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 вычислительного эксперимента на данных, полученных в ходе эксперимента по определению концентраций компонентов смеси по суммарной дисперсии.

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

Введение

В ряде прикладных областей возникает задача восстановления зависимости показателя преломления прозрачной для света среды от длины волны. И задаваемая в ходе эксперимента длина волны, и измеряемый коэффициент преломления, соответствующий этой длине волны, определяются с достаточно высокой точностью в широком диапазоне длин волн.

TODO описать вкратце ход эксперимента.

Физические соображения [1, 2] определяют некоторый общий вид зависимости коэффициента преломления прозрачной среды от длины волны, которой должны описываться экспериментальные данные. Для эксперта представляют интерес конкретные параметры этой зависимости, поэтому требуется экспертная интерпретация полученной модели.

Кроме того, каждое измерение имеет свою допустимую погрешность, задаваемую экспериментатором TODO попробовать сослаться на методику оценки погрешности в подобных экспериментах, поэтому необходимо оценить устойчивость полученных результатов. Иными словами, требуется понять, насколько сильно меняются коэффициенты результирующей суперпозиции с фиксированной структурой при небольшом изменении экспериментальных данных.

В настоящей работе исследуется возможность применения алгоритма, предложенного в [3] для восстановления искомой зависимости, и результаты его работы сравниваются с результатами применения SVM-регрессии. Кроме того, исследуется влияние штрафа за сложность на качество и структурную сложность получающихся суперпозиций. Также исследуется возможность оценки устойчивости решения и вклада различных элементов результирующей модели.

В первой части данной работы формально поставлена задача восстановления дисперсии жидкости. Во второй части вкратце описывается алгоритм [3], 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 случайных ошибок. 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 коэффициентам 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 $\boldsymbol{\omega}_f$ 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 $\acute{\mathcal{D}}_N = \{\acute{D}_1, \ldots, \acute{D}_N\}$. 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 $\acute{\mathcal{D}}_N$ будем называть устойчивостью $T_f(i)$ коэффициента i (или устойчивостью параметра i) суперпозиции f, а вектор устойчивости $\mathbf{T}_f = \{T_f(i)\}$ — устойчивостью суперпозиции f.

Требуется оценить устойчивость решения $\hat{f}$, исследуя зависимость устойчивости $\mathbf{T}_{\hat{f}}$ от σ_n и σ_λ.

2 Алгоритм индуктивного порождения суперпозиций

Опишем предложенный в [3] 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:

$$Q_f = \frac{1}{1 + S(f)} \left(\alpha + \frac{1 - \alpha}{1 + \exp(C_f - \tau)} \right), \tag{5}$$

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 (5) 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.

3 Метод исследования стабильности решения

Для оценки устойчивости $\mathbf{T}_{\hat{f}}$ решения $\hat{f}$ задачи (5) 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 $\mathbf{T}_f$.

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 структурной формуле и могут являться свидетельством переобучения.

4 Вычислительный эксперимент

В вычислительном эксперименте используются данные, полученные в ходе физического эксперимента по изучению 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. Характерное значение $n \approx 1.33$, $\lambda \in (400 - 950)$ нм.

Предполагается, что свойства веществ описываются одной и той же формулой, так как подчиняются одним и тем же законам. Поэтому сначала получена формула регрессионной зависимости по данным для первого вещества, а затем для каждого из трех веществ оценена стабильность полученного решения и найдены конкретные коэффициенты.

В силу замечания о невозможности подбора структурных параметров α и τ, 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 [2], что зависимость коэффициента преломления от длины волны 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}, принудительно заменялись на 0.

В результате применения описанного выше алгоритма со значениями $\hat{Q} = 0.95$, $\alpha = 0.05$, $\tau = 10$ получена следующая формула (константы округлены до третьей значащей цифры для удобства чтения):

$$n(\lambda) = 1.35 + \frac{5.82}{\lambda^2} + \frac{3.58 \cdot 10^{-5}}{\lambda^4}, \tag{6}$$

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.

4.1 Влияние штрафа за сложность

Исследуем, как влияет добавление нечетных степеней на получающийся результат, 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 (6).

Увеличим τ до 25. Получим следующую формулу:

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

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.

4.2 SVM

В качестве базового алгоритма используется SVM-регрессия с RBF-ядром [4]. 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.

4.3 Исследование стабильности решения

Для оценки стабильности решения фиксировалась формула (6) 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 (6) для первого полимера для некоторых значений $(\sigma_\lambda, \sigma_n)$ приведены в таблицах 1, 2 и 3 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 1: 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 2: 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 3: Значения матожидания и стандартного отклонения для третьего коэффициента первого полимера

В таблице 4 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

Материал 1	Материал 2	Материал 3

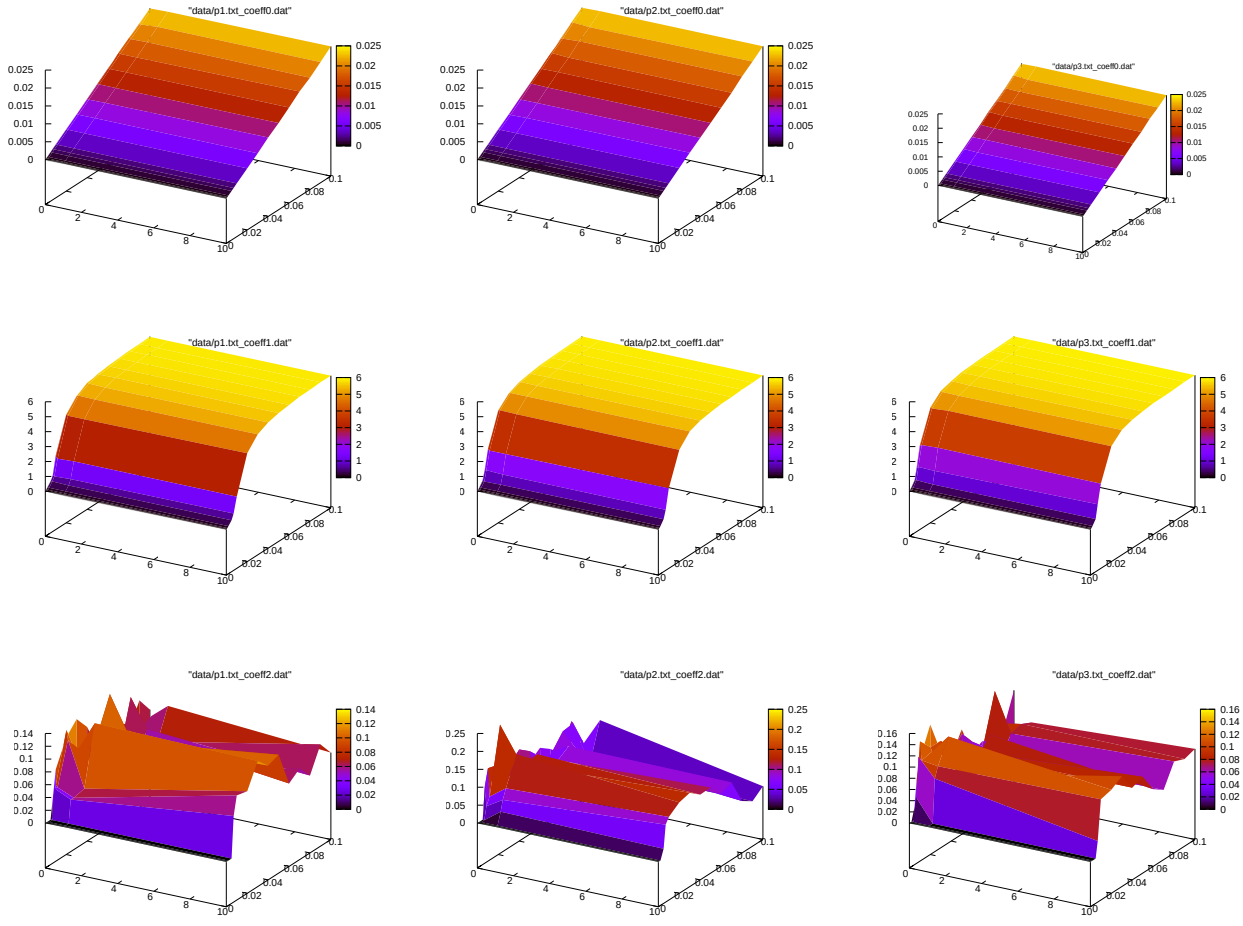


Таблица 4: Поверхности дисперсии для формулы (6).

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.

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, гра-

фики 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.

4.4 Стабильность некорректного решения

Аналогично исследуем стабильность решения (7). Для данных значений дисперсии приведем только графики зависимости, см. таблицу 5.

Из графиков видно, что в случае формулы (7) дисперсия соответствующих параметров существенно превышает таковую для (6). 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.

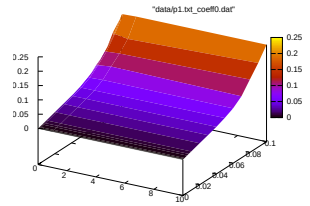
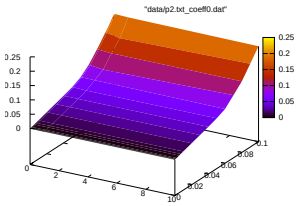
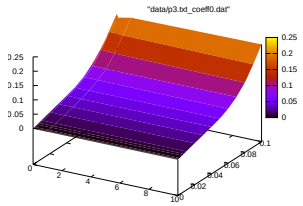
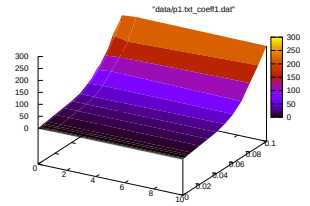
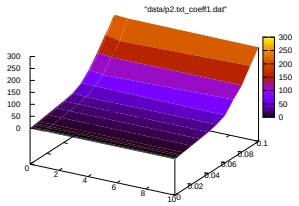
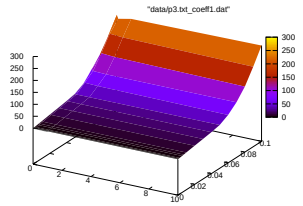
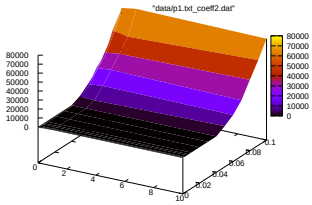
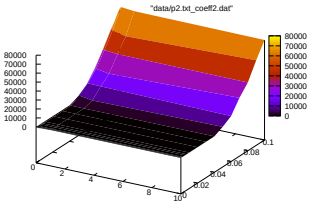
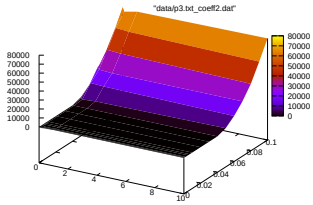
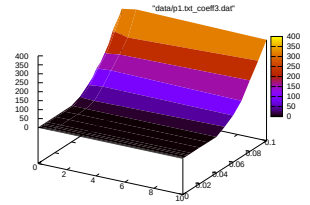
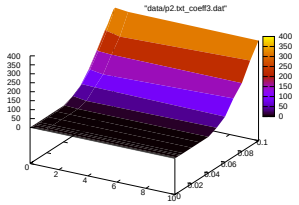
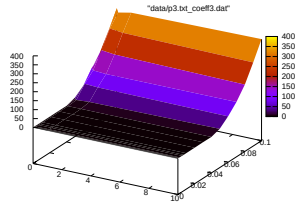
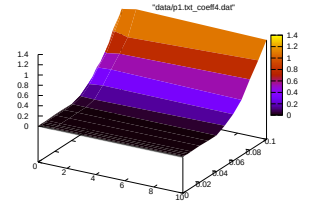
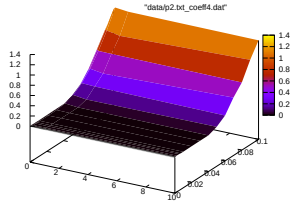
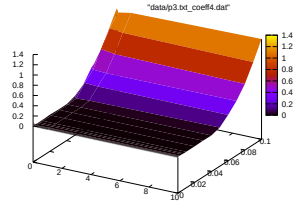
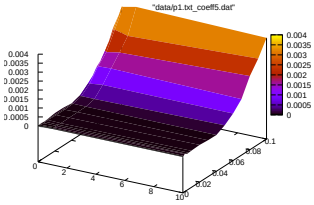
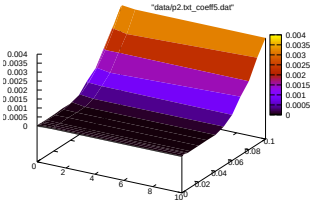
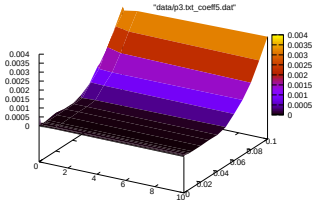
Материал 1	Материал 2	Материал 3
		
		
		
		
		
		

Таблица 5: Поверхности дисперсии для формулы (7).

4.5 Исследование экспертного предположения

Экспертом предположено, что формула так же может иметь вид

$$n(\lambda) = a + \frac{b}{c - \frac{1}{\lambda^2}}, \tag{8}$$

если измерения находятся вблизи точки резонанса.

Результаты нахождения параметров a, b и c методом Левенберга-Марквардта приведены в таблице 6.

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 6: Значения коэффициентов формулы (8).

Коэффициент c в формуле (8) 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 7.

Отметим, что характерная дисперсия первого коэффициента на порядок больше, чем для формулы (6), 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 (6).

Поверхности дисперсии также не являются настолько же гладкими, как для формулы (6).

Все это позволяет заключить, что, хотя экспериментальные данные хорошо описываются формулой (8), они не являются корректными с экспертно-физической точки зрения. Это, в частности, подтверждается экспертным соображением об ограничениях на коэффициенты формулы (8), которые не выполняются в полученной модели.

5 Заключение

Предложенный в [3] 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

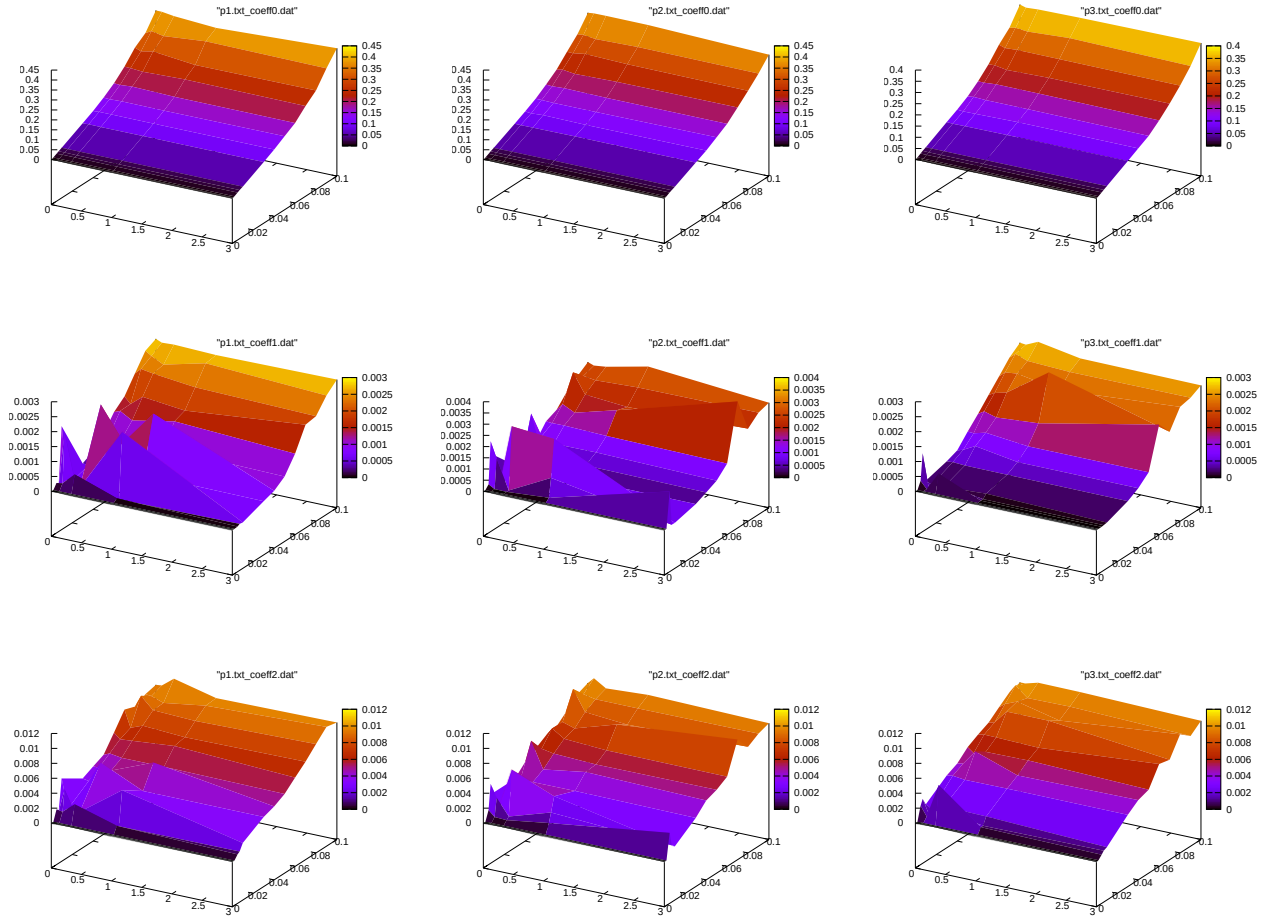


Таблица 7: Поверхности дисперсии для формулы (8).

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] Сивухин, Д. В.: *Оптика.* ФИЗМАТЛИТ, 3 редакция, 2005.

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

[3] Рудой, Г. И. и Стрижов, В. В.: *Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных.* Информатика и ее применения, 7(1):44–53, 2013.

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