BALANCED IDENTIFICATION: TESTING HYPOTHESES ABOUT THE FUNCTIONING OF AN OBJECT BY EXPERIMENTAL DATA PROCESSING

Anonymous authors
Paper under double-blind review

ABSTRACT

The general formulation of the identification problem is considered as an optimization task. The issues related to changing the set of feasible solutions when adding additional (or modifying existing) hypotheses are discussed. A balanced identification technology is used to connect mathematical models with data. The dynamics of some statistical estimates of modeling accuracy were investigated when adding "correct and incorrect" hypotheses on both methodological (damped pendulum) and real examples (plant physiology and pollution by heavy metals in the Kola Peninsula).

1 INTRODUCTION

The same object, process, or phenomenon can be described by various mathematical models. Choosing a suitable model among possible candidates remains one of the pressing problems in applied mathematics. The mathematical description depends both on qualitative knowledge about the object (which can be formalized as mathematical statements or hypotheses) and on the availability of quantitative data (measurements), their volume, detail, reliability, and accuracy. The more complex the phenomenon under consideration and the corresponding mathematical model, the more detailed and reliable the measurements should be.

This study considers the step-by-step construction of models based on balanced identification technology (see Appendix A or Sokolov & Voloshinov (2018; 2020)). At each step, an attempt is made to modify the model—by adding, removing, or changing the hypothesis about the functioning of the object, formalized as a mathematical description. The criterion for the success of the modification is the cross-validation mean square error—its reduction or insignificant increase can serve as an argument in favor of accepting the hypothesis.

This approach is illustrated by three examples: the step-by-step construction of a pendulum model based on artificially prepared data, the construction of a photosynthesis model for pine based on real data Sokolov & Bolondinskii (2020), and a model of lake pollution in the Kola Peninsula by heavy metals based on real data Sokolov et al. (2023).

An analytical study of the settings considered below is not possible, so they are replaced by numerical analogs. For this purpose, functions are replaced by piecewise or polynomial approximations, integrals by sums, and derivatives by differences. The results were obtained from numerical models.

The special SvF-software[1] used in this work, program implementation of some examples and corresponding data[2] are available in open access. This allows the calculations to be replicated.

2 METHOD OF SVF-TECHNOLOGY

Balanced identification technology (see Appendix A or Sokolov & Voloshinov (2020)) is a means of constructing (selecting) mathematical models whose complexity corresponds to the quantity and

[1] https://github.com/distcomp/SvF
[2] https://github.com/distcomp/SvF/tree/main/Examples/1-Oscillator

quality (errors) of experimental data. A quantitative assessment of model complexity and the specification of the norm for the closeness of model characteristics to experimental data allow the formulation of the problem of choosing an optimally balanced solution based on minimizing the cross-validation mean square error (CV).

Let us illustrate the application of the technology with a simple example. Suppose there is a set of experimental data (points x-data in Figure 1):

$$D : \{x_i, t_i\}, i \in I, I = 1, i_{max} = 21 \tag{1}$$

and it is required to construct a model in the form of a bounded (by rectangular Region in Figure 1) twice continuously differentiable function $x(t)$ describing these data:

$$M[x] = \begin{cases} x \in C^2[-1, 2.5] \\ -0.1 \leq x(t) \leq 2.2 \end{cases} \tag{2}$$

The mathematical model (2) defines a set of feasible solutions in the function space (as is customary in optimization problems):

$$S = \{x : M[x]\}. \tag{3}$$

SvF-method has been inspired by Tikhonov regularisation method Tikhonov (1980) and a rule to characterize the smoothness of functions Green & Silverman (1993), Sec. 7. As a rule, the set of feasible solutions is too broad, and to choose an appropriate solution, an optimization criterion must be set, which should include a measure of the reproducing of experimental data and impose some desirable properties on the solution. Here, the criterion used is

$$F(x, I, \alpha) = \frac{1}{|I|} \sum_{i \in I} (x_i - x(t_i))^2 + \alpha \int_{-1}^{2.5} (x'')^2 dt. \tag{4}$$

The first term accounts for the closeness of the solution to the data, and the second term can be considered a measure of model complexity expressed through the curvature of the function. The parameter α, which we call the complexity penalty, allows the terms to be "mixed" in various proportions and thus choose solutions with different properties from the set of feasible solutions S:

$$x_I^\alpha(t) = \arg\min_{x \in S} F(x, I, \alpha). \tag{5}$$

In Figure 1, three solutions of problem (5) are shown for different values of α. At small α, the first term suppresses the second one, and the overfitted solution passes through all data points but also repeats measurement errors. At large α, the second term straightens the solution (minimizes curvature), and its closeness to the data becomes secondary (underfitted model). Finally, the optimally balanced solution is presented in the figure. It is obtained by minimizing the cross-validation mean square error with respect to the parameter α Sokolov & Voloshinov (2020).

Here, we use the simplest cross-validation variant, where the entire dataset (I) except for one point (i) is used for training. The obtained solution $(x^\alpha_{I\setminus\{i\}}(t)$ is used for comparison with the test sample, where is the value at the removed point (i). This procedure is performed for all (i). Summing the discrepancies across all points and normalizing by the root mean square deviation of the initial data $\sigma(D)$, we obtain the cross-validation mean square error:

$$CV(\alpha) = \sqrt{\frac{1}{|I|} \sum_{i \in I} (x_i - x^\alpha_{I\setminus\{i\}}(t_i))^2} \frac{100\%}{\sigma(D)}. \tag{6}$$

Additionally, we use the root mean square deviation of the model from the data:

$$SD(\alpha) = \sqrt{\frac{1}{|I|} \sum_{i \in I} (x_i - x^\alpha_I(t_i))^2} \frac{100\%}{\sigma(D)}. \tag{7}$$

The values of the introduced indicators of the solutions depicted in Figure 1 are given in Table 1. As expected, as the penalty for complexity (curvature) α increases, the standard deviation SD increases, while the cross-validation error CV first decreases, reaches a minimum (at α=0.1166), and then begins to increase. The minimum value is used as the modeling error for data (1) with model (2). It determines the optimally balanced solution. Thus, the balanced identification method

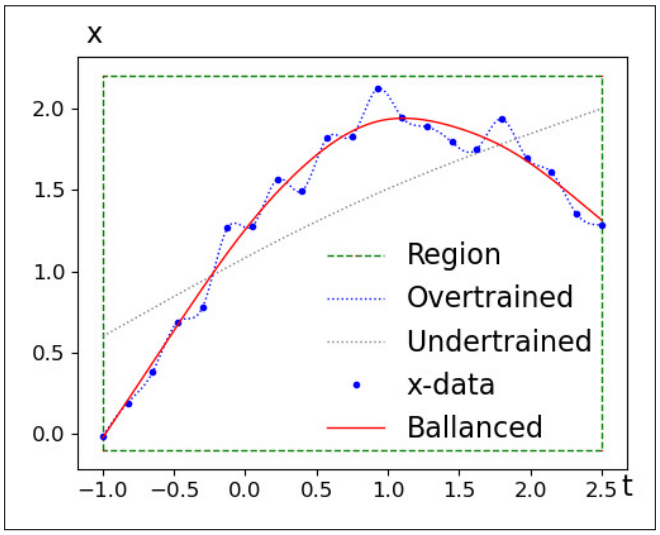


Figure 1: Three data approximation variants: underfitted, overfitted, and optimally balanced.

Table 1: The value of the penalty α determines the model type and errors CV and SD.

$\boldsymbol{\alpha}$	**MODEL**	**CV%**	**SD%**
0.001	Overtrained	34.67	2.95e-05
0.1166	Ballanced	18.60	14.78
0.5	Undertrained	72.33	63.31

consists in finding the penalty (for complexity) α^* minimizing $CV(\alpha)$ (considering (1)-(6)):

$$\alpha^* = \arg\min_{\alpha \geq 0} CV(\alpha). \tag{8}$$

Let's show how the balanced identification method can be used to check additional hypotheses during stepwise model modification. Suppose we have a solution of problem (1)-(6),(8) for some model $M[x]$. Assume we have reasons to enhance the model by additional statement (hypothesis) $H[x]$, represented as a system of constraints on x:

$$M_H[x] = \begin{cases} M[x] \\ H[x] \end{cases} \tag{2'}$$

$$S = \{x : M[x]; H[x]\}. \tag{3'}$$

Obviously, adding additional constraints reduces the set of feasible solutions: $S_H \subset S$.

Substitute S_H into (3) and solve (1)-(7) with the fixed α^* obtained by solving the original problem. Denote the resulting errors as CV_H and SD_H. Compare them with the errors of the original problem.

Analyzing the dynamics of the cross-validation error CV is the main way to test the validity of a hypothesis. If the hypothesis successfully refines the model, then the discarded part of the set of feasible solutions $S \backslash S_H$ is insignificant, and the cross-validation error does not increase. If the discarded part is significant and the modified model becomes too narrow to approximate the data, then an increase in CV is observed, and the hypothesis is rejected.

Another important parameter is the standard deviation SD. As a rule, $SD_H \geqslant SD$. Indeed, reducing the set of feasible solutions by $S \backslash S_H$ while keeping the selection criterion (4) unchanged (α=α^* fixed) leads to an increase (not a decrease) in the minimum of the optimization problem. In this case, both terms (4) usually increase, and the first term is the square of SD (multipled by $(\sigma(D)/100)^2$). So we can expect some increase in SD.

If the error SD increased significantly, it is likely that the discarded part of the set of feasible solutions $S \backslash S_H$ contains significant parts of the model, complicating data replication. In this case, accepting the hypothesis seems inappropriate.

3 DAMPED PENDULUM

To illustrate the proposed approach for hypothesis testing, we'll consider a methodological example using the dataset (1) "prepared" (using a random number generator) by the following formula:

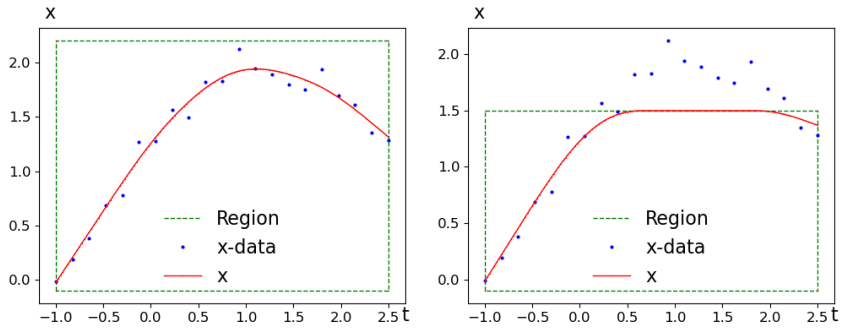
$$x(t) = \sin(K{\cdot}t){\cdot}\exp^{\frac{\mu}{2}{\cdot}t} + x_r + \varepsilon,$$

which corresponds to the model of a pendulum with viscous friction:

$$x'' = -K{\cdot}(x - x_r) - \mu{\cdot}v; \quad v = x'.$$

3.1 FUNCTION: $x(t) \leqslant 1.5$

A clear and obvious example of hypothesis testing is shown in Figure 2.



a) initial model b) with hypothesis

Figure 2: The initial model and the model with the additional hypothesis $x(t)\leqslant 1.5$.

The additional hypothesis $x(t) \leq 1.5$ significantly narrows the set of feasible solutions. The set of functions within the rectangle (Region) does not satisfactorily describe the data. Errors (see Table 1) increased significantly. This hypothesis is rejected.

3.2 FIRST-ORDER ODE: $x'{=}f(x)$

Suppose that data (1) are described by an autonomous first-order differential equation (ODE).

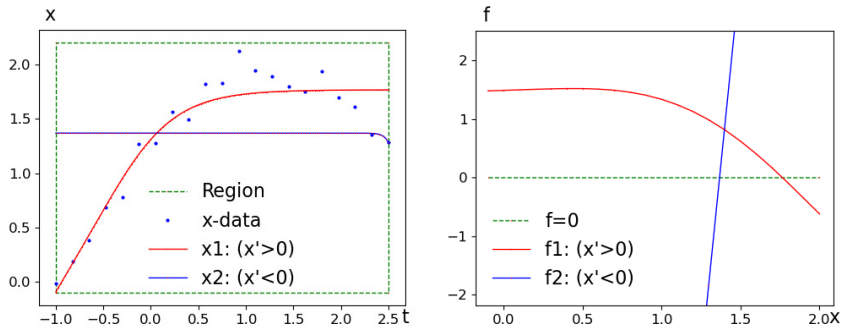
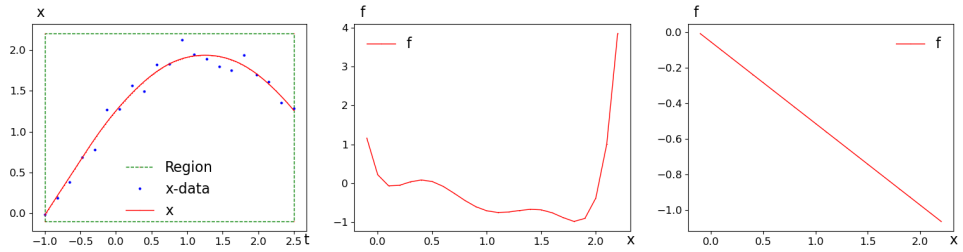


Figure 3: Two solutions of the first-order ODE model with positive and negative derivatives.

It is easy to see that the additional hypothesis imposes the monotonicity property on the solution. Thus, the set of feasible solutions (the set of monotonic functions within the rectangle Region) is not sufficiently extensive, leading to solutions far from the data (see Figure 3) and resulting in a significant increase in errors (see Table 2). The hypothesis is rejected.

3.3 SECOND-ORDER ODE: $x''=f(x)$

Let's try an autonomous second-order ODE with the right-hand side depending on x.



A B C

Figure 4: Solutions of the second-order ODE model (A, B) and the right-hand side with an additional penalty (C).

An error analysis (see Table 2) shows a decrease in CV and a moderate increase in SD. The hypothesis is not rejected. Note the somewhat chaotic behavior of function $f(x)$ in Figure 4B. To regularize it, an additional term (penalty for the complexity of function $f(x)$) is added to the functional:

$$F(x, I, \alpha) = \frac{1}{|I|}\sum_{i\in I}(x_i - x(t_i))^2 + \alpha_1 \int_{-1}^{2.5}(x''_{tt}(t))^2 dt + \alpha_2 \int_{-0.1}^{2.1}(f''_{xx}(x))^2 dx. \tag{9}$$

The solution and errors are calculated for this case. A decrease in CV and an acceptable increase in SD support the hypothesis, and examining the solution shown in Figure 4C (where f(x) is a straight line) leads to a new hypothesis about the parametrization of the right-hand side (oscillator equation): $x'' = -K\cdot(x - x_r)$.

Using this parametrization practically does not change the errors, which supports this hypothesis.

3.4 SECOND-ORDER ODE: $x''=f(x); v=x'$

Consider an autonomous second-order ODE where the right-hand side depends not only on x but also on its rate of change v. Similar to the previous example, a penalty for the complexity of the right-hand side was added to the functional to obtain the solution (not shown here). Figure 5 shows the obtained solution for the right-hand side of the ODE.

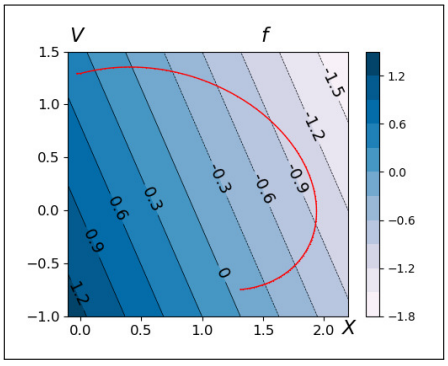


Figure 5: The right-hand side of the ODE. Contour lines of the function $f(x, v)$.

A significant decrease in CV and almost the same SD support the hypothesis, and examining the solution shown in Figure 5 (the contour map of $f(x, v)$) leads to a new hypothesis about linearity of the right-hand side (oscillator equation with viscous friction): $x'' = -K\cdot(x - x_r) - \mu\cdot v$.

Using this parametrization practically does not change the errors, which supports this hypothesis.

Thus, we have built the model:

$$M[x] = \begin{cases} x \in C^2[-1, 2.5] \\ -0.1 \leq x(t) \leq 2.2 \\ x'' = -K \cdot (x - x_r) - \mu \cdot v \end{cases} \tag{10}$$

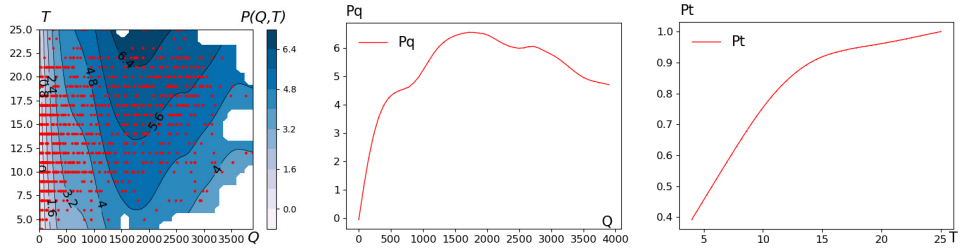
that best describes the dataset (1).

Table 2: CV&SD for various models and their folder names with problem statements in Examples

MODEL	CV%	SD%	FOLDER NAME
$x(t), x \leqslant 2.2$	18.60	14.78	0d-x(t)-x.Comp-opt=1166_x<2.2
$x(t), x \leqslant 1.5$	43.60	42.58	0d-x(t)-x.Comp-opt=1166_x<1.5
$x'=f(x), x' \geqslant 0$		31.82	1a-x'=f(x)-Poly(x'><0)
$x' \leqslant 0$	103.03	97.54	1a-x'=f(x)-Poly(x'><0)
$x''=f(x)$	18.29	15.43	21a-x″=f(x)_x.Comp-1166
$x''=f(x), f.Comp$	18.06	15.67	21b-x″=f(x)_x.Comp-1166-f.Comp
$x''=-K \cdot (x - x_r)$	18.06	15.67	21c-x″=-K_(x-xr)_x.Comp-1166
$x''=f(x, v); v=x'$	17.99	15.20	22b-x″=f(x,v)_x.Comp-1166_f.Comp-opt
$x''=-K \cdot (x - x_r) - \mu \cdot \nu$	17.99	15.16	22c-x″=-K_(x-xr)-mu_v-x.Comp-1166

4 PHOTOSYNTHESIS OF PINE

The results presented below are based on real experimental data (time series) obtained in the summer of 1977 in Karelia Sokolov & Bolondinskii (2020). The accuracy of the data is not high, and many important factors (e.g., stomatal regulation) are not taken into account, leading to significant modeling errors. Figure 6A shows the results of modeling pine photosynthesis P(Q,T), where Q is solar radiation and T is temperature. The errors are: CV = 50.80, SD = 48.80.



A B C

Figure 6: A - model $P(Q, T)$. B, C - model $Pq(Q) \cdot Pt(T)$.

Suppose the multiplicative representation $P(Q, T) = Pq(Q) \cdot Pt(T)$ holds. The results are shown in Figures 6B and 6C. The errors (CV = 50.26, SD = 47.95) noticeably decreased. The hypothesis is accepted.

Suppose in the multiplicative model the function $Pq(Q)$ does not decrease with increasing Q. See results in Figure 7. The errors are: CV=50.26 and SD=47.95.

CV did not actually change, while SD noticeably increased. The hypothesis requires further investigation or modification.

5 NICKEL POLLUTION IN THE LAKES OF THE KOLA PENINSULA

Below are the results of modeling the Ni concentration in the water of the lakes of the Kola Peninsula, based on processing a unique data set on the state of about 100 lakes from 1990 to 2019, with an interval of 4-5 years Sokolov et al. (2023). Figure 8 presents the results of using two models: (A)

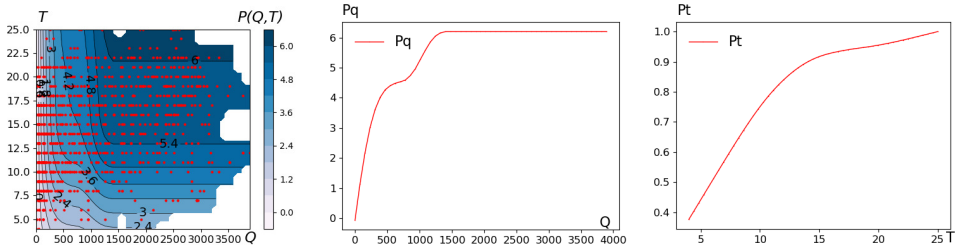
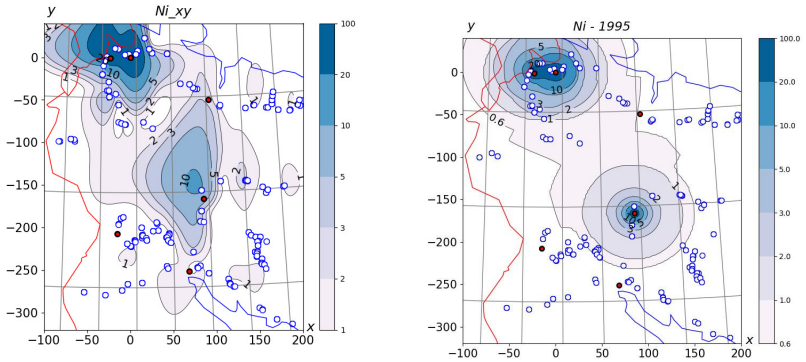


Figure 7: Model $Pq(Q) \cdot Pt(T)$ with the condition that $Pq(Q)$ does not decrease.

a simple function of coordinates X and Y, $\mathsf{Ni}(X, Y)$ and (B) a more complex model (not shown here), with the additional assumption that pollution sources are located at three points.



A B

Figure 8: Nickel pollution (Ni) in the lakes of the Kola Peninsula: A - a simple function of coordinates $\mathsf{Ni}(X, Y)$, B - model with three pollution sources.

The cross-validation error for model B (CV = 42.5) turned out to be significantly lower than for model A (CV = 54.4). Thus, the hypothesis of three pollution sources can be accepted.

6 Conclusion

In studying complex objects, the number of models describing their properties can be quite large. If we build a mathematical model in an evolutionary way - from simple to complex - the task can be represented as choosing a model from a tree of candidate models, which can be multi-level with many branches. To formalize such a choice, it is proposed to use the technology of balanced identification Sokolov & Voloshinov (2018; 2020), and to assess the quality of the model and justify its modification using the mean squared cross-validation error. At the same time, a decrease in CV with each successive modification of the model (e.g., when adding an additional hypothesis) is an argument in favor of such a change.

The presented results demonstrate the effectiveness of using the mean squared cross-validation error and standard deviation to investigate the acceptability of additional hypotheses on relatively simple illustrative examples. This approach can be generalized to more complex cases with heterogeneous data and a large number of unknown parameters and functions. In such cases, the model and both parts of the selection criterion (4) become more complex. Examples of setting up and step-by-step solving such problems can be found in Sokolov & Bolondinskii (2020), Sokolov et al. (2019) and Sokolov et al. (2023).

ACKNOWLEDGMENTS

The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation (theme No. 122040800139-4)

REFERENCES

Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1993.

A.V. Sokolov and V.K. Bolondinskii. Data-Driven Modeling of Stomatal, Mesophyll, and Biochemical Regulation of Scots Pine Photosynthesis. *Geochemistry International*, 58:1145–1158, 2020.

A.V. Sokolov and V.V. Voloshinov. Choice of mathematical model: balance between complexity and proximity to measurements. *International Journal of Open Information Technologies*, 6(9), 2018.

A.V. Sokolov and V.V. Voloshinov. Model selection by balanced identification: the interplay of optimization and distributed computing. *Open Computer Science*, 10(1):283–295, 2020.

A.V. Sokolov, V.K. Bolondinsky, and V.V. Voloshinov. Technology of balanced identification for selection of pine transpiration mathematical model. *Matematicheskaya Biologiya i Bioinformatika*, 14(2):665–682, 2019.

A.V. Sokolov, T.I. Moiseenko, N.A. Gashkina, and Yu.G. Tatsiy. Modeling and Predicting the Environment State in the Impact Area of a Copper–Nickel Plant: A Balanced Model of the Transformations of Atmospheric Deposition at the Catchment and in Lake. *Geochemistry International*, 61(7):768–779, 2023.

A.N. Tikhonov. On mathematical methods for automating the processing of observations. In *Problems of Computational Mathematics*, pp. 3–17, 1980.

A ADDITIONAL INFORMATION ABOUT SVF-TECHNOLOGY

The SvF balanced identification technology (Simplicity vs Fitting) allows the selection of mathematical models based on a balance between some measure of model complexity and accuracy of compliance with available experimental data. The technology is based on the following procedures.

Procedure 1. Setting a set of model families that depend on unknown parameters and possibly unknown functional dependencies. Among them, it is necessary to find a model that corresponds to the quality (accuracy) and quantity of experimental information and meets assumptions about the functioning of the object. In the future, an optimally balanced model and corresponding modeling accuracy indicators are found for each family. Based on these indicators and the opinion of experts on the object of research, a decision is made on the choice of a model or the direction of its further modification. As a rule, the process is iterative: the set of model families is replenished after analyzing the previously considered ones.

Procedure 2. Setting the selection criteria. For each family of models, it is necessary to set a selection criterion containing a weighted sum of the measure of the model's proximity to experimental data and the measure of the model's complexity.

The deviation of the model from the data is determined by the measurement error (depends on measuring instruments) and the error in describing the studied object by this model, which occurs when replacing a complex real object with a simplified mathematical description. The latter is usually unknown and its evaluation is one of the goals of SvF technology. Usually, the standard error is chosen to estimate the deviation from the data.

The complexity measure of the model is selected based on the specifics of the problem and of the object under study. It is necessary to identify the key function (or functions) of the model that largely determines the behavior of the model. For this function, it is necessary to formalize the concept of complexity, for example, use the function curvature. The complexity of the model can be

considered as a regularizing additive that should make the identification problem statement correct. An unsuccessful choice of complexity measure can greatly distort the solution. An analysis of such a solution (for example, an analysis of residuals) can indicate in favor of changing the method of describing the complexity.

The selected identification criterion contains some parameter(s) defining the weight of the parts responsible for proximity to the data and the model complexity. Depending on its value, the resulting model may be under-trained (the complexity weight in the criterion is too high), over-trained (the complexity weight is too low), or optimally balanced - have a minimum mean square cross-validation error.

Procedure 3. Cross-validation procedure. The cross-validation procedure consists of (repeatedly) dividing the data set into two subsets: the training subset and the testing subset. The solution of the optimization problem obtained on the training set is evaluated (the error is calculated) on the testing set. Then another division is taken and the procedure is repeated. The obtained errors are summed up. The result is the mean square error of cross-validation. The implementation specific of the cross-validation procedure depends on the specifics of the data: when studying the pendulum, one point was discarded for testing, when studying pine photosynthesis - data for one day, when studying nickel pollution of lakes on the Kola Peninsula - one lake (all measurements related to it).

Procedure 4. Search for an optimally balanced solution. The search for an optimal compromise between the complexity of the model and the proximity to the measurements is carried out by minimizing the mean square cross-validation error.

Here is a formal description of the described procedures.

Let's denote the selected mathematical description (model), in a form of equations (including differential and/or integral if any), inequalities etc.:

$$M\left[x\right], x \in X,$$

where x is a vector of corresponding vector space X (unknown parameters and functions, if any)

The model defines the corresponding of feasible domain:

$$Q = \{x \in X : M\left[x\right]\}$$

Let's denote the set of experimental data, where d_i are vector of measurements and $P_i(\cdot)$ are projectors linking measurements to model variables:

$$D = \left\{d_i, P_i(x) : i{\in}I\right\}, I{=}1{:}i_{max}.$$

Let's denote the criterion for a balanced choice of model parameters:

$$F(x, I, \alpha) = F_F(x, I) + F_S(\alpha, x),$$

where the first term is responsible for the proximity of the model to the data, the second is a measure of its complexity, and α is the penalties vector, e.g. regularization coefficients. As an example see (4) or (9).

For a given experimental data set I and the penalty α, it is possible to obtain a solution of the identification problem:

$$x_I^\alpha = \arg\min_{x \in Q} F(x, I, \alpha).$$

The search for the optimal α for the selected model is the main computational procedure of the SvF-technology. It is based on minimizing the mean square error of cross-validation. Let's select two subsets in the dataset I: the training subset E and the testing subset T:

$$E \subset I, T \subset I, E \cap T{=}\varnothing.$$

Let's solve the following optimization problem on the training dataset E for a given α:

$$x_E^\alpha = \arg\min_{x \in Q} F(x, E, \alpha),$$

and use the found solution to estimate the error of the model on the testing dataset T:

$$CV^{\alpha}_{E,T} = \sqrt{\frac{1}{|T|}\sum_{i\in T}(d_i - P_i(x^{\alpha}_E))^2}$$

This procedure is usually repeated several times (N) with different selections of training and testing sets:

$$\{E_n, T_n\}\,; n{=}1{:}N.$$

Note that optimization problems to find $x^{\alpha}_{E_n}$ are independent and may be solved in parallel on available computing resources.

The resulting errors can be combined to obtain a mean square error of cross-validation:

$$CV(\alpha) = \sqrt{\frac{1}{\sum\limits_{n=1:N}|T_n|}\sum_{n=1:N}\sum_{i\in T_n}(d_i - P_i(x^{\alpha}_{E_n}))^2}.$$

Then we modify penalty vecor a to minimize $CV(\alpha)$ (by α). Thus, for a given model SvF-technology solves bilevel optimization problem, with minimization by α on the upper level. Minimization leads to a balanced model that optimally combines proximity to measurements and simplicity. The obtained minimum value of the mean square error of cross-validation will be considered as the quality of the model, as the error of data modeling.