Multi-Penalty Regularization for Detecting Relevant Variables

Kateřina Hlaváčková-Schindler, Valeriya Naumova, and Sergiy Pereverzyev Jr.

1 Introduction and Description of Approach

Natural and social phenomena usually emerge from the behavior of complex systems consisting of interacting components or variables. In practice, we do not have a direct access to the "laws" governing the underlying relationships between them; instead, we are faced with a dataset recorded from the possibly interacting variables. How can we tell from these given data whether there exists any relationship between two or more variables?

This question can be made precise by considering a dataset

$$Z_N = \{ (x_1^i, x_2^i, \ldots, x_p^i, y^i) \}_{i=1}^N$$

of the observed values y^i, $i = 1, 2, \ldots, N$, of a variable of interest y paired with the simultaneously observed values x_ν^i, $\nu = 1, 2, \ldots, p$, of variables $x_1, x_2, \ldots, x_p$ that

K. Hlaváčková-Schindler (✉)
Data Mining Group, Faculty of Computer Science, University of Vienna, Vienna, Austria

Department of Adaptive Systems, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, Czech Republic
e-mail: katerina.schindler@gmail.com

V. Naumova
Simula Research Laboratory, Lysaker, Norway
e-mail: valeriya@simula.no

S. Pereverzyev Jr.
Department of Mathematics, Applied Mathematics Group, University of Innsbruck, Innsbruck, Austria
e-mail: sergiy.pereverzyev@uibk.ac.at

I. Pesenson et al. (eds.), *Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-55556-0_15

possibly interact with y. Then the set Z_N is used to quantify how strong is the effect of $x = (x_1, x_2, \ldots, x_p)$ on y.

An instance of this situation is the problem of reconstructing from the set Z_N a multivariate function $y = f(x_{\nu_1}, x_{\nu_2}, \ldots, x_{\nu_l})$ that depends only on a subset $\{x_{\nu_i}\}_{i=1}^{l}$ of the variables $\{x_\nu\}_{\nu=1}^{p}$ (very often, l is much smaller than p). The variables in this subset $\{x_{\nu_i}\}_{i=1}^{l}$ are called relevant, and they exhibit an effect on the variable y in contrast to the remaining variables $\{x_\nu\}_{\nu=1}^{p} \setminus \{x_{\nu_i}\}_{i=1}^{l}$. In this work, we are interested in detecting these relevant variables x_{ν_i} from the given data Z_N.

Note that the above problem has been extensively studied under the assumption that the target function f depends linearly on the relevant variables such that it admits the representation

$$f(x) = \sum_{j=1}^{p} \beta_j x_j$$

with only a few non-zero coefficients β_j for $j = \nu_1, \nu_2, \ldots, \nu_l$. Under such assumption the problem of detecting relevant variables from the data set Z_N can be reduced to the linear regression with a so-called sparsity constraint. The latter one is now fairly well understood and can be solved efficiently by means of l_1-regularization. For comprehensive treatments of this subject, we refer the reader to the classical work [9] and some more recent ones [12, 13, 15, 22, 28] (see also the references therein).

Despite the computational benefit of the linear regression, it should be noted that this model is too simple to be always appropriately matched to the underlying dynamics and may sometimes lead to a misspecification (we defer this discussion to the last section). A more realistic situation, where the target function f depends nonlinearly on the relevant variables, is much less understood, and in the literature it is mostly restricted to the so-called *additive models* [7, 16, 24, 29]. In this model, the target function is assumed to be the sum

$$f(x) = \sum_{j=1}^{p} f_j(x_j) \tag{1}$$

of nonlinear univariate functions f_j in some Reproducing Kernel Hilbert Spaces (RKHS) $\mathscr{H}_j$ such that $f_j \equiv 0$ for $j \notin \{\nu_i\}_{i=1}^{l}$. For the sake of brevity, we omit the discussion on Reproducing Kernel Hilbert Spaces, and refer the reader to the seminal paper [2] on a comprehensive theory of RKHS.

In [3], it has been observed that the detection of relevant variables in model (1) can be performed by using a technique from the *multiple kernel learning* [4, 20]. Then, an estimator of the target function (1) can be constructed as the sum $\sum_{j=1}^{p} f_j^{\lambda}(x_j)$ of the minimizers of the functional

$$T_\lambda^q(f_1,f_2,\ldots,f_p;\ Z_N) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - \sum_{j=1}^{p} f_j(x_j^i)\right)^2 + \sum_{j=1}^{p}\lambda_j \|f_j\|_{\mathscr{H}_j}^q, \tag{2}$$

i.e.,

$$T_\lambda^q(f_1^\lambda,f_2^\lambda,\ldots,f_p^\lambda;\ Z_N) = \min\{\ T_\lambda^q(f_1,f_2,\ldots,f_p;\ Z_N),\ f_j \in \mathscr{H}_j,\ j = 1,2,\ldots,p\ \},$$

where $\lambda = (\lambda_1,\lambda_2,\ldots,\lambda_p)$ is a vector of the regularization parameters, and $q > 0$.

Note that functional (2) can be seen as a Tikhonov-type functional (see, e.g., [10, 34]). The first term in (2) represents the quality of data fitting, while the second term is called the regularization term. The purpose of regularization is to avoid functions with big norms that lead to overfitting. Regularization can be also seen as a penalty on the complexity of the involved functions.

A different approach has been recently proposed in [24]. This approach is based on the idea that the importance of a variable can be captured by the partial derivatives. Then in [24], the target function is estimated as the minimizer of the functional

$$\hat{T}_\lambda(f;Z_N) = \frac{1}{N}\sum_{i=1}^{N}(y^i - f(x^i))^2 + \lambda_1\|f\|_{\mathscr{H}}^2 + \lambda_2\sum_{j=1}^{p}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial f(x^i)}{\partial x_j}\right)^2\right)^{1/2}, \tag{3}$$

where $x^i = (x_1^i, x_2^i,\ldots,x_p^i)$, and $\mathscr{H}$ is a RKHS of functions $f = f(x_1,x_2,\ldots,x_p)$.

Note that the choice of the regularization parameters λ_j is an open issue in the both above-mentioned approaches. For the multiple kernel learning scheme of type (2), an *a priori* parameter choice strategy has been proposed in [20]. In this strategy, the choice of λ_j depends only on the kernels generating RKHS $\mathscr{H}_j$ and on the distribution of the points x_j^i in Z_N. It is clear that such a strategy may not be suitable for detecting relevant variables because the functions in (1) depending on different variables x_j may be associated with the same $\mathscr{H}_j$ and x_j^i. As to the scheme based on (3), no recipe for choosing the parameters λ_1, λ_2 was given.

Observe also that the numerical implementation of the above-mentioned approaches can be nontrivial. For example, the functional (3), as well as the functional (2) with $q \in (0,1]$, is not differentiable and, hence, its minimization cannot be done by simple gradient methods. Moreover, the minimizers of the functionals can only be computed in an iterative fashion requiring the solution of a system of $M = O(Np)$ equations at each step, and this can be computationally expensive for large N and/or p.

In the present paper, we propose a new approach attempting to detect relevant variables one by one such that the dimension of the corresponding system of equations increases only when it is necessary. The first step of our approach consists in constructing the minimizers $f_j = f_j^{\lambda_j}(x_j)$ of the functionals $T_{\lambda_j}^q(f_j;Z_N)$ defined by (2) with $q = 2$, $p = 1$, $\lambda_1 = \lambda_j$, $x_1^i = x_j^i$, $\mathscr{H}_1 = \mathscr{H}_j$, $j = 1,2,\ldots$. From

the representer theorem [18, 31], it follows that such minimization is reduced to solving systems of N linear equations. Then the minimizers $f_j^{\lambda_j}(x_j)$ are used to rank the variables x_j according to the values of the discrepancies

$$\mathscr{D}(f_j^{\lambda_j}(x_j);\, Z_N) = \left(\frac{1}{N}\sum_{i=1}^{N}\left(y^i - f_j^{\lambda_j}(x_j^i)\right)^2\right)^{1/2},\ j = 1, 2, \ldots,$$

as follows: the smaller the value of $\mathscr{D}(f_j^{\lambda_j}(x_j); Z_N)$, the higher the rank of x_j. This step can be seen as an attempt to interpret the data Z_N by using only a univariate function, and the variable with the highest rank is considered as the first relevant variable x_{ν_1}.

The next step consists in testing the hypothesis that a variable with the second highest rank, say x_μ, is also relevant. For such a testing we compute the minimizers $f_{\nu_1}^{\lambda_{\nu_1}}, f_\mu^{\lambda_\mu}$ of the functional

$$\begin{aligned} T_\lambda^2(f_{\nu_1}, f_\mu;\, Z_N) = &\frac{1}{N}\sum_{i=1}^{N}\left(y^i - f_{\nu_1}(x_{\nu_1}^i) - f_\mu(x_\mu^i)\right)^2 + \\ &\lambda_{\nu_1}\|f_{\nu_1}\|_{\mathscr{H}_{\nu_1}}^2 + \lambda_\mu\|f_\mu\|_{\mathscr{H}_\mu}^2. \end{aligned} \tag{4}$$

Our idea is based on the observation [25] that in multi-penalty regularization with a component-wise penalization, such as (4), one needs to use small as well as large values of the regularization parameters $\lambda_{\nu_1}, \lambda_\mu$, i.e., both λ_{ν_1} and $\lambda_\mu << 1$, and $\lambda_\mu > 1$, respectively. Therefore, in the proposed approach the variable x_μ is considered as the relevant one if for $\{\lambda_{\nu_1}, \lambda_\mu\} \subset (0, 1)$, the values of the discrepancy

$$\mathscr{D}(f_{\nu_1}^{\lambda_{\nu_1}}, f_\mu^{\lambda_\mu};\, Z_N) = \left(\frac{1}{N}\sum_{i=1}^{N}\left(y^i - f_{\nu_1}^{\lambda_{\nu_1}}(x_{\nu_1}^i) - f_\mu^{\lambda_\mu}(x_\mu^i)\right)^2\right)^{1/2} \tag{5}$$

are essentially smaller than the ones for $\lambda_{\nu_1} \in (0, 1)$, $\lambda_\mu > 1$. If it is not the case, then the above-mentioned hypothesis is rejected, and in the same way we test the variable with the third highest rank and so on. In the next section, we provide a theoretical justification for the use of discrepancies values corresponding to the regularization parameters from different intervals for detecting relevant variables.

On the other hand, if the variable x_μ has been accepted as the second relevant variable, i.e., $x_{\nu_2} = x_\mu$, then to test whether or not the variable with the third highest rank, say x_ν, can be taken as the third relevant variable, i.e., whether or not $x_{\nu_3} = x_\nu$, we compute the minimizers $f_{\nu_1}^{\lambda_{\nu_1}}, f_{\nu_2}^{\lambda_{\nu_2}}, f_\nu^{\lambda_\nu}$ of the functional

$$T_\lambda^2(f_{\nu_1},f_{\nu_2},f_\nu;\ Z_N) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - f_{\nu_1}(x^i_{\nu_1}) - f_{\nu_2}(x^i_{\nu_2}) - f_\nu(x^i_\nu)\right)^2 + \lambda_{\nu_1}\|f_{\nu_1}\|^2_{\mathscr{H}_{\nu_1}} + \lambda_{\nu_2}\|f_{\nu_2}\|^2_{\mathscr{H}_{\nu_2}} + \lambda_\nu\|f_\nu\|^2_{\mathscr{H}_\nu}, \tag{6}$$

where, with a little abuse of notation, we use the same symbols $f_{\nu_1}, f_{\nu_1}^{\lambda_{\nu_1}}$ as in (4),(5). Then, as above, the variable x_ν is considered as relevant if for $\{\lambda_{\nu_1}, \lambda_{\nu_2}, \lambda_\nu\} \subset (0,1)$, the values of the discrepancy

$$\mathscr{D}(f_{\nu_1}^{\lambda_{\nu_1}}, f_{\nu_2}^{\lambda_{\nu_2}}, f_\nu^{\lambda_\nu};\ Z_N) = \left(\frac{1}{N}\sum_{i=1}^{N}\left(y^i - f_{\nu_1}^{\lambda_{\nu_1}}(x^i_{\nu_1}) - f_{\nu_2}^{\lambda_{\nu_2}}(x^i_{\nu_2}) - f_\nu^{\lambda_\nu}(x^i_\nu)\right)^2\right)^{1/2} \tag{7}$$

are essentially smaller than the corresponding values of (7) for $\{\lambda_{\nu_1}, \lambda_{\nu_2}\} \subset (0,1)$, $\lambda_\nu > 1$. Otherwise, the variable with the next highest rank is tested in the same way.

If the discrepancy (7) does exhibit the above-mentioned behavior, then for testing the variable with the next highest rank in accordance with the proposed approach, we need to add to (6) one more penalty term corresponding to that variable, so that the functional $T_\lambda^2(f_1,f_2,\ldots,f_p;Z_N)$ of the form (2) containing the whole set of penalties may appear only at the end of the testing procedure.

Below, we present an algorithmic realization of the above presented approach.

Task: Find a subset I of given variables that exhibit an effect on the variable y. The effect is modeled as follows: $y = \sum_{j\in I} f_j^\lambda(x_j)$.

Tuning parameters: N_{MC} — number of Monte-Carlo simulations (see Section 2, p. 902). C — tolerance level for determining the essential variability of the discrepancy values (see Section 2, p. 900).

$I = \emptyset; J = \{1,2,\ldots,p\}$.

for $j = 1$ to p **do**

$$f_j^{\lambda_j}(x) = \arg\min_{f\in\mathscr{H}_j} \frac{1}{N}\sum_{i=1}^{N}\left(y^i - f_j(x_j^i)\right)^2 + \lambda_j\|f_j\|^2_{\mathscr{H}_j}.$$

The regularization parameters λ_j can be chosen using the quasi-optimality criterion. {See Section 2, p. 903 for details.}

end for

Choose the first relevant variable as

$$x_k = \arg\min_{x_j\in J}\left(\frac{1}{N}\sum_{i=1}^{N}\left(y^i - f_j^{\lambda_j}(x_j^i)\right)^2\right)^{1/2}.$$

$I = \{k\}; J = J \setminus \{k\}$.

while $J \neq \emptyset$ **do**

Select the candidate for the next relevant variable as

$$x_k = \arg\min_{x_j \in J} \left(\frac{1}{N} \sum_{i=1}^{N} \left(y^i - f_j^{\lambda_j}(x_j^i) \right)^2 \right)^{1/2}.$$

Define the functional

$$T_\lambda^2 \left(f_j,\, j \in I;\, f_k;\, Z_N \right) = \frac{1}{N} \sum_{i=1}^{N} \left(y^i - \sum_{j \in I} f_j(x_j^i) - f_k(x_k^i) \right)^2 + \sum_{j \in I} \lambda_j \|f_j\|_{\mathcal{H}_j}^2 + \lambda_k \|f_k\|_{\mathcal{H}_k}^2.$$

Denote the minimizers of the above functional as $\left(f_j^{\lambda_j},\, j \in I;\, f_k^{\lambda_k} \right)$, and let the corresponding discrepancy be defined as

$$\mathscr{D}\left(f_j^{\lambda_j},\, j \in I;\, f_k^{\lambda_k};\, Z_N \right) = \left(\frac{1}{N} \sum_{i=1}^{N} \left(y^i - \sum_{j \in I} f_j^{\lambda_j}(x_j^i) - f_k^{\lambda_k}(x_{\nu_k}^i) \right)^2 \right)^{1/2}.$$

$i_C = 0$.

for $i_{\mathrm{MC}} = 1$ to N_{MC} **do**

Select randomly $\left(\lambda_j, j \in I;\, \lambda_k \right) \in \left(\Lambda_{50}^{\mathrm{small}} \right)^{|I|+1}$. { $\Lambda_{50}^{\mathrm{small}}$ is defined in (27). }

Compute corresponding minimizers $\left(f_j^{\lambda_j}, j \in I;\, f_k^{\lambda_k} \right)$ and discrepancy $\mathscr{D}_1 = \mathscr{D}\left(f_j^{\lambda_j},\, j \in I;\, f_k^{\lambda_k};\, Z_N \right)$.

Select randomly $\left(\lambda_j, j \in I;\, \lambda_k \right) \in \left(\Lambda_{50}^{\mathrm{small}} \right)^{|I|} \times \Lambda_{50}^{\mathrm{large}}$. { $\Lambda_{50}^{\mathrm{large}}$ is defined in (28). }

Compute corresponding minimizers $\left(f_j^{\lambda_j}, j \in I;\, f_k^{\lambda_k} \right)$ and discrepancy $\mathscr{D}_2 = \mathscr{D}\left(f_j^{\lambda_j},\, j \in I;\, f_k^{\lambda_k};\, Z_N \right)$.

if $|\mathscr{D}_1 - \mathscr{D}_2| \geq C$ **then**

$i_C = i_C + 1$.

end if

end for

if $i_C > N_{\mathrm{MC}}/2$ **then**

$I = I \cup \{k\}$.

end if

$J = J \setminus \{k\}$.

end while

In the next sections, after presenting the theoretical background, we will illustrate the application of the proposed approach to the recovery of causal relationships in a gene regulatory network, and compare it with the results known from the literature.

2 Theoretical Background

At first, we shall write a system of necessary conditions for the minimizers of the functional (2), where, according to the proposed approach, p may take values $1, 2, \ldots$, and $q = 2$.

Let $\mathbb{R}^N$ be the N-dimensional Euclidean space of vectors $u = (u^1, u^2, \ldots, u^N)$ equipped with the norm $\|u\|_{\mathbb{R}^N} := \left(\frac{1}{N}\sum_{i=1}^{N}(u^i)^2\right)^{1/2}$ and the corresponding inner product $\langle\cdot,\cdot\rangle_{\mathbb{R}^N}$.

Consider the sampling operators $S_{N,j}$ mapping RKHSs $\mathscr{H}_j$ generated by the kernels $K_j = K_j(x_j, v_j), j = 1, 2, \ldots, p$, into $\mathbb{R}^N$ such that for $f \in \mathscr{H}_j$,

$$S_{N,j}f = \left(f(x_j^1), f(x_j^2), \ldots, f(x_j^N)\right) \in \mathbb{R}^N.$$

Let us shorty recall that a RKHS $\mathscr{H}$ [2, 6, 8] is defined by a symmetric positive definite function $K(x, \tilde{x}) : X \times X \to \mathbb{R}$, which is called the kernel. Examples of the kernels are the polynomial kernel $K(x, \tilde{x}) = (x\tilde{x} + 1)^d$ of degree $d \in \mathbb{N}$, and the Gaussian kernel $K(x, \tilde{x}) = e^{-(x-\tilde{x})^2}$. Also, let us note that for functions $f \in \mathscr{H}$, the so-called reproducing property holds: $\langle f(\cdot), K(x, \cdot)\rangle_{\mathscr{H}} = f(x)$ for all $x \in X$.

In view of the above-mentioned reproducing property, we can write the adjoints $S_{N,j}^* : \mathbb{R}^N \to \mathscr{H}_j$ of the sampling operators as follows:

$$\left(S_{N,j}^* u\right)(x_j) = \frac{1}{N}\sum_{i=1}^{N} K_j(x_j^i, x_j)u^i. \tag{8}$$

In terms of $S_{N,j}$, the functional (2) has the form

$$T_\lambda^2(f_1, f_2, \ldots, f_p;\ Z_N) = \left\| Y - \sum_{j=1}^{p} S_{N,j}f_j \right\|_{\mathbb{R}^N}^2 + \sum_{j=1}^{p} \lambda_j \|f_j\|_{\mathscr{H}_j}^2, \tag{9}$$

where $Y = (y^1, y^2, \ldots, y^N)$. Then, using the standard technique of the calculus of variations, we obtain the following system of equations for the minimizers $f_j^{\lambda_j}$

$$\lambda_j f_j^{\lambda_j} + \sum_{\nu=1}^{p} S_{N,j}^* S_{N,\nu} f_\nu^{\lambda_\nu} = S_{N,j}^* Y, \quad j = 1, 2, \ldots, p. \tag{10}$$

From (8) and (10), it is clear that $f_j^{\lambda_j}$ can be represented as

$$f_j^{\lambda_j}(x_j) = \sum_{i=1}^{N} \gamma_i^j K_j(x_j^i, x_j), \tag{11}$$

where $\{\gamma_i^j\} \subset \mathbb{R}$. Note that (11) can be seen as an analog of the well-known representer theorem [18, 31] for the case of the regularization with a component-wise penalization in RKHS. This allows the reduction of the minimization of (9) to solving systems of Np linear equations with respect to γ_i^j. Recall that in the approach described above, p will successively take the values $1, 2, \ldots$, such that the dimension of the corresponding system (10) increases only when it is necessary.

Now, for the sake of definiteness and simplicity of the presentation, suppose that

$$Y = S_{N,1} f_1 + S_{N,2} f_2 + \varepsilon, \tag{12}$$

where $f_1 = f_1(x_1)$, $f_2 = f_2(x_2)$, and the vector $\varepsilon \in \mathbb{R}^N$ may represent a noise in measurements, as well as a contribution to the data Y coming from functions of other relevant variables. Note that (12) means that x_1, x_2 are relevant variables. Below we analyze the behavior of the discrepancy (5) for $\nu_1 = 1$, $\mu = 2$, and $Y = (y^1, y^2, \ldots, y^N)$ given by (12). This means that we consider the second step of the proposed approach when the variables x_1, x_2 have already received the ranks 1 and 2, respectively. The analysis of other steps and possibilities can be done similarly, but it is too technical and is omitted here for brevity. For simplicity, let us denote

$$\mathscr{D} := \mathscr{D}(f_1^{\lambda_1}, f_2^{\lambda_2};\, Z_N) = \left\| Y - S_{N,1} f_1^{\lambda_1} - S_{N,2} f_2^{\lambda_2} \right\|_{\mathbb{R}^N}. \tag{13}$$

It is easy to check that for $p = 2$, the solutions of the system (10) can be written as follows:

$$\begin{aligned} f_1^{\lambda_1} &= \left(\frac{\lambda_1}{\lambda_2} \mathbb{I}_{K_1} + S_{N,1}^* \left(\lambda_2 \mathbb{I}_N + S_{N,2} S_{N,2}^* \right)^{-1} S_{N,1} \right)^{-1} \\ &\quad S_{N,1}^* \left(\lambda_2 \mathbb{I}_N + S_{N,2} S_{N,2}^* \right)^{-1} Y, \\ f_2^{\lambda_2} &= \left(\frac{\lambda_2}{\lambda_1} \mathbb{I}_{K_2} + S_{N,2}^* \left(\lambda_1 \mathbb{I}_N + S_{N,1} S_{N,1}^* \right)^{-1} S_{N,2} \right)^{-1} \\ &\quad S_{N,2}^* \left(\lambda_1 \mathbb{I}_N + S_{N,1} S_{N,1}^* \right)^{-1} Y, \end{aligned} \tag{14}$$

where $\mathbb{I}_N$ is the identity matrix of size $N \times N$, and $\mathbb{I}_{K_j}$ is the identity operator on RKHS $\mathscr{H}_j$ generated by the kernel $K_j(x_j, v_j)$, $j = 1, 2$.

Note that for determining $f_1^{\lambda_1}, f_2^{\lambda_2}$ in practice, we can use the representation for $f_1^{\lambda_1}, f_2^{\lambda_2}$ in (11). The system of linear equations for the coefficients $\left\{ \gamma_i^1, \gamma_i^2 \right\}_{i=1}^N$ in

this representation can be determined from (10) by equating the factors near the functions $K_j\left(x_j^i,\cdot\right)$. This system of linear equations has a simple but rather bulky form, and therefore, we don't present it here.

Now, we introduce the first assumption used in our theoretical analysis. This assumption is formulated in terms of the elements of the singular-value decomposition (SVD) of the sampling operators

$$S_{N,j}=\sum_{i=1}^{N} a_{ij}h_{ij}\langle\kappa_{ij},\cdot\rangle_{\mathscr{H}_j},\quad j=1,2, \tag{15}$$

where $\{h_{ij}\}$, $\{\kappa_{ij}\}$ are some orthonormal systems in $\mathbb{R}^N$ and $\mathscr{H}_j$, respectively, and $a_{ij}\geq 0$.

Assumption 2. The sampling operators $S_{N,j}$ share the same system of $\{h_{ij}\}$, i.e.,

$$\{h_{i,1}\}=\{h_{i,2}\}=\{h_i\}. \tag{16}$$

Assumption 2 is in fact an assumption on the distribution of the sampling points $\{x_j^i\}$. We illustrate it in the following simple example.

Example 1. Let $N=2$, and $x_1^1=x_1^2=t$, $x_2^1=\tau_1$, $x_2^2=\tau_2$. This means that the sampling points belong to a line parallel to the x_2-axis. If x_1 is already accepted as the relevant variable, then such sampling points allow an easy test whether or not x_2 should be accepted as the relevant variable. Indeed, if $|y^1-y^2|$ is essentially large, then one really needs one more variable to explain the given data $Y=(y^1,y^2)$.

In the considered case, the sampling operators have the following representations

$$S_{N,1}f=(f(t),f(t)),\ S_{N,2}f=(f(\tau_1),f(\tau_2)).$$

Assume that both RKHS are generated by the same Gaussian kernel $K(x,v)=e^{-(x-v)^2}$. Then

$$\begin{aligned}
S_{N,2}f=&(1,1)\langle(K(\tau_1,\cdot)+K(\tau_2,\cdot))/2,f\rangle_{\mathscr{H}_2}+(1,-1)\langle(K(\tau_1,\cdot)-K(\tau_2,\cdot))/2,f\rangle_{\mathscr{H}_2},\\
S_{N,1}f=&(1,1)\langle K(t,\cdot),f\rangle_{\mathscr{H}_1},
\end{aligned}$$

and it is easy to check that these operators admit the decomposition (15) with

$$\begin{aligned}
h_{1,1}=h_{1,2}=(1,1)/\sqrt{2},\ h_{2,1}=h_{2,2}=(1,-1)/\sqrt{2},\\
\kappa_{1,2}=\frac{1}{\sqrt{2}}(K(\tau_1,\cdot)+K(\tau_2,\cdot))/(1+e^{-(\tau_1-\tau_2)^2})^{1/2},\\
\kappa_{2,2}=\frac{1}{\sqrt{2}}(K(\tau_1,\cdot)-K(\tau_2,\cdot))/(1-e^{-(\tau_1-\tau_2)^2})^{1/2},\ \kappa_{1,1}=K(t,\cdot),\\
a_{1,2}=(1+e^{-(\tau_1-\tau_2)^2})^{1/2},\ a_{2,2}=(1-e^{-(\tau_1-\tau_2)^2})^{1/2},\ a_{1,1}=\sqrt{2},\ a_{2,1}=0.
\end{aligned}$$

Thus, in the considered case Assumption 2 is satisfied. □

We would like to stress that Assumption 2 is only of the theoretical nature. At the same time, it is clear that a successful detection of relevant variables cannot be done from the data sampled at poorly distributed points $\{x_j^i\}$. Therefore, some restrictions on the sampling operators are unavoidable, and the condition (16) is just one of them.

Theorem 1. *Assume that Assumption 2 holds true. Consider the data Y from* (12) *and the discrepancy* $\mathscr{D}$ *from* (13)*. Then,*

(a) $\mathscr{D} \leq \frac{1}{2}\left(\sqrt{\lambda_1}\|f_1\|_{\mathscr{H}_1} + \sqrt{\lambda_2}\|f_2\|_{\mathscr{H}_2}\right) + \|\varepsilon\|_{\mathbb{R}^N}$.
(b) $\mathscr{D}$ *is an increasing function of* λ_1 *and* λ_2.

Proof: (a) From (14)-(16), it follows that

$$S_{N,1}f_1^{\lambda_1} + S_{N,2}f_2^{\lambda_2} = \sum_{i=1}^{N} \frac{\lambda_1 a_{i,2}^2 + \lambda_2 a_{i,1}^2}{\lambda_1\lambda_2 + \lambda_1 a_{i,2}^2 + \lambda_2 a_{i,1}^2} h_i \langle h_i, Y\rangle_{\mathbb{R}^N}. \tag{17}$$

Then in view of (12), we have

$$Y - S_{N,1}f_1^{\lambda_1} - S_{N,2}f_2^{\lambda_2} = \Sigma_1 + \Sigma_2 + \Sigma_3, \tag{18}$$

where

$$\Sigma_1 = \sum_{i=1}^{N} \frac{\lambda_1\lambda_2 a_{i,1}}{\lambda_1\lambda_2 + \lambda_1 a_{i,2}^2 + \lambda_2 a_{i,1}^2} h_i \langle \kappa_{i,1}, f_1\rangle_{\mathscr{H}_1}, \tag{19}$$

$$\Sigma_2 = \sum_{i=1}^{N} \frac{\lambda_1\lambda_2 a_{i,2}}{\lambda_1\lambda_2 + \lambda_1 a_{i,2}^2 + \lambda_2 a_{i,1}^2} h_i \langle \kappa_{i,2}, f_2\rangle_{\mathscr{H}_2}, \tag{20}$$

$$\Sigma_3 = \sum_{i=1}^{N} \frac{\lambda_1\lambda_2}{\lambda_1\lambda_2 + \lambda_1 a_{i,2}^2 + \lambda_2 a_{i,1}^2} h_i \langle h_i, \varepsilon\rangle_{\mathbb{R}^N}.$$

Observe now that

$$\begin{aligned}
\|\Sigma_1\|_{\mathbb{R}^N} &= \left(\sum_{i=1}^{N}\left(\frac{\lambda_1\lambda_2 a_{i,1}}{\lambda_1\lambda_2 + \lambda_1 a_{i,2}^2 + \lambda_2 a_{i,1}^2}\right)^2 \langle \kappa_{i,1}, f_1\rangle_{\mathscr{H}_1}^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^{N}\left(\frac{\lambda_1 a_{i,1}}{\lambda_1 + a_{i,1}^2}\right)^2 \langle \kappa_{i,1}, f_1\rangle_{\mathscr{H}_1}^2\right)^{1/2} \\
&\leq \sup_t \left|\frac{\lambda_1 t}{\lambda_1 + t^2}\right| \|f_1\|_{\mathscr{H}_1} = \frac{\sqrt{\lambda_1}}{2}\|f_1\|_{\mathscr{H}_1}.
\end{aligned}$$

Moreover, in the same way, we obtain that

$$\begin{aligned}\|\Sigma_2\|_{\mathbb{R}^N} &\leq \frac{\sqrt{\lambda_2}}{2}\|f_2\|_{\mathscr{H}_2},\\ \|\Sigma_3\|_{\mathbb{R}^N} &= \left(\sum_{i=1}^{N}\left(\frac{\lambda_1\lambda_2}{\lambda_1\lambda_2+\lambda_1 a_{i,2}^2+\lambda_2 a_{i,1}^2}\right)^2\langle h_i,\varepsilon\rangle_{\mathbb{R}^N}^2\right)^{1/2}\\ &\leq \left(\sum_{i=1}^{N}\langle h_i,\varepsilon\rangle_{\mathbb{R}^N}^2\right)^{1/2} = \|\varepsilon\|_{\mathbb{R}^N}. \end{aligned} \tag{21}$$

Combining these bounds with (18), we obtain the asserted statement.

(b) Since $Y = \sum_{i=1}^{N} h_i\langle h_i, Y\rangle_{\mathbb{R}^N}$, using (17), we obtain that

$$\mathscr{D}^2 = \sum_{i=1}^{N}\left(\frac{\lambda_1\lambda_2}{\lambda_1\lambda_2+\lambda_1 a_{i,2}^2+\lambda_2 a_{i,1}^2}\right)^2\langle h_i, Y\rangle_{\mathbb{R}^N}^2.$$

One can show that in the above sum, the (λ_1,λ_2)-dependent coefficients are monotonically increasing functions of λ_1 and λ_2. Therefore, the discrepancy $\mathscr{D}$ is also a monotonically increasing function of λ_1 and λ_2. □

Now what happens if x_1 and x_2 are not relevant variables, that is in (12) $f_1 \equiv f_2 \equiv 0$. In order to analyze the behavior of the discrepancy $\mathscr{D}$ in this case, we need to introduce additional assumptions.

First of all, it is natural to assume that the number of the nonzero singular values a_{ij} in (15) is very small compared to the number of the sampling points N.

Assumption 3. Let a_{ij} be the singular values of $S_{N,j}$. Denote $A_j := \{\, i \mid a_{ij} > 0 \,\}$. Then

$$\#A_j \ll N,$$

where $\#A_j$ denotes the number of elements in the set A_j.

It should be clear that the above assumption is violated, when, as the following example demonstrates, the distribution of the sampling points $\{x_j^i\}$ may not allow a conclusion about the relevance of the variables.

Example 2. Let the sampling points $\{x_j^i\}$ be such that $x_2^i = c\,x_1^i$, where $c \in \mathbb{R}$ is some constant, and $x_j^{i_1} \neq x_j^{i_2}$ for $i_1 \neq i_2$. Then, $\operatorname{rank}(S_{N,1}) = \operatorname{rank}(S_{N,2}) = N$, and $\#A_1 = \#A_2 = N$, so that Assumption 3 is violated. At the same time, the corresponding data Y can be interpolated well by univariate functions $f(x_1)$ or $f(x_2)$ that does not allow to conclude which of the variables x_1 or x_2 can be selected as the relevant variable.

Another assumption, which we need for the analysis of the situation when in (12) either $f_1 \equiv f_2 \equiv 0$, or $f_2 \equiv 0$, is related to the structure of the noise $\varepsilon \in \mathbb{R}^N$ in (12). Noise vector $\varepsilon \in \mathbb{R}^N$ can be represented as follows:

$$\varepsilon = \sum_{i=1}^{N} h_i \langle h_i, \varepsilon \rangle_{\mathbb{R}^N} = \sum_{i \in A_1 \cup A_2} h_i \langle h_i, \varepsilon \rangle_{\mathbb{R}^N} + \sum_{i \notin A_1 \cup A_2} h_i \langle h_i, \varepsilon \rangle_{\mathbb{R}^N}.$$

Denote the first term in the last formula as $\varepsilon_{1,2}$, and the last term as $\bar{\varepsilon}_{1,2}$. Since $S^*_{N,j} \bar{\varepsilon}_{1,2} = 0$, this part of the noise ε has little influence on the minimizers $f_1^{\lambda_1}$ and $f_2^{\lambda_2}$ in (14). For another part of the noise $\varepsilon_{1,2}$, we assume the following.

Assumption 4. Let $\varepsilon_{1,2} \in \mathbb{R}^N$ be a part of the noise $\varepsilon \in \mathbb{R}^N$ in (12), defined as above. Then,

$$\|\varepsilon_{1,2}\|^2_{\mathbb{R}^N} = \sum_{i \in A_1 \cup A_2} \langle h_i, \varepsilon \rangle^2_{\mathbb{R}^N} \ll \|\varepsilon\|^2_{\mathbb{R}^N}.$$

In view of Assumption 3, the above noise assumption is not so restrictive. This assumption allows a quantification of the behavior of $\mathscr{D}$ with respect to the non-relevant variables. Of course, the interpretation of the symbol $\ll$ depends on a particular application. In the sequel, we say that the discrepancy $\mathscr{D}$ does not essentially change if the differences in values of $\mathscr{D}$ deviate within the interval $[-C_1 \|\varepsilon\|_{\mathbb{R}^N}, C_1 \|\varepsilon\|_{\mathbb{R}^N}]$, where $C_1 > 0$ is an application dependent constant. Moreover, we say that some quantity takes values around $\|\varepsilon\|_{\mathbb{R}^N}$ if these values appear in the interval $[(1 - C_2) \|\varepsilon\|_{\mathbb{R}^N}, (1 + C_2) \|\varepsilon\|_{\mathbb{R}^N}]$, where $0 < C_2 < 1$ is another application dependent constant.

Using Assumptions 3 and 4, we can obtain the following statement about the behavior of the discrepancy $\mathscr{D}$, when the variables x_1 and x_2 are not relevant.

Theorem 2. *Assume that Assumptions 2–4 hold true. If x_1 and x_2 are not relevant variables, i.e., if in (12) $f_1 \equiv f_2 \equiv 0$, then the discrepancy $\mathscr{D}$ does not essentially change with λ_1, λ_2, and may take values around $\|\varepsilon\|_{\mathbb{R}^N}$.*

Proof. Under conditions of the theorem, the representation of the discrepancy vector (18) becomes

$$Y - S_{N,1} f_1^{\lambda_1} - S_{N,2} f_2^{\lambda_2} = \Sigma_3.$$

Then,

$$\mathscr{D}^2 = \sum_{i \notin A_1 \cup A_2} \langle h_i, \varepsilon \rangle^2_{\mathbb{R}^N} + \sum_{i \in A_1 \cup A_2} \left(\frac{\lambda_1 \lambda_2}{\lambda_1 \lambda_2 + \lambda_1 a_{i,2}^2 + \lambda_2 a_{i,1}^2} \right)^2 \langle h_i, \varepsilon \rangle^2_{\mathbb{R}^N}.$$

In view of Assumption 4, the second sum in the above representation is negligible, and this gives us the statement of the Theorem. □

In the case, when one of the variables, say x_1, is relevant, whereas another one is not, the following statement about the behavior of the discrepancy $\mathscr{D}$ can be derived.

Theorem 3. *Assume that Assumptions 2–4 hold true. Assume further that x_1 is the relevant variable, and x_2 is not, i.e., in (12) $f_2 \equiv 0$. If*

$$\sum_{i \in A_1 \cap A_2} a_{i,1}^2 \langle \kappa_{i,1}, f_1 \rangle_{\mathscr{H}_1}^2 \leq \|\varepsilon_{1,2}\|_{\mathbb{R}^N}^2, \tag{22}$$

then the discrepancy $\mathscr{D}$ does not essentially change with λ_2.

Proof. Since $f_2 \equiv 0$, the discrepancy vector (18) has the following representation:

$$Y - S_{N,1} f_1^{\lambda_1} - S_{N,2} f_2^{\lambda_2} = \Sigma_1 + \Sigma_3.$$

Since Assumptions 2–4 hold true, the same argument as in the proof of Theorem 2 tells us that the norm $\|\Sigma_3\|$ does not essentially change with λ_1, λ_2. As to the term Σ_1, it can be written as follows:

$$\Sigma_1 = \sum_{i \in A_1 \setminus A_2} \frac{\lambda_1 a_{i,1}}{\lambda_1 + a_{i,1}^2} h_i \langle \kappa_{i,1}, f_1 \rangle_{\mathscr{H}_1} + \sum_{i \in A_1 \cap A_2} \frac{\lambda_1 \lambda_2 a_{i,1}}{\lambda_1 \lambda_2 + \lambda_1 a_{i,2}^2 + \lambda_2 a_{i,1}^2} h_i \langle \kappa_{i,1}, f_1 \rangle_{\mathscr{H}_1}.$$

In view of (22) and the inequality

$$\frac{\lambda_1 \lambda_2}{\lambda_1 \lambda_2 + \lambda_1 a_{i,2}^2 + \lambda_2 a_{i,1}^2} < 1,$$

the second summand is negligible, while the first one does not depend on λ_2. This allows the conclusion of the theorem. □

A typical example of the behavior of the discrepancy $\mathscr{D}$ described by Theorems 2 and 3 has been observed in our numerical tests below and is displayed in Fig. 2.

The above theorems allow us a conclusion that if there is a contribution to the data Y that comes from functions of variables, say x_1, x_2, then the values of the discrepancy corresponding to the small values of the regularization parameters $\{\lambda_1, \lambda_2\} \subset (0, 1)$ are expected to be essentially dominated by the ones corresponding to at least one large parameter.

Using similar arguments, we can extend the statements of the above theorems to any number of variables. Then the above conclusion can also be made for more than two variables, and it is the reason behind the use of the values of the discrepancy corresponding to large and small values of the regularization parameters for detecting relevant variables as it has been described in Introduction. Thus, if the discrepancy

$$\mathscr{D}\left(f_{\nu_1}^{\lambda_{\nu_1}}, f_{\nu_2}^{\lambda_{\nu_2}}, \ldots, f_{\nu_l}^{\lambda_{\nu_l}};\ Z_N\right) = \left\| Y - \sum_{j=1}^{l} S_{N,\nu_j} f_{\nu_j}^{\lambda_{\nu_j}} \right\|_{\mathbb{R}^N} \tag{23}$$

as a function of $(\lambda_{\nu_1}, \lambda_{\nu_2}, \ldots, \lambda_{\nu_l})$ exhibits a substantial growth in each variable, then the variables $x_{\nu_1}, x_{\nu_2}, \ldots, x_{\nu_l}$ are considered as the relevant ones.

Since in applications it is usually difficult to check the values of (23) for all $\lambda_{\nu_1}, \lambda_{\nu_2}, \ldots, \lambda_{\nu_l}$, one can realize the above-mentioned approach by using Monte-Carlo-type simulations. Namely, if $x_{\nu_1}, x_{\nu_2}, \ldots, x_{\nu_{l-1}}$ have been already accepted as relevant variables, then the values of (23) for the randomly chosen $(\lambda_{\nu_1}, \lambda_{\nu_2}, \ldots, \lambda_{\nu_l}) \in (0,1)^l$ are compared to the ones for the randomly chosen $(\lambda_{\nu_1}, \lambda_{\nu_2}, \ldots, \lambda_{\nu_l}) \in (0,1)^{l-1} \times [1, B]$, $B > 1$, and x_{ν_l} is accepted as the relevant variable if in the above simulations the values of (23) for $(\lambda_{\nu_1}, \lambda_{\nu_2}, \ldots, \lambda_{\nu_l}) \in (0,1)^l$ are essentially dominated by the ones for $(\lambda_{\nu_1}, \lambda_{\nu_2}, \ldots, \lambda_{\nu_l}) \in (0,1)^{l-1} \times [1, B]$.

Remark 1. Note that the conclusion about the ordered behavior of the discrepancy made on the basis of Theorem 1 can be seen as an extension of the following interpretation of the values of the discrepancy $\|S_{N,1} f_j^{\lambda_j} - Y\|_{\mathbb{R}^N}$ for the single penalty regularization. From [36, Lemma 3.1], it follows that

$$\lim_{\lambda_j \to 0} \left\| S_{N,j} f_j^{\lambda_j} - Y \right\|_{\mathbb{R}^N} = \inf_{f \in \mathscr{H}_j} \left\| S_{N,j} f - Y \right\|_{\mathbb{R}^N},$$

$$\lim_{\lambda_j \to \infty} \left\| S_{N,j} f_j^{\lambda_j} - Y \right\|_{\mathbb{R}^N} = \|Y\|_{\mathbb{R}^N}.$$

Then it is clear that if $\mathscr{H}_j$ is dense in the corresponding space of continuous functions, and

$$Y = S_{N,j} f_j + \varepsilon, \quad \|\varepsilon\|_{\mathbb{R}^N} < \|Y\|_{\mathbb{R}^N},$$

then for small λ_j and large $\bar{\lambda}_j$, one can expect

$$\left\| S_{N,j} f_j^{\lambda_j} - Y \right\|_{\mathbb{R}^N} < \left\| S_{N,j} f_j^{\bar{\lambda}_j} - Y \right\|_{\mathbb{R}^N}.$$

On the other hand, if $Y \in (\mathrm{Range}(S_{N,j}))^{\perp}$ such that there is no contribution to Y allowing a representation in terms of the values of $f_j \in \mathscr{H}_j$ at the points $\{x_j^i\}_{i=1}^N$, then the discrepancies $\left\| S_{N,j} f_j^{\lambda_j} - Y \right\|_{\mathbb{R}^N}$ do not behave in the ordered way.

Of course, in the case of the single variable and penalty, no additional assumptions, for example, (16) are needed to justify the ordered behavior of the discrepancy $\left\| S_{N,j} f_j^{\lambda_j} - Y \right\|_{\mathbb{R}^N}$ for $Y = S_{N,j} f_j + \varepsilon$. □

At the end of this theoretical section, we illustrate the above approach on the example from [24], where for $p = 40$ and $N = 100$, the data set

$Z_N = \{ (x_1^i, x_2^i, \ldots, x_p^i;\ y^i) \}_{i=1}^N$ is simulated in such a way that the values x_j^i are sampled uniformly at random from the interval $[-2, 2]$, and

$$y^i = \sum_{j=1}^{4} \left(x_j^i\right)^2 + \varepsilon^i, \tag{24}$$

where ε^i are zero-mean Gaussian random variables with variances chosen so that the signal-to-noise ratio is 15 : 1.

The input (24) means that in this example the target function (1) depends only on the first 4 variables. Recall that in our approach, at first, we need to rank the variables $x_1, x_2, \ldots, x_{40}$ according to the values of the discrepancies $\mathscr{D}(f_j^{\lambda_j}(x_j);\ Z_N)$, $j = 1, 2, \ldots, 40$, where $f_j^{\lambda_j}$ is the minimizer of the Tikhonov functional

$$T_\lambda(f;\ Z_N) = \frac{1}{N} \sum_{i=1}^{N} \left(y^i - f(x_j^i)\right)^2 + \lambda \|f\|_{\mathscr{H}}^2. \tag{25}$$

In our experiments, we choose in (25) $\lambda = \lambda_j = \lambda^{(k_j)}$ from the set

$$\Lambda_{50} = \{ \lambda = \lambda^{(k)} = 10^{-4} \cdot (1.3)^k,\ k = 1, 2, \ldots, 50 \}$$

according to the quasi-optimality criterion (see, e.g., [5, 19, 35]). Moreover, in (25) the space $\mathscr{H}$ is chosen to be RKHS generated by the polynomial kernel K of degree 2, i.e., $K(x, \tilde{x}) = (x\tilde{x} + 1)^2$. This choice is made according to [24], where the same kernel has been used in the approach (3) for dealing with the data (24).

For the considered simulation of the data (24) the sequence of the variables ordered according to their ranks looks as follows:

$$x_2, x_4, x_3, x_1, x_{33}, x_6, \ldots, x_{18}. \tag{26}$$

Then as it is described above, the next step consists in testing whether the values of the discrepancy

$$\mathscr{D}(f_2^{\lambda_2}, f_4^{\lambda_4};\ Z_N) = \left(\frac{1}{N} \sum_{i=1}^{N} \left(y^i - f_2^{\lambda_2}(x_2^i) - f_4^{\lambda_4}(x_4^i) \right)^2 \right)^{1/2}$$

corresponding to the small values λ_2, λ_4 are dominated by the ones corresponding to the small λ_2 and the large λ_4. Here and below we use the convention that in the notation $\mathscr{D}\left(f_{\mu_1}^{\lambda_{\mu_1}}, f_{\mu_2}^{\lambda_{\mu_2}}, \ldots, f_{\mu_l}^{\lambda_{\mu_l}};\ Z_N\right)$, the symbols $f_{\mu_j}^{\lambda_{\mu_j}}$ mean the minimizers of the functional

$$T_\lambda^2\left(f_{\mu_1}, f_{\mu_2}, \ldots, f_{\mu_l};\ Z_N\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - \sum_{j=1}^{l} f_{\mu_j}\left(x^i_{\mu_j}\right)\right)^2 + \sum_{j=1}^{l}\lambda_{\mu_j}\left\|f_{\mu_j}\right\|^2_{\mathscr{H}_{\mu_j}}.$$

In our experiments the small values of the regularization parameters are randomly chosen within the set

$$\Lambda_{50}^{\text{small}} = \{\ \lambda = \lambda^{(k)} = 10^{-4}\cdot(1.3)^k,\ k = 1, 2, \ldots, 15\ \}, \tag{27}$$

while the large values are selected at random from

$$\Lambda_{50}^{\text{large}} = \{\ \lambda = \lambda^{(k)} = 10^{-4}\cdot(1.3)^k,\ k = 40, 41, \ldots, 50\ \}. \tag{28}$$

Moreover, in all experiments the random choice of the regularization parameters from $\Lambda_{50}^{\text{small}}$ and $\Lambda_{50}^{\text{large}}$ is performed 15 times.

For the considered simulations of the data (24) and randomly chosen λ_2, λ_4, the values of the discrepancy $\mathscr{D}(f_2^{\lambda_2}, f_4^{\lambda_4};\ Z_N)$ are displayed in Fig. 1 (top). Note that in Fig. 1 and in some other figures below, the curves displaying the values of the discrepancy for the regularization parameters from $\Lambda_{50}^{\text{small}}$ look like straight lines. In view of Theorem 1, the fluctuations in the values of the discrepancy corresponding to the small values of the regularization parameters are indeed small. They are not so much visible because of the vertical axis scaling used in the figures.

According to our approach, the behavior of the discrepancy displayed in Fig. 1 (top) means that the corresponding variables x_2, x_4 have to be accepted as the relevant ones. Then taking into account the ranking (26), we need to check the behavior of the discrepancy $\mathscr{D}(f_2^{\lambda_2}, f_4^{\lambda_4}, f_3^{\lambda_3};\ Z_N)$ for $\{\lambda_2, \lambda_4, \lambda_3\} \subset \Lambda_{50}^{\text{small}}$, and $\{\lambda_2, \lambda_4\} \subset \Lambda_{50}^{\text{small}}$, $\lambda_3 \in \Lambda_{50}^{\text{large}}$. This behavior is displayed in Fig. 1 (middle), and it allows the acceptance of x_3 as the next relevant variable.

In view of Fig. 1 (bottom) displaying the behavior of the discrepancy

$$\mathscr{D}(f_2^{\lambda_2}, f_4^{\lambda_4}, f_3^{\lambda_3}, f_1^{\lambda_1};\ Z_N),$$

the same conclusion can be made regarding the variable x_1.

At the same time, further testing along the ranking list (26) shows that the discrepancies $\mathscr{D}(f_2^{\lambda_2}, f_4^{\lambda_4}, f_3^{\lambda_3}, f_1^{\lambda_1}, f_j^{\lambda_j};\ Z_N)$ with $j = 33, 6, \ldots, 18$ do not exhibit a substantial growth for $\lambda_j \in \Lambda_{50}^{\text{large}}$. Typical examples are displayed in Fig. 2, and they correspond to the behavior described by Theorems 2 and 3. Therefore, our approach does not allow the acceptance of $x_{33}, x_6, \ldots, x_{18}$ as the relevant variables.

Thus, for the considered simulation of the data (24) all relevant variables are correctly detected by the proposed approach.

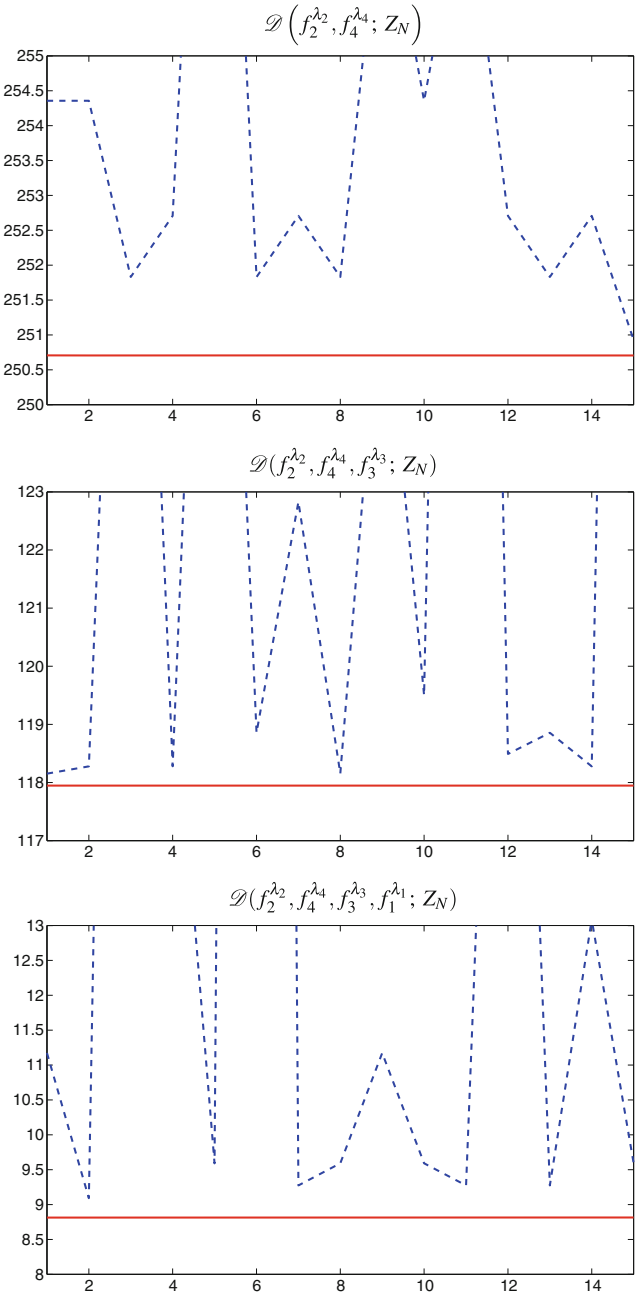

Fig. 1 Behavior of the discrepancies in the experiment with the data (24); x-axis corresponds to the simulation number, y-axis — to the discrepancy value. Red solid line depicts the values of the discrepancy when all regularization parameters are randomly chosen from $\Lambda_{50}^{\text{small}}$. Blue dashed line depicts the values of the discrepancy when the last regularization parameter is randomly chosen from $\Lambda_{50}^{\text{large}}$, and other regularization parameters are randomly chosen from $\Lambda_{50}^{\text{small}}$.

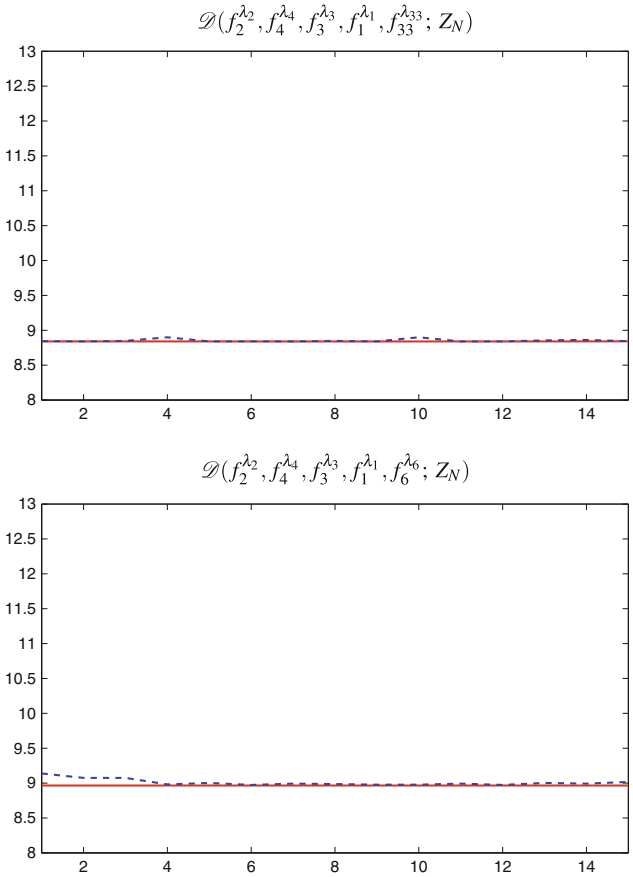

Fig. 2 Behavior of the discrepancies in the experiment with the data (24); x-axis corresponds to the simulation number, y-axis — to the discrepancy value. Red solid line depicts the values of the discrepancy when all regularization parameters are randomly chosen from $\Lambda_{50}^{\text{small}}$. Blue dashed line depicts the values of the discrepancy when the last regularization parameter is randomly chosen from $\Lambda_{50}^{\text{large}}$, and other regularization parameters are randomly chosen from $\Lambda_{50}^{\text{small}}$.

3 Application to the Reconstruction of a Causality Network

In this section we discuss the application of our approach based on multi-penalty regularization to the inverse problem of detecting causal relationships between genes from the time series of their expression levels.

Considering each gene in a genome as a distinct variable, say u_v, associated to the rate of gene expression, the value $u_v^t = u_v(t)$ of this variable at the time moment t can be influenced by the values $u_j^\tau = u_j(\tau), j = 1, \ldots, p$, at the time moments preceding t, i.e., $\tau < t$. This influence is realized through the regulatory proteins produced by genes. Moreover, gene expression levels u_j^τ are often interpreted and

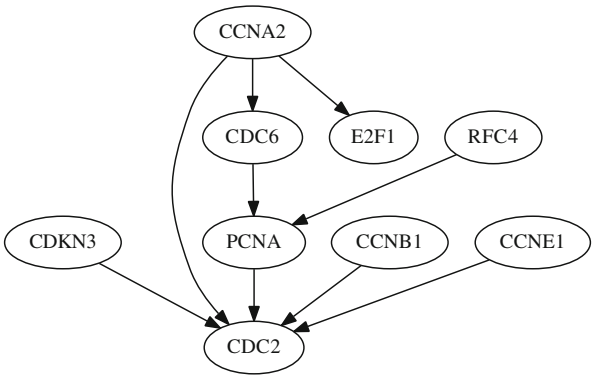

Fig. 3 Causality network of the human cancer cell line HeLa from the BioGRID database (www.thebiogrid.org).

measured in terms of levels or amounts of such proteins. Therefore, time series gene expression data can be used for detecting causal relationships between genes and constructing gene regulatory networks allowing better insights into the underlying cellular mechanisms.

A gene regulatory network or, more generally, a causality network is a directed graph with nodes that are variables u_ν, $\nu = 1, 2, \ldots, p$, and directed edges representing causal relations between variables. We write $u_\nu \leftarrow u_j$ if the variable u_j has the causal influence on the variable u_ν. An example of such a network is presented in Fig. 3. This network contains genes that are active in the human cancer cell line HeLa [37]. This network was derived from the biological experiments in [21], and then, it was used for testing several algorithms devoted to the causality detection [23, 27, 30, 32]. Using the same data as in the above papers, we discuss an applicability of our approach in reconstructing the causalities within this network.

A causality network can be characterized by the so-called adjacency matrix $A = \{A_{\nu,j}\}_{\nu,j=1}^{p}$ with the following elements $A_{\nu,j} = 1$ if $u_\nu \leftarrow u_j$, otherwise, $A_{\nu,j} = 0$. In Fig. 4 we present the adjacency matrix $A = A^{\text{true}}$ corresponding to the causality network displayed in Fig. 3. Adjacency matrices allow a convenient comparison of different reconstruction methods of causality networks.

Note that causality networks arise in various scientific contexts. A detailed overview of the approaches for measuring a causal influence can be found in [17]. A concept of causality in the analysis of time series data has been proposed by Clive W. J. Granger [14], who was awarded the Nobel Prize in Economic Sciences in 2003.

The concept of causality in the Granger approach is based on the assumption that (i) the cause should precede its effect, and (ii) the cause contains information about the effect that is in no other variable. A consequence of these assumptions is that the causal variable u_j can help to forecast the effect variable u_ν. In this restricted sense of causality, referred to as Granger causality, the variable u_j is said to cause another variable u_ν if future values u_ν^t, $t = L+1, L+2, \ldots, T$, of u_ν can be better predicted using the past values u_j^τ, u_ν^τ, $\tau = t-1, t-2, \ldots, t-L$, of u_j and u_ν rather

than using only the past values of u_v. Here L is the maximum lag allowed in the past observations, and we assume that the available time series data are $\{u_j^t\}_{t=1}^T$, $\{u_v^t\}_{t=1}^T$.

The notion of Granger causality was originally defined for a pair of time series and was based on linear regression models. If we are interested in cases in which p time series variables are presented, and we wish to determine causal relationships between them, then we naturally turn to the Graphical Granger modeling [1] based on the linear multivariate regression of the form

$$u_v^t \approx \sum_{j=1}^{p} \sum_{l=1}^{L} \beta_j^l u_j^{t-l}, \quad t = L+1, L+2, \ldots, T. \tag{29}$$

Then, u_j is said to be Granger-causal for u_v if the corresponding coefficients β_j^l, $l = 1, 2, \ldots, L$, are in some sense significant. Thus, we are interested in selecting the most important coefficients. For this purpose, a particular relevant class of methodologies is those that combine regression with variable selection, such as the Lasso [33, 41], which minimizes the squared discrepancy plus a penalty on the sum, or the weighted sum of the absolute values of the regression coefficients β_j^l.

Lasso-type estimates have been used for discovering graphical Granger causality by a number of researchers, including [1, 27, 32]. Note that in regularization theory Lasso is known as the l_1-Tikhonov regularization. It has been extensively studied in the framework of the reconstruction of the sparse structure of an unknown signal. It should be also mentioned that the sparsity enforcing regularization techniques, such as Lasso, are viewed now as a methodology for the quantitative inverse problems in systems biology [11].

At the same time, as it is mentioned in [23], the Lasso estimate of the graphical Granger causality may result in a model (29) in which the large (significant) coefficients β_j^l appear in many sums $\sum_{l=1}^{L} \beta_j^l u_j^{t-l}$. Such a model is hard to interpret, because of natural groupings existing between time series variables $\{u_j^{t-l}\}_{l=1}^L$, $j = 1, 2, \ldots, p$. We mean that the time series variables $\{u_j^{t-l}\}_{l=1}^L$ with the same index, say $j = j_1$, should be either selected or eliminated as a whole. The group Lasso procedure [39, 40] was invented to address this issue, and it was used in [23] in order to obtain the corresponding Granger graphical model of gene regulatory networks. According to this model, a gene u_{j_1} causes a gene u_v if in (29) the coefficients $\beta_{j_1}^l$, $l = 1, 2, \ldots, L$, are significant components of the vector $\beta = (\beta_j^l)$ solving the minimization problem

$$\sum_{t=L+1}^{T} \left(u_v^t - \sum_{j=1}^{p} \sum_{l=1}^{L} \beta_j^l u_j^{t-l} \right)^2 + \lambda \sum_{j=1}^{p} \left(\sum_{l=1}^{L} (\beta_j^l)^2 \right)^{1/2} \to \min_{\beta}. \tag{30}$$

Here note that similarly to (3) by using the square root $(\cdot)^{1/2}$ in the penalty term, one encourages the coefficients associated with each particular gene to be similar in

amplitude, as contrary to using the l_1-norm, for example. The opposite side of this is that the procedures of minimizing (30) are nonlinear and require the solution of $O(pL)$ equations on each iteration step. This can be computationally expensive for large number p of genes.

On the other hand, the above-mentioned natural groupings between the values u_j^τ of variables u_j can be introduced already in the multivariate regression by considering instead of (29) the following form

$$u_\nu^t \approx \sum_{j=1}^{p} f_j \left(\sum_{l=1}^{L} \beta_j^l u_j^{t-l} \right), \quad t = L+1, L+2, \ldots, T, \tag{31}$$

where f_j are univariate functions in some Reproducing Kernel Hilbert Spaces $\mathscr{H}_j$. Note that (31) can be seen as a particular form of structural equation models discussed in [26]. Then a conclusion that the gene u_{j_1} causes the gene u_ν can be drawn by determining that the variable x_{j_1} is a relevant variable of a function of the form (1) whose values at the points

$$x_j^i = \sum_{l=1}^{L} \beta_j^l u_j^{L+i-l}, \; i = 1, 2, \ldots, T-L, \; j = 1, 2, \ldots, p, \tag{32}$$

are equal to

$$y^i = u_\nu^{L+i}, \; i = 1, 2, \ldots, T-L. \tag{33}$$

Of course, the latter conclusion can be drawn only when in (32) some values of the coefficients β_j^l have been already set. For example, these regression coefficients can be precomputed in (29) by some inexpensive algorithm such as the ordinary or regularized least squares (OLS or RLS). Note that such a precomputation step is also required in Adaptive Lasso [41] that has been discussed in the context of the regulatory networks discovery in [23], and where an auxiliary vector estimator of the coefficients in (29) is usually obtained by OLS or Ridge Regression.

Another possibility of determining the coefficients β_j^l in (32) is to use the output vector of any of the graphical Granger models based on (29) such as [23, 32]. In this case, the discussed approach provides an opportunity of additional evaluation of these models in the sense that causal relationships detected by them and confirmed in the discussed approach can be considered as more certain.

After specifying the coefficients β_j^l in (32), the values (32), (33) can form the data set $Z_N = \{(x_1^i, x_2^i, \ldots, x_p^i \; ; \; y^i)\}_{i=1}^N$, $N = T - L$. Then, the detection of the relevant variables from the data Z_N follows the approach described in Sect. 1 and analyzed in Sect. 2. The only adjustment is that in view of the idea of Granger causality (comparison of the accuracy of regressing for u_ν in terms of its own past values with that of regressing in terms of the values u_ν and the values of a possible cause), we start the ranking list of variables with the variable x_ν when looking for the genes causing the gene u_ν.

Below we present the results of the application of the proposed approach to the data of the gene expressions for the network of genes displayed in Fig. 3. These data is taken as in [23, 30, 32]. In (9), (10), (31) all univariate functions f_j are assumed to be in the same RKHS generated by the Gaussian kernel $K(x, v) = e^{-(x-v)^2}$. Moreover, the standard RLS-algorithm has been used for precomputing the regression coefficients in (31), (32). The regularization parameter in RLS has been chosen according to the quasi-optimality criterion. As in [23, 30, 32] the gene expressions $\{u_j^t\}$ are observed for $t = 1, 2, \ldots, 47$, and, as in [23, 27], the maximum lag was chosen as $L = 4$. Then, we follow the same steps as in the illustrating example in Sect. 2. In particular, we use the same sets $\Lambda_{50}^{\text{small}}$, $\Lambda_{50}^{\text{large}}$.

The application of the proposed approach to the above-mentioned data results in the adjacency matrix A^{MP} displayed in Fig. 4. The corresponding causality network can be found in Fig. 5.

As it has been already mentioned, the data corresponding to the causality network in Fig. 3 was used for testing several methods devoted to the regulatory networks modeling. First, it was used in [30], where the authors developed a search-based algorithm, called CNET, and applied it to this set of data. Then, the same set of nine genes was also analyzed in [23] by means of group Lasso (GL) algorithm based on the minimization of the functionals of the form (30). In [32] the authors pointed out some limitations of GL-algorithm and proposed to overcome them by means of the so-called truncating Lasso (TL) penalty algorithm. Fig. 4 presents the adjacency

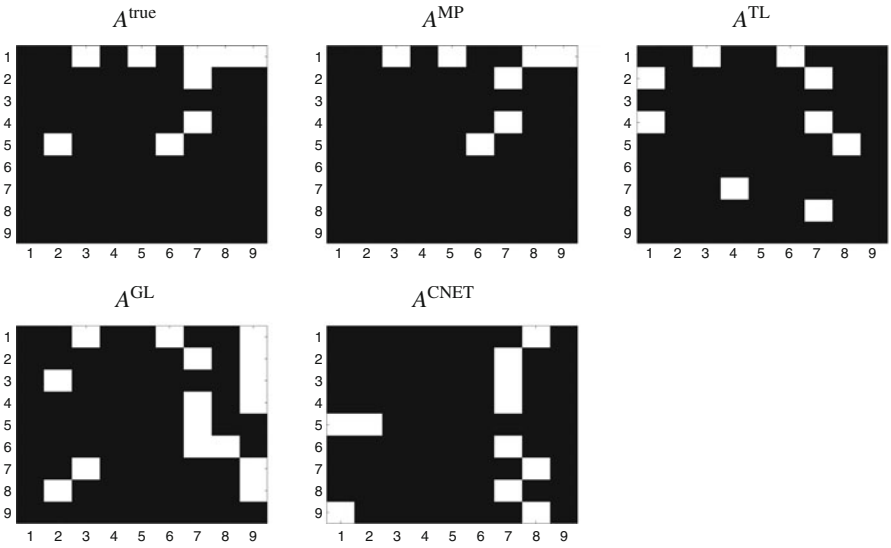

Fig. 4 The adjacency matrix A^{true} for the causality network in Fig. 3 and its various estimations. The white squares correspond to $A_{i,j} = 1$; the black squares — to the zero-elements. The genes are numbered in the following order: CDC2, CDC6, CDKN3, E2F1, PCNA, RFC4, CCNA2, CCNB1, CCNE1.

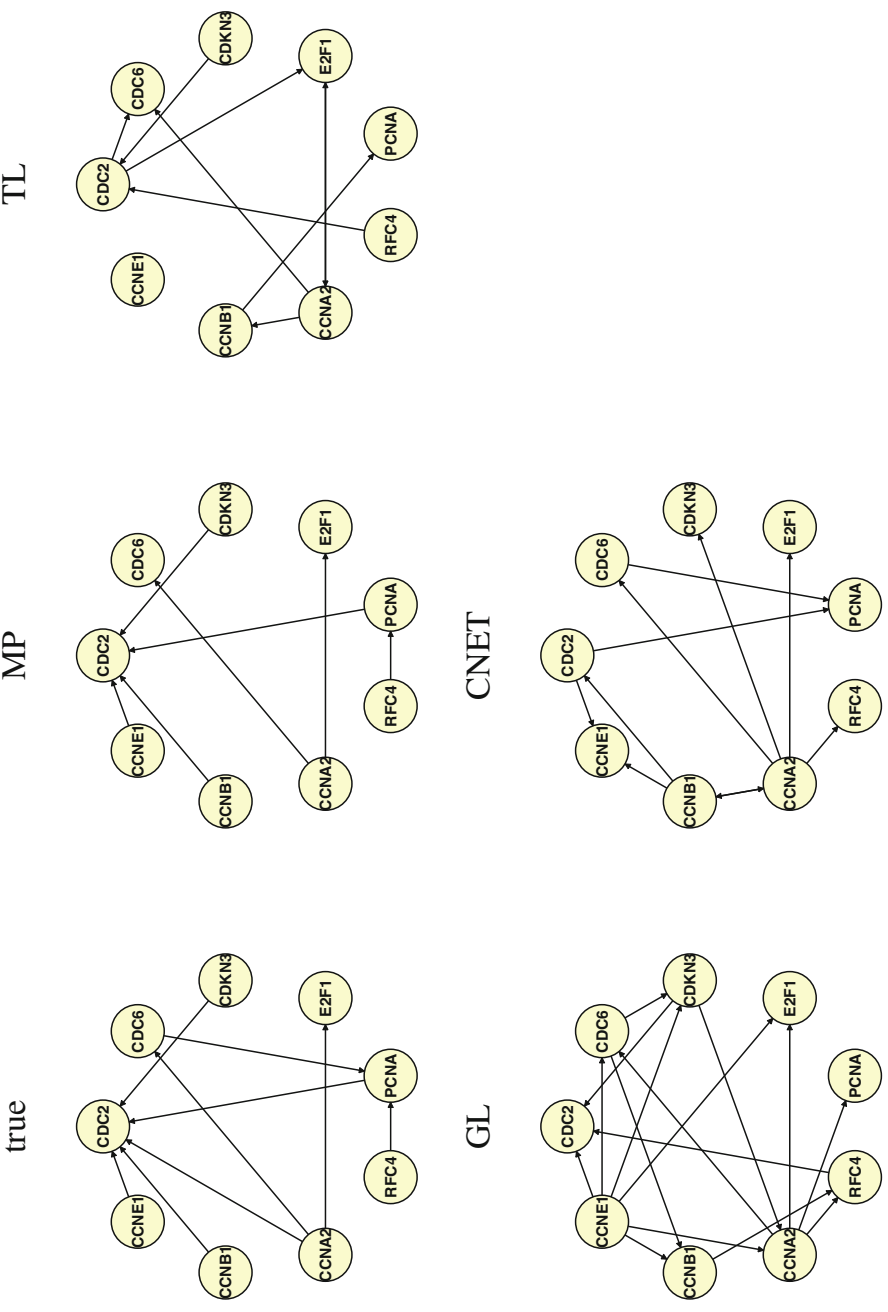

Fig. 5 The causality network in Fig. 3 and its various estimations.

Table 1 The values of the performance measures for the adjacency matrices in Fig. 4.

	P	R	F_1
A^{MP}	1	0.78	0.88
A^{CNET}	0.36	0.44	0.4
A^{GL}	0.24	0.44	0.3
A^{TL}	0.3	0.33	0.32

matrices A^{CNET}, A^{GL}, A^{TL} of the estimated causality network with the genes from Fig. 3 obtained, respectively, by the algorithms from [23, 30, 32]. The corresponding causality networks are presented in Fig. 5.

As in [32] to assess the performance of the discussed algorithms, we use three well-known performance measures: precision (P), recall (R), and their harmonic mean (F_1) (see, e.g., [38]). Table 1 contains the values of these measures for the adjacency matrices given by the discussed methods and displayed in Fig. 4. This table shows that the best performance is achieved by our approach.

To illustrate the steps of our approach in reconstructing the network from Fig. 3, we present Fig. 6 displaying the behavior of the discrepancies, which in the present context play the role of the indicators for the causal relationships. This figure is related to gene CDC2 numbered as x_1. We take this gene as an example because its causing genes are poorly detected by the CNET, GL, and TL algorithms.

Using the data for this gene and transforming them into (32),(33) with $\nu = 1$, we receive the following sequence of the variables ordered according to their ranks

$$x_1, x_3, x_7, x_5, x_4, x_9, x_6, x_2, x_8.$$

Fig. 6 displays the behavior of the discrepancies

$$\mathscr{D}\left(f_1^{\lambda_1}, f_3^{\lambda_3}, f_5^{\lambda_5};\ Z_N\right),$$

$$\mathscr{D}\left(f_1^{\lambda_1}, f_3^{\lambda_3}, f_5^{\lambda_5}, f_9^{\lambda_9}, f_6^{\lambda_6};\ Z_N\right),$$

$$\mathscr{D}\left(f_1^{\lambda_1}, f_3^{\lambda_3}, f_5^{\lambda_5}, f_9^{\lambda_9}, f_8^{\lambda_8};\ Z_N\right)$$

considered, respectively, at 3th, 6th, and 8th steps of our approach. The reason to present these steps as examples is explained below.

The behavior of the discrepancy displayed in Fig. 6 (top) indicates that according to our approach, the variable x_5, which corresponds to gene PCNA, should be considered as the cause for CDC2. From Fig. 3 one can see that this causal relationship is true, but it has not been detected by any other considered algorithms.

According to our approach, the interpretation of the erratic behavior of the discrepancies in Fig. 6 (middle) is that x_6 is not the relevant variable, and therefore, the corresponding gene RFC4 does not cause CDC2. This conclusion is also in agreement with Fig. 3. At the same time, the relationship RFC4 $\rightarrow$ CDC2 is wrongly detected by both Lasso-based algorithms GL and TL.

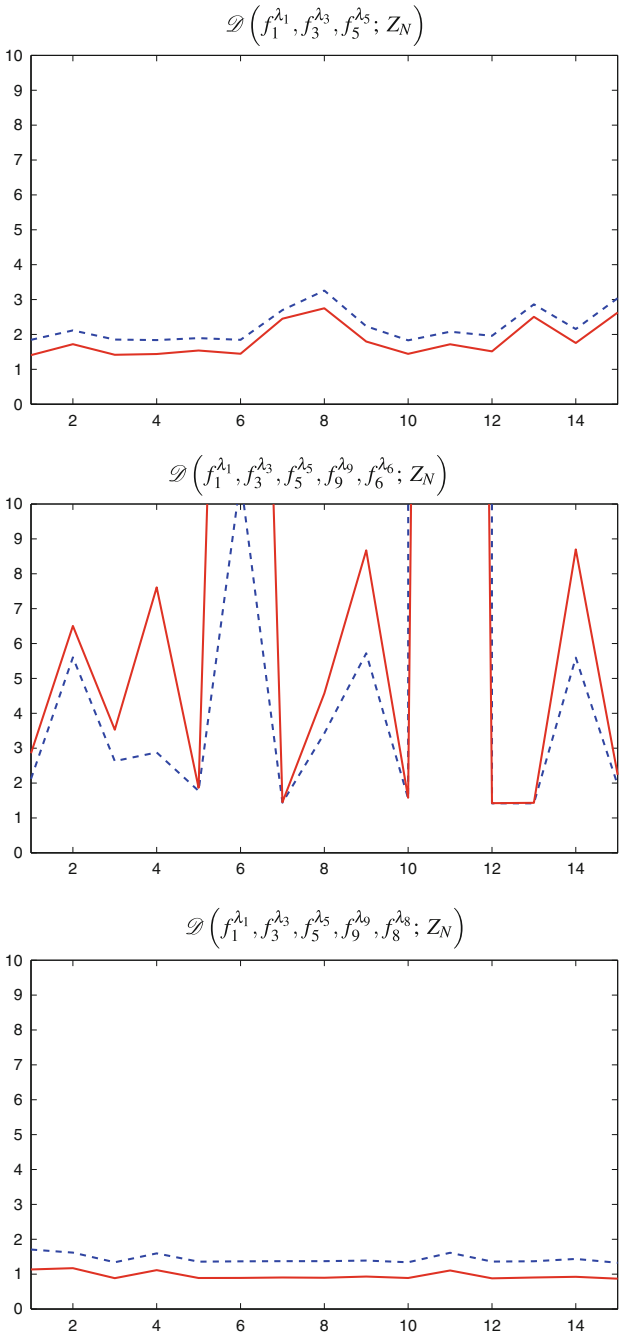

Fig. 6 Behavior of the discrepancies in the experiment with the gene expressions data; x-axis corresponds to the simulation number, y-axis — to the discrepancy value. Red solid line depicts the values of the discrepancy when all regularization parameters are randomly chosen from $\Lambda_{50}^{\text{small}}$. Blue dashed line depicts the values of the discrepancy when the last regularization parameter is randomly chosen from $\Lambda_{50}^{\text{large}}$, and other regularization parameters are randomly chosen from $\Lambda_{50}^{\text{small}}$.

The situation in Fig. 6 (bottom) is opposite. According to our approach, the behavior displayed in this Fig. means that x_8 is the relevant variable and, thus, CCNB1 $\rightarrow$ CDC2. This relationship is true, but it was not detected by the Lasso-based algorithms.

Therefore, in our opinion, Table 1 and Fig. 6 can be seen as an evidence of the reliability of the proposed approach in the application to the real data.

4 Conclusion

We have proposed a new method for detecting relevant variables. The method is based on the inspection of the behavior of discrepancies of multi-penalty regularization with a component-wise penalization for small and large values of the regularization parameters. An ordered behavior suggests the acceptance of the hypothesis that the corresponding variable is the relevant one, while an erratic behavior of discrepancies is the signal for the rejection of the hypothesis.

We provided a justification of the proposed method under the condition that the corresponding sampling operators share a common singular system in $\mathbb{R}^n$. We also demonstrated the applicability of the method on the inverse problem of the reconstruction of a gene regulatory network.

The promising performance of the method in the mentioned application calls for its further investigation. In particular, it is interesting to study the conditions on the sampling points/operators guaranteeing or preventing the detection of relevant variables. It is also interesting to study the application of the proposed approach to the detection of the cause-effect relationships in various scientific contexts. As it was mentioned, the approach can be realized on the top of different techniques for discovering Granger causality. Therefore, the coupling of the known techniques with the presented approach is a further interesting point for detailed investigations.

Acknowledgements The first author gratefully acknowledges the partial support by the research grant GA16-09848S of the Grant Agency of the Czech Republic (Czech Science Foundation). The major part of this work has been prepared, when the second author was staying at RICAM as a PostDoc. She gratefully acknowledges the partial support by the Austrian Science Fund (FWF): project P 25424, "Data-driven and problem-oriented choice of the regularization space." The third author gratefully acknowledges the support by the Austrian Science Fund (FWF): project P 29514-N32, "Regularization techniques in learning with Big Data."

References

1. A. Arnold, Y. Liu, N. Abe, Temporal causal modeling with graphical Granger methods, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2007), pp. 66–75
2. N. Aronszajn, Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)

3. F. Bach, Exploring large feature spaces with hierarchical multiple kernel learning, in *Advances in Neural Information Processing Systems 21*, ed. by D. Koller et al. (2009), pp. 105–112
4. F. Bach, G.R.G. Lanckriet, M.I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm, in *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004
5. F. Bauer, M. Reiß, Regularization independent of the noise level: an analysis of quasi-optimality. Inverse Prob. **24**(5), 16 pp. (2008)
6. C. Berg, J.P.R. Christensen, P. Ressel, *Harmonic Analysis on Semigroups. Theory of Positive Definite and Related Functions* (Springer, New York, 1984)
7. A. Christmann, R. Hable, Consistency of support vector machines using additive kernels for additive models. Comput. Stat. Data Anal. **56**(4), 854–873 (2012)
8. F. Cucker, S. Smale, On the mathematical foundations of learning. Bull. Am. Math. Soc. New Ser. **39**, 1–49 (2002)
9. I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004)
10. H.W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems* (Kluwer Academic Publishers, Dordrecht, 1996)
11. H.W. Engl, C. Flamm, J. Lu, P. Kügler, S. Müller, P. Schuster, Inverse problems in systems biology. Inverse Prob. **25**(12), 123014 (2009)
12. M. Fornasier (ed.), *Theoretical Foundations and Numerical Methods for Sparse Recovery* (de Gruyter, Berlin, 2010)
13. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Springer, New York, 2013)
14. C. Granger, Investigating causal relations by econometric models and crossspectral methods. Econometrica **37**, 424–438 (1969)
15. M. Grasmair, M. Haltmeier, O. Scherzer, Sparse regularization with l^q penalty term. Inverse Prob. **24**(5), 13 (2008)
16. T. Hastie, R. Tibshirani, *Generalized Additive Models* (Chapman and Hall, London, 1990)
17. K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, J. Bhattacharya, Causality detection based on information-theoretic approaches in time series analysis. Phys. Rep. **441**, 1–46 (2007)
18. G.S. Kimeldorf, G. Wahba, A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Stat. **41**(2), 495–502 (1970)
19. S. Kindermann, A. Neubauer, On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization. Inverse Prob. Imag. **2**(2), 291–299 (2008)
20. V. Koltchinskii, M. Yuan, Sparsity in multiple kernel learning. Ann. Stat. **38**(6), 3660–3695 (2010)
21. X. Li et al., Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. BMC Bioinformatics **7**(26) (2006). doi:10.1186/1471-2105-7-26
22. D.A. Lorenz, P. Maass, Q.M. Pham, Gradient descent for Tikhonov functionals with sparsity constraints: theory and numerical comparison of step size rules. Electron. Trans. Numer. Anal. **39**, 437–463 (2012)
23. A.C. Lozano, N. Abe, Y. Liu, S. Rosset, Grouped graphical Granger modeling for gene expression regulatory networks discovery. Bioinformatics **25**, 110–118 (2009)
24. S. Mosci, L. Rosasco, M. Santoro, A. Verri, S. Villa, Nonparametric sparsity and regularization. Technical Report 41, MIT, CSAIL, Cambridge (2011)
25. V. Naumova, S. Pereverzyev, Multi-penalty regularization with a component-wise penalization. Inverse Prob. **29**(7), 15 (2013)
26. J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, 2000)
27. S. Pereverzyev, K. Hlaváčková-Schindler, Graphical Lasso Granger method with 2-levels-thresholding for recovering causality networks, in *System Modeling and Optimization: 26th IFIP TC 7 Conference, Klagenfurt, 2013*, Revised Selected Papers, ed. by C. Pötzsche et al. (Springer, Berlin, 2014), pp. 220–229

28. R. Ramlau, G. Teschke, A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints. Numer. Math. **104**(2), 177–203 (2006)
29. P. Ravikumar, J. Lafferty, H. Liu, L. Wasserman, Sparse additive models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **71**(5), 1009–1030 (2009)
30. F. Sambo, B.D. Camillo, G. Toffolo, CNET: an algorithm for reverse engineering of causal gene networks, in *NETTAB2008*, Varenna (2008)
31. B. Schölkopf, R. Herbrich, A.J. Smola, A generalized representer theorem, in *Computational Learning Theory*. Lecture Notes in Computer Science, vol. 2111 (Springer, Berlin, 2001), pp. 416–426
32. A. Shojaie, G. Michailidis, Discovering graphical Granger causality using the truncating lasso penalty. Bioinformatics **26**, i517–i523 (2010)
33. R. Tibshirani, Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)
34. A.N. Tikhonov, V.Y. Arsenin, *Solutions of Ill-Posed Problems* (Winston, New York, 1977)
35. A.N. Tikhonov, V.B. Glasko, Use of the regularization method in non-linear problems. USSR Comput. Math. Math. Phys. **5**, 93–107 (1965)
36. G.M. Vainikko, A.Y. Veretennikov, *Iteration Procedures in Ill-Posed Problems* (Nauka, Moscow, 1986) [in Russian]
37. M.L. Whitfield et al., Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell **13**, 1977–2000 (2002)
38. Wikipedia, Precision and recall — Wikipedia, The Free Encyclopedia, 2014. Online Accessed 3 Jan 2014
39. M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. B **68**, 49–67 (2006)
40. P. Zhao, G. Rocha, B. Yu, The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat. **37**(6A), 3468–3497 (2009)
41. H. Zou, The adaptive Lasso and its oracle properties. J. Am. Stat. Assoc. **101**(476), 1418–1429 (2006)

http://www.springer.com/978-3-319-55555-3

Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science
Novel Methods in Harmonic Analysis, Volume 2
Pesenson, I.; Le Gia, Q.T.; Mayeli, A.; Mhaskar, H.; Zhou, D.-X. (Eds.)
2017, X, 515 p. 81 illus., 25 illus. in color., Hardcover
ISBN: 978-3-319-55555-3
A product of Birkhäuser Basel