ВОССТАНОВЛЕНИЕ ДИСПЕРСИИ ПРОЗРАЧНОЙ СРЕДЫ ПО ЭКСПЕРИМЕНТАЛЬНЫМ ДАННЫМ

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 метод оценки устойчивости полученного решения. Приводятся результаты численного моделирования на данных, полученных в ходе эксперимента по определению состава смеси по ммарной дисперсии.

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

Введение

В подавляющем большинстве случаев анализа физического эксперимента требуется восстановление функциональной зависимости, описывающей соотношение измеряемых параметров. 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 сложные нелинейные модели. Как правило, различные приближения сравниваются по методу наименьших квадратов, при этом оптимизация варьируемых параметров проводится, например, с помощью алгоритма Левенберга-Марквардта [6, 7].

Однако при анализе физического эксперимента важное значение имеет не только величина самих параметров искомой функциональной зависимости, но и погрешность их определения, обусловленная погрешностями измеряемых в эксперименте величин. Для задачи линейной регрессии соответствующая задача решена в простом варианте, 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 техника 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 ее экспериментально определенной зависимости $n(\lambda)$. Другими словами, для случая двух полимеров, заранее определив зависимости $n_1(\lambda)$ и $n_2(\lambda)$, необходимо определить суммарную зависимость $n(\lambda) = \alpha n_1(\lambda) + (1-\alpha)n_2(\lambda)$ и по ней определить коэффициент α, имеющий смысл концентрации первого полимера в смеси.

Поскольку показатели преломления для прозрачных полимеров близкого химического состава различаются незначительно, то 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 связана с 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 заведомо узкополосных источников света (лазеров, известным атомных переходов типа триплета ртути или дублета натрия, и т. д.). Типичная относительная погрешность определения длины волны в рассматриваемой задаче составляет $0.03 \div 3\%$. Отметим также, что погрешность определения длины волны, как правило, меняется с изменением самой длины волны. Экспериментальная погрешность показателя преломления n зависит от выбранного способа его измерения и для часто используемого варианта определения 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 очень высокого уровня до $(1 \div 10) \cdot 10^{-4}$. Для рассматриваемых в настоящей работе задач существенно то, что величины погрешностей могут считаться известными.

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. Поставлена задача и предложен метод оценки устойчивости решения к погрешностям эксперимента и изучена зависимость этих характеристик от параметров модели.

В первой части данной работы формально поставлена задача восстановления дисперсии жидкости. Во второй части вкратце описывается алгоритм [5], используемый для порождения аналитической функции-суперпозиции, аппроксимирующей данные. В третьей части описывается метод, позволяющий учитывать и анализировать вклад различных членов порожденной суперпозици и их зависимость от погрешности экспе-

римента. В четвертой части приводятся результаты вычислительного эксперимента на реальных данных, полученных в ходе физического эксперимента по изучению возможности определения состава смеси прозрачных веществ по суммарной дисперсионной зависимости. Рассматривается три прозрачных для света вещества, для каждого из которых имеется 18 экспериментальных точек, соответствующих величине коэффициента преломления при различных значениях длины волны.

1 Постановка задачи

Дана выборка $\tilde{D}$ из ℓ результатов измерений коэффициента преломления для некоторого вещества: $\tilde{D} = \{\tilde{\lambda}_i, \tilde{n}_i\}$, где $\tilde{\lambda}_i$ — длина волны, а $\tilde{n}_i$ — измеренный коэффициент преломления в i-ом измерении.

Требуется найти функцию $\hat{f} = \hat{f}(\lambda)$, минимизирующую 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 $D = \{\lambda_i, n_i \mid i \in \{1, \ldots, \ell\}\}$, а $\mathcal{F}$ — некоторое множество суперпозиций, из которого выбирается оптимальная.

Иными словами,

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

Введем понятие устойчивости суперпозиции f. Рассмотрим вектор параметров $\boldsymbol{\omega}_f = \{\omega_i^f \mid i \in \{1, \ldots, l_f\}\}$ суперпозиции f: $f(\lambda) = f(\lambda, \boldsymbol{\omega}_f)$. Пусть для некоторой выборки $D = \{\lambda_i, n_i\}$ функция f_D с вектором параметров $\hat{\boldsymbol{\omega}}_{f_D}$ минимизирует функционал (1). Рассмотрим выборку

$$\acute{D}(\sigma_n, \sigma_\lambda) = \{\lambda_i + \xi_i^\lambda, n + \xi_i^n \mid i \in 1, \ldots, \ell; \xi_i^n \in \mathcal{N}(0, \sigma_n); \xi_i^\lambda \in \mathcal{N}(0, \sigma_\lambda)\}. \tag{3}$$

Для этой выборки найдем оптимальный вектор $\acute{\hat{\boldsymbol{\omega}}}_{f_D,\sigma_n,\sigma_\lambda}$ параметров суперпозиции f_D, минимизирующий функционал (1):

$$\acute{\hat{\boldsymbol{\omega}}}_{f_D,\sigma_n,\sigma_\lambda} = \arg\min_{\boldsymbol{\omega}_{f_D} \in R^{|\hat{\boldsymbol{\omega}}_{f_D}|}} S(f_D(\cdot, \boldsymbol{\omega}_{f_D}), \acute{D}). \tag{4}$$

Понятно, что $\acute{\hat{\boldsymbol{\omega}}}_{f_D,\sigma_n,\sigma_\lambda}$ — векторная случайная величина, и, следовательно, $\hat{\boldsymbol{\omega}}_{f_D} - \acute{\hat{\boldsymbol{\omega}}}_{f_D,\sigma_n,\sigma_\lambda}$ — также векторная случайная величина.

Пусть дан набор выборок $\acute{\mathcal{D}}_N = \{\acute{D}_1, \ldots, \acute{D}_N\}$, и пусть $\overline{\sigma}_i$ — эмпирическое стандартное отклонение i-ой компоненты векторной случайной величины $\hat{\boldsymbol{\omega}}_{f_D} - \acute{\hat{\boldsymbol{\omega}}}_{f_D,\sigma_n,\sigma_\lambda}$ на множестве $\acute{\mathcal{D}}_N$. *Относительной устойчивостью* (или просто *устойчивостью*) параметра ω_i относительно вектора $\boldsymbol{\alpha}$ будем называть вектор

$$\mathbf{T}_f^{\boldsymbol{\alpha}}(i) = \Big\{ \frac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{\frac{\overline{\sigma}_{\alpha_j}}{\alpha_j}} \Big\}. \tag{5}$$

В частности, в искомой задаче восстановления дисперсионной зависимости,

$$\mathbf{T}_f^{\{n,\lambda\}}(i) = \{\frac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{\frac{\sigma_n}{n}}, \frac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{\frac{\sigma_\lambda}{\lambda}}\}.$$

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

Требуется исследовать зависимость устойчивости $\mathbb{T}_{\hat{f}}^{\{n,\lambda\}}$ относительно вектора $\{n, \lambda\}$ от σ_n и σ_λ. В дальнейшем там, где это не приведет к неоднозначностям, $\mathbb{T}_{\hat{f}}^{\{n,\lambda\}}$ будет обозначаться как $\mathbb{T}_f$,

Кроме того, отдельный интерес представляет и вектор первых производных устойчивости $\mathbb{T}_f^{\boldsymbol{\alpha}}$ по $\boldsymbol{\alpha}$:

$$\frac{\partial \mathbb{T}_f^{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}},$$

являющийся характеристикой изменения устойчивости $\mathbb{T}$ при изменении стандартных отклонений погрешностей, накладываемых на обучающую выборку.

2 Алгоритм индуктивного порождения суперпозиций

Опишем предложенный в [5] алгоритм.

Пусть задано некоторое множество $G = \{g_1, \ldots, g_k\}$ 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:

$$Q_f = \frac{1}{1 + S(f)} \left(\alpha + \frac{1 - \alpha}{1 + \exp(C_f - \tau)} \right), \tag{6}$$

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.

Второй множитель в (6) 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 (6):

$$Q_f = \frac{1}{1+S(f)}\left(\alpha + \frac{1-\alpha}{1+\exp(C_f-\tau)}\right) \to \min_{f\in\mathcal{F}}. \tag{7}$$

3 Метод исследования стабильности решения

Для оценки устойчивости $\mathbb{T}_{\hat{f}}$ решения $\hat{f}$ задачи (6) предлагается следующий подход. 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 (3). Для этой выборки вычисляются значения коэффициентов суперпозиции $\hat{f}$, минимизирующие функционал (1) согласно (4), 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.

Кроме того, сравнение различных суперпозиций может также производиться по критерию устойчивости в дополнение к сравнению по сложности и по значению функционала (1).

4 Вычислительный эксперимент

В вычислительном эксперименте используются данные, полученные в ходе изучения возможности определения состава смеси прозрачных веществ по суммарной дисперсионной зависимости, если известна экспериментальная зависимость дисперсии для

каждого конкретного веществ. Рассматривается три вещества, для каждого из которых имеется 18 экспериментальных точек, соответствующих коэффициенту преломления при разных значениях длины волны. Значения приведены в таблице 1.

λ, нм	Первый материал	Второй материал	Третий материал
435.8	1.36852	1.35715	1.34850
447.1	1.36745	1.35625	1.34767
471.3	1.36543	1.35449	1.34620
486.1	1.36446	1.35349	1.34542
501.6	1.36347	1.35275	1.34461
546.1	1.36126	1.35083	1.34294
577.0	1.3599	1.34968	1.34191
587.6	1.3597	1.34946	1.34174
589.3	1.35952	1.34938	1.34158
656.3	1.35767	1.34768	1.34005
667.8	1.35743	1.34740	1.33987
706.5	1.35652	1.34664	1.33917
750	1.35587	1.34607	1.33855
800	1.35504	1.34544	1.33794
850	1.3544	1.34487	1.33741
900	1.35403	1.34437	1.33685
950	1.35364	1.34407	1.33652

Таблица 1: Экспериментальные значения коэффициентов преломления для трех разных материалов.

Предполагается, что дисперсионные свойства веществ описываются одной и той же функциональной зависимостью, так как подчиняются одним и тем же законам. Поэтому сначала получена суперпозиция $\hat{f}$, минимиризующая (6) для первого вещества, а затем для каждого из трех веществ 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 [11], что зависимость коэффициента преломления n от длины волны λ должна выражаться суммой отрицательных четных степеней дисперсии, 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 третьей значащей цифры для удобства чтения):

$$f(\lambda) = 1.35 + \frac{5.82}{\lambda^2} + \frac{3.58 \cdot 10^{-5}}{\lambda^4}, \tag{8}$$

со сложностью 13, среднеквадратичной ошибкой $2.2 \cdot 10^{-5}$ и значением $Q_f \approx 0.0475$.

Отметим, что обычно в приложениях учитывают только квадратичный член, а более высокими степенями пренебрегают. Коэффициент при $\frac{1}{\lambda^4}$ указывает на полное согласие полученных результатов с принятой практикой.

Влияние штрафа за сложность. Исследуем, как влияет добавление нечетных степеней на результат решения задачи (7), 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 (8).

Увеличим τ до 25. Получим следующую формулу:

$$n(\lambda) = 1.34 + \frac{11.6}{\lambda} + \frac{17.37}{\lambda^2} + \frac{0.0866}{\lambda^3} + \frac{2.95 \cdot 10^{-4}}{\lambda^4} + \frac{8.54 \cdot 10^{-7}}{\lambda^5}, \tag{9}$$

сложность которой составляет 31, и для которой среднеквадратичная ошибка на выборке составляет $\approx 3.9 \cdot 10^{-7}$, а значение $Q_f \approx 0.0498$.

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.

То есть, 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 (8) 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 $(\sigma_\lambda, \sigma_n)$.

Численные значения эмпирического матожидания и дисперсии для первого, второго и третьего коэффициентов формулы (8) для первого полимера для некоторых значений $(\sigma_\lambda, \sigma_n)$ приведены в таблицах 2, 3 и 4 соответственно. По строкам указаны значения для σ_n, по столбцам — для σ_λ.

	0	10^{-5}	10^{-4}	0.001	0.01
0	$(1.36; 5.97 \cdot 10^{-7})$	$(1.36; 2.34 \cdot 10^{-6})$	$(1.36; 2.39 \cdot 10^{-5})$	$(1.36; 2.34 \cdot 10^{-4})$	$(1.36; 0.00238)$
0.01	$(1.36; 0)$	$(1.36; 2.37 \cdot 10^{-6})$	$(1.36; 2.38 \cdot 10^{-5})$	$(1.36; 2.37 \cdot 10^{-4})$	$(1.36; 0.00231)$
0.1	$(1.36; 0)$	$(1.36; 2.37 \cdot 10^{-6})$	$(1.36; 2.35 \cdot 10^{-5})$	$(1.36; 2.34 \cdot 10^{-4})$	$(1.36; 0.00237)$
1	$(1.36; 1.03 \cdot 10^{-7})$	$(1.36; 2.36 \cdot 10^{-6})$	$(1.36; 2.34 \cdot 10^{-5})$	$(1.36; 2.39 \cdot 10^{-4})$	$(1.36; 0.00235)$
10	$(1.36; 2.74 \cdot 10^{-7})$	$(1.36; 2.37 \cdot 10^{-6})$	$(1.36; 2.35 \cdot 10^{-5})$	$(1.36; 2.35 \cdot 10^{-4})$	$(1.36; 0.00236)$

Таблица 2: Значения матожидания и стандартного отклонения для первого коэффициента первого полимера

	0	10^{-5}	10^{-4}	0.001	0.01
0	$(5.82; 0)$	$(5.82; 9.12 \cdot 10^{-4})$	$(5.82; 0.00913)$	$(5.82; 0.0871)$	$(5.55; 1.87)$
0.01	$(5.82; 5.41 \cdot 10^{-5})$	$(5.82; 9.17 \cdot 10^{-4})$	$(5.82; 0.00904)$	$(5.82; 0.0867)$	$(5.56; 1.81)$
0.1	$(5.82; 5.37 \cdot 10^{-4})$	$(5.82; 0.00105)$	$(5.82; 0.00907)$	$(5.82; 0.0873)$	$(5.56; 1.81)$
1	$(5.82; 0.00538)$	$(5.82; 0.00549)$	$(5.82; 0.0106)$	$(5.82; 0.0866)$	$(5.56; 1.82)$
10	$(5.82; 0.0511)$	$(5.82; 0.0516)$	$(5.82; 0.0.20)$	$(5.82; 0.103)$	$(5.59; 1.75)$

Таблица 3: Значения матожидания и стандартного отклонения для второго коэффициента первого полимера

	0	10^{-5}	10^{-4}	0.001	0.01
0	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 6.21 \cdot 10^{-7})$	$(0; 0.0657)$
0.01	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 6.23 \cdot 10^{-7})$	$(5.7 \cdot 10^{-4}; 0.0536)$
0.1	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 6.21 \cdot 10^{-7})$	$(0; 0.106)$
1	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 1.16 \cdot 10^{-7})$	$(3.58 \cdot 10^{-5}; 6.32 \cdot 10^{-7})$	$(0; 0.185)$
10	$(3.59 \cdot 10^{-5}; 9.71 \cdot 10^{-7})$	$(3.59 \cdot 10^{-5}; 9.83 \cdot 10^{-7})$	$(3.59 \cdot 10^{-5}; 9.89 \cdot 10^{-7})$	$(3.59 \cdot 10^{-5}; 1.18 \cdot 10^{-6})$	$(2.4 \cdot 10^{-4}; 0.0882)$

Таблица 4: Значения матожидания и стандартного отклонения для третьего коэффициента первого полимера

В таблице 5 представлены поверхности уровня дисперсии для первого, второго и третьего коэффициентов каждого из полимеров соответственно.

Из графиков видно, что от шума, накладываемого на значения длины волны, дисперсия значений первого и второго коэффициентов практически не зависит. 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 теряется. Кроме того, дисперсия третьего коэффициента формулы не имеет явной гладкой зависимости от дисперсии шума длины волны или коэффициента преломления, что еще раз подтверждает его достаточно случайную природу.

Материал 1	Материал 2	Материал 3

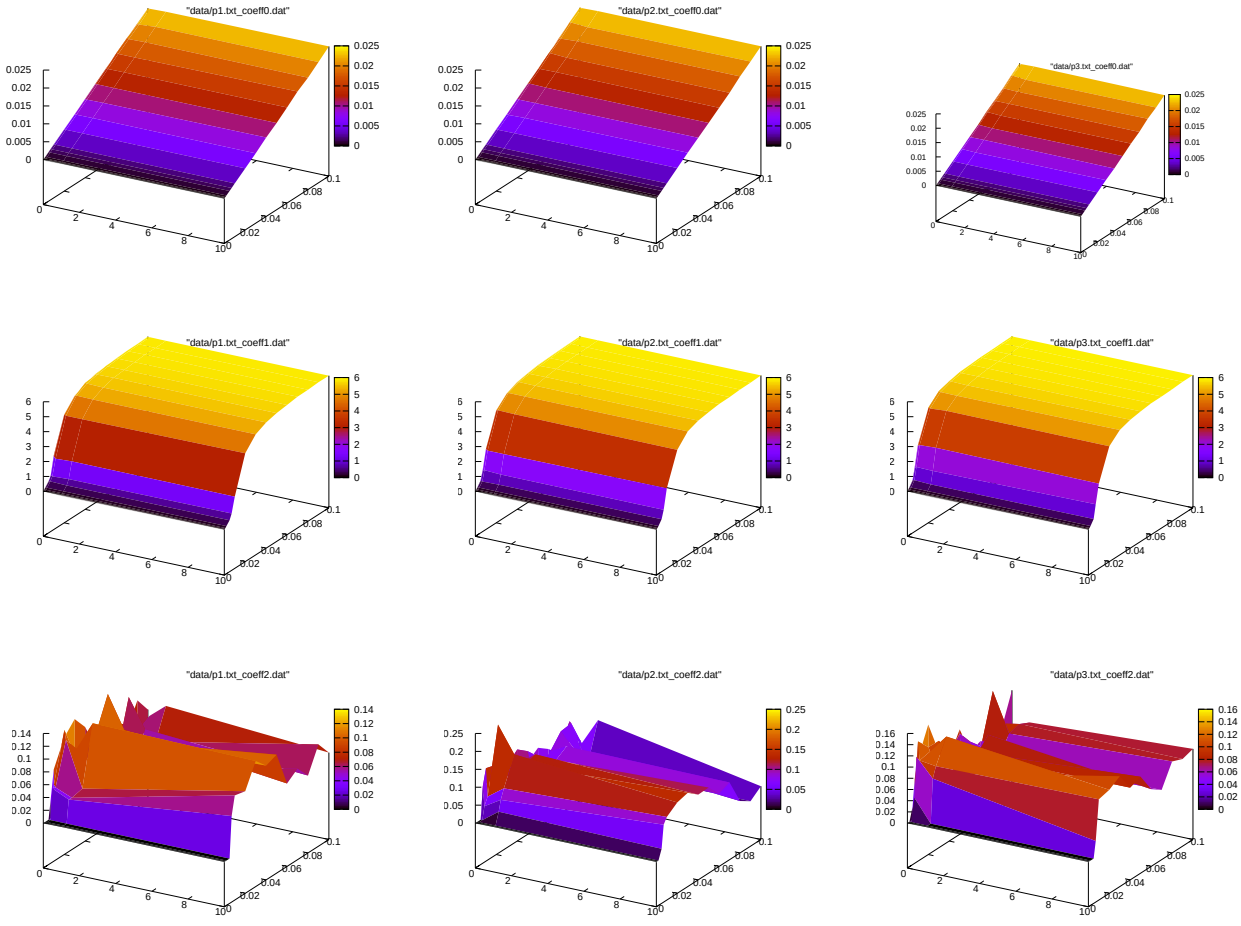


Таблица 5: Поверхности дисперсии для формулы (8).

Физическая интерпретация этих результатов — при построении прибора для измерения дисперсии сред значительное внимание следует уделять точности измерения коэффициента преломления, тогда как измерения длины волны могут быть неточны вплоть до нескольких процентов. Кроме того, предложенный метод прямо указывает, на каких интервалах шума какой будет выигрыш в точности предсказания от небольшого увеличения точности.

Отметим так же, что для коэффициентов, стоящих на одних и тех же местах, графики дисперсии похожи даже для разных веществ. Из этого можно сделать вывод, что полученная формула действительно описывает наблюдаемое физическое явление — именно такая зависимость и «должна» была бы получиться, и что полученная зависимость носит универсальный характер, не являясь переобученной моделью.

Кроме того, значения дисперсии не превосходят настоящие значения параметров по порядку величины, что означает, в частности, что вещества могут быть различены

даже не очень точным рефрактометром.

Стабильность некорректного решения. Аналогично исследуем стабильность решения (9). Для данных значений дисперсии приведем только графики зависимости, см. таблицу 6.

Из графиков видно, что в случае формулы (9) дисперсия соответствующих параметров существенно превышает таковую для (8). 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.

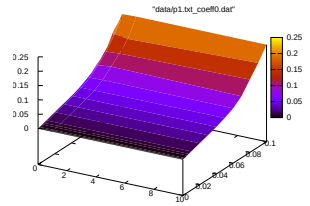
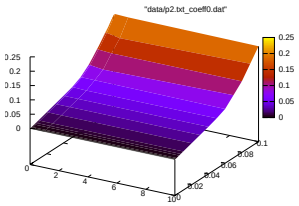
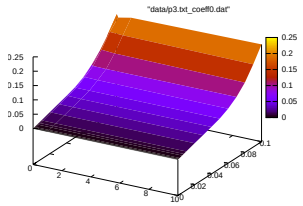
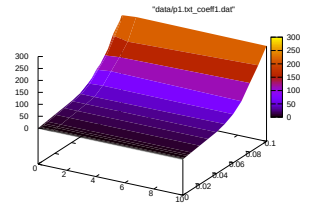
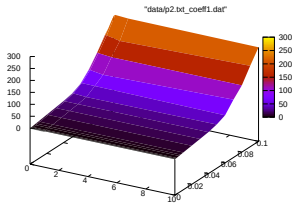
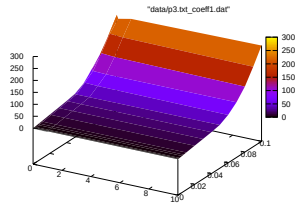
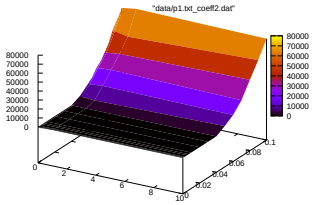
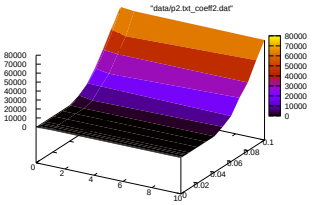
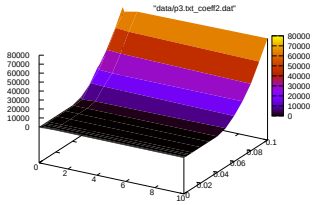
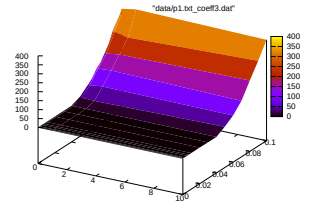
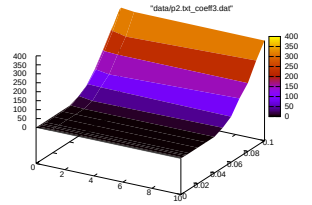
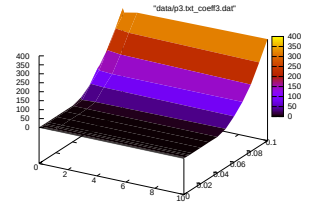
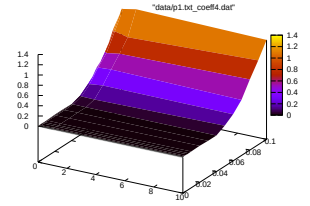
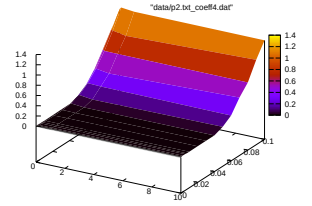
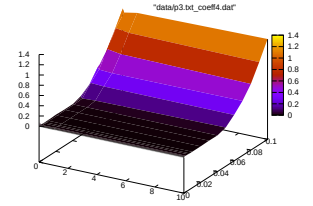
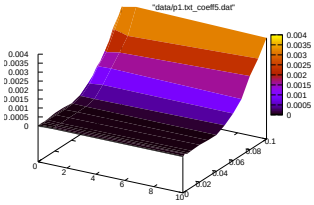
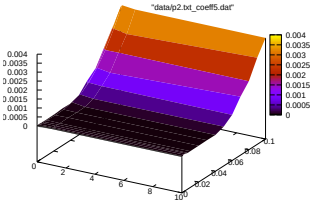
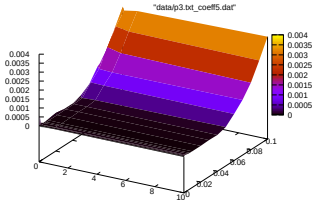
Материал 1	Материал 2	Материал 3
		
		
		
		
		
		

Таблица 6: Поверхности дисперсии для формулы (9).

Исследование экспертного предположения. Экспертом предположено, что формула так же может иметь вид

$$n(\lambda) = a + \frac{b}{c - \frac{1}{\lambda^2}}, \tag{10}$$

если измерения находятся вблизи точки резонанса.

Результаты нахождения параметров a, b и c методом Левенберга-Марквардта приведены в таблице 7.

Материал	a	b	c
1	1.385	$1.79 \cdot 10^{-7}$	$-4.49 \cdot 10^{-6}$
2	1.372	$1.62 \cdot 10^{-7}$	$-4.59 \cdot 10^{-6}$
3	1.361	$1.21 \cdot 10^{-7}$	$-3.75 \cdot 10^{-6}$

Таблица 7: Значения коэффициентов формулы (10).

Коэффициент c в формуле (10) имеет смысл резонансной частоты, приближение к которой описывается этой формулой, поэтому коэффициент c должен быть неотрицательным. Ввиду этого полученные результаты не имеют физического смысла.

Тем не менее, исследуем стабильность данного решения тем же методом, что и в предыдущих случаях. Поверхности дисперсии приведены в таблице 8.

Отметим, что характерная дисперсия первого коэффициента на порядок больше, чем для формулы (8), что затрудняет различение веществ в смеси при достаточно большой погрешности измерения λ, однако дисперсии второго и третьего коэффициента примерно на порядок меньше, чем для формулы (8).

Поверхности дисперсии также не являются настолько же гладкими, как для формулы (8).

Все это позволяет заключить, что, хотя экспериментальные данные хорошо описываются формулой (10), они не являются корректными с экспертно-физической точки зрения. Это, в частности, подтверждается экспертным соображением об ограничениях на коэффициенты формулы (10), которые не выполняются в полученной модели.

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 прибегания к методам 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

Материал 1	Материал 2	Материал 3

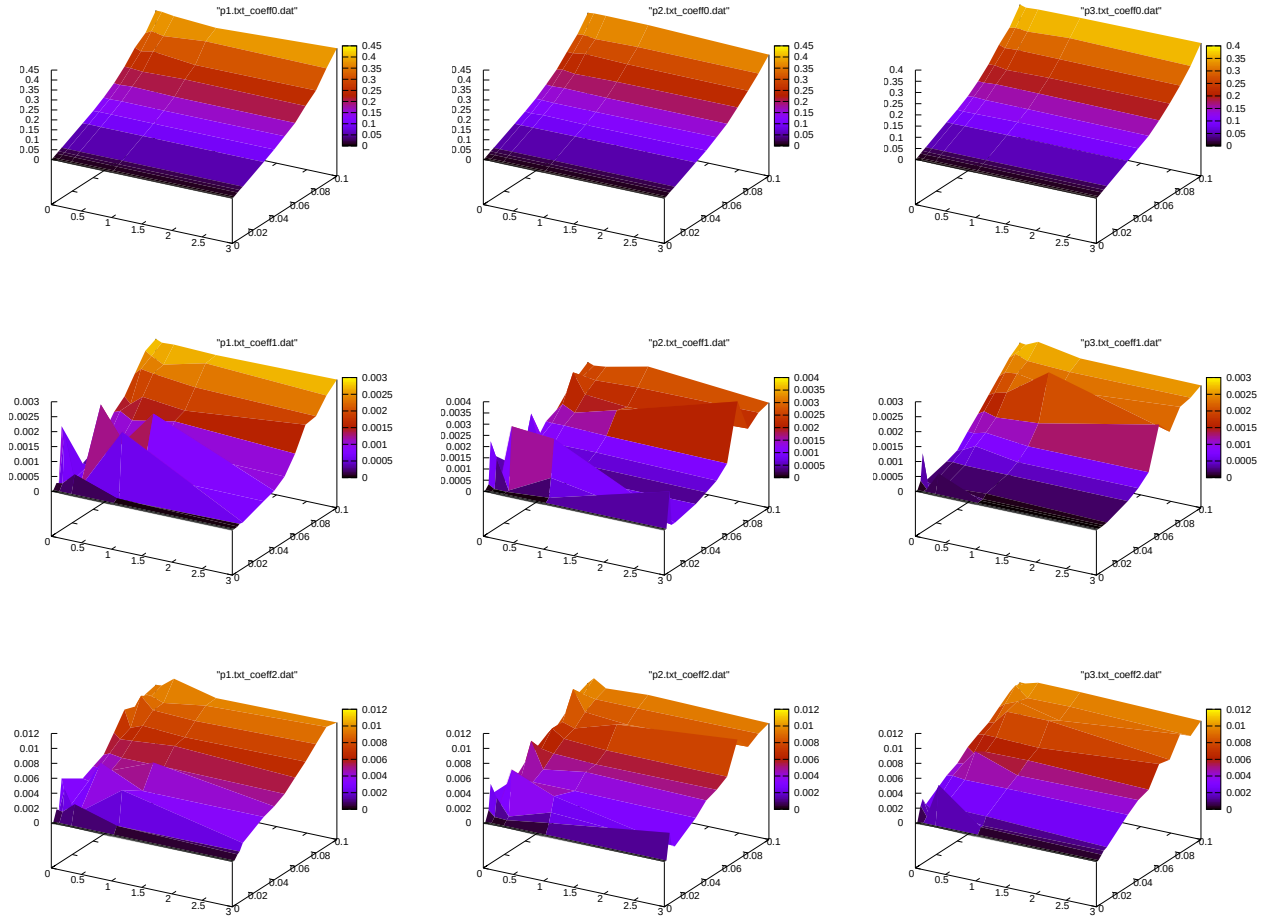


Таблица 8: Поверхности дисперсии для формулы (10).

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 в решение, 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): *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.