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 error in determining the parameters of the generated models on the variation of the data in the learning set. This criteria is used directly to determine the error of the model parameters, 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 errors 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 error in determining the $n(\lambda)$ function parameters and its dependency on the measurement errors 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 error is around $0.03 \div 0.5\%$, thus absolute measurement error depends on the wavelength itself. Refraction index error 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 angle measurement error and so on. The error 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 errors 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 stability of model parameters 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 various wavelengths.

1 PROBLEM STATEMENT

Regression problem. Let D be the data set of ℓ refraction index measurements for a polymer: $D = \{\lambda_i, n_i \mid i \in \{1, \dots, \ell\}\}$, 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 parameters $\boldsymbol{w}$ 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 $\boldsymbol{\omega}_f = \{\omega_i^f \mid i \in \{1, \dots, l_f\}\}$ 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 $\boldsymbol{\sigma}^y$:

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

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):

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

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:

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

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 $|\mathbf{x}| + 1$:

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

where $r(\boldsymbol{\alpha}, \mathbf{a}) = r(\frac{\alpha_1}{a_1}, \dots, \frac{\alpha_{|\boldsymbol{\alpha}|}}{a_{|\mathbf{a}|}})$ 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 ratio is greater than one, then the error in determining the $\hat{\omega}_i$ parameter is bigger than the measurement error 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 $\mathbf{T}_f(i) \mid i \in \{1, \dots, l_f\}$ 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 $G = \{g_1, \dots, g_k\}$ 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 STABILITY ESTIMATION METHOD

In order to estimate the stability $\mathbb{T}_{\hat{f}}$ of some model $\hat{f}$ which a solution of (7), the structure of the model $\hat{f}$ is fixed and the standard deviation of its parameters is studied as the function of the standard deviation of a noise in the learning set, as proposed in section 1.

In other words, some values for σ_λ and σ_n are chosen, then the modified learning set $\acute{D}(\sigma_n, \sigma_\lambda)$ is generated for the chosen values according to (4). The new parameters vector is then calculated which minimizes (1) on the modified learning set $\acute{D}(\sigma_n, \sigma_\lambda)$ according to (5).

This procedure is repeated multiple times for each given pair of σ_λ and σ_n until some stop condition is reached (like the number of iterations), after which empirical value for $\mathbb{T}_{\hat{f}}$ is computed.

By performing the above steps for different σ_λ and σ_n, it is possible to estimate the dependency of the superposition parameters standard deviation of the parameters σ_λ and σ_n of the noise.

It is physically sensible to expect this dependency to be smooth, while extremely non-smooth dependency means an erroneously chosen superposition and can also be a sign of overfitting: the less the parameters depend on the random error in the data, the better generalization is.

Moreover, different superpositions can be compared according to the proposed stability criteria in addition to the complexity and error criteria. In some applications the stability can even be more important than the error on the data set.

4 COMPUTATIONAL EXPERIMENT

The data used in this section are the measurements of the refraction index of transparent polymers as a function of wavelength. Two different polymers are considered, each of them having 17 data points corresponding to the refraction index at different wavelengths. The values of the measurements are shown in table 1.

The dispersion of both polymers is assumed to be described by the functional dependency of the same structure, as it obeys the same physical laws. Because of this, firstly the model $\hat{f}$ is chosen which minimizes (7) for the first polymer, and then for each of the polymers optimal

Таблица 1: Measured refraction indexes at different wavelengths.

λ, nm	Polymer 1	Polymer 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

parameter vectors $\hat{\boldsymbol{\omega}}_{\hat{f}}$ are found for the given model, and their stability is estimated.

The data set was not splitted to learning set and control set, as overfitting was mitigated by the complexity penalty.

Physical considerations show [11] that dispersion should be a sum of even powers of the wavelength, so the elementary function set consists of the functions

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

in addition to standard addition and multiplication operations:

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

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

During the experiment constants with absolute value less than 10^{-7} were zeroed.

The algorithm described above generated the following superposition at $\alpha = 0.05$, $\tau = 10$:

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

The complexity of this model is 13, MSE is $2.4 \cdot 10^{-8}$ and $Q_f \approx 0.095$. Wavelengths are assumed to be in nanometers.

It is worth noting that only two first terms are considered in practical applications, while higher powers are neglected. The value of the last term in (9) agrees with this practice.

Complexity penalty. The effect of adding odd powers to the elementary function set is studied here by replacing g_3 with

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

With the same parameters $\alpha = 0.05$ and $\tau = 10$ the resulting function is still (9).

Increasing τ up to 30 results in the following model:

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

Its complexity is 31, MSE is $\approx 3.9 \cdot 10^{-9}$ and $Q_f \approx 0.31$.

In other words, bigger target complexity (expressed by τ) leads to selecting more complex (and, in this case, physically incorrect) model with smaller mean square error.

Naturally, excessive values of τ lead to overfitting.

SVM. SVM with RBF kernel [12] was used as baseline algorithm. The γ kernel parameter was selected by crossvalidation. The best result was a combination of 15 support vectors with $\gamma \approx 2 \cdot 10^{-6}$. MSE during 2-fold crossvalidation was $8.96 \cdot 10^{-8}$. The resulting model, though, is uninterpretable.

Model stability. In order to estimate the stability, the structure of (9) had been fixed as

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

and the dependency of standard deviation of ω_1, ω_2 and ω_3 of standard deviation of a gaussian noise was studied by the method proposed above. The stop criteria was reaching 10^4 iterations for each pair of $(\sigma_\lambda, \sigma_n)$.

Таблица 2: Standard deviation for (9).

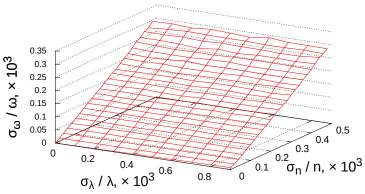
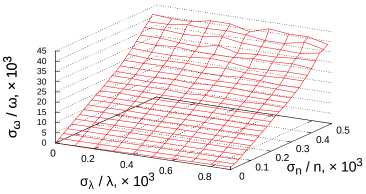
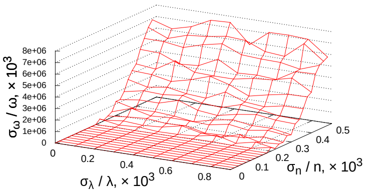
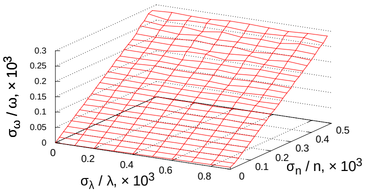
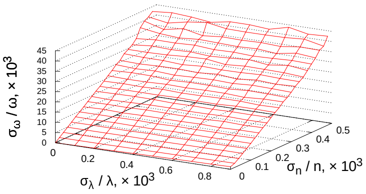
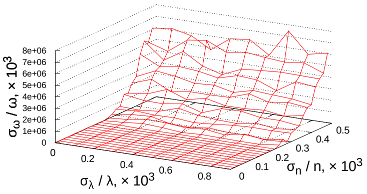
	ω_1	ω_2	ω_3
Polymer 1			
Polymer 2			

Таблица 3: Coefficients of model (9) and their relative residual.

	ω_1	ω_2	ω_3	MSE
Polymer 1	1.34946	3558.95	1924.33	$2.2 \cdot 10^{-8}$
Polymer 2	1.34047	3118.84	1578.59	$1.4 \cdot 10^{-8}$
Residual	$6.71 \cdot 10^{-3}$	$1.41 \cdot 10^{-1}$	$2.2 \cdot 10^{-1}$	

The standard deviation surfaces of each parameter ω_i are shown in table 2.

The graphs show that the wavelength measurement error does not significantly affect first and second parameters in the region of interest. At the same time, their standard deviation depends on the standard deviation of the refraction index almost linearly.

These results can be interpreted in the following way: during experiment planning most attention should be paid to maximizing the certainity in measuring the refraction index, while the wavelength can be measured quite inaccurately with errors up to few nanometers. Moreover, the suggested method directly shows how the parameters error depends on the measurement errors of different variables.

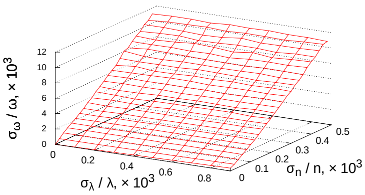
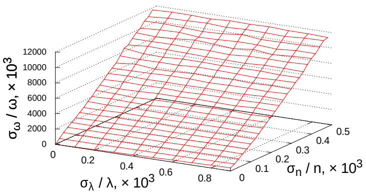
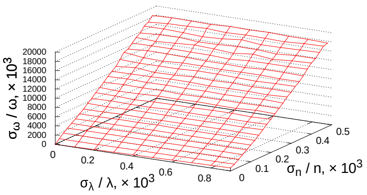
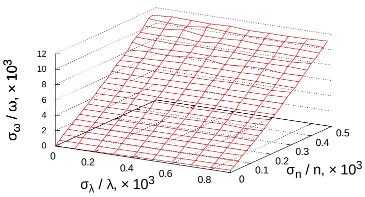
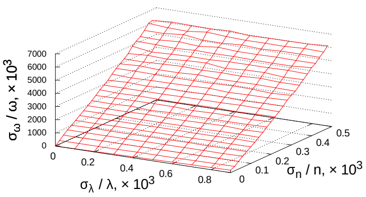
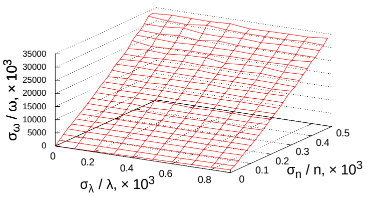
It is fundamentally important that the standard deviation of the parameters of the model (9) are considerably smaller than the difference between the parameters for two polymers (as shown by tables 3 and 4), which means that the polymers can be separated by such measurements even by an imprecise refractometer.

Таблица 4: Standard deviation of (9) parameters for the first polymer for selected noise parameters.

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

Stability of the overfitted model. The stability of the model (10) is studied analogously. Standard deviation graphs for the first three parameters are shown in table 5.

Таблица 5: Standard deviation for (10).

	ω_1	ω_2	ω_3
Polymer 1			
Polymer 2			

The graphs show that standard deviation values for (10) are considerably higher than the corresponding ones for (9). Particularly the second, third and fourth parameters have standard deviation orders of magnitude higher than their corresponding values.

These results may be a sign of overfitting, and that the resulting model can't be used to describe the physical process, not to mention separating two polymers in their mixture.

5 CONVERGENCE TO THE LINEAR CASE.

The case of linear regression is considered:

$$y = ax + b.$$

Taking the measurement errors into account:

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

where the errors ξ_i are independent, and $E(\xi_i) = 0; D(\xi_i) = \sigma^2$ [8]. In other words, the error doesn't depend on the measurement, and the dependent variable is measured precisely.

Transitioning to the following presentation:

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

results in a and b being independent normally distributed random variables, and their dispersions can be calculated according to [8]:

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

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

Next the results obtained by the proposed method are compared to the ones resulting from(11) and (12). For this, the relative difference between these values and empiric standard deviations is considered as a function of number of iterations N:

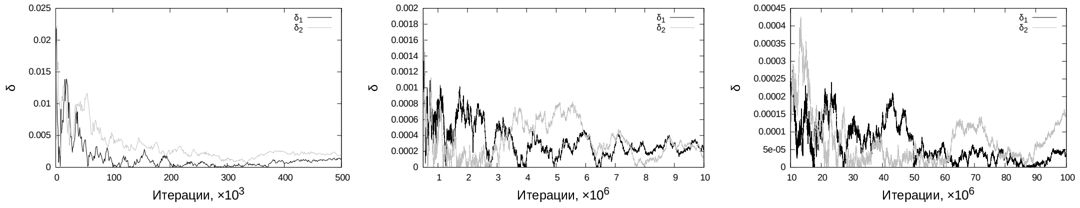
$$\delta_1 = \frac{|\mathbf{T}_y^N(1) - D(a)|}{D(a)},$$

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

Corresponding graphs for the function $y = 2x + 1 + \xi_i$ with $x \in [0; 10]$, $n = 10$ sample points and $D(\xi_i) = 10$ are shown on fig. 1. Particularly, the fig. 1a shows the initial part of the graph for N smaller than $5 \cdot 10^5$, the fig. 1б shows the part for N between $5 \cdot 10^5$ and 10^7, and fig. 1в represents the convergence on big N (from 10^7 to 10^8).

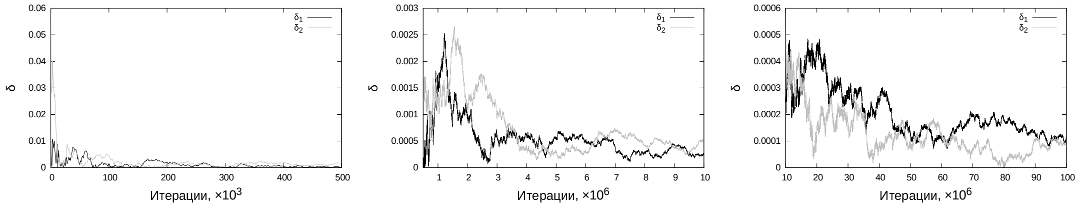
Analogous graphs are also shown for $n = 10$ and $D(\xi_i) = 1$, and $n = 50$ and $D(\xi_i) = 1$, on fig. 2 and 3 respectively.

The graphs show that the relative difference stabilizes around $(1.5 \div 3) \cdot 10^6$ iterations and doesn't explicit dependence on the number of sample points or standard deviation of the error.



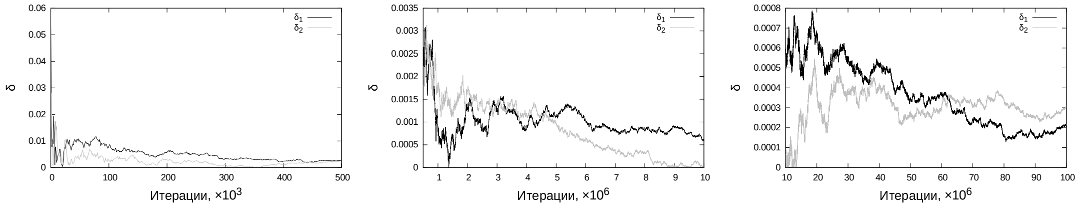
(а) $N \in [0;\ 5 \cdot 10^5]$ (б) $N \in [5 \cdot 10^5;\ 10^7]$ (в) $N \in [10^7;\ 10^8]$

Фиг. 1: Dependence of δ on N with $D(\xi) = 10$ and $n = 10$.



(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. 2: Dependence of δ on N with $D(\xi) = 1$ and $n = 10$.



(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: Dependence of δ on N with $D(\xi) = 1$ and $n = 50$.

6 CONCLUSION

The algorithm proposed in [5] allows generating interpretable analytic model describing the dependency of refraction index on the wavelength. The complexity penalty introduced in the algorithm mitigates overfitting without resorting to methods like crossvalidation.

Though other algorithms like SVM regression can learn models with lower mean square error, the learned models are uninterpretable and prone to overfitting. Moreover, their structural parameters need to be estimated according to, for example, cross-validation, while the proposed method's hyperparameters can be chosen directly according to expert considerations.

The stability criteria proposed in this paper allows studying the contribution of each term of the resulting superposition and the relation between measurement errors and errors in determining the superposition parameters. Particularly, this method also allows detecting which components of feature vectors are the least susceptible to noise in the learning data. Moreover, expertly correct models tend to be more stable than incorrect ones.

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.