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

G. Rudoy

Abstract

A set of various non-linear regression models is considered to select an optimal one describing a given physical experiment. For this, a new model selection criteria is proposed, which we will call model *stability*. This criteria shows the dependency of the model parameters on the variation of samples in the learning set. The proposed stability criteria is also used to estimate the error of determining the model parameters, which is of interest to the experts. Experimental data for refraction index of transparent polymers at different wavelengths is used as illustration of the criteria, studying several different expert-proposed models. The criteria is also illustrated for the case of a linear model, where a known theoretical estimate of parameters errors exists.

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

Introduction

Analysis of a physical experiment results often requires finding a functional dependency between the measured data. For example, given a set of measurements of wavelength and the corresponding refraction index of a substance, a dependency between them should be derived. It is also very desirable for such dependency to be interpretable by an expert in the corresponding field.

In many cases some theoretical assumptions about the structure of the functional dependency are available, or a choice should be made between various proposed models. For instance, in the above case of refraction index measurements the data can be described either by a sum of even powers of the wavelength in the common case, or by a known physical formula that is valid near the resonance wavelength.

Different models (either suggested by experts or, for example, inductively generated(Davidson et al., 2001; Рудой and Стрижов, 2013)) are usually compared by their respective errors on the measured data, and their numeric parameters are found, for instance, using the Levenberg-Marquardt algorithm(Marquardt, 1963; More, 1978).

On the other hand, in addition to the model parameters themselves the errors in determining their values resulting from the intrinsic measurement inaccuracies are also of interest to experts. The errors determine whether the physical experiment and the selected model make sense, whether its results can be used in particular applications, and they also define the requirements for the experimental devices and their precision.

This naturally leads to another model selection criteria, suggested in this paper, which we will call *model stability*, which is to be used in addition to mean square error and various kinds of model complexity. Model stability describes the dependency of the change of model parameters due to a slight variation of the data in the learning set.

For linear regression this problem has a theoretical solution(Vatunin et al., 2005) in the particular case of independent variable being measured exactly and the dependent variable having the same Gaussian distribution of the error at all measured samples. The case of non-linear regression with independent variables measured inexactly, as well as all samples having different error distributions (like varying standard deviations in case of gaussian error distribution), has not been considered as far as we know.

In this paper a few non-linear regression models are considered, describing the dependency of a liquid polymer refraction index $n(\lambda) = n(\lambda, \boldsymbol{\omega})$ where λ is the wavelength and n is parametrized by the parameter vector $\boldsymbol{\omega}$, describing a concrete polymer. The frequencies where the polymer is transparent, including visible and near infrared fields, are considered. The goal of the experimenters was to, firstly, find the dispersion for each polymer, and then derive the concentration of each polymer in their mixture, assuming the refraction index of the mixture of polymers to be a weighted sum of their refraction indexes. In other words, for two polymers characterized by model parameters $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ respectively, knowing the functions $n(\lambda, \boldsymbol{\omega}_1)$ and $n_2(\lambda, \boldsymbol{\omega}_2)$ and the mixture dispersion dependency $n(\lambda)$, the concentration α of the first polymer should be derived, since $n(\lambda) = \alpha n(\lambda, \boldsymbol{\omega}_1) + (1 - \alpha) n(\lambda, \boldsymbol{\omega}_2)$.

The refraction indexes for transparent polymers of a similar chemical composition differ only slightly. Thus, the error in determining parameters $\boldsymbol{\omega}$ of $n(\lambda) = n(\lambda, \boldsymbol{\omega})$ and their dependencies on the measurement errors of the wavelength λ and refraction index n must be considered, since if the errors in determining $\boldsymbol{\omega}$ in $n(\lambda, \boldsymbol{\omega})$ are of the same magnitude (or even bigger) as the difference between the corresponding parameters for different polymers, then the polymers are effectively indistinguishable. These dependencies are also important because they define requirements for devices and, consecutively, largely affect the cost and duration of the experiment.

Typically broad spectrum sources are used in refractometers, and the tolerance of single wavelength extraction is defined by the hardware function of the monochromator being used and is thoroughly considered, for example, in (Malishev, 1979; Zaidel, 1972). 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 angle of total internal refraction, 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 varying for each sample.

The problem of determining the stability of model parameters in the general non-linear case of multivariate models is formally stated, a method for evaluating model stability is proposed, and their dependency on the model selection parameters is studied for the given case of determining the dispersion of transparent polymers.

In the first part of this paper the problem of recovering the refraction index dependency model is formally stated, and the stability criteria is proposed. In the second part the exact numerical method for stability estimation is described. In the third part the results of the computational experiment are shown, where two polymers are considered, for each of them 17 samples are given, corresponding to the refraction index at various wavelengths.

1 Problem statement

We first consider the general case of multivariate regression problem and define stability for this general case.

Let $D = \{\mathbf{x}_i, y_i \mid i \in \{1, \dots, \ell\}\ \}$ be the sample set of ℓ measurements, where $\mathbf{x}_i \in \mathcal{R}^n$ 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 function $\hat{f} = \hat{f}(\mathbf{x}_i)$ is selected minimizing the standard loss function, assuming Gaussian error distribution:

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

where $\mathcal{F}$ is a superpositions set from which an optimal one must be found.

In other words,

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

The stability describes the variance of the parameters $\boldsymbol{\omega}$ of the model f during slight random variation of the source sample set D,

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.

Consider the parameter 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 parameter vector minimizing the functional (1) for some sample set $D = \{\mathbf{x}_i, y_i\}$ and parametric function f:

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

Let $\Sigma^{\mathbf{x}} = [\sigma_{ij}^{\mathbf{x}}], i \in \{1, \dots, \ell\}, j \in \{1, \dots, n\}$ be the matrix of standard deviations of independend variables, where $\sigma_{ij}^{\mathbf{x}}$ is the standard deviation of the j-th component of the feature vector $\mathbf{x}_i$ of the i-th object of the sample set. Let $\boldsymbol{\sigma}^y = [\sigma_1^y, \dots, \sigma_\ell^y]$ be the vector of standard deviations of the dependent variable, where σ_i^y is the standard deviation of the dependent variable for the i-th object. The modified sample set $\acute{D}$ is then considered, which is derived from the source data set D by summing its components with some realizations of the random variables from the Gaussian distribution with zero mean and deviations corresponding to $\Sigma^{\mathbf{x}}$ and $\boldsymbol{\sigma}^y$:

$$\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}^y) = \{\mathbf{x}_i + \boldsymbol{\xi}_i^{\mathbf{x}}, y_i + \xi_i^y \mid i \in 1, \dots, \ell; \boldsymbol{\xi}_i^{\mathbf{x}} \sim \mathcal{N}(0; \boldsymbol{\sigma}_{i\cdot}^{\mathbf{x}}); \xi_i^y \sim \mathcal{N}(0; \sigma_i^y)\}. \tag{3}$$

For this new sample set $\acute{D}$ the new parameter vector $\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y))$ is found for the superposition f minimizing (6):

$$\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)) = \arg\min_{\boldsymbol{\omega}_{f_D}} S(f_D(\cdot, \boldsymbol{\omega}_{f_D}), \acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)). \tag{4}$$

Let $\hat{\boldsymbol{\omega}}_f(\acute{D}(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y))$ be

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

Let $\acute{\mathcal{D}}_N$ be a set of N such modified sample 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 $\overline{\sigma}_k$ be the sample standard deviation of the k-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.

We will call the vector obtained by appending the target value y_i to the end of the corresponding feature vector $\mathbf{x}_i$ the *combined feature vector*.

Определение 1. *Relative stability* of the parameter ω_k to j-th component of the combined feature vector, given $\acute{\mathcal{D}}_N(\Sigma^{\mathbf{x}}, \boldsymbol{\sigma}_y)$ and source sample set D is the following value:

$$T_{kj}(f) = \begin{cases} \dfrac{\frac{\overline{\sigma}_k}{\hat{\omega}_k}}{r\left(\{\frac{\sigma^{\mathbf{x}}_{ij}}{x_{ij}}\}_{i \in \{1,\dots,\ell\}}\right)} & j \leq n \\ \dfrac{\frac{\overline{\sigma}_k}{\hat{\omega}_k}}{r\left(\{\frac{\sigma^{y}_{i}}{y_{i}}\}_{i \in \{1,\dots,\ell\}}\right)} & j = n+1 \end{cases} \tag{5}$$

where r is a function mapping a vector of quotients to a single scalar value, and n is the dimensionality of the feature space.

The function r maps a set of (perhaps different) ratios of standard deviation of a measured variable to the value of that variable to a single scalar value. The mapped scalar can be viewed as 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, for instance the first one: $r(\alpha_1, \alpha_2, \dots) = \alpha_1$.

$T_{ij}(f)$ describes the ratio between the standard deviation of the $\hat{\omega}_i$ parameter (normalized by the value of that parameter) and the some characteristic (defined by r) standard deviation of the corresponding j-th feature vector component (again, normalized by the value of that component). 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 corresponding variable.

In the particular case of the dispersion regression considered in this paper, (1) is

$$S(f, D) = \sum_{i=1}^{\ell} (f(\lambda_i) - n_i)^2 \to \min_{f \in \mathcal{F}}, \tag{6}$$

and, taking into account the constant relative measurement error:

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

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

In case of the optical dispersion regression, it is required to study the dependency of stability $T_{i0}(f)$ and $T_{i1}(f)$ as function of σ_n and σ_λ.

For simple cases of regression functions depending on a single scalar parameter (like the optical dispersion case illustrating this approach) it is probably more interpretable to study the dependency of *absolute stability* $\frac{\overline{\sigma}_i}{\hat{\omega}_i}$ on the normalized slight variations in sample set, $\frac{\sigma_n}{n}$ and $\frac{\sigma_\lambda}{\lambda}$.

2 The algorithm for inductive models generation

In this section we briefly describe the algorithm proposed in (Рудой and Стрижов, 2013).

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. Each superposition in $\mathcal{F}$ may contain free variables corresponding to the components of the feature vectors from the sample set $\mathcal{D}$, as well as constants, which are considered as the superposition parameters $\boldsymbol{\omega}_f$. The algorithm then works in lockstep: the parameters $\boldsymbol{\omega}_f$ are optimized by the Levenberg-Marquardt procedure according to (6), and then each superposition can be modified 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 sample set D 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 (6) on the sample set D;

C_f is the complexity of the superposition f, equal to 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 sample set. The primary hypothesis is that simpler superpositions (even with slightly bigger sample set errors) generalize better. It is worth noting that α and τ are chosen by experts.

The initial problem of minimizing the functional (6) is replaced by the problem of finding the superposition $\hat{f}$ minimizing (7):

$$\hat{f} = \arg\min_{f \in \mathcal{F}} Q_f, \tag{8}$$

and minimizing (6) becomes an intermediate step.

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 sample set, as proposed in section 1.

For the case of transparent polymers dispersion curve considered, some values for σ_λ and σ_n are chosen, then the modified sample set $\acute{D}(\sigma_n, \sigma_\lambda)$ is generated for the chosen values according to (3). The new parameter vector is then calculated which minimizes (6) on the modified sample set $\acute{D}(\sigma_n, \sigma_\lambda)$ according to (4).

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 on the parameters σ_λ and σ_n of the noise.

It is reasonable to expect that smooth dependency of superposition coefficients on the noise means stable (in expert sense) model, 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.

4 Computational experiment

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

Table 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

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 parameter vectors $\hat{\boldsymbol{\omega}}_{\hat{f}}$ are found for the given model $\hat{f}$, and their stability T 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 (Serova, 2011) that dispersion should be a sum of even powers of the wavelength, so the elementary function set consists of the function

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

and 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 computational 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.

Thus, excessive values of τ lead to overfitting.

Support vector machine. SVM with RBF kernel (Vapnik, 1979) 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 proposed method. The stop criteria was reaching 10^4 iterations for each pair of $(\sigma_\lambda, \sigma_n)$.

The standard deviation surfaces for parameters ω_i for both polymers 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.

Table 2: Standard deviation for (9).

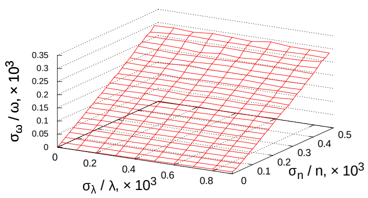
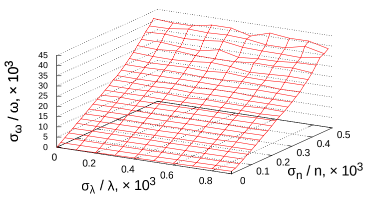
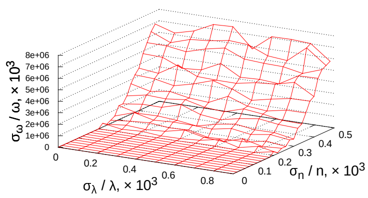
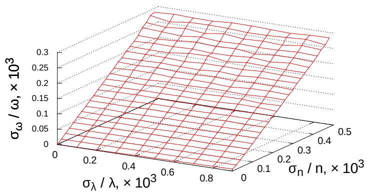
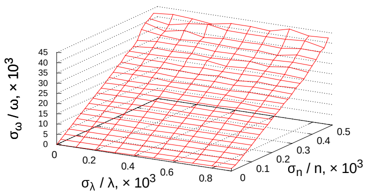
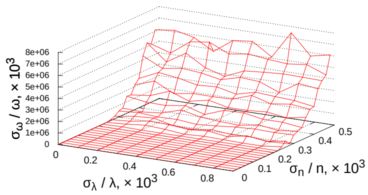
	ω_1	ω_2	ω_3
Polymer 1			
Polymer 2			

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

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 with an imprecise refractometer.

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

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, and can not be used to separate two polymers in their mixture.

5 Convergence to the linear case.

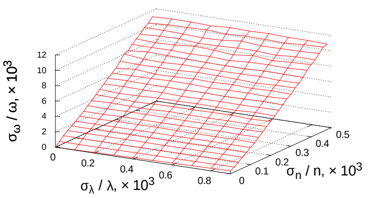
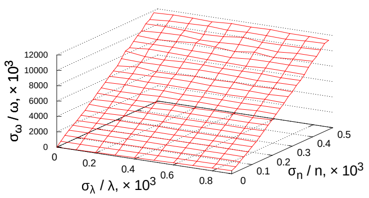
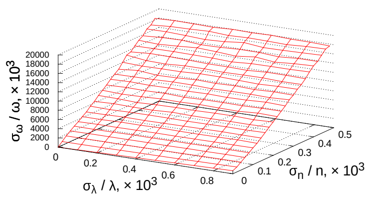
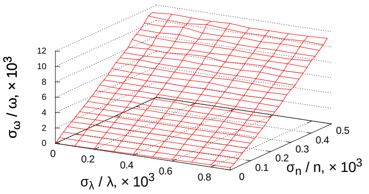
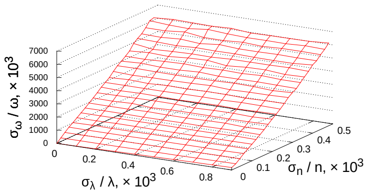
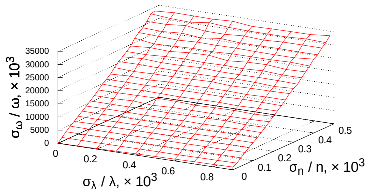
The case of linear regression is considered:

$$y = ax + b.$$

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

Table 5: Standard deviation for (10).

	ω_1	ω_2	ω_3
Polymer 1			
Polymer 2			

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 $E(\xi_i) = 0; D(\xi_i) = \sigma^2$ (Vatunin et al., 2005). In other words, the error doesn't depend on the measurement, and the independent variable is measured precisely.

According to (Vatunin et al., 2005) for the following presentation:

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

a and b are independent normally distributed random variables, and their dispersions can be exactly calculated:

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

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

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

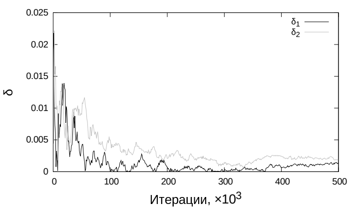
$$\delta_1 = \frac{|\overline{\sigma}_a^2 - D(a)|}{D(a)},$$

$$\delta_2 = \frac{|\overline{\sigma}_b^2 - D(b)|}{D(b)}.$$

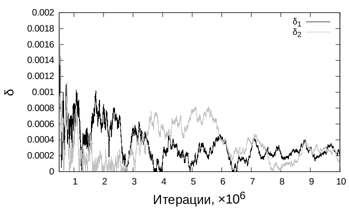
Corresponding graphs for the function $y = 2x + 1 + \xi_i$ with $x \in [0; 10]$, $n = 10$ samples and $D(\xi_i) = 10$ are shown on fig. 1. Particularly, the fig. 1a shows the initial part of the graph for N smaller than $5 \cdot 10^5$, the fig. 1b shows the part for N between $5 \cdot 10^5$ and 10^7, and fig. 1c 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.

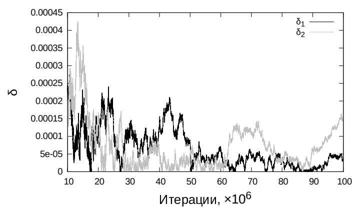
The graphs show that the relative difference stabilizes around $(1.5 \div 3) \cdot 10^6$ iterations and doesn't explicit dependence on the number of samples or standard deviation of the error. The resulting relative difference is close to zero, but is not equal to it probably to rounding errors and other similar computational effects.



(a) $N \in [0;\ 5 \cdot 10^5]$

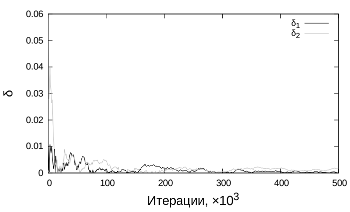


(b) $N \in [5 \cdot 10^5;\ 10^7]$

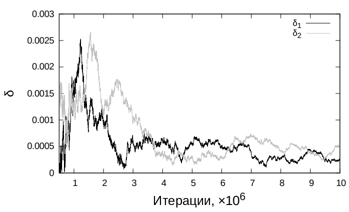


(c) $N \in [10^7;\ 10^8]$

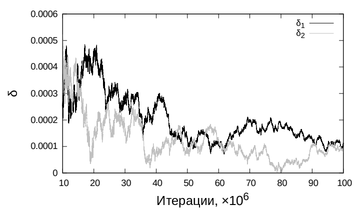
Figure 1: Dependence of δ on N with $D(\xi) = 10$ and $n = 10$.



(a) $N \in [0;\ 5 \cdot 10^5]$

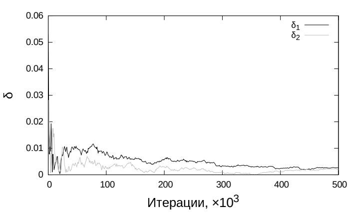


(b) $N \in [5 \cdot 10^5;\ 10^7]$

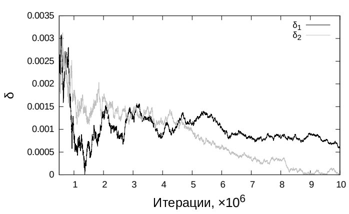


(c) $N \in [10^7;\ 10^8]$

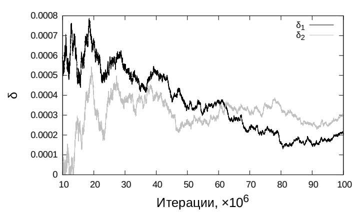
Figure 2: Dependence of δ on N with $D(\xi) = 1$ and $n = 10$.



(a) $N \in [0;\ 5 \cdot 10^5]$



(b) $N \in [5 \cdot 10^5;\ 10^7]$



(c) $N \in [10^7;\ 10^8]$

Figure 3: Dependence of δ on N with $D(\xi) = 1$ and $n = 50$.

6 Conclusion

The algorithm proposed in (Рудой and Стрижов, 2013) 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, such 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 in the overall error and the relation between measurement errors and errors in determining the superposition parameters. Particularly, the offered 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.

References

J. W. Davidson, D. A. Savic, and G. A. Walters. Symbolic and numerical regression: experiments and applications. In Robert John and Ralph Birkenhead, editors, *Developments in Soft Computing*, pages 175–182, De Montfort University, Leicester, UK, 29-30 June 2000. 2001. Physica Verlag. ISBN 3-7908-1361-3.

V. I. Malishev. *Introduction to experimental spectroscopy.* Nauka, 1979.

D. W. Marquardt. An algorithm for least-squares estimation of non-linear parameters. *Journal of the Society of Industrial and Applied Mathematics*, 11(2):431–441, 1963.

J. J. More. 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`.

C. Sammut and G. I. Webb. Symbolic regression. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, page 954. Springer, 2010. ISBN 978-0-387-30768-8. URL `http://dx.doi.org/10.1007/978-0-387-30164-8`.

N. V. Serova. *Polymeric optical materials.* Scientific fundamentals and technologies, 2011.

V. Strijov and G.-W. Weber. Nonlinear regression model generation using hyperparameter optimization. *Computers & Mathematics with Applications*, 60(4):981–988, 2010. URL `http://dx.doi.org/10.1016/j.camwa.2010.03.021`.

V. N. Vapnik. *Dependencies recovering from empirical data.* Nauka, 1979.

V. A. Vatunin, G. I. Ivchenko, Yu. I. Medvedev, and V. P. Chistyakov. *Probability theory and mathematical statistics in problems.* Drofa, 3 edition, 2005.

I. N. Zaidel. *Spectroscopy technics and practice.* Nauka, 1972.

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

В. В. Стрижов. Методы индуктивного порождения регрессионных моделей. Препринт ВЦ РАН им. А. А. Дородницына. — М., 2008.