Kuznetsov M.P., Tokmakova A.A., Strijov V.V. Analytic and stochastic methods of structure parameter estimation // Informatica, 2016, 607-624.

Structure parameter estimation algorithms for model selection[☆],[☆☆]

V. V. Strijov[a,**], M. P. Kuznetsov[b,*], A. A. Tokmakova[b]

[a]*Dorodnicyn Computing Center of Russian Academy of Sciences, Vavilov st. 40, 119333 Moscow, Russia*
[b]*Moscow Institute of Physics and Technology, Institutskiy lane 9, Dolgoprudny city, Moscow region, 141700, Russia*

Abstract

This paper presents deterministic and stochastic algorithms of the structure parameters estimation for the model selection problem. Structure parameters optimization for linear and non-linear models is investigated. The optimized error function is inferred from statistical hypothesis on the model parameter distributions. Analytic algorithms are based on the error function derivatives estimation with respect to the model parameters. Stochastic algorithms are based on the model parameters sampling and on the data cross-validation. The algorithms are tested and compared on model and real data.

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

1. Introduction

The model complexity estimation is an important problem of model selection. The problem is to find a regression function [1, 2, 3, 4] modeling measured data and to estimate regression model parameters [5]. The measured data are dependent and independent variable measurements.

To estimate model parameters one must optimize the error function over the set of parameters [6, 7]. The error function is inferred from some algebraic or statistic approaches. This paper considers the statistical approach of the data generation.

According to this approach, the dependent variable and the model parameters are considered to be random values and identified by their probability

[☆] This document is a collaborative effort.

[☆☆] This project was supported by the Russian Foundation for Basic Research, grants 12-07-31095, 13-07-00709.

[*] Corresponding author

[**] Principal corresponding author

Email addresses: strijov@ccas.ru (V. V. Strijov), mikhail.kuznecov@phystech.edu (M. P. Kuznetsov), aleksandra-tok@yandex.ru (A. A. Tokmakova)

distribution functions. In this case, an error function is a form of the likelihood [8, 9] which should be optimized.

To optimize the error function we use the Bayesian model comparison method [10, 1, 11]. According to this method, the error function contains so-called structure parameters which indicate model complexity. The error function should be optimized over both set of parameters and set of structure parameters to find the optimal model.

The structure parameters are regularization parameters and penalize elements of model parameters vector [12, 13, 14]. The main goal of this paper is to estimate the structure parameters [15, 16, 17]. To do this we maximize the model evidence [18, 19].

One of the methods of model evidence maximization is the Laplace approximation [20, 21]. The dependent variable and parameters vector are considered to be multivariate normal vectors. Covariation matrices of this vectors are the structure parameters. We propose various estimations of the structure parameters depending on types of the covariance matrices.

An alternative method considered in this paper is the Monte Carlo approximation of the model evidence [22, 23]. The parameters vector is sampled according to the given distribution. We maximize the sum over the set of sampled parameters approximating the model evidence.

To validate the proposed methods we use the cross-validation method of the structure parameters estimation [7, 24]. This method is based on the sample splitting into roughly equal-size parts. The model parameters should be estimated on the each part of the sample.

As a special case we consider linear regression models [5]. For this type of models we derive explicit values of parameters vector and Hessian matrix [25].

2. Structural parameters estimation problem

The measured data consist of measured data of a dependent variable y and an independent variable $\mathbf{x}$. Let this dependence be statistical such that

$$\mathsf{E}(y|\mathbf{x}) = f(\hat{\mathbf{w}}, \mathbf{x}).$$

Denote by $D = \big\{(\mathbf{x}_i, y_i)\big\}_{i=1}^m$ a regression sample set that consists of pairs of the independent vectors $\mathbf{x}_i = [x_{ij}]_{j=1}^n$, $\mathbf{x} \in X \subseteq R^n$ and corresponding values of the dependent variable y_i, $y \in Y \subseteq R^1$.

Let index i of a sample and index j of an independent variable be elements of finite unordered sets $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. By $D = (\mathbf{X}, \mathbf{y})$ denote a set such that $\mathbf{y} = [y_1, \dots, y_m]^\mathsf{T}$ is a vector of dependent variable measurements and $\mathbf{X} = [\mathbf{x}_1^\mathsf{T}, \dots, \mathbf{x}_m^\mathsf{T}]^\mathsf{T}$ is a design $m \times n$ matrix.

Suppose the elements of sample are related by

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

with additive random noise $\varepsilon = \varepsilon(\mathbf{x})$.

Let $f : W \times X \to Y$ be the regression model mapping Cartesian product of model parameters space W and independent variables space X to dependent variables space Y. In other words, the regression model is a map $f : (\mathbf{w}, \mathbf{x}) \mapsto y$, where $\mathbf{w} \in W$ is the parameters vector, $\mathbf{x} \in X$ is the independent variable, and $y \in Y$ is the dependent variable. By

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

denote the vector function

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

2.1. Model evidence

We use the coherent Bayesian inference method to estimate the parameters $\mathbf{w}$ and the structure parameters $\mathbf{A}, \mathbf{B}$ of the model f.

The first level of Bayesian inference estimates $\mathbf{w}$ by maximizing posterior distribution

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

Elements of this equation and the corresponding parameters are as follows.

$p(\mathbf{w}|D, \mathbf{A}, \mathbf{B})$ — parameters posterior distribution,

$\mathbf{w}_{\text{MP}} = \arg \max\limits_{w \in R^n} p(\mathbf{w}|D, \mathbf{A}, \mathbf{B})$ — most probable parameters,

$p(D|\mathbf{w}, \mathbf{B})$ — marginal likelihood function,

$p(\mathbf{w}|\mathbf{A})$ — prior distribution of parameters,

$p(D|\mathbf{A}, \mathbf{B})$ — model evidence.

Matrices $\mathbf{A}$ and $\mathbf{B}$ are called the *structure parameters*. In particular, matrices $\mathbf{A}$ and $\mathbf{B}$ are the parameters of the prior distribution $p(\mathbf{w})$ and the conditional distribution $p(D|\mathbf{w})$, respectively. Below we will consider special types of this distributions. Let us remark that a model type can also be a structure parameter but in this paper we will fix a model type.

The second level of the bayesian inference selects the best model from the set of competitive models F by maximizing a posterior probability

$$p(\mathbf{A}, \mathbf{B}|D)$$

over structure parameters A and B. To do this, we will take into account the Bayes' theorem:

$$p(\mathbf{A}, \mathbf{B}|D) \propto p(D|\mathbf{A}, \mathbf{B})p(\mathbf{A}, \mathbf{B}),$$

where $p(D|\mathbf{A}, \mathbf{B})$ is called the model evidence, and $p(\mathbf{A}, \mathbf{B})$ — the prior distribution over the set of models. In this paper we will consider uniform distribution over the set of models, i. e.

$$p(\mathbf{A}, \mathbf{B}|D) \propto p(D|\mathbf{A}, \mathbf{B}).$$

Table 1: Data generation hypothesis: dependent variable $\mathbf{y}$ and model parameters $\mathbf{w}$.

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

Therefore we must maximize the model evidence to estimate the structure parameters $\mathbf{A}, \mathbf{B}$

$$p\left(D|\mathbf{A}, \mathbf{B}\right) = \int\limits_{\mathbf{w}\in W} p(D|\mathbf{w}, \mathbf{B})p(\mathbf{w}|\mathbf{A})d\mathbf{w} \to \max_{\mathbf{A}\in M^n, \mathbf{B}\in M^m}, \tag{1}$$

where M^n is the set of positive semi-definite $n \times n$ matrices.

2.2. Data generation hypothesis

Let vectors $\mathbf{y}$ and $\mathbf{w}$ have the multivariate normal distribution with covariance matrices $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$, respectively. To estimate parameters $\mathbf{A}, \mathbf{B}, \mathbf{w}$ let us make some assumptions about distributions $p(D|\mathbf{w}, \mathbf{B})$ and $p(\mathbf{w}|\mathbf{A})$. Table 1 shows various cases of the data generation hypothesis for the dependent variable $\mathbf{y}$ and the model parameters $\mathbf{w}$. We consider matrices $\mathbf{A}$ and $\mathbf{B}$ of a scalar, diagonal, and full type, independently.

The methods considered in this paper allow to estimate structure parameters only for the scalar-type $\mathbf{B} = \beta\mathbf{I}$. Different types of the $\mathbf{A}$ matrix are considered.

3. Laplace approximation method

In this section we use the Laplace approximation of the model evidence to estimate structure parameters $\mathbf{A}, \mathbf{B}$ and model parameters $\mathbf{w}$.

To estimate structure parameters $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ transform the optimization problem (1) according to the data generation hypothesis:

$$\frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})}\frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})} \int\limits_{\mathbf{w}\in W} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y}-\mathbf{f})\right) \exp\left(-\frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w}\right) d\mathbf{w} \to \max_{\mathbf{A}\in M^n, \mathbf{B}\in M^m}. \tag{2}$$

By the error function $S(\mathbf{w}, \mathbf{A}, \mathbf{B})$ denote the exponent of the expression (2) with a negative sign:

$$S(\mathbf{w}, \mathbf{A}, \mathbf{B}) = \frac{1}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y}-\mathbf{f}) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w} \tag{3}$$

and the optimization problem (2) will be as follows:

$$\frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})}\frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})} \int\limits_{\mathbf{w}\in W} \exp\bigl(-S\left(\mathbf{w}, \mathbf{A}, \mathbf{B}\right)\bigr)d\mathbf{w} \to \max_{\mathbf{A}, \mathbf{B}}.$$

Suppose that parameters $\hat{\mathbf{w}}$ maximize the posterior distribution of parameters, or minimize the error function; then this parameters are optimal:

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

where $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ are estimations of the structure parameters maximizing (2).

The Laplace approximation method uses the error function $S(\mathbf{w})$ expansion near the optimal solution $S(\hat{\mathbf{w}})$ to approximate the expression

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

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

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

at $\mathbf{w} = \hat{\mathbf{w}}$. Denote by $\|\mathbf{w}\|$ the Euclidean norm $\|\mathbf{w}\| = \|\mathbf{w}\|_2$. Instead of optimizing (2) let us optimize the approximated function

$$\frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})}\frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})}\exp\bigl(S(\hat{\mathbf{w}})\bigr)\int\limits_{\mathbf{w}\in W}\exp\left(-\frac{1}{2}\Delta\mathbf{w}^{\mathsf{T}}\mathbf{H}\Delta\mathbf{w}\right)d\mathbf{w} \to \max_{\mathbf{A},\mathbf{B}}. \tag{4}$$

Let us remark that the integrand of (4) is a part of the normal distribution, hence we can substitute an integral in (4) for normalization and obtain:

$$g(\mathbf{A},\mathbf{B}) = \frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})}\frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})}\exp\bigl(S(\hat{\mathbf{w}})\bigr)\frac{(2\pi^{\frac{n}{2}})}{|\mathbf{H}|^{\frac{1}{2}}} \to \max_{\mathbf{A},\mathbf{B}}. \tag{5}$$

Taking the logarithm of (5), we obtain the optimization problem:

$$-\ln g(\mathbf{A},\mathbf{B}) = -\frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{A}| + \frac{1}{2}\ln|\mathbf{B}| - S(\mathbf{w}_0) - \frac{1}{2}\ln|\mathbf{H}| \to \max_{\mathbf{A},\mathbf{B}}. \tag{6}$$

Let us to 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 (3) is given by

$$S(\mathbf{w},\mathbf{A},\beta) = \frac{\beta}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}(\mathbf{y}-\mathbf{f}) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w} = \beta S_D(\mathbf{w}) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w}, \tag{7}$$

where

$$S_D(\mathbf{w}) = \frac{1}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}(\mathbf{y}-\mathbf{f}), \tag{8}$$

and Hessian is given by

$$\mathbf{H} = \beta\mathbf{H}_D + A,$$

where $\mathbf{H}_D$ is a Hessian of the function $S_D(\mathbf{w})$ at $\mathbf{w} = \hat{\mathbf{w}}$.

The function (6) is given by

$$-\ln g(\mathbf{A},\beta) = -\frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{A}| + \frac{m}{2}\ln\beta - \frac{\beta}{2}\big(\mathbf{y}-\mathbf{f}\,(\hat{\mathbf{w}},\mathbf{X})\big)^{\mathsf{T}}\big(\mathbf{y}-\mathbf{f}\,(\hat{\mathbf{w}},\mathbf{X})\big) - \\ -\frac{1}{2}\hat{\mathbf{w}}^{\mathsf{T}}\mathbf{A}\hat{\mathbf{w}} - \frac{1}{2}\ln|\beta\mathbf{H}_D + A| \to \max_{\mathbf{A},\mathbf{B}}. \tag{9}$$

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

$$\frac{1}{2}\ln|\beta\mathbf{H}_D + \mathbf{A}| \tag{10}$$

of the function (9).

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 (10) equals

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

where h_j is an eigenvector of $\mathbf{H}_D$.

Equating derivatives of (9) with respect to α and β tending to zero, we will estimate 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

$$\gamma = \beta\sum_{j=1}^{n}\frac{h_j}{\beta h_j + \alpha}, \tag{11}$$

then

$$\alpha = \frac{\gamma}{\|\hat{\mathbf{w}}\|^2}. \tag{12}$$

Similarly, equating a derivative of (9) with respect to β to zero, we obtain

$$\beta = \frac{m-\gamma}{\|\mathbf{y}-\mathbf{f}(\hat{\mathbf{w}},\mathbf{X})\|^2}. \tag{13}$$

Since γ is a function of β, α and optimal model parameters $\hat{\mathbf{w}}$ we solve equations (11), (12) and (13) iteratively for the fixed $\hat{\mathbf{w}}$.

3.2. Diagonal type of matrix $\mathbf{A}$

In the case of diagonal matrix $\mathbf{A} = \mathrm{diag}(\alpha_j)$ the obtained results are comparable with equations (11), (12) and (13). In particular, instead of (11), by definition 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 elements of the matrix $\mathbf{A} = \mathrm{diag}(\alpha_j)$ 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 case

In the case of linear model

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

we can obtain the explicit form of some optimization problems solutions. For example, the integral of the error function exponent function equals

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

where $\hat{\mathbf{w}}$ is a unique global minimum of the unimodal error function $S(\mathbf{w})$. Whereas Hessian

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

In this case, the most probable parameters

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

equal

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

In particular, for the case of diagonal matrix $\mathbf{A} = \mathrm{diag}(\alpha_j)$ we can 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

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

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

4. Hessian computation

In the non-linear case we must apply a numerical method to determine Hessian values. To do this we use a method of approximation of the error function second derivatives with 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, 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.

5. Monte Carlo approximation method

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

$$\int\limits_{\mathbf{w}\in W} p(D|\mathbf{w}, \mathbf{B}) p(\mathbf{w}|\mathbf{A}) d\mathbf{w} \to \max_{\mathbf{A}\in M^n, \mathbf{B}\in M^m}. \tag{14}$$

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 [16]

$$\mathbf{A}^{-1} = \mathbf{R}^{\mathsf{T}}\mathbf{R}; \tag{15}$$

here $\mathbf{R}$ is an upper triangular matrix with positive diagonal elements. 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 [16] is unique for the matrix $\mathbf{A}$, let us find the optimal solution of (14) as

$$\int\limits_{\mathbf{w}\in W} p(D|\mathbf{w}, \mathbf{B}) p(\mathbf{w}|\mathbf{R}) d\mathbf{w} \to \max_{\mathbf{R}, \mathbf{B}}.$$

In this section let matrix $\mathbf{B}$ be constant, $\mathbf{B} = \mathbf{B}^0$. The the optimization problem (14) will be as follows:

$$\int\limits_{\mathbf{w}\in W} p(D|\mathbf{w}, \mathbf{B}^0) p(\mathbf{w}|\mathbf{R}) d\mathbf{w} \to \max_{\mathbf{R}}. \tag{16}$$

Since the integral (16) cannot be computed analytically, we will use a stochastic method of integration over the parameters space W. Note that the expression (16) equals the expected value of the likelihood

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

and according to the law of large numbers

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

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

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

$$\mathcal{E}(\mathbf{R}) \approx \frac{1}{K}\sum_{\mathbf{w}\in\mathcal{W}(\mathbf{R})} p(D|\mathbf{w},\mathbf{B}^0) \to \max_{\mathbf{R}}. \tag{17}$$

To estimate the optimal parameters $\mathbf{R}$ of the optimization problem (17) it is necessary to carry out 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 $\mathbf{R}^{\mathsf{T}}\mathbf{R}$.

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

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

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

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

5.1. Metropolis-Hastings sampling algorithm

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

The basic idea of the algorithm is to generate a sample such that the sample forms 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:

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

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

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

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

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

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

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

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

6. Cross-validation estimation method

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 will estimate the expected loss

$$L(\mathbf{w}) = \mathsf{E}_D\big(S_D(\mathbf{w})\big),$$

where

$$S_D(\mathbf{w}) = \frac{1}{2}(\mathbf{y} - \mathbf{f})^{\mathsf{T}}(\mathbf{y} - \mathbf{f})$$

according to the (8). Note that the function $S_D(\mathbf{w})$ is the part of the first summand of the error function $S(\mathbf{w})$ in (7):

$$S(\mathbf{w}) = \beta S_D(\mathbf{w}) + \frac{1}{2}\mathbf{w}^{\mathsf{T}} A\mathbf{w},$$

where the second summand $\frac{1}{2}\mathbf{w}^{\mathsf{T}} A\mathbf{w}$ is corresponded with the prior distribution of the model parameters $\mathbf{w}$.

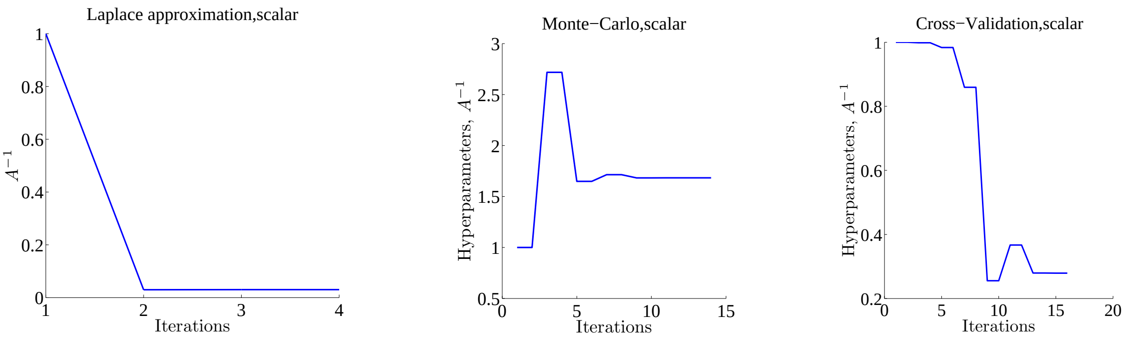
According to [7] we split the sample D into Q roughly equal-sized parts to estimate the expected loss $L(\mathbf{w})$,

$$D = D_1^{l_1} \sqcup \ldots \sqcup D_Q^{l_Q}.$$

By $\hat{\mathbf{w}}_{D\backslash D_q}(\mathbf{A})$ denote an estimation of the parameters vector $\mathbf{w}$ such that $\hat{\mathbf{w}}$ minimizes the error function (7) over the subsample $D\backslash D_q$ for the constant matrix $\mathbf{A}$. We minimize the expected loss estimation (CV — Cross-Validation)

$$\mathrm{CV}(D, \mathbf{A}) = \frac{1}{m}\sum_{i=1}^{m} S_{D_q}(\hat{\mathbf{w}}_{D\backslash D_q}(\mathbf{A})) \to \min_{\mathbf{A}\in M^n},$$

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



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

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

7. Computational experiment

Table 2: Error analysis: estimations relative bias.

	Scalar		Diag		Full	
	$\frac{\lVert\hat{\mathbf{w}}-\mathbf{w}^*\rVert}{\lVert\mathbf{w}^*\rVert}$	$\frac{\lVert\hat{\mathbf{A}}-\mathbf{A}^*\rVert}{\lVert\mathbf{A}^*\rVert}$	$\frac{\lVert\hat{\mathbf{w}}-\mathbf{w}^*\rVert}{\lVert\mathbf{w}^*\rVert}$	$\frac{\lVert\hat{\mathbf{A}}-\mathbf{A}^*\rVert}{\lVert\mathbf{A}^*\rVert}$	$\frac{\lVert\hat{\mathbf{w}}-\mathbf{w}^*\rVert}{\lVert\mathbf{w}^*\rVert}$	$\frac{\lVert\hat{\mathbf{A}}-\mathbf{A}^*\rVert}{\lVert\mathbf{A}^*\rVert}$
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**

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 $\mathbf{w}^*, \mathbf{A}^*$.

Consider the sample set generated by the linear polynomial model

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

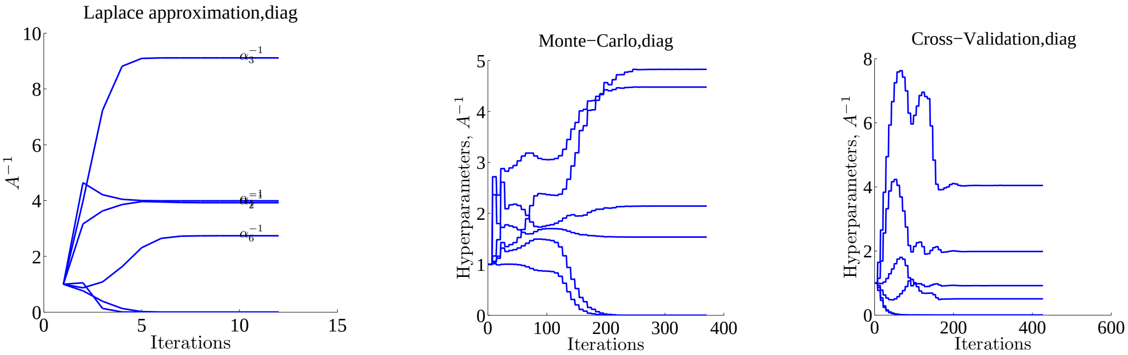
where

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^*), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{B}^*) = \mathcal{N}(\mathbf{0}, \beta^*\mathbf{I});$$

here matrices $\mathbf{A}^*$ and $\mathbf{B}^*$ are given.

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 $\frac{\lVert\hat{\mathbf{w}}-\mathbf{w}^*\rVert}{\lVert\mathbf{w}^*\rVert}$ is a norm of the estimation relative bias from parameters true value. Similarly, $\frac{\lVert\hat{\mathbf{A}}-\mathbf{A}^*\rVert}{\lVert\mathbf{A}^*\rVert}$ is a norm of the estimation relative bias from structure parameters $\mathbf{A}^*$ true value. The first row of the table

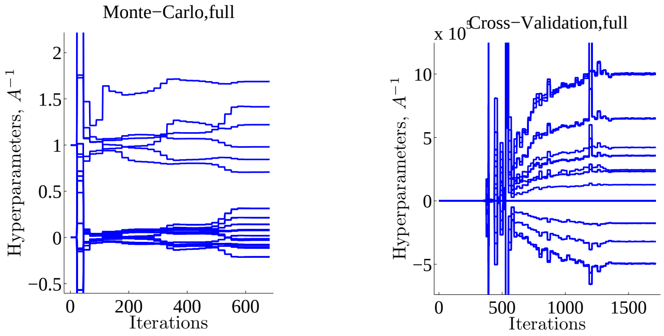


(a) Laplace approximation

(b) Monte-Carlo

(c) Cross validation

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



(a) Monte-Carlo

(b) Cross validation

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

shows the results for ordinal least squares method of the parameters estimation. The best fitted parameters are marked bold. Table shows that algorithms return comparable results.

Figures 1, 2, 3 illustrate iterative parameters convergence for the real data. The real data are 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 and figure 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.

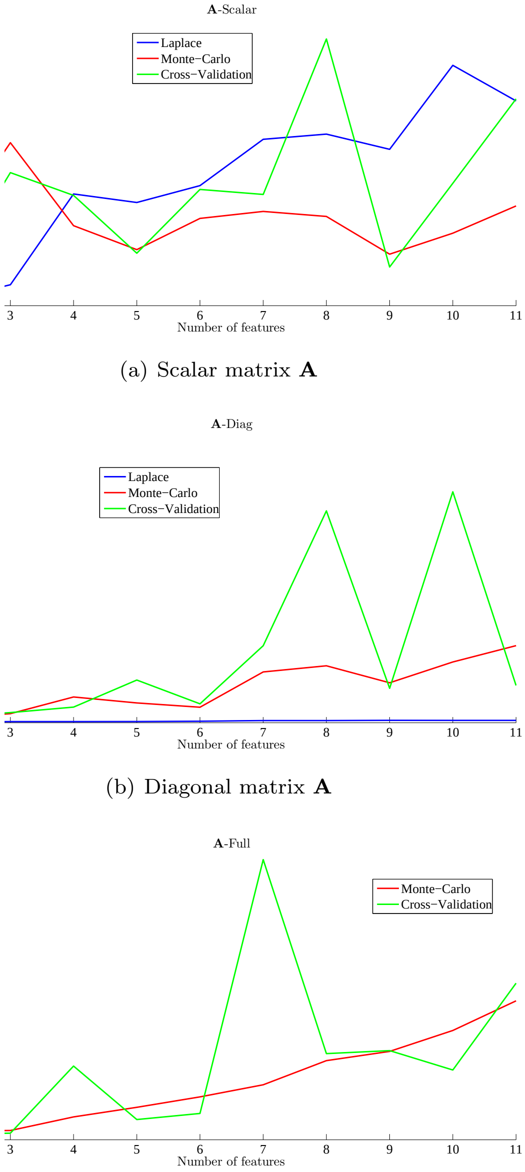
Figure 4 shows that computational times of the algorithms are comparable in the case of the scalar matrix $\mathbf{A}^* = \alpha\mathbf{I}$ since there is only one parameter α for optimization. Figure 4 shows that the Laplace approximation method works much more faster because it solves n independent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

8. Conclusion

In this paper we presented the algorithms of the regression model structure parameters estimation. We proposed the Laplace approximation of the error function for the model evidence estimation to estimate diagonal covariance matrix. Also we proposed the Monte Carlo method for model evidence approximation and cross-validation method for the model parameters estimation to estimate full covariance matrix. The paper illustrates features of this methods: convergence and computational time. Model and real data were used to illustrate the results. The error analysis and proposed methods comparison are shown.

References

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



(a) Scalar matrix **A**

(b) Diagonal matrix **A**

(c) Full matrix **A**

Figure 4: Computational time.

[2] N. R. Draper, H. Smith, Applied regression analysis, John Wiley and Sons, 1998.

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

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

[5] P. McCullagh, J. A. Nelder, Generalized Linear Models, Chapman and Hall, 1989.

[6] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[7] T. Hastie, R. Tibshirani, J. Firedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer, 2009.

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

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

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

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

[12] J. Lampe, H. Voss, Large-scale tikhonov regularization of total least squares, Journal of Computational and Applied Mathematics 238 (2013) 95108.

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

[14] J. Gillard, Asymptotic variancecovariance matrices for the linear structural model, Statistical Methodology 8 (2010) 291–301.

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

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

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

[18] F. Pascal, H. Harari-Kermadec, P. Larzabal, The empirical likelihood method applied to covariance matrix estimation, Signal Processing 90 (2010) 566–578.

[19] T. Ando, R. Tsay, Predictive likelihood for bayesian model selection and averaging, International Journal of Forecasting 26 (2010) 744–763.

[20] A. T. W. Ronald W. Butler, Laplace approximation for bessel functions of matrix argument, Journal of Computational and Applied Mathematics 155 (2003) 359–382.

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

[22] A. Alessandri, C. Cervellera, D. Maccio, M. Sanguineti, Optimization based on quasi-monte carlo sampling to design state estimators for non-linear systems, Optimization 59 (2010) 963–984.

[23] B. Betro, C. Vercellis, Bayesian nonparametrie inference and monte carlo optimization, Optimization 17 (2007) 681–694.

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

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