_i, \boldsymbol{\omega}))^2 \sigma_{x_i}^2}.$$

Отметим, что числитель в этом выражении равен соответствующему i-ой точке члену в функционале (1).

Таким образом, получаем следующий функционал, который необходимо минимизировать:

$$\dot{S}(f, D) = \sum_{i=1}^{\ell} \frac{(y_i - f(x_i, \boldsymbol{\omega}))^2}{\sigma_{y_i}^2 + (\frac{\partial f}{\partial x}(x_i, \boldsymbol{\omega}))^2 \sigma_{x_i}^2}. \tag{7}$$

Для решения задачи

$$\hat{\omega} = \arg\min_{\boldsymbol{\omega}} \sum_{i=1}^{\ell} \frac{(y_i - f(x_i, \boldsymbol{\omega}))^2}{\sigma_{y_i}^2 + (\frac{\partial f}{\partial x}(x_i, \boldsymbol{\omega}))^2 \sigma_{x_i}^2}$$

предлагается следующая итеративная процедура:

1. Для каждой точки (x_i, y_i) в обучающей выборке оптимизируемая функция $f(x, \boldsymbol{\omega})$ линеаризуется в окрестности этой точки. Иными словами, рассматривается частная производная $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial x}(x, \boldsymbol{\omega})$.
2. Каждый i-й член функционала (1) нормируется на соответствующую величину

   $$\sigma_{y_i}^2 + (\frac{\partial f}{\partial x}(x_i, \boldsymbol{\omega}))^2 \sigma_{x_i}^2,$$

   зависящую, в том числе, от текущего значения вектора параметров $\boldsymbol{\omega}$. Таким образом, получается новый функционал $\dot{S}(\boldsymbol{\omega})$.
3. Выполняется одна итерация стандартного алгоритма, оптимизирующая функционал $\dot{S}$, в результате получается новое приближение вектора параметров $\boldsymbol{\omega}$.

4. Шаги 1-3 выполняются до тех пор, пока не будет выполняться какое-либо условие останова (например, по числу итераций или по норме изменения вектора $\boldsymbol{\omega}$).

Подробное обоснование таких свойств предложенной процедуры, как сходимость, корректность и статистическая оптимальность не является предметом настоящей работы.

Отметим, что на шаге 3 предложенной процедуры при достаточной гладкости частной производной $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial x}(x, \boldsymbol{\omega})$ представляется возможным выполнять сразу несколько итераций исходного немодифицированного алгоритма Левенберга-Марквардта без пересчёта параметров на шаге 1.

Кроме того, предложенный алгоритм естественным образом обобщается на случай функций f многих переменных.

3 МЕТОД ИССЛЕДОВАНИЯ СТАБИЛЬНОСТИ РЕШЕНИЯ

Для оценки устойчивости $\mathbb{T}_{\hat{f}}$ решения $\hat{f}$ задачи (1), 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 $\acute{D}(\sigma_n, \sigma_\lambda)$ 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.

Таблица 1: Экспериментальные значения коэффициентов преломления.

λ, нм	Полимер 1	Полимер 2
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

Для получения суперпозиций используется алгоритм, предложенный в [5] и решающий следующую задачу:

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

где:

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

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

α — $0 \ll \alpha < 1$, 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.

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 (8) для первого полимера, а затем для каждого из полимеров находятся соответствующие векторы параметров $\hat{\boldsymbol{\omega}}_{\hat{f}}$ и оценивается устойчивость полученного решения.

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

Из физических соображений следует [12], 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):

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

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):

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

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 [13]. 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) в виде

$$f(\lambda) = \omega_1 + \frac{\omega_2}{\lambda^2} + \frac{\omega_3}{\lambda^4},$$

и исследовалась зависимость стандартного отклонения ее коэффициентов ω_1, ω_2 и ω_3 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).

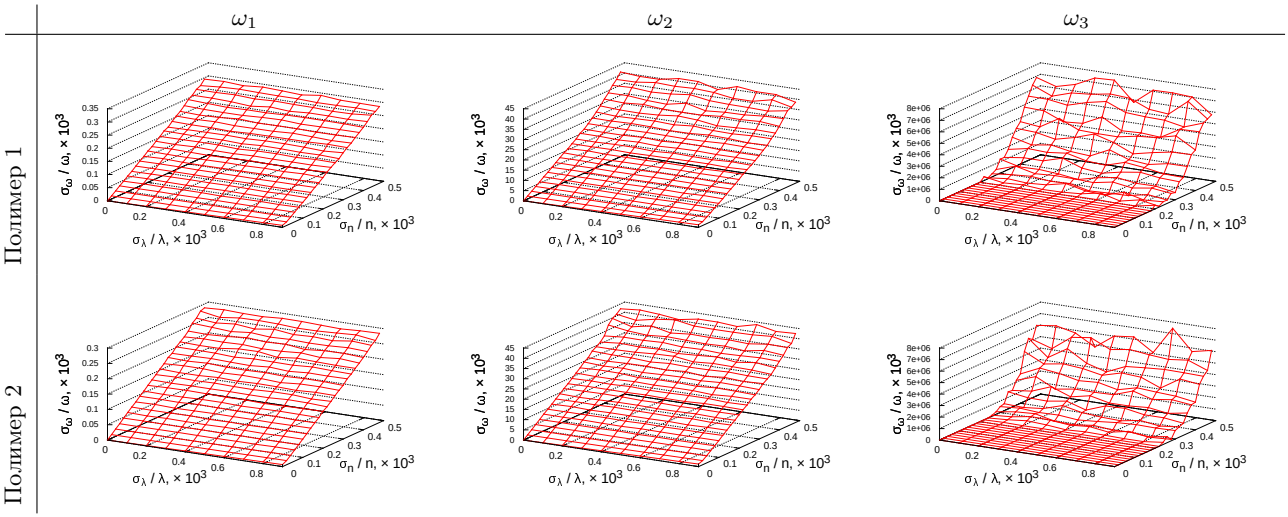


Таблица 3: Значения коэффициентов для формулы (9) и их относительная разность.

	ω_1	ω_2	ω_3	MSE
Полимер 1	1.34946	3558.95	1924.33	$2.2 \cdot 10^{-8}$
Полимер 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 4: Значения стандартного отклонения для коэффициентов формулы (9) для первого полимера в зависимости от относительных дисперсий.

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

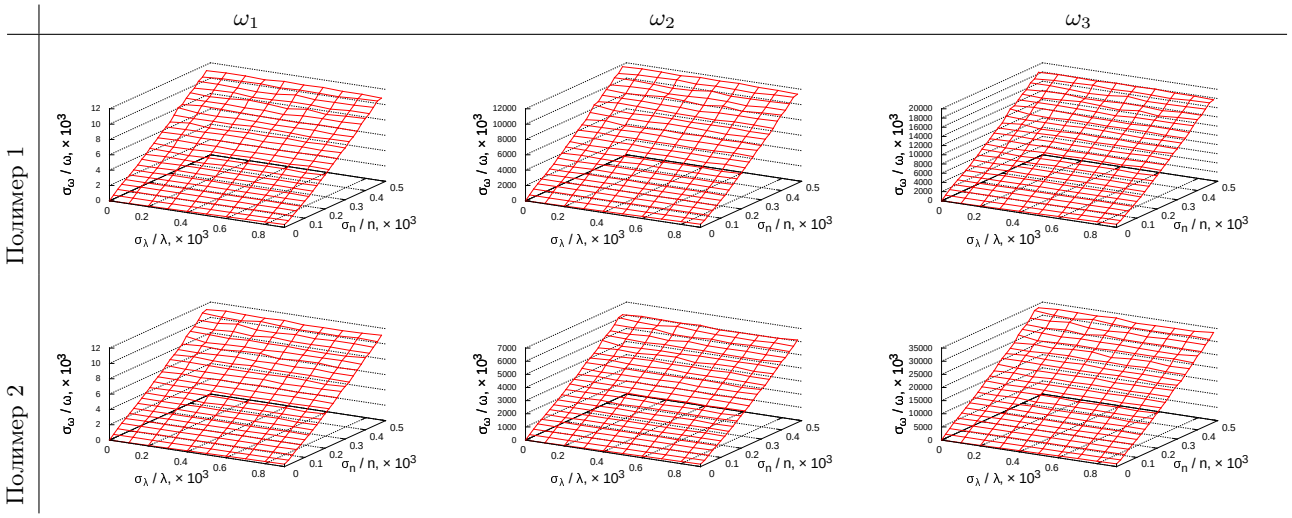
Из графиков видно, что от шума, накладываемого на значения длины волны, дисперсия значений первого и второго коэффициентов практически не зависит в достаточно широком диапазоне точности определения длины волны, представляющем практический интерес. В то же время дисперсия значений первого коэффициента 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.

Использование модифицированного алгоритма Левенберга-Марквардта. Сравним предложенный модифицированный алгоритм Левенберга-Марквардта, предложенный в настоящей работе, с немодифицированным. Для этого выполним оптимизацию параметров порождённой модели (9) на исходных данных для первого полимера. Получившиеся результаты в сравнении с результатами для немодифицированного алгоритма приведены в таблице 6.

Таблица 6: Значения коэффициентов для формулы (9), полученные согласно различным алгоритмам оптимизации.

	ω_1	ω_2	ω_3
АЛМ	1.34946	3558.95	1924.33
Модифицированный АЛМ	1.34948	3551.97	0.0230
Относительная разность	$1.48 \cdot 10^{-5}$	$1.963 \cdot 10^{-3}$	0.9999880

Отметим, что полученные результаты для модифицированного алгоритма отличаются от стандартных незначительно (кроме третьего коэффициента), что объясняется достаточно медленно убывающей регрессионной моделью. Соответственно, расстояние в направлении перпендикуляра к линеаризованной прямой отличается от расстояния по вертикали незначительно. Кроме того, в данной задаче погрешность измерения независимой переменной (длины волны) достаточно мала, а погрешность измерения зависимой переменной (коэффициента преломления) практически одинакова для всех точек в обучающей выборке, поэтому нормирование на соответствующую величину приводит к практически равному уменьшению каждого из членов функционала (1) и не имеет существенного виляния. Однако, регрессионные модели другого вида демонстрируют более существенные различия.

Значительное отличие третьего коэффициента ω_3 объясняется несущественным вкладом соответствующего члена в сумму ввиду высокой степени при λ в знаменателе. Как следствие, он носит скорее случайный характер.

5 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.

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): *Devel-*

opments 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] Strutz, Tilo: *Data fitting and uncertainty: a practical introduction to weighted least squares and beyond.* Vieweg + Teubner, Wiesbaden, 2011, ISBN 9783834810229. http://www.worldcat.org/isbn/9783834810229.

[12] Серова, Н. В.: *Полимерные оптические материалы.* Научные основы и технологии, 2011.

[13] Вапник, В. Н.: *Восстановление зависимостей по эмпирическим данным.* М.: Наука, 1979.