INFORMATICA, 2016, Vol. 27, No. 3, 607–624

DOI: http://dx.doi.org/10.15388/Informatica.2016.102

Analytic and Stochastic Methods of Structure Parameter Estimation

Mikhail KUZNETSOV*, Aleksandra TOKMAKOVA, Vadim STRIJOV
Moscow Institute of Physics and Technology
Institutskiy lane 9, Dolgoprudny, Moscow region, 141700, Russia
e-mail: mikhail.kuznecov@phystech.edu, aleksandra-tok@yandex.ru, strijov@phystech.edu

Received: November 2014; accepted: April 2015

Abstract. The paper presents analytic and stochastic methods of structure parameters estimation for a model selection problem. Structure parameters are covariance matrices of parameters of linear and non-linear regression models. To optimize model parameters and structure parameters we maximize a model evidence, a convolution of a data likelihood with a prior distribution of model parameters. The analytic methods are based on the derivatives computation of the approximated model evidence. The stochastic methods are based on the model parameters sampling and data cross-validation. The proposed methods are tested and compared on the synthetic and real data.

Key words: structure parameters optimization, regression model, error function, Laplace approximation, Monte-Carlo estimation, cross-validation.

1. Introduction

Structure parameter estimation is the key problem of model complexity estimation and model selection. The problem is to find a regression model (Strijov and Weber, 2010; Draper and Smith, 1998; Kutner *et al.*, 2004; Bishop and Tipping, 2003) approximating measured data and to estimate regression model parameters (McCullagh and Nelder, 1989). The measured data consists of the dependent and independent variable measurements.

To estimate the model parameters one should find a minimum of the error function over the set of parameters (Bishop, 2006; Hastie *et al.*, 2009). The error function is inferred from the algebraic or statistic approaches.

The statistical approach to the parameter estimation problem assumes the dependent variable and the model parameters to be multidimensional random variables described by their probability distribution functions. According to the statistical framework, the optimized error function is regarded as the form of the transformed likelihood (Eidsvik *et al.*, 2011; Packalen and Wirjanto, 2012).

To construct and optimize the error function we use the Bayesian model comparison method (Zellner, 1995; Strijov and Weber, 2010; Liang and Yang, 2009). According to this method, a prior distribution is defined over the set of regression model parameters.

*Corresponding author.

We consider the case of multivariate normal prior and dependent variable distributions. In this case the structure parameters are the covariance matrices of the distributions.

The structure parameters are also regarded as the regularization parameters that penalize the elements of the model parameters vector (Lampe and Voss, 2013; Cawley and Talbot, 2007; Gillard, 2010). In this study the structure parameters are unknown and should be estimated to find the optimal model parameters. To estimate them we compare several methods of the *model evidence* optimization (Hu, 2005; Chang and Tsay, 2010; Lopes *et al*., 1999).

The first method is the maximization of the simplified form of model evidence called *Laplace approximation* (Wood and Butler, 2003; Mackay, 1998). The method expands the data likelihood function around an optimum value of the parameter vector and allows to obtain an exact form of the approximated evidence. We consider different types of the covariance matrices structure and obtain different forms of the evidence approximation for each type.

The second method is based on the Monte-Carlo approximation of the model evidence (Alessandri *et al*., 2010; Betro and Vercellis, 2007). The parameter vector is sampled according to the given distribution function. The optimum structure parameters maximize the model evidence calculated over the set of sampled parameters.

To validate and compare the proposed methods we use the cross-validation schema of the structure parameters estimation (Hastie *et al*., 2009; An *et al*., 2007). The method is based on the sample splitting into parts of nearly equal size. The model parameters are estimated on the each part of the sample set.

In a special case we consider linear regression models (McCullagh and Nelder, 1989). For this type of models we derive explicit values of the model parameters and of the Hessian matrix (Zhang and Leithead, 2005).

2. Structure Parameter Estimation Problem

Consider a data set $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ consisting of the independent variables $\mathbf{x}_i = [x_{ij}]_{j=1}^n$, $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ and the dependent targets y_i, $y \in \mathbb{Y} \subseteq \mathbb{R}^1$.

Let $f : \mathbb{W} \times \mathbb{X} \to \mathbb{Y}$ be the regression model mapping Cartesian product of the model parameter space $\mathbb{W}$ and the independent variable space $\mathbb{X}$ to the dependent variable space $\mathbb{Y}$. The elements of the data set are related by the following equation,

$$y_i = f(\mathbf{w}, \mathbf{x}_i) + \varepsilon(\mathbf{x}_i), \tag{1}$$

with a normal random noise $\varepsilon = \varepsilon(\mathbf{x})$. By $\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X})$ we denote a vector-valued function

$$\mathbf{f}(\mathbf{w}, \mathbf{X}) = \big[f(\mathbf{w}, \mathbf{x}_1), \ldots, f(\mathbf{w}, \mathbf{x}_m)\big]^T.$$

According to the regression equation (1) we treat a vector $\mathbf{y}$ as a multivariate normally distributed random variable,

$$\mathbf{y} \sim \mathcal{N}\big(\mathbf{f}, \mathbf{B}^{-1}\big),$$

such that $\mathbf{B}$ is an inverse covariance matrix of the noise vector $\boldsymbol{\varepsilon}$.

To determine the parameter space $\mathbb{W}$ we use the Bayesian framework. The parameter vector $\mathbf{w}$ is the multivariate random vector defined by *a prior* distribution $p(\mathbf{w}|\mathbf{A})$. In this paper we consider multivariate normal priors with different types of covariance matrices described in Section 2.2.

2.1. *Model Evidence*

We use the coherent Bayesian inference method to estimate the parameters $\mathbf{w}$ of the model f. According to this approach, we define a prior probability distribution $p(\mathbf{w}|\mathbf{A})$ of the parameters $\mathbf{w}$, where $\mathbf{A}$ is an inverse covariance matrix of the parameters vector $\mathbf{w}$.

Matrices $\mathbf{A}$ and $\mathbf{B}$ are referred to as *structure parameters*. In particular, matrices $\mathbf{A}$ and $\mathbf{B}$ are the parameters of the prior distribution $p(\mathbf{w})$ and the conditional distribution $p(\mathbf{y}|\mathbf{w})$, respectively. Below we will consider special types of this distributions.

The first level of the Bayesian inference selects the best model from the set of competitive models $\mathfrak{F}$ by maximizing a model evidence $(\mathfrak{D}|\mathbf{A}, \mathbf{B})$ over the structure parameters $\mathbf{A}$ and $\mathbf{B}$. To calculate the model evidence, we integrate the data likelihood $p(\mathfrak{D}|\mathbf{w}, \mathbf{B})$ over the set of parameters $\mathbb{W}$ defining by the prior distribution $p(\mathbf{w}|\mathbf{A})$:

$$p(\mathfrak{D}|\mathbf{A}, \mathbf{B}) = \int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w}, \mathbf{B})\, p(\mathbf{w}|\mathbf{A}) d\mathbf{w},$$

so that

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{M}^n, \mathbf{B}\in\mathbb{M}^m}{\arg\max} \left(\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w}, \mathbf{B})\, p(\mathbf{w}|\mathbf{A}) d\mathbf{w} \right), \tag{2}$$

where $\mathbb{M}^n$ is the set of positive-definite symmetric $n \times n$ matrices.

The second level of Bayesian inference estimates $\mathbf{w}$ by maximizing posterior distribution of the parameters,

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbb{W}} p(\mathbf{w}|\mathfrak{D}, \hat{\mathbf{A}}, \hat{\mathbf{B}}).$$

Here the posterior distribution $p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B})$ equals

$$p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}) = \frac{p(\mathfrak{D}|\mathbf{w}, \mathbf{B})\, p(\mathbf{w}|\mathbf{A})}{p(\mathfrak{D}|\mathbf{A}, \mathbf{B})},$$

so that optimal parameters $\hat{\mathbf{w}}$ maximize the integrand of (2):

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbb{W}} \left(p(\mathfrak{D}|\mathbf{w}, \hat{\mathbf{B}})\, p(\mathbf{w}|\hat{\mathbf{A}}) \right).$$

Table 1
Distribution of dependent variable **y** and model parameters **w**.

	Dependent variable **y**	Model parameters **w**	Notations
1)	$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma^2(\mathbf{y})\mathbf{I}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{f}, \beta^{-1}\mathbf{I})$	$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \sigma^2(\mathbf{w})\mathbf{I}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$	$\mathbf{A} = \alpha\mathbf{I}$
2)	$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \text{diag}^{-1}(\beta_1, \ldots, \beta_m)\mathbf{I})$	$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \text{diag}^{-1}(\alpha_1, \ldots, \alpha_n)\mathbf{I})$	$\mathbf{A} = \text{diag}(\alpha_i)\mathbf{I}$
3)	$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{B}^{-1})$	$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \mathbf{A}^{-1})$	$\mathbf{A} \in \mathbb{M}^n$

2.2. *Types of Structure Parameters*

To estimate the parameters $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{w}}$ let us make some assumptions about types of the structure parameters. We will consider normal priors: let vectors **w** and **y** have the multivariate normal distribution with covariance matrices $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$, respectively.

We will distinguish three various types of the matrices **A** and **B**: scalar, diagonal and full type. Table 1 shows various cases of the distribution of the dependent variable **y** and the model parameters **w** according to the possible types of matrices.

In this paper we will focus on different types of the matrix **A** and consider only a scalar type of the matrix **B**, $\mathbf{B} = \beta\mathbf{I}$.

To estimate structure parameters $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ we transform the optimization problem (2) according to the normal distribution of the dependent variable **y** and the model parameters **w**:

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{M}^n, \mathbf{B}\in\mathbb{M}^m}{\arg\max} \left(\frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})} \frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})} \int\limits_{\mathbf{w}\in\mathbb{W}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^T\mathbf{B}(\mathbf{y}-\mathbf{f}) \right) \times \exp\left(-\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} \right) d\mathbf{w} \right). \tag{3}$$

By the *error function* $S(\mathbf{w}, \mathbf{A}, \mathbf{B})$ denote the exponent of the expression (3) with a negative sign,

$$S(\mathbf{w}, \mathbf{A}, \mathbf{B}) = \frac{1}{2}(\mathbf{y}-\mathbf{f})^T\mathbf{B}(\mathbf{y}-\mathbf{f}) + \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}, \tag{4}$$

and the optimization problem (3) will be formulated as follows,

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{M}^n, \mathbf{B}\in\mathbb{M}^m}{\arg\max} \left(\frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})} \frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})} \int\limits_{\mathbf{w}\in\mathbb{W}} \exp\big(-S\big(\mathbf{w}, \mathbf{A}, \mathbf{B}\big)\big) d\mathbf{w} \right).$$

The optimal parameters $\hat{\mathbf{w}}$ maximize the posterior distribution of parameters, or, equivalently, minimize the error function,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|\hat{\mathbf{A}}, \hat{\mathbf{B}}),$$

where $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ are the estimations of the structure parameters maximizing (3).

3. Model Evidence Approximation

In this section we use the Laplace approximation of the model evidence to estimate the structure parameters $\mathbf{A}$, $\mathbf{B}$ and the model parameters $\mathbf{w}$. The method uses Taylor expansion of the error function $S(\mathbf{w})$ near the optimal solution $S(\hat{\mathbf{w}})$ to approximate the expression

$$S(\mathbf{w}) = S(\hat{\mathbf{w}} + \Delta\mathbf{w}) \approx S(\hat{\mathbf{w}}) + \frac{1}{2}\Delta\mathbf{w}^T\mathbf{H}\Delta\mathbf{w},$$

where $\mathbf{H}$ is Hessian of the error function,

$$\mathbf{H} = \nabla\nabla S(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}},$$

at the optimum point, $\mathbf{w} = \hat{\mathbf{w}}$. Denote by $\|\mathbf{w}\|$ the Euclidean norm $\|\mathbf{w}\| = \|\mathbf{w}\|_2$. Instead of optimizing (3) let us optimize its approximated form,

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{M}^n, \mathbf{B}\in\mathbb{M}^m}{\arg\max} \left(\frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})} \frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})} \exp\bigl(-S(\hat{\mathbf{w}})\bigr) \int\limits_{\mathbf{w}\in\mathbb{W}} \exp\left(-\frac{1}{2}\Delta\mathbf{w}^T\mathbf{H}\Delta\mathbf{w}\right) d\mathbf{w} \right). \tag{5}$$

Let us remark that the integrand of (5) is a part of the normal distribution, hence we can substitute an integral in (5) for normalization and obtain:

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{M}^n, \mathbf{B}\in\mathbb{M}^m}{\arg\max} \bigl(g(\mathbf{A}, \mathbf{B})\bigr),$$

$$g(\mathbf{A}, \mathbf{B}) = \frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})} \frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})} \exp\bigl(-S(\hat{\mathbf{w}})\bigr) \frac{(2\pi^{\frac{n}{2}})}{|\mathbf{H}|^{\frac{1}{2}}}. \tag{6}$$

Taking the logarithm of (6), we obtain the following optimization problem,

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{M}^n, \mathbf{B}\in\mathbb{M}^m}{\arg\max} \bigl(\ln g(\mathbf{A}, \mathbf{B})\bigr),$$

$$\ln g(\mathbf{A}, \mathbf{B}) = -\frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{A}| + \frac{1}{2}\ln|\mathbf{B}| - S(\hat{\mathbf{w}}) - \frac{1}{2}\ln|\mathbf{H}|. \tag{7}$$

Let us define a type of the matrices $\mathbf{A}$, $\mathbf{B}$ to simplify the function $\ln g(\mathbf{A}, \mathbf{B})$. In particular, below we will consider the scalar-type $\mathbf{B}$ matrix, $\mathbf{B} = \beta\mathbf{I}$. In this case, the error function (4) is given by

$$S(\mathbf{w}, \mathbf{A}, \beta) = \frac{\beta}{2}(\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} = \beta S_{\mathfrak{D}}(\mathbf{w}) + \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}, \tag{8}$$

where

$$S_{\mathfrak{D}}(\mathbf{w}) = \frac{1}{2}(\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}), \tag{9}$$

and Hessian is given by

$$\mathbf{H} = \beta \mathbf{H}_{\mathfrak{D}} + A,$$

where $\mathbf{H}_{\mathfrak{D}}$ is a Hessian of the function $S_{\mathfrak{D}}(\mathbf{w})$ at the optimum point $\mathbf{w} = \hat{\mathbf{w}}$.

The function (7) is given by the expression

$$\begin{aligned} \ln g(\mathbf{A}, \beta) \; = \; & -\frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{A}| + \frac{m}{2}\ln\beta - \frac{\beta}{2}\big(\mathbf{y} - \mathbf{f}\left(\hat{\mathbf{w}}, \mathbf{X}\right)\big)^T \big(\mathbf{y} - \mathbf{f}\left(\hat{\mathbf{w}}, \mathbf{X}\right)\big) \\ & - \frac{1}{2}\hat{\mathbf{w}}^T \mathbf{A}\hat{\mathbf{w}} - \frac{1}{2}\ln|\beta\mathbf{H}_{\mathfrak{D}} + A|. \end{aligned} \tag{10}$$

Below we will consider scalar and diagonal types of matrix $\mathbf{A}$ to differentiate a summand

$$\frac{1}{2}\ln|\beta\mathbf{H}_{\mathfrak{D}} + \mathbf{A}| \tag{11}$$

of the function (10).

3.1. *Scalar Type of Matrix* $\mathbf{A}$

In this section, let $\mathbf{A}$ be scalar, $\mathbf{A} = \alpha\mathbf{I}$. By this assumption, the expression (11) equals to

$$\frac{1}{2}\ln|\beta\mathbf{H}_{\mathfrak{D}} + \alpha\mathbf{I}| = \frac{1}{2}\sum_{j=1}^{n}\ln(\beta h_j + \alpha),$$

where h_j is an eigenvector of $\mathbf{H}_{\mathfrak{D}}$.

Equating derivatives of (10) with respect to α and β tending to zero, we get an estimation of the structure parameters α and β:

$$\frac{\partial(-\ln g(\alpha, \beta))}{\partial\alpha} = \frac{n}{2\alpha} - \frac{\|\hat{\mathbf{w}}\|^2}{2} - \frac{1}{2}\sum_{j=1}^{n}\frac{1}{\beta h_j + \alpha} = 0,$$

$$\alpha\|\hat{\mathbf{w}}\|^2 = n - \sum_{j=1}^{n}\frac{\alpha}{\beta h_j + \alpha} = \beta\sum_{j=1}^{n}\frac{h_j}{\beta h_j + \alpha}.$$

By definition, put

$$\gamma = \beta\sum_{j=1}^{n}\frac{h_j}{\beta h_j + \alpha}, \tag{12}$$

then

$$\alpha = \frac{\gamma}{\|\hat{\mathbf{w}}\|^2}. \tag{13}$$

Similarly, equating a derivative of (10) with respect to β to zero, we obtain

$$\beta = \frac{m - \gamma}{\|\mathbf{y} - \mathbf{f}(\hat{\mathbf{w}}, \mathbf{X})\|^2}. \tag{14}$$

Since γ is a function of β, α and optimal model parameters $\hat{\mathbf{w}}$, we solve equations (12), (13) and (14) iteratively for the fixed $\hat{\mathbf{w}}$.

3.2. *Diagonal Type of Matrix* **A**

In the case of diagonal matrix $\mathbf{A} = \operatorname{diag}(\alpha_j)$ the obtained results are comparable with equations (12), (13) and (14). In particular, instead of (12), by definition we put

$$\rho = \beta \sum_{j=1}^{n} \frac{h_j}{\beta h_j + \alpha_j},$$

and β is given by

$$\beta = \frac{m - \rho}{\|\mathbf{y} - \mathbf{f}(\hat{\mathbf{w}}, \mathbf{X})\|^2}.$$

To compute the elements of the matrix $\mathbf{A} = \operatorname{diag}(\alpha_j)$ we must solve n independent equations

$$\alpha_j = \frac{\beta h_j}{2}\left(-1 + \sqrt{1 + \frac{4}{\beta h_j \|\hat{\mathbf{w}}\|^2}} \right).$$

3.3. *Linear Model*

In the special case of linear model,

$$\mathbf{f}(\mathbf{w}, \mathbf{X}) = \mathbf{X}\mathbf{w},$$

we obtain an explicit form for the part of optimization procedure. In particular, the integral of the error function exponent function equals

$$\int \exp\left(-S(\mathbf{w})\right) d\mathbf{w} = S(\hat{\mathbf{w}})(2\pi)^{\frac{n}{2}} \left(\det \mathbf{H}^{-1}\right)^{\frac{1}{2}},$$

where $\hat{\mathbf{w}}$ is a unique global minimum of the unimodal error function $S(\mathbf{w})$, and the Hessian matrix equals to

$$\mathbf{H} = \mathbf{A} + \beta \mathbf{X}^T \mathbf{X}.$$

In this case, the most probable parameters,

$$\hat{\mathbf{w}} = \arg\max p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}),$$

equal

$$\hat{\mathbf{w}} = \left(\mathbf{A} + \beta \mathbf{X}^T \mathbf{X}\right)^{-1} \beta \mathbf{X}^T \mathbf{y}.$$

In particular, for the case of diagonal matrix $\mathbf{A} = \mathrm{diag}(\alpha_j)$ we write explicit estimations of the structure parameters:

$$\beta = \frac{m - \rho}{\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|^2},$$

where

$$\rho = \sum_{j=1}^{n} \frac{\beta h_j}{\alpha_j + \beta h_j},$$

and

$$\alpha_j = \frac{\beta h_j}{2}\left(-1 + \sqrt{1 + \frac{4}{\beta h_j \|\hat{\mathbf{w}}\|^2}} \right),$$

where h_j is a j-th eigenvalue of the matrix $\mathbf{X}^T\mathbf{X}$.

3.4. *Hessian Computation*

In the non-linear case we must apply a numerical method to determine the Hessian values. To do this we use a method of approximation of the error function second derivatives with the finite differences. The element h_{jk} of the Hessian $\mathbf{H}$ at $\mathbf{w} = \hat{\mathbf{w}}$ can be computed as

$$h_{jk} = \frac{\partial^2 S}{\partial w_j \partial w_k} = \frac{S(\hat{\mathbf{w}} + (\mathbf{e}_j + \mathbf{e}_k)r) - S(\hat{\mathbf{w}} + \mathbf{e}_j r) - S(\hat{\mathbf{w}} + \mathbf{e}_k r) + S(\hat{\mathbf{w}})}{r^2},$$

where $\mathbf{e}_j$, $\mathbf{e}_k$ are unit vectors, and r is a small parameter. An error of this method is of the order $O(r)$. The method requires computation of the error function in the $\frac{n(n+1)}{2}$ points and is computationally efficient.

4. Monte-Carlo Approximation Method

From the Bayesian inference it follows that to estimate structure parameters we must maximize integral

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{M}^n, \mathbf{B}\in\mathbb{M}^m}{\arg\max} \left(\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w}, \mathbf{B})\, p(\mathbf{w}|\mathbf{A}) d\mathbf{w} \right). \tag{15}$$

In this section let $\mathbf{A}$ be the matrix inverse to the covariance matrix $\mathbf{\Sigma}$ of the random vector $\mathbf{w}$, $\mathbf{A} = \mathbf{\Sigma}^{-1}$. Without loss of generality it can be assumed that $\mathsf{E}(\mathbf{w}) = \mathbf{0}$. This generalizes the hypothesis of the normal distribution of the parameters vector $\mathbf{w}$, given in the previous section.

Let us remark that under this conditions $\mathbf{A}^{-1}$ is a Gramian matrix of Euclidean space of random vectors $\mathbf{w}$. Since matrix $\mathbf{A}$ is a positive definite matrix it follows that matrix $\mathbf{A}$ has a unique Cholesky decomposition (Chang and Tsay, 2010),

$$\mathbf{A}^{-1} = \mathbf{R}^T\mathbf{R}, \quad \mathbf{R} \in \mathbb{U}^n, \tag{16}$$

where $\mathbf{R}$ is an upper triangular matrix with positive diagonal elements, and $\mathbb{U}^n$ is the set of all those matrices of the size $n \times n$. Note that $\mathbf{R}$ is a transformation matrix from Euclidean space of random vectors $\mathbf{w}$ with the Gramian matrix $\mathbf{\Sigma}^0 = \mathbf{I}$ to Euclidean space of vectors $\mathbf{w}$ with the Gramian matrix $\mathbf{\Sigma}$.

Since the Cholesky decomposition (Chang and Tsay, 2010) is unique for the matrix $\mathbf{A}$, let us find the optimal solution of (15) as follows,

$$(\hat{\mathbf{R}}, \hat{\mathbf{B}}) = \underset{\mathbf{R}\in\mathbb{U}^n, \mathbf{B}\in\mathbb{M}^m}{\arg\max} \left(\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w}, \mathbf{B})\, p(\mathbf{w}|\mathbf{R})\, d\mathbf{w} \right).$$

In this section let matrix $\mathbf{B}$ be constant, $\mathbf{B} = \mathbf{B}^0$. Then the optimization problem (15) will be reformulated as follows,

$$\hat{\mathbf{R}} = \underset{\mathbf{R}\in\mathbb{U}^n}{\arg\max} \left(\int\limits_{\mathbf{w}\in\mathbb{W}} p\big(\mathfrak{D}|\mathbf{w}, \mathbf{B}^0\big)\, p(\mathbf{w}|\mathbf{R})\, d\mathbf{w} \right). \tag{17}$$

Since the integral (17) cannot be computed analytically, we use a stochastic method of the integration over the parameter space $\mathbb{W}$. Note that the expression (17) equals the expected value of the likelihood,

$$\int\limits_{\mathbf{w}\in\mathbb{W}} p\big(\mathfrak{D}|\mathbf{w}, \mathbf{B}^0\big)\, p(\mathbf{w}|\mathbf{R})\, d\mathbf{w} = \mathsf{E}\big(p\big(\mathfrak{D}|\mathbf{w}, \mathbf{B}^0\big)\big),$$

and according to the law of large numbers

$$\int\limits_{\mathbf{w}\in\mathbb{W}} p\big(\mathfrak{D}|\mathbf{w}, \mathbf{B}^0\big)\, p(\mathbf{w}|\mathbf{R})\, d\mathbf{w} \approx \frac{1}{K} \sum_{\mathbf{w}\in\mathcal{W}(\mathbf{R})} p\big(\mathfrak{D}|\mathbf{w}, \mathbf{B}^0\big),$$

where $\mathcal{W}(\mathbf{R})$ is a set of vectors $\mathbf{w}$ with the covariance matrix $\mathbf{R}^T\mathbf{R}$. The set $\mathcal{W}(\mathbf{R})$ of cardinality K can be computed through sampling.

Denote by $\mathcal{E}(\mathbf{R})$ a model evidence approximation that should be maximized over $\mathbf{R}$,

$$\hat{\mathbf{R}} = \underset{\mathbf{R}\in\mathbb{U}^n}{\arg\max} \big(\mathcal{E}(\mathbf{R})\big),$$

$$\mathcal{E}(\mathbf{R}) \approx \frac{1}{K} \sum_{\mathbf{w} \in \mathcal{W}(\mathbf{R})} p\big(\mathfrak{D}\big|\mathbf{w}, \mathbf{B}^0\big). \tag{18}$$

To estimate the optimal parameters $\mathbf{R}$ of the optimization problem (18) it is necessary to perform the sampling procedure of the parameters $\mathcal{W}(\mathbf{R})$ for each $\mathbf{R}$. However, it is readily seen that the matrix $\mathbf{R}$ is the transformation matrix for the map from Euclidean space with the Gramian matrix $\mathbf{I}$ to Euclidean space with the Gramian matrix $\mathbf{R}^T\mathbf{R}$.

This means that it is sufficient to perform the sampling procedure once before optimization algorithm starts. Doing this we obtain the set

$$\mathcal{W}^0 = \mathcal{W}(\mathbf{I}) = \big\{\mathbf{w}^0\big|\mathbf{w}^0 \sim p\big(\mathbf{w}^0\big|\mathbf{I}\big)\big\}.$$

Then we compute the set $\mathcal{W}(\mathbf{R})$ on each iteration of the algorithm by rescaling the set $\mathcal{W}^0$:

$$\mathcal{W}(\mathbf{R}) = \big\{\mathbf{R}^T\mathbf{w}^0\big|\mathbf{w}^0 \in \mathcal{W}^0\big\}.$$

4.1. *Metropolis-Hastings Sampling Algorithm*

To generate the sample $\mathcal{W}^0 = \{\mathbf{w}|\mathbf{w} \sim p(\mathbf{w}|\mathbf{I})\}$ we use the Metropolis-Hastings algorithm.

The basic idea of the algorithm is to generate a sample constituting a Markov chain. Each element $\mathbf{w}_{t+1}$ of the sample correlates only with the previous element $\mathbf{w}_t$ of the sample.

Denote by $Q(\mathbf{w}|\mathbf{w}')$ an auxiliary distribution $Q(\mathbf{w}|\mathbf{w}')$, choose an initial element $\mathbf{w}_0$ and assign $\mathcal{W}^0 = \{\mathbf{w}_0\}$. Then let an element $\mathbf{w}_t$ be chosen according to the distribution $Q(\mathbf{w}'|\mathbf{w}_t)$. The next element $\mathbf{w}'$ is generated randomly. Then the algorithm computes the acceptance ratio a:

$$a = \min_{\mathbf{w}' \in \mathbb{R}^n} \left(\frac{p(\mathfrak{D}|\mathbf{w}', \mathbf{B}^0)\, Q(\mathbf{w}_t|\mathbf{w}')}{p(\mathfrak{D}|\mathbf{w}_t, \mathbf{B}^0)\, Q(\mathbf{w}'|\mathbf{w}_t)}, 1 \right).$$

The algorithm accepts the element $\mathbf{w}'$ with probability a, $\mathbf{w}_{t+1} = \mathbf{w}'$, $\mathcal{W}^0 := \mathcal{W}^0 \cup \mathbf{w}'$. Otherwise, the algorithm rejects the candidate and puts $\mathbf{w}_{t+1} = \mathbf{w}_t$.

$$\mathbf{w}_{t+1} = \begin{cases} \mathbf{w}', & \text{with probability } a, \\ \mathbf{w}_t, & \text{with probability } 1 - a. \end{cases}$$

Let the auxiliary distribution $Q(\mathbf{w}|\mathbf{w}')$ be normal:

$$Q\big(\mathbf{w}|\mathbf{w}'\big) = Q\big(\mathbf{w}'|\mathbf{w}\big) = \frac{1}{(2\pi\alpha^{-1})^{\frac{n}{2}}} \exp\left(-\frac{\alpha}{2}\big(\mathbf{w} - \mathbf{w}'\big)^T\big(\mathbf{w} - \mathbf{w}'\big) \right).$$

That is, the function $Q(\mathbf{w}|\mathbf{w}')$ is symmetric and

$$a = \frac{p(\mathfrak{D}|\mathbf{w}', \mathbf{B}^0)}{p(\mathfrak{D}|\mathbf{w}_t, \mathbf{B}^0)}.$$

The initial element $\mathbf{w}_0$ is chosen randomly from the distribution $P(\mathbf{w}|\mathbf{I})$.

5. Cross-Validation Estimation Method

The cross-validation method assumes realizations of the random vector $\mathbf{w}$ to be defined by the regression sample elements. Each realization is the optimal value of the parameters vector $\mathbf{w}$ on the corresponding subsample. We estimate the expected loss,

$$L(\mathbf{w}) = \mathsf{E}_{\mathfrak{D}}\big(S_{\mathfrak{D}}(\mathbf{w})\big),$$

where

$$S_{\mathfrak{D}}(\mathbf{w}) = \frac{1}{2}(\mathbf{y}-\mathbf{f})^T(\mathbf{y}-\mathbf{f}),$$

according to (9). Note that the function $S_{\mathfrak{D}}(\mathbf{w})$ is the part of the first summand of the error function $S(\mathbf{w})$ in (8):

$$S(\mathbf{w}) = \beta S_{\mathfrak{D}}(\mathbf{w}) + \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w},$$

where the second summand $\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}$ corresponds to the prior distribution of the model parameters $\mathbf{w}$.

According to Hastie *et al*. (2009) we split the sample $\mathfrak{D}$ into Q roughly equal-sized parts to estimate the expected loss $L(\mathbf{w})$,

$$\mathfrak{D} = \mathfrak{D}_1^{l_1} \sqcup \cdots \sqcup \mathfrak{D}_Q^{l_Q}.$$

By $\hat{\mathbf{w}}_{\mathfrak{D}\setminus\mathfrak{D}_q}(\mathbf{A})$ denote an estimation of the parameters vector $\mathbf{w}$ such that $\hat{\mathbf{w}}$ minimizes the error function (8) over the subsample $\mathfrak{D}\setminus\mathfrak{D}_q$ for the constant matrix $\mathbf{A}$. We minimize the expected loss estimation (CV — Cross-Validation),

$$\hat{\mathbf{A}} = \arg\max_{\mathbf{A}\in\mathbb{M}^n}\big(\mathrm{CV}(\mathfrak{D},\mathbf{A})\big),$$

$$\mathrm{CV}(\mathfrak{D},\mathbf{A}) = \frac{1}{m}\sum_{i=1}^{m} S_{\mathfrak{D}_q}\big(\hat{\mathbf{w}}_{\mathfrak{D}\setminus\mathfrak{D}_q}(\mathbf{A})\big),$$

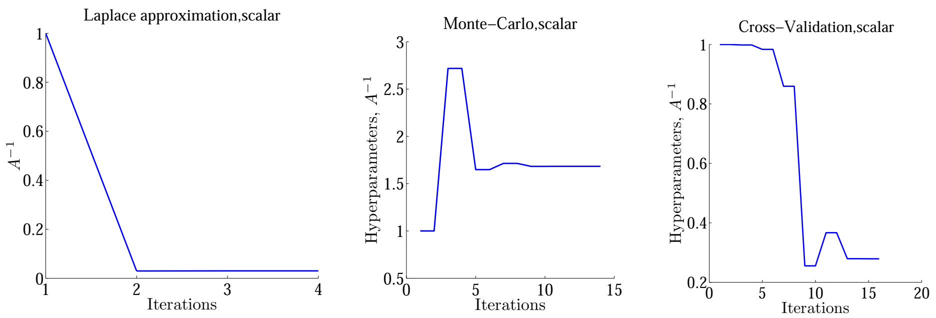
where $S_{\mathfrak{D}_q}(\hat{\mathbf{w}}_{\mathfrak{D}\setminus\mathfrak{D}_q}(\mathbf{A}))$ is estimated on the validation subsample $\mathfrak{D}_q$ with the parameter vector $\hat{\mathbf{w}}$ such that $\hat{\mathbf{w}}$ is estimated on the learn subsample $\mathfrak{D}\setminus\mathfrak{D}_q$. Note that the matrix $\mathbf{B}$ is fixed, $\mathbf{B} = \mathbf{B}_0$, and the algorithm computes only an estimation of the matrix $\mathbf{A}$.

6. Computational Experiment

The proposed algorithms were tested on synthetic and real data. Figures below illustrate convergence of the structure parameters estimations $\hat{\mathbf{w}}$, $\hat{\mathbf{A}}$. The results are compared with the true values $\mathbf{w}^*$, $\mathbf{A}^*$.

Table 2
Error analysis: estimations relative bias.

	Scalar		Diag		Full	
	$\frac{\|\hat{\mathbf{w}}-\mathbf{w}^*\|}{\|\mathbf{w}^*\|}$	$\frac{\|\hat{\mathbf{A}}-\mathbf{A}^*\|}{\|\mathbf{A}^*\|}$	$\frac{\|\hat{\mathbf{w}}-\mathbf{w}^*\|}{\|\mathbf{w}^*\|}$	$\frac{\|\hat{\mathbf{A}}-\mathbf{A}^*\|}{\|\mathbf{A}^*\|}$	$\frac{\|\hat{\mathbf{w}}-\mathbf{w}^*\|}{\|\mathbf{w}^*\|}$	$\frac{\|\hat{\mathbf{A}}-\mathbf{A}^*\|}{\|\mathbf{A}^*\|}$
OLS	0.3	–	0.67	–	0.37	–
LA	0.095	**0.14**	0.54	1.09	–	–
MK	0.078	0.16	**0.52**	**0.36**	**0.34**	0.57
CV	**0.041**	0.39	0.53	0.42	0.36	**0.55**



(a) Laplace approximation (b) Monte-Carlo (c) Cross validation

Fig. 1. Structural parameters convergence for the scalar matrix $\mathbf{A}^{-1}$, $\mathbf{A} = \alpha\mathbf{I}$.

Consider the sample set generated by the linear polynomial model

$$y = \sum_{j=0}^{n} w_j x^j + \varepsilon,$$

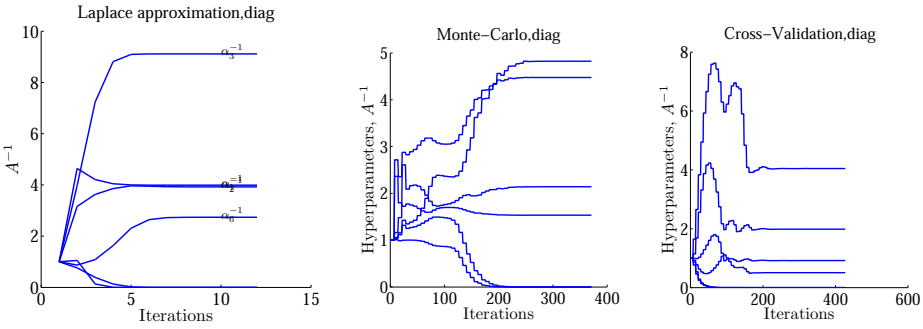
where

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{*-1}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{B}^{*-1}) = \mathcal{N}\left(\mathbf{0}, \frac{1}{\beta^*}\mathbf{I}\right).$$

The proposed algorithms estimated the matrix $\hat{\mathbf{A}}$ and the corresponding optimal parameters vector $\hat{\mathbf{w}}$. The Laplace approximation also estimated the matrix $\hat{\mathbf{B}}$.

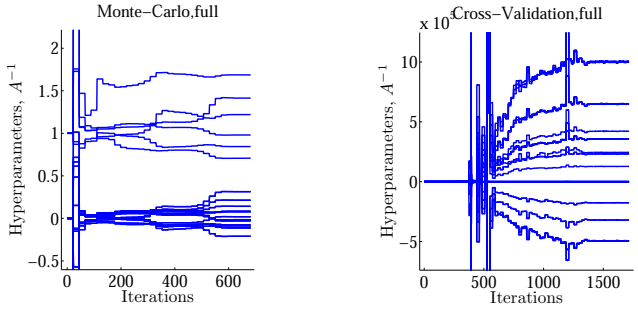
Table 2 shows the results; here $\frac{\|\hat{\mathbf{w}}-\mathbf{w}^*\|}{\|\mathbf{w}^*\|}$ is a norm of the estimation relative bias from the parameters ground truth. Similarly, $\frac{\|\hat{\mathbf{A}}-\mathbf{A}^*\|}{\|\mathbf{A}^*\|}$ is a norm of the estimation relative bias from the structure parameters $\mathbf{A}^*$ ground truth. The first row of the table shows the results for ordinal least squares method of the parameters estimation. The best fitted parameters are marked bold. The table shows that algorithms return comparable results.

Figures 1, 2, 3 illustrate iterative parameters convergence for the real data. The real data are the bread prices data with time as the independent variable and price as the dependent variable. Additional columns of the matrix $\mathbf{X}$ are polynoms of the time variable. Figure 1 illustrates convergence for scalar type of the matrix $\mathbf{A}$. Figure 2 illustrates diagonal type



(a) Laplace approximation (b) Monte-Carlo (c) Cross validation

Fig. 2. Structural parameters convergence for the diagonal matrix $\mathbf{A}^{-1}$, $\mathbf{A} = \alpha\mathbf{I}$.



(a) Monte-Carlo (b) Cross validation

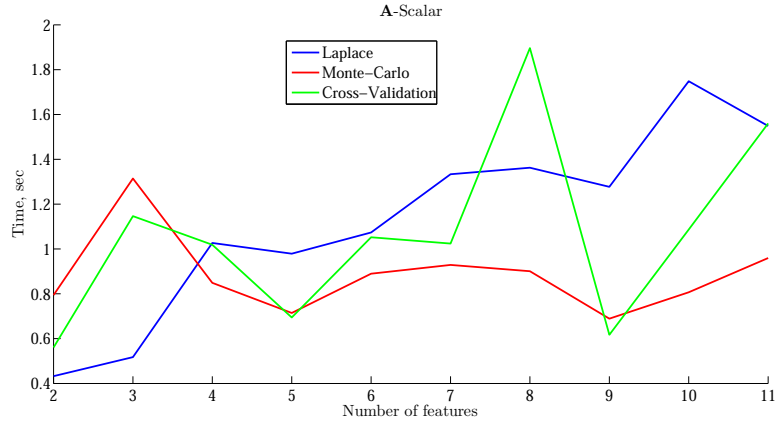
Fig. 3. Structural parameters convergence for the full matrix $\mathbf{A}^{-1}$, $\mathbf{A} = \alpha\mathbf{I}$.

and Fig. 3 illustrates full type of the matrix $\mathbf{A}$. x-axis shows iterations number, y-axis shows value of the elements of the matrix $\mathbf{A}$.

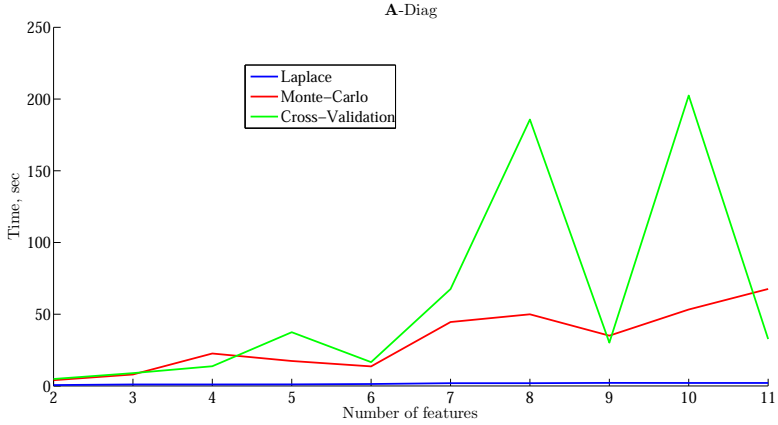
Figure 1 shows that in the scalar case convergence appears after 10–20 iterations. Algorithms need more iterations for the diagonal (Fig. 2) and for the full (Fig. 3) cases. Figure 2 shows zero diagonal elements of the matrix $\hat{\mathbf{A}}^{-1}$. The zero element follows that the corresponding feature is non-informative due to the large penalty in the error function. All three algorithms show that two features (fourth and fifth polynomial degrees) are non-informative.

Figure 4 shows computational time of the algorithms. x-axis shows maximum polynomial degree which grows from 2 to 11. Size of the generated sample equals 400. The cardinality of the set $|\mathcal{W}(\mathbf{R})|$ equals 1000, blocks number Q for cross-validation equals 100.

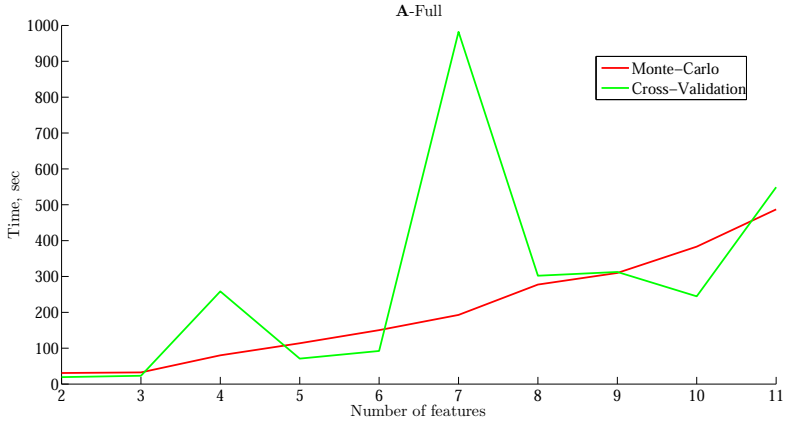
Figure 4 shows that computational times of the algorithms are comparable in the case of the scalar matrix $\mathbf{A}^* = \alpha\mathbf{I}$ since there is only one parameter α for optimization. Figure 4 shows that the Laplace approximation method works much more faster because it solves n independent equations. Note that computational time function is not monotonic for the Monte-Carlo and cross-validation algorithms due to randomization of the initial values.



(a) Scalar matrix **A**



(b) Diagonal matrix **A**



(c) Full matrix **A**

Fig. 4. Computational time.

7. Summary and Discussion

In this study we presented the algorithms of the regression model structure parameters estimation. We proposed the method of the approximated model evidence optimization to estimate the distribution covariance. To obtain exact computation formulas we use the Laplace approximation of the evidence. Furthermore, we proposed the Monte-Carlo method for model evidence approximation and cross-validation method for the model parameters estimation to estimate the covariance. The computational experiment section provides the analysis of the methods convergence and computational time comparison. To illustrate the results we used simulated and real data.

The proposed theoretical study and experimental framework allow us to highlight the pros and cons of the proposed methods. First, due to the exact derivative computation formulas the Laplace approximation method outperforms other methods by both accuracy and time complexity. However, it has a limited scope: at present we have derived the exact formulas only for the scalar and diagonal cases of the prior distribution covariance. The full-covariance case must involve the development of new numerical methods to solve the system of equations for the Hessian computation that is a promising further study. Second, we have discovered that in the full-covariance case the Monte-Carlo approach works better. The weights rescaling trick allowed us to avoid weights resampling and to make the method more accurate and significantly more faster. In further study we are going to investigate the multistart optimization strategy using the different initial simulations for the Monte-Carlo method to improve its stability.

Another further direction is the investigation of non-Gaussian distributions of a target variable and model parameters. In particular, the binomial distribution of a target leads to a classification problem. The analysis of hyperparameters for classification was explored by Bishop (2006) but requires a significant development. In practical point of view, the structure parameter analysis for classification can be very useful for the credit scoring problem that requires a profound analysis of model structure. Besides that, the expansion of the prior distribution class (e.g. using the Laplace-kind distributions) leads to the Bayesian elastic net problem (Zou and Hastie, 2005), a state-of-the-art variable selection method. The target model evidence becomes non-differentiable for this case, and the development of numerical methods for the evidence estimation is a great promise.

A developing alternative to the proposed parametric regression is the non-parametric CMARS-like methods (Weber *et al*., 2012). The important problem for those methods is to choose the minimum number of splines to reduce a model complexity. The proposed methodology can be expanded to estimate the general form of evidence for the non-parametric problems.

The developed tool is used by the authors as a part of the data analysis framework for the several real-world applications. First, the evidence estimation is used for feature selection in the option volatility forecasting problem (Strijov and Weber, 2010). Second, the evidence maximization helps to select an accurate robust model for the credit scoring (Motrenko *et al*., 2014). In addition, the authors are going to apply the framework in the human behaviour recognition problem to select the minimum set of features encoding each type of behaviour.

Acknowledgements. This project was supported by the Ministry of Education and Science of the Russian Federation, RFMEFI60414X0041.

References

Alessandri, A., Cervellera, C., Maccio, D., Sanguineti, M. (2010). Optimization based on quasi-Monte Carlo sampling to design state estimators for non-linear systems. *Optimization*, 59, 963–984.

An, S., Liu, W., Venkatesh, S. (2007). Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognition*, 40, 2154–2162.

Betro, B., Vercellis, C. (2007). Bayesian nonparametric inference and monte carlo optimization. *Optimization*, 17, 681–694.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bishop, C.M., Tipping, M.E. (2003). Bayesian regression and classification. *Advances in Learning Theory: Methods, Models and Applications*, 190, 267–285.

Cawley, G.C., Talbot, N.L.C.. (2007). Preventing over-fitting during model selection via bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research*, 8, 841–861.

Chang, C., Tsay, R.S. (2010). Estimation of covariance matrix via the sparse cholesky factor with lasso. *Journal of Statistical Planning and Inference*, 140, 3858–3873.

Draper, N.R., Smith, H. (1998). *Applied Regression Analysis*. Wiley, New York.

Eidsvik, J., Finley, A.O., Banerjee, S., Ru, H. (2011). Approximate bayesian inference for large spatial datasets using predictive process models. *Computational Statistics & Data Analysis*, 1362–1380.

Gillard, J. (2010). Asymptotic variance-covariance matrices for the linear structural model. *Statistical Methodology*, 8, 291–301.

Hastie, T., Tibshirani, R., Firedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second ed. Springer, New York.

Hu, H. (2005). Ridge estimation of a semiparametric regression model. *Journal of Computational and Applied Mathematics*, 176, 215–222.

Kutner, M.H., Nachtsheim, C.J., Neter, J. (2004). *Applied Linear Regression Models*. McGraw-Hill/Irwin Series Operations and Decision Sciences.

Lampe, J., Voss, H. (2013). Large-scale tikhonov regularization of total least squares. *Journal of Computational and Applied Mathematics*, 238, 95–108.

Liang, C.F., Yang, Y.F. (2009). A bayesian inference approach to identify a robin coefficient in one-dimensional parabolic problems. *Journal of Computational and Applied Mathematics*, 231, 840–850.

Lopes, H.F., Moreirac, A.R.B., Schmidt, A.M. (1999). Hyperparameter estimation in forecast models. *Computational Statistics & Data Analysis*, 29, 387–410.

Mackay, D.J.C. (1998). Choice of basis for laplace approximation. *Machine Learning*, 33, 77–86.

McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models*, Chapman and Hall.

Motrenko, A., Strijov, V., Weber, G.-W. (2014). Weber Bayesian sample size estimation for logistic regression. *Journal of Computational and Applied Mathematics*, 255, 743–752.

Packalen, M., Wirjanto, T.S. (2012). Inference about clustering and parametric assumptions in covariance matrix estimation. *Computational Statistics & Data Analysis*, 56, 1–14.

Strijov, V., Weber, G.W. (2010). Nonlinear regression model generation using hyperparameter optimization. *Computers & Mathematics with Applications*, 60, 981–988.

Weber, G.W., Batmaz, I., Koksal, G., Taylan, P., Yerlikaya-Ozkurt, F. (2012). CMARS: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. *Inverse Problems in Science and Engineering*, 20(3), 371–400.

Wood, A.T.A., Butler, R.W. (2003). Laplace approximation for bessel functions of matrix argument. *Journal of Computational and Applied Mathematics*, 155, 359–382.

Zellner, A. (1995). Bayesian and non-bayesian approaches to statistical inference and decision-making. *Journal of Computational and Applied Mathematics*, 64, 3–10.

Zhang, Y., Leithead, W.E. (2005). Exploiting hessian matrix and trust-region algorithm in hyperparameters estimation of gaussian process. *Applied Mathematics and Computation*, 171, 1264–1281.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

M.P. Kuznetsov was awarded the master of applied mathematical sciences degree at Moscow Institute of Physics and Technology in 2013 (Faculty of Department of Control and Applied Mathematics). He is a PhD student of Moscow Institute of Physics and Technology at the Intelligent Systems department. His research interests include the problems of machine learning, data analysis and statistics.

A.A. Tokmakova was awarded the master of mathematical sciences degree at Moscow Institute of Physics and Technology in 2014. At present she is a developer at Yandex. Her research interests are the problems of multivariate statistics and machine learning.

V.V. Strijov was awarded the candidate in 2002 and doctor of physics and mathematics in 2014 at Computing Center of the Russian Academy of Sciences. At present he is a principal investigator at the Computing Center of RAS. His research interests are machine learning, data analysis, time series forecasting and algebraic approach to pattern recognition.

Analitiniai ir stochastiniai struktūrų parametrų vertinimo metodai

Mikhail KUZNETSOV, Aleksandra TOKMAKOVA, Vadim STRIJOV

Straipsnyje pateikiami analitiniai ir stochastiniai struktūrų parametrų vertinimo metodai, taikomi spręsti modelio parinkimo problemai. Struktūros parametrais laikomos tiesinių ir netiesinių regresijų modelių kovariacijų matricos. Modelio struktūra ir parametrai optimizuojami maksimizuojant modelio akivaizdumą, išreiškiamą per duomenų tikėtinumo sąsuką su modelio parametrų apriorinių skirstiniu. Analitiniai metodai grindžiami aproksimuoto akivaizdumo išvestinių apskaičiavimu. Stochastiniai metodai konstruojami per modelio parametrų imitavimą ir duomenų kryžminį tikrinimą. Pasiūlyti metodai yra testuojami ir lyginami su imituotais bei realiais duomenimis.