STABILITY OF NON-LINEAR REGRESSION MODELS WITH RESPECT TO VARIATIONS IN THE MEASURED DATA

G. Rudoy.

Аннотация

A set of inductively generated non-linear regression models is considered to find an optimal one. Firstly, a previously suggested model generation method taking the complexity of the models into account is applied. Then, a new model selection criteria is proposed, called model stability, which shows the dependency of the inaccuracy in determining the coefficients of the generated models on the variation of the data in the learning set. This criteria is used directly to determine the inaccuracy of the model coefficients, which is of interest to the experts, as well as to select the optimal model amongst different ones generated using different algorithm hyperparameters. The data obtained during experiments on optical dispersion of transparent polymers is used to illustrate the method.

Keywords: *symbolic regression, non-linear models, inductive generation, model stability, transparent polymers dispersion.*

INTRODUCTION

Analysing the results of a physical experiment typically requires finding a functional dependency between the measured data. It is also very desirable for the dependency to be interpretable by an expert in the corresponding physical field. In many cases some theoretical assumptions about the structure of the functional dependency are available, or a choice should be made between different proposed models.

One of the methods allowing to find interpretable models is symbolic regression [1]-[5], which generates structurally complex non-linear models. Different models can be compared by their respective errors on the measured data, and the optimization of their numeric parameters is performed, for example, using the Levenberg-Marquardt algorithm [6], [7].

On the other hand, during physical experiment analysis not only the model parameters themselves are important for the expert, but the uncertainities in determining their values resulting from the intrinsic measurement inaccuracies. For the linear regression this problem is known to have a theoretical solution [8] in the particular case of the independent variable measured exactly and the dependent variable having the same Gaussian distribution of the error at all measured points. More complex case of non-linear regression, including the case of independent variables measured inexactly, as well as all points having different error distributions, has not been considered as far as we know.

In this paper the non-linear symbolic regression method is applied to find the dependency of the refraction index n of a polymer as a function of the wavelength λ for those frequencies where the considered polymer is transparent, including visible and near infrared light. The goal of the experimenters was to, first, find the dispersion for each polymer, and then derive the concentration of each polymer in their mixture, given that the dispersion of the mixture of polymers is a weighted sum of their respective dispersions. In other words, in case of two polymers, knowing the functions $n_1(\lambda)$ and $n_2(\lambda)$, the mixture dispersion dependency $n(\lambda)$ should be measured and, since $n(\lambda) = \alpha n_1(\lambda) + (1 - \alpha)n_2(\lambda)$, the concentration of the first polymer α should be derived.

The refraction indexes for the transparent polymers of a similar chemical composition differ only slightly. Thus, the uncertainity in determining the $n(\lambda)$ function coefficients and its dependency on the inaccuracies of measurements of the wavelength λ and refraction index n must be considered. This dependency is also important because it defines the requirements for the precision of the devices and, consecutively, it largely affects the cost and duration of the experiment.

Typically broad spectrum sources are used in refractometers, and the inaccuracy in extracting a single wavelength is defined by the hardware function of the monochromator being used and is thorougly considered, for example, in [9], [10]. In most cases the inaccuracy of λ can be computed as well as determined experimentally using narrow light sources like lasers, known atomic transitions like the mercury triplet or sodium dublet. Typical relative wavelength measurement inaccuracy is around $0.03 \div 0.5\%$, thus absolute inaccuracy depends on the wavelength itself. Inaccuracy of the refraction index n depends on the measurement method and, for example, in case of using the total internal refraction angle, is defined by the degree of non-parallelism of the light beams used, the inaccuracy in angle measurement and so on. The inaccuracy ranges from $(1 \div 2) \cdot 10^{-5}$ for high-class devices to $(1 \div 10) \cdot 10^{-4}$ for simpler devices. Thus, it is important for this paper that the inaccuracies can be considered to be known and perhaps different for each data point.

In this paper the dependency of the refraction index of the wavelength is used to illustrate the model generation algorithm proposed in [5]. Its results are compared with the SVM regression. Moreover, the impact of the complexity penalty on the quality and complexity of the generated models is investigated. The problem of determining the model coefficients stability in the general case of multivariate models is formally stated, a method for evaluating solution stability is proposed, and the dependency of these characteristics of the model hyperparameters is studied for the given case of determining the dispersion of transparent polymers.

In the first part of this paper the dispersion problem is formally stated, and the stability criteria is proposed. In the second part the algorithm proposed in [5], used to generate the regression model, is briefly described. In the third part a numeric method for stability estimation is proposed. In the fourth part the results of the computational experiment are shown. In the experiment two polymers are considered, for each of them 17 data points are given, corresponding to the refraction index at different wavelengths.

1 PROBLEM STATEMENT

Regression problem. Let D be the data set of ℓ refraction index measurements for a polymer: $D = \{\lambda_i, n_i \mid i \in \{1, \ldots, \ell\}\}$, where λ_i is the wavelength, and n_i is the measured refraction index.

It is required to find a function $\hat{f} = \hat{f}(\lambda)$, minimizing the standard loss function, assuming Gaussian error:

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

where $\mathcal{F}$ is some set of superpositions from which an optimal one is to be found.

In other words,

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

Stability estimation. We define the notion of the stability of some superposition f in general case. The stability describes the behavior of the coefficients of the superposition f during slight random variation of the source learning data set $D = \{\mathbf{x}_i, y_i\}$, where $\mathbf{x}_i$ is the feature vector of i-th object measured during the experiment, and y_i is the corresponding measured value of the target function to be recovered.

The loss function (1) in this case is:

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

We denote the matrix representing the data set as $X = \|x_{ij}\|$, where rows are feature vectors of the objects in D. In other words, x_{ij} is the j-th component of the feature vector of the i-th object.

We consider the parameters vector $\boldsymbol{\omega}_f = \{\omega_i^f \mid i \in \{1, \ldots, l_f\}\}$ of some superposition f: $f(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\omega}_f)$. Let $\hat{\boldsymbol{\omega}}_f(D)$ be the parameters vector minimizing the functional (3) for some learning set $D = \{\mathbf{x}_i, y_i\}$ and function f with fixed structure:

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

Let $\Sigma^{\mathbf{x}} = \|\sigma_{ij}^{\mathbf{x}}\|$ be the matrix of standard deviations of independend variables, where $\sigma_{ij}^{\mathbf{x}}$ is the standard deviation of the j-th element of the feature vector $\mathbf{x}_i$ of the i-th object of the learning set. Let $\boldsymbol{\sigma}^y$ be the vector of standard deviations of the dependent variable, where σ_i^y is the standard deviation of the measured variable for the i-th object. We then consider the modified learning set $\acute{D}$ derived from the source data set D by adding to its components some realizations of the random variables from the Gaussian distribution with zero mean and deviations corresponding to $\Sigma^{\mathbf{x}}$ and $\boldsymbol{\sigma}^y$:

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

For this new learning set $\acute{D}$ we find the new parameters vector $\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y))$ for the superposition f minimizing the functional (1):

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

Thus, $\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y))$ is a random vector, and, consecutively

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

is also a random vector.

Let $\acute{\mathcal{D}}_N$ be a set of N such modified learning sets, where each set is obtained by adding a separate realization of the corresponding random variables to the source data set:

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

Let $\overline{\sigma}_i$ be the sample standard deviation of the i-th component of the $\Delta\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y))$ random vector on the $\acute{\mathcal{D}}_N(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)$ set.

Определение 1. *Relative stability* (or simply *stability*) of the parameter ω_i given $\acute{\mathcal{D}}_N(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)$ and source learning set D is the following vector of length $|\mathbf{x}| + 1$:

$$\mathbf{T}_f^N(i) = \left\{ \frac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{r(\boldsymbol{\sigma}_{\cdot 1}^{\mathbf{x}}, \mathbf{x}_{\cdot 1})}, \ldots, \frac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{r(\boldsymbol{\sigma}_{\cdot |\mathbf{x}|}^{\mathbf{x}}, \mathbf{x}_{\cdot |\mathbf{x}|})}, \frac{\frac{\overline{\sigma}_i}{\hat{\omega}_i}}{r(\boldsymbol{\sigma}^y, \mathbf{y})} \right\}, \tag{6}$$

where $r(\boldsymbol{\alpha}, \mathbf{a}) = r(\frac{\alpha_1}{a_1}, \ldots, \frac{\alpha_{|\boldsymbol{\alpha}|}}{a_{|\mathbf{a}|}})$ is a function mapping a vector (comprised from quotients of the corresponding elements of vectors $\boldsymbol{\alpha}$ and $\mathbf{a}$), to a scalar value.

The function r maps to a single scalar value a set of (perhaps different) ratios of standard deviation of a measured variable to the value of that variable. The mapped scalar can be viewed as some kind of a characteristic of those ratios. The function r is chosen by the experts based on the assumptions about the error distribution characteristics. For example, in the case of polymers dispersion data the relative measurement error is constant as was described in the introduction, thus the r function may just choose any argument.

Each component of the $\mathbf{T}_f^N(i)$ vector describes the ratio between the standard deviation of the $\hat{\omega}_i$ parameter (normalized by the value of that parameter) and the standard deviation of the corresponding feature vector element (again, normalized by the value of that element). For instance, if this ration is greater than one, then the uncertainity in determining the coefficient grows faster than the measurement inaccuracy of the corresponding variable.

In the particular case of the dispersion regression considered in this paper, taking into account the constant relative measurement error:

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

Matrix comprised of vector columns $\mathbf{T}_f(i) \mid i \in \{1, \ldots, l_f\}$ is called the *stability* of the function f and is denoted as $\mathbb{T}_f$.

In case of the dispersion regression, it is required to study the dependency of stability $\mathbb{T}_{\hat{f}}$ as function of σ_n and σ_λ.

2 THE ALGORITHM FOR INDUCTIVE MODELS GENERATION

In this section we briefly describe the algorithm proposed in [5].

Let $G = \{g_1, \ldots, g_k\}$ be the set of some elementary functions. The set $\mathcal{F} = \{f\}$ of generated models is first initialized by random admissible superpositions of functions $g \in G$, taking their arity, domain and codomain into account. Superpositions in $\mathcal{F}$ contain free variables corresponding to the components of the feature vectors from the learning set, as well as constants which are subject to optimization by the Levenberg-Marquardt procedure (according to functional (1)) on each algorithm step. Each superposition can also be modified on each iteration in order to improve the quality Q_f of best superpositions in the set $\mathcal{F}$.

The quality Q_f of the model f is defined by the error on the learning set as well as the structural complexity of the superposition according to the following:

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

where:

$S(f)$ is the value of the loss functional (1) on the learning set D;

C_f is the complexity of the superposition f defined by the number of elementary functions, free variables and constants;

$\alpha, 0 \ll \alpha < 1$ adjusts the penalty of excessive model complexity (bigger α values prefer more complex but more precise models, while smaller choose simpler ones);

τ defines the desired complexity of the model, after which it is considered excessive.

The second multiplier in (7) is the penalty for excessive model complexity, which mitigates overfitting and allows obtaining simpler superpositions at the cost of bigger error on the learning set. The primary hypothesis is that simpler superpositions with slightly bigger learning set errors generalize better.

It is worth noting that α and τ are chosen by experts.

So, the initial problem of minimizing the functional (1) is replaced by the problem of minimizing (7):

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

3 МЕТОД ИССЛЕДОВАНИЯ СТАБИЛЬНОСТИ РЕШЕНИЯ

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

Иными словами, выбираются значения σ_λ и σ_n, затем для этих значений генерируется выборка $\acute{D}(\sigma_n, \sigma_\lambda)$ 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

Предполагается, что дисперсионные свойства полимеров описываются одной и той же функциональной зависимостью, так как подчиняются одним и тем же физическим закономерностям. Поэтому сначала получена суперпозиция $\hat{f}$, минимизирующая (7) для первого полимера, а затем для каждого из полимеров находятся соответствующие векторы параметров $\hat{\boldsymbol{\omega}}_{\hat{f}}$ 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 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 [12]. Параметр γ ядра подбирался по методу скользящего контроля, наилучшим результатом является комбинация из 15 опорных векторов с $\gamma \approx 2 \cdot 10^{-6}$, при этом среднеквадратичная ошибка при кросс-валидации с тестовой выборкой, содержащей по 2 объекта, составляет $8.96 \cdot 10^{-8}$. Однако, проинтерпретировать полученную решающую функцию не представляется возможным.

Исследование стабильности решения. Для оценки стабильности решения фиксировалась формула (9) 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 $(\sigma_\lambda, \sigma_n)$.

Таблица 2: Поверхности дисперсии для формулы (9).

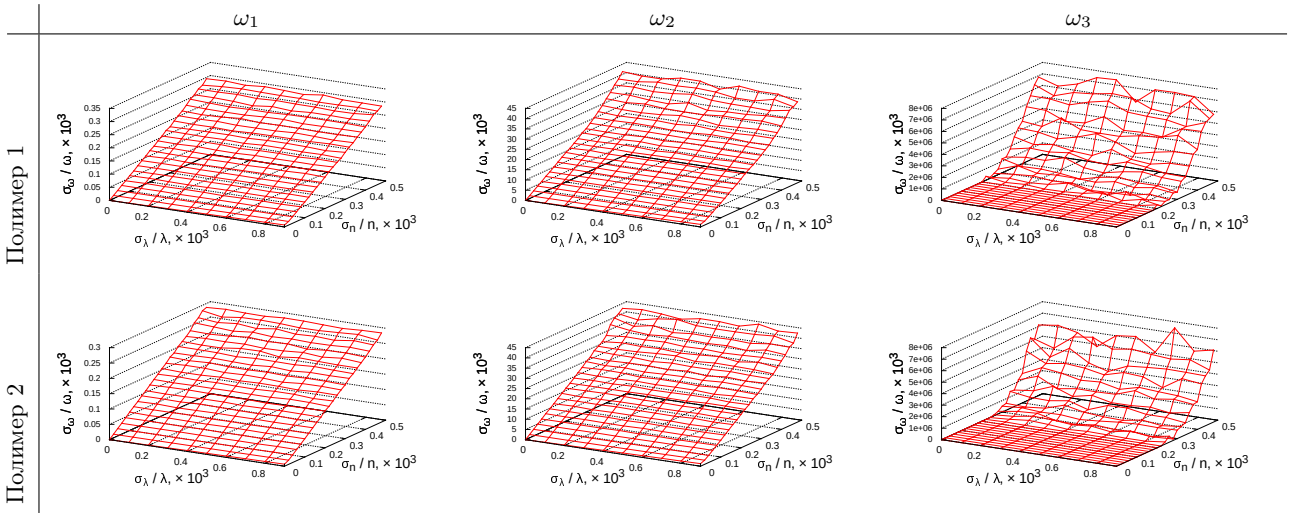


Таблица 3: Значения коэффициентов для формулы (9) и их относительная разность.

	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	$6.71 \cdot 10^{-3}$	$1.41 \cdot 10^{-1}$	$2.2 \cdot 10^{-1}$	

В таблице 2 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 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.

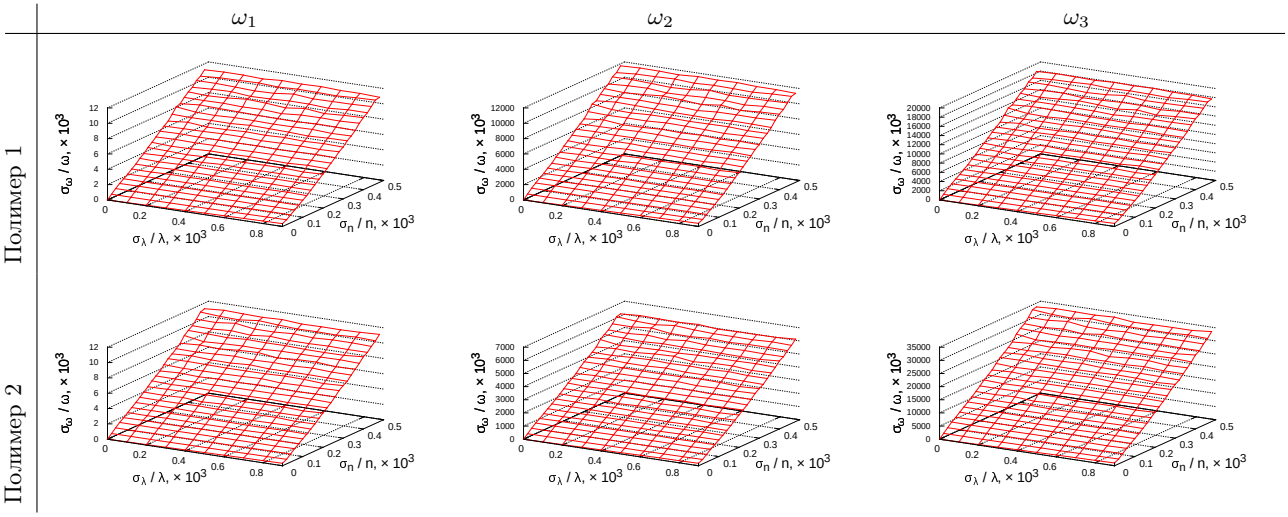
Принципиально важно, что значения стандартного отклонения параметров регрессионной модели существенно меньше разности между самими значениями этих параметров по порядку величины (см. таблицы 3 и 4), что означает, в частности, что полимеры могут быть различены даже не очень точным рефрактометром.

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

Таблица 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	$1.48 \cdot 10^{-3}$	$4.38 \cdot 10^{-3}$	$1.44 \cdot 10^{-2}$

Таблица 5: Поверхности дисперсии для формулы (10).



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

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

5 СХОДИМОСТЬ К КЛАССИЧЕСКОМУ СЛУЧАЮ

Рассмотрим случай, когда зависимость линейна:

$$y = ax + b,$$

и с учетом ошибок измерений представима в виде

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

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

Перейдем к представлению

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

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

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

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

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

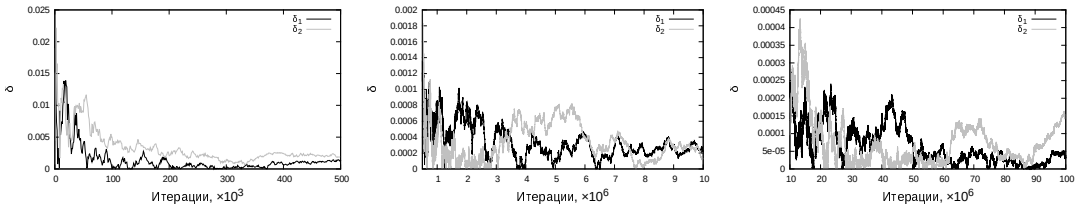
$$\delta_1 = \frac{|\mathbf{T}_y^N(1) - D(a)|}{D(a)},$$

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

Соответствующие графики для функции $y = 2x + 1 + \xi_i$ на интервале $x \in [0; 10]$ при $n = 10$ порождённых точках и $D(\xi_i) = 10$ приведены на фиг. 1. В частности, на фиг. 1а представлена начальная часть графика при количестве итераций N, меньшем $5 \cdot 10^5$, на фиг. 1б — средняя часть (при N от $5 \cdot 10^5$ до 10^7), а на фиг. 1в — характер сходимости при больших N (от 10^7 до 10^8).

Аналогичные графики приведены для $n = 10$ и $D(\xi_i) = 1$ и $n = 50$ и $D(\xi_i) = 1$ соответственно на фиг. 2 и 3.

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



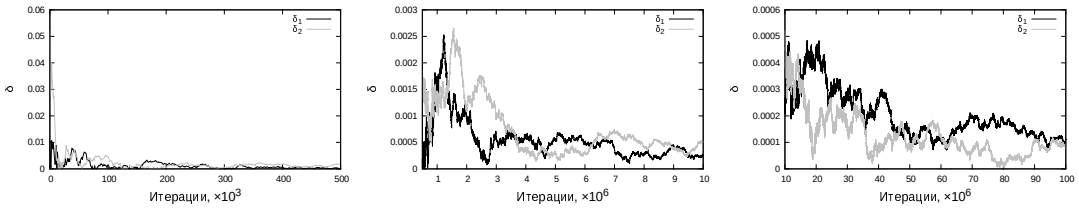
(а) $N \in [0;\ 5 \cdot 10^5]$ (б) $N \in [5 \cdot 10^5;\ 10^7]$ (в) $N \in [10^7;\ 10^8]$

Фиг. 1: Зависимость δ от числа итераций N при $D(\xi) = 10$ и $n = 10$.

6 ЗАКЛЮЧЕНИЕ

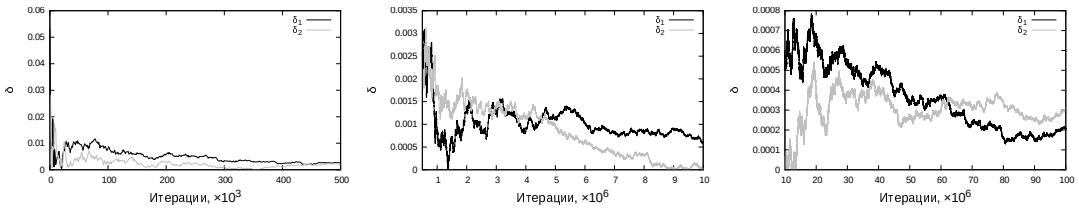
Предложенный в [5] алгоритм позволяет получить интерпретируемую аналитическую формулу, описывающую зависимость коэффициента преломления среды от длины волны. Введенный штраф за сложность позволяет избежать переобучения без использования методов вроде скользящего контроля, и, таким образом, отпадает необходимость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

Хотя другие алгоритмы, такие как SVM-регрессия, могут демонстрировать более высокое качество приближения данных, их результаты неинтерпретируемы и не защищены



(а) $N \in [0;\ 5 \cdot 10^5]$ (б) $N \in [5 \cdot 10^5;\ 10^7]$ (в) $N \in [10^7;\ 10^8]$

Фиг. 2: Зависимость δ от числа итераций N при $D(\xi) = 1$ и $n = 10$.



(а) $N \in [0;\ 5 \cdot 10^5]$ (б) $N \in [5 \cdot 10^5;\ 10^7]$ (в) $N \in [10^7;\ 10^8]$

Фиг. 3: Зависимость δ от числа итераций N при $D(\xi) = 1$ и $n = 50$.

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): *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.