KYBERNETIKA — VOLUME *55* (2019), NUMBER 1, PAGES 1–11

BAYESIAN STOPPING RULE IN DISCRETE PARAMETER SPACE WITH MULTIPLE LOCAL MAXIMA

Miroslav Kárný

The paper presents the stopping rule for random search for Bayesian model-structure estimation by maximising the likelihood function. The inspected maximisation uses random restarts to cope with local maxima in discrete space. The stopping rule, suitable for any maximisation of this type, exploits the probability of finding global maximum implied by the number of local maxima already found. It stops the search when this probability crosses a given threshold. The inspected case represents an important example of the search in a huge space of hypotheses so common in artificial intelligence, machine learning and computer science.

Keywords: global maximum, model structure, Bayesian estimation

Classification: 62L15,62P99

1. INTRODUCTION

Global optimisation is a demanding research area with many sub-problems [7] and an extreme applicability width. Generally, it tries to find absolute extremes of multi-modal functions. Maximisation of a multi-modal function defined on a huge discrete grid is common to feature extraction, hypotheses testing, structure estimation etc, [4, 13, 15]. In computer science, similar problems are met when constructing hash functions that randomly map a big space into a much smaller space [11]. Size of the grid, the lack of smoothness and function-evaluation costs prevent the exhaustive search and makes the use of Gaussian processes [12] unsuitable.

No free lunch theorem [16] indicates that no universal maximising algorithm can be gained without additional assumptions on the optimised function. Typically, a sort of "continuity", relating mutual positions of function arguments to mutual positions function values, is to be assumed. The motivating structure estimation does not meet such an assumption.

In outlined cases, an initial guess of the maximising argument is chosen and via a local search a local maximum is found and the procedure repeats with another randomly chosen start. Then, the maximum of local maxima is taken as an estimate of the global maximum. Its quality depends on the sufficient number of trials, which have to stop

DOI: 10.14736/kyb-2019-1-0001

well before making an exhaustive search. A rational choice of the number of restarts has to be made. This choice becomes vital if the evaluation of the function values is computationally expensive. The choice is the decision-making task under uncertainty [3] and as such it is addressed here.

The paper shows that a very straightforward use of Bayesian methodology provides an efficient solution. The results concern a specific practically-important problem of the model-structure estimation [8, 10], which fits the above framework and which has lacked a suitable stopping rule. Many other problems are solvable by using the same search strategy with their specific locally-maximising algorithms. Their users and developers may directly adopt the proposed stopping rule.

Section 2 formulates and solves the central stopping problem. The formulation and solution are made in the vein of [9]. Its behavior is illustrated in Section 3 and conclusions drawn in Section 4. An evaluation of a normalisation constant determining the stopping rule is in Appendix 5. Appendix 6 recalls main steps of the motivating structure-estimation algorithm.

Throughout, $\boldsymbol{x}$ is a set of x-values, $|\boldsymbol{x}|$ is cardinality of $\boldsymbol{x}$. Sequences $x(t) = (x_1, \ldots, x_t)$ are inspected in discrete time $\tau, t \in \boldsymbol{t}$, $|\boldsymbol{t}| < \infty$. Time index is the first one and separated by semicolon if it is used together with another subscript. The symbol f denotes probability (density) function (pdf). Its arguments are used for distinguishing them. No formal distinction is made between random variable, its realisation and an argument of a pdf. Context provides the correct meaning. Basic properties of pdfs are only used, see e. g. [14].

2. PROBLEM FORMULATION AND SOLUTION

A huge set of competitive structures[1] $s \in \boldsymbol{s}$, $|\boldsymbol{s}| < \infty$, is inspected. Within the adopted Bayesian paradigm, each structure is characterised by the posterior probability. Due to the impossibility to evaluate its normalising factor the corresponding a posteriori real-valued likelihood function $\mathcal{L}(s)$ is to be considered.

The likelihood function $\mathcal{L}(s)$ has an unknown number $|\boldsymbol{m}| \leq |\boldsymbol{s}| < \infty$ of local maxima. A fixed deterministic algorithm $\mathcal{A}$, as in [8], is at disposal. It assigns to any initial guess $s \in \boldsymbol{s}$ an argument $a = \mathcal{A}(s) \in \boldsymbol{s}$ locally maximising $\mathcal{L}(s)$. In the motivating problem, a local maximum is searched for structures gained by adding to or by removing one entry from to the regressors specified by the initial guess, see [8, 10] and Appendix 6. The global maximum of $\mathcal{L}(s)$, $s \in \boldsymbol{s}$, is searched for, by sequentially evaluating mutually-independent, uniformly-distributed initial guesses $s_t \in \boldsymbol{s}$, $t \in \boldsymbol{t}$. For each $s_t \in \boldsymbol{s}$, the algorithm $\mathcal{A}$ finds the argument $a_t = \mathcal{A}(s_t) \in \boldsymbol{s}$ giving the local maximum of the likelihood function. The *sequential search is stopped when the probability that all local maxima were inspected is high enough*[2]. The design of such a rule is formulated and solved here as the problem of decision making under uncertainty.

[1] For instance, 100 potential regressors creates $|\boldsymbol{s}| = 2^{100}$ hypotheses, which cannot be fully inspected.

[2] This is a generalised form of secretary or marriage problems [5].

2.1. Parametric model

First, a parametric model relating the observed locally-maximising arguments to the unknown $|\boldsymbol{m}|$ is constructed. The domain $\boldsymbol{s}$ splits into $|\boldsymbol{m}|$ mutually disjoint subsets $\boldsymbol{s}_m \subset \boldsymbol{s}$, $m \in \boldsymbol{m} = \{1, \dots, |\boldsymbol{m}|\}$, such that a common mth structure $a_m = \mathcal{A}(s)$ is found for all $s \in \boldsymbol{s}_m$. The probability that the structure a_m is found in tth independent trial is

$$f\left(a_t = a_m|\Theta\right) = f\left(a_t = a_m|\boldsymbol{s}_1, \dots, \boldsymbol{s}_{|\boldsymbol{m}|}, |\boldsymbol{m}|\right) = \frac{|\boldsymbol{s}_m|}{|\boldsymbol{s}|} = \alpha_m^{|\boldsymbol{m}|}. \tag{1}$$

The unknown multivariate parameter is $\Theta = \left(\alpha^{|\boldsymbol{m}|}, |\boldsymbol{m}|\right)$, where

$$\alpha^{|\boldsymbol{m}|} = \left(\alpha_1^{|\boldsymbol{m}|}, \dots, \alpha_{|\boldsymbol{m}|}^{|\boldsymbol{m}|}\right) \in \boldsymbol{\alpha}^{|\boldsymbol{m}|} = \left\{ \alpha_m^{|\boldsymbol{m}|} \geq 0, \quad \sum_{m \in \boldsymbol{m}} \alpha_m^{|\boldsymbol{m}|} = 1 \right\}, \ |\boldsymbol{m}| \in \{1, \dots, |\boldsymbol{s}|\}. \tag{2}$$

Thus, the parametric model relating the observed data $a(t) = (a_1, \dots, a_t)$ to the unknown Θ is

$$f(a(t)|\Theta) = \prod_{\tau \leq t} \prod_{m \in \boldsymbol{m}} \left(\alpha_m^{|\boldsymbol{m}|}\right)^{\delta(a_\tau, a_m)} \chi(|\boldsymbol{m}| - m_\tau) = \prod_{m \in \boldsymbol{m}} \left(\alpha_m^{|\boldsymbol{m}|}\right)^{\tilde{\kappa}_{t;m}} \chi(|\boldsymbol{m}| - m_t), \tag{3}$$

where Kronecker's δ equals one for equal arguments and is zero otherwise. Heaviside's function $\chi(\cdot)$ equals one on non-negative arguments and zero otherwise. The *value* $\tilde{\kappa}_{t;m}$ *counts* how many times the structure a_m occurred during t evaluations, cf. (8). The probability that the *observed number* m_t *of different maxima* is greater than the number of maxima $|\boldsymbol{m}|$ is zero.

2.2. Prior pdf

For a fixed number of local maxima $|\boldsymbol{m}|$, the adopted model belongs to exponential family [2] and thus it has conjugated prior pdf. It is Dirichlet's pdf[3]

$$\begin{aligned} f\left(\alpha^{|\boldsymbol{m}|}||\boldsymbol{m}|\right) &= \mathcal{D}_{\alpha^{|\boldsymbol{m}|}}\left(\kappa_0^{|\boldsymbol{m}|}\right) = \frac{\prod_{m \in \boldsymbol{m}} \left(\alpha_m^{|\boldsymbol{m}|}\right)^{\kappa_{0;m}^{|\boldsymbol{m}|} - 1}}{\mathcal{B}\left(\kappa_0^{|\boldsymbol{m}|}\right)} \\ \kappa_0^{|\boldsymbol{m}|} &= \left[\kappa_{0;1}^{|\boldsymbol{m}|}, \dots, \kappa_{0;|\boldsymbol{m}|}^{|\boldsymbol{m}|}\right], \quad \kappa_{0;m}^{|\boldsymbol{m}|} > 0, \quad \mathcal{B}\left(\kappa_0^{|\boldsymbol{m}|}\right) = \frac{\prod_{m \in \boldsymbol{m}} \Gamma\left(\kappa_{0;m}^{|\boldsymbol{m}|}\right)}{\Gamma\left(\sum_{m \in \boldsymbol{m}} \kappa_{0;m}^{|\boldsymbol{m}|}\right)}. \end{aligned} \tag{4}$$

It remains to choose marginal prior probability of $|\boldsymbol{m}|$. The uniform pdf on $|\boldsymbol{m}| = \{1, \dots, |\boldsymbol{s}|\}$ is the most simple option ($\propto$ is proportionality)

$$f(|\boldsymbol{m}|) \propto \chi(|\boldsymbol{m}| - 1)\chi(|\boldsymbol{s}| - |\boldsymbol{m}|). \tag{5}$$

[3]The pdf (4) is zero out of the set $\boldsymbol{\alpha}^{|\boldsymbol{m}|}$, see (2). Gamma function $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x)\, \mathrm{d}x$, $z > 0$, is used, see [1].

It is, however, unreasonable as it a priori expect that the probability that each structure reaches the local maximum is the same as the probability that the maximised likelihood is unimodal. Thus, it is wise to assign lower probabilities to higher values of $|\boldsymbol{m}|$. We are choosing prior probability reciprocal to polynomial of the order k. The next heavy-tailed type distribution [6] respects this

$$f(|\boldsymbol{m}|) \propto \big(\Gamma(|\boldsymbol{m}|-k)\chi(|\boldsymbol{m}|-k-1)\Gamma^{-1}(|\boldsymbol{m}|)+\chi(k-|\boldsymbol{m}|)\chi(|\boldsymbol{m}|-1)\big)\chi(|\boldsymbol{s}|-|\boldsymbol{m}|). \quad (6)$$

The accumulated experience indicates the reasonable range $[1,2]$ of its real positive parameter k. The normalisation constant for (6) is derived in Appendix 5.

2.3. Estimation and prediction

The pdf $f(\alpha^{|\boldsymbol{m}|}|a(t),|\boldsymbol{m}|)$ preserves the conjugated prior form

$$f(\alpha^{|\boldsymbol{m}|}|a(t),|\boldsymbol{m}|) \quad \propto \quad \mathcal{D}_{\alpha^{|\boldsymbol{m}|}}\left(\kappa_t^{|\boldsymbol{m}|}\right)\chi(|\boldsymbol{m}|-m_t)\chi(|\boldsymbol{s}|-|\boldsymbol{m}|), \quad \kappa_t^{|\boldsymbol{m}|}=\tilde{\kappa}_t+\kappa_0^{|\boldsymbol{m}|}, \quad (7)$$

where $|\boldsymbol{m}| \in \{1,\dots,|\boldsymbol{s}|\}$ and mth entry $\tilde{k}_{t;m}$ of $\tilde{\kappa}_t$ is defined

$$\tilde{k}_{t;m} \quad = \quad \sum_{\tau\le t}\delta(a_\tau,a_m). \quad (8)$$

Further on, m_t denotes the number of positive entries in $\tilde{\kappa}_{t;m}$.

The probabilities $\alpha^{|\boldsymbol{m}|}$ form nuisance parameter of the problem. We only need the parametric model relating $a(t)$ to $|\boldsymbol{m}|$. It is the predictive pdf conditioned on $|\boldsymbol{m}| \in \{m_t,\dots,|\boldsymbol{s}|\}$. It is proportional to the ratio of the normalising factor

$$f(a(t)||\boldsymbol{m}|) \propto \frac{\mathcal{B}\left(\kappa_t^{|\boldsymbol{m}|}\right)}{\mathcal{B}\left(\kappa_0^{|\boldsymbol{m}|}\right)}.\chi(|\boldsymbol{m}|-m_t). \quad (9)$$

The sufficient statistic for any decision making about $|\boldsymbol{m}|$ is the posterior probability of the unknown $|\boldsymbol{m}|$. With uniform prior on $|\boldsymbol{m}| \in \boldsymbol{s}$, i. e. the prior pdf (6) for $k=0$, it is proportional to the expression (9). The appealing form arises for the uniform prior pdf on unknown probabilities $\alpha^{|\boldsymbol{m}|}$ given by $\kappa_{0;m}^{|\boldsymbol{m}|}=1$, $m \le |\boldsymbol{m}|$ and $|\boldsymbol{m}| \le |\boldsymbol{s}|$. This special case gives

$$f(|\boldsymbol{m}||a(t)) \propto f(|\boldsymbol{m}|)\prod_{m\in\boldsymbol{m}}\Gamma\left(\kappa_{t;m}^{|\boldsymbol{m}|}\right)\frac{\Gamma(|\boldsymbol{m}|)}{\Gamma(t+|\boldsymbol{m}|)}\chi(|\boldsymbol{m}|-m_t)\chi(|\boldsymbol{s}|-|\boldsymbol{m}|). \quad (10)$$

For constant $\kappa_{0;m}^{|\boldsymbol{m}|}=1$, the product over m in (10) is independent of $|\boldsymbol{m}|$ and

$$f(|\boldsymbol{m}||a(t)) \propto f(|\boldsymbol{m}|)\frac{\Gamma(|\boldsymbol{m}|)}{\Gamma(t+|\boldsymbol{m}|)}\chi(|\boldsymbol{m}|-m_t)\chi(|\boldsymbol{s}|-|\boldsymbol{m}|). \quad (11)$$

The more sound general prior pdf (6) gives $f(|\boldsymbol{m}||a(t)) \propto$

$$\left(\frac{\Gamma(|\boldsymbol{m}|-k)}{\Gamma(t+|\boldsymbol{m}|)}\chi(|\boldsymbol{s}|-|\boldsymbol{m}|)\chi(|\boldsymbol{m}|-k-1)+\frac{\Gamma(|\boldsymbol{m}|)}{\Gamma(t+|\boldsymbol{m}|)}\chi(k-|\boldsymbol{m}|)\right)\chi(|\boldsymbol{m}|-m_t). \quad (12)$$

With this, the final form of the posterior probability is $f(|\boldsymbol{m}||a(t)) =$

$$\frac{\left(\frac{\Gamma(|\boldsymbol{m}|-k)}{\Gamma(t+|\boldsymbol{m}|)}\chi(|\boldsymbol{s}|-|\boldsymbol{m}|)\chi(|\boldsymbol{m}|-k-1)+\frac{\Gamma(|\boldsymbol{m}|)}{\Gamma(t+|\boldsymbol{m}|)}\chi(k-|\boldsymbol{m}|)\right)\chi(|\boldsymbol{m}|-m_t)}{\underbrace{\chi(k-m_t)\underbrace{\sum_{|\boldsymbol{m}|=m_t}^{k}\frac{\Gamma(|\boldsymbol{m}|)}{\Gamma(t+|\boldsymbol{m}|)}}_{\mathcal{I}(m_t,k,t)}+\underbrace{\sum_{|\boldsymbol{m}|=\max(m_t,k+1)}^{|\boldsymbol{s}|}\frac{\Gamma(|\boldsymbol{m}|-k)}{\Gamma(t+|\boldsymbol{m}|)}}_{\mathcal{I}(=\max(m_t,k+1)-k,|\boldsymbol{s}|-k,t+k)}}_{D=\chi(k-m_t)\mathcal{I}(m_t,k,t)+\mathcal{I}(\max(m_t,k+1)-k,|\boldsymbol{s}|-k,t+k)}}. \tag{13}$$

Using evaluations presented in Appendix 5, the denominator D in (13) can be expressed in a simple form. It exploits the definition of beta function and gives

$$\begin{aligned} D &= \chi(k-m_t)\mathcal{I}(m_t,k,t)+\mathcal{I}(\max(m_t,k+1)-k,|\boldsymbol{s}|-k,t+k) \\ &= \frac{\chi(k-m_t)}{t-1}\left(\frac{\Gamma(m_t)}{\Gamma(m_t+t-1)}-\frac{\Gamma(k+1)}{\Gamma(k+t)}\right) \\ &+ \frac{1}{t+k-1}\left(\frac{\Gamma(m-k)}{\Gamma(m+t-1)}-\frac{\Gamma(|\boldsymbol{s}|-k+1)}{\Gamma(|\boldsymbol{s}|+t)}\right), \end{aligned} \tag{14}$$

where $m = \max(m_t, k+1)$. For simplicity, $|\boldsymbol{s}| = \infty$ is assumed from now on. It is realistic approximation for the target application and, if need be, it can be avoided. For $|\boldsymbol{s}| >> |\boldsymbol{m}|$, the last member $\frac{\Gamma(|\boldsymbol{s}|-k+1)}{\Gamma(|\boldsymbol{s}|+t)}$ in denominator D is dominated by the other members anyway, so it can certainly be neglected.

If m_t is above k the formula (13) can be further simplified

$$f(|\boldsymbol{m}||a(t)) = \frac{(t+k-1)\Gamma(|\boldsymbol{m}|-k)\Gamma(m_t+t-1)}{\Gamma(m_t-k)\Gamma(|\boldsymbol{m}|+t)}\chi(|\boldsymbol{m}|-m_t). \tag{15}$$

2.4. Sequential choice of the number of experiments

The decision task formulated at the section beginning can be now solved formally.

After generating tth random initial structure s_t and evaluating the local maximiser $a_t = \mathcal{A}(s_t)$, we have to decide whether to generate a new sample or not. The sufficient statistic available to this is the number of experiments t, the parameter k of the prior pdf (6) and, most importantly, the number m_t of different observed values of local maxima. With them, the stopping is recommended if

$$f(|\boldsymbol{m}| = m_t|a(t)) = f(|\boldsymbol{m}| = m_t|m_t,t,k) \geq \lambda, \text{ where } \lambda \in (0,1) \text{ is a chosen threshold.} \tag{16}$$

By introducing expression (15) into (16), we get the final stopping rule in which the optional parameter $\lambda \in (0,1)$ balances reliability of the guess $|\boldsymbol{m}| = m_t$ and costs of the local maximum evaluation. The *final stopping rule* is

$$\begin{aligned} &\text{For } m_t > k, \text{ stop if} && \frac{(t+k-1)\Gamma(m_t-k)\Gamma(m_t+t-1)}{\Gamma(m_t-k)\Gamma(m_t+t)} = \frac{t+k-1}{m_t+t-1} \geq \lambda \\ &\text{For } m_t \leq k, \text{ stop if} && \frac{1}{\frac{m_t+t-1}{t-1}+\frac{\Gamma(t+m_t)}{\Gamma(m_t)\Gamma(t+k)}\left(\frac{1}{t+k-1}-\frac{\Gamma(k+1)}{t-1}\right)} \geq \lambda. \end{aligned} \tag{17}$$

The final stopping rule (17) is computationally cheap and may significantly spare computational budget of the overall learning to which the maximisation serves.

3. ILLUSTRATIVE EXAMPLES

This section indicates usability of the stopping rule (17).

3.1. Experiment 1: $k = 2$ (6) and varying λ (17)

$|\boldsymbol{m}|$ and $\alpha^{|\boldsymbol{m}|}$ were sampled from the considered prior pdf. For each sample, search for the global maximum in randomly generated space with the given Θ was performed and the stopping rule was applied with a given λ. With $k = 2$ and $|\boldsymbol{s}| = 32768$, the number of Monte-Carlo experiments was 10^5 for each tested λ.

Results The runs in which all local maxima were found was taken as successful. Their portion in all runs is denoted $\hat{\lambda}$ and provides a bound on λ, to which it should converge. It was indeed so as the results in Table 1 show.

λ	0.7000	0.7500	0.8000	0.8500	0.9000	0.9500	0.9700	0.9800	0.9900
$\hat{\lambda}$	0.7477	0.7728	0.8087	0.8429	0.8821	0.9307	0.9532	0.9670	0.9809

Tab. 1. Used λ and its estimated lower bound $\hat{\lambda}$.

3.2. Experiment 2: Maxima in the model-structure space

Maximum a posteriori likelihood (probability) of the structure of autoregressive-regressive model was searched. Scalar outputs y_τ, stimulated by scalar normal, zero mean white-noise inputs u_τ and unobserved noise e_τ, were generated

$$y_\tau = 1.4183y_{\tau-1} - 1.5894y_{\tau-2} + 1.3161y_{\tau-3} - 0.8864u_{\tau-3} + 0.2826u_{\tau-4} + 0.5067e_\tau.$$

The richest inspected structure for the structure selection was

$$y_\tau = \sum_{i=1}^{6} a_i y_{\tau-i} + \sum_{i=0}^{6} b_i u_{\tau-i} + c + \sigma e_\tau \text{ with unknown } a_i,\ b_i,\ c,\ \sigma > 0.$$

The richest structure has dimension 14. The number of its substructures is $|\boldsymbol{s}| = 16384$. Maximum of the posterior likelihood $\mathcal{L}(s)$ is searched by the algorithm $\mathcal{A}(s)$ proposed in [8]. It acts on randomly chosen prior guess of the structure $s \in \boldsymbol{s}$ and returns $a_m \in \boldsymbol{s}$ corresponding to a local maximum of $\mathcal{L}$, cf. Section 2. A small number of data was simulated in order to get the high number $|\boldsymbol{m}| = 245$ of local maxima. Stopping rule (15) was used with $\lambda = 0.7$ and $k = 2$.

Results. Figure 1 shows the increasing number of local maxima found when the number of restarts t increases. The search stopped at $t = 400$ when the number of found local maxima was $m_t = 168$. This is quite a good result, because $m_t/|\boldsymbol{m}| = 168/245 = 0.686 \approx \lambda = 0.7$. It indicates applicability of the prior probability (6) chosen irrespectively of the problem specificity. Figure 2 shows the cumulative number of local maxima found in experiment without a stopping rule. It confirms that the stopping rule spares a lot of the computational effort.

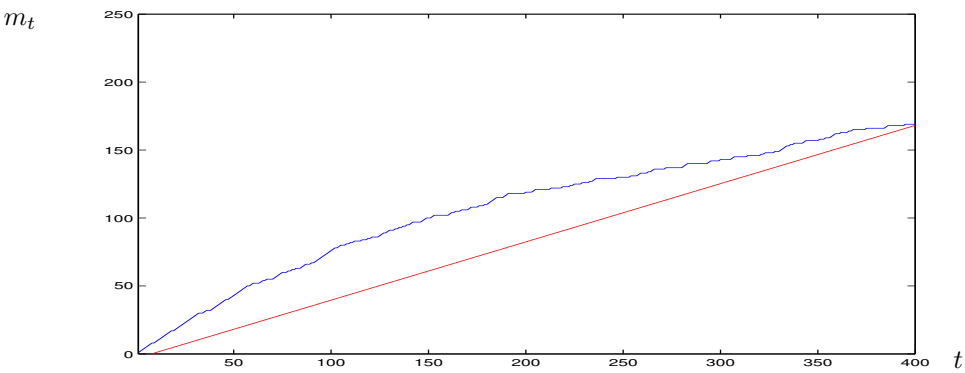

Fig. 1. The number of local maxima in Experiment 2 *with* the stopping rule applied. The straight line is the limit given by the stopping rule (17) on the probability (15). The search was stopped after $t = 400$ restarts when the number of found local maxima was $m_{400} = 168$.

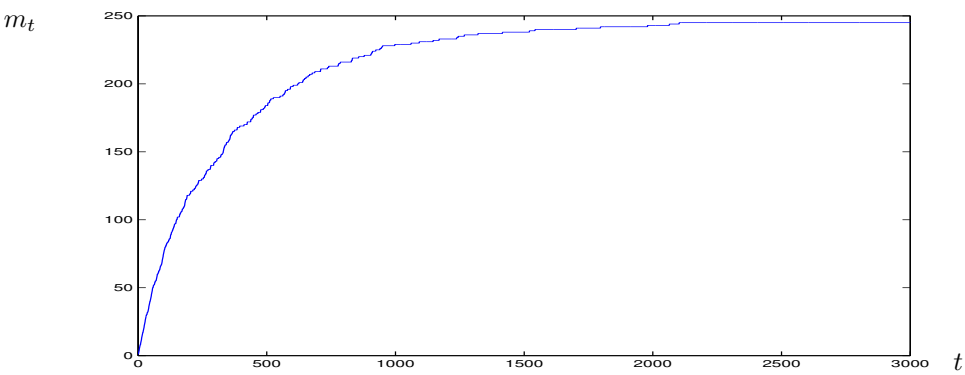

Fig. 2. The number of local maxima in Experiment 2 *without* the stopping rule applied. The search found all of them for $t > 2200$.

4. CONCLUSIONS

Bayesian sequential rule (17) was designed to stop the search when there is a high probability that all local maxima were found. Its design was motivated by the quest for an enhanced efficiency of the structure estimation algorithm [8, 10], which relies on a sufficient number of random restarts. The same problem is, however, common in many sub-domains of machine learning, artificial intelligence and wherever a global optimum is searched for. This made us to elaborate the stopping rule without entering details of the local search employed. The experimental evidence whose illustrative sample is presented indicates a high efficiency of the stopping rule and its wide applicability. Quest for a wider use, testing and analysing made us to present the rule and its technical details to a wider audience.

The desirable additional extensive Monte-Carlo type experiments concerning parameters k and λ and optimisation variants will be published elsewhere. Our accumulated experience with various real-data problems indicates robustness of the stopping rule with respect to their choice. The rule simplicity offers the reader the possibility to test it on his/her specific cases.

5. APPENDIX: NORMALISATION CONSTANT

Calculation of normalisation constants and generating the random variable $|\boldsymbol{m}|$ distributed as (6) need to evaluate the sum

$$\mathcal{I}(A,B,C) \;=\; \sum_{i=A}^{B} \frac{\Gamma(i)}{\Gamma(C+i)}.$$

For $C \neq 0$ and $C \neq 1$,

$$\mathcal{I}(A,B,C) \;=\; \frac{1}{\Gamma(C)} \sum_{i=A}^{B} B(i,C) = \frac{1}{\Gamma(C)} \int_0^1 \sum_{i=A}^{B} x^{i-1}(1-x)^{C-1}\,\mathrm{d}x.$$

Identity $\quad \displaystyle\sum_{i=A}^{B} x^i = \frac{(x^A - x^{B+1})}{1-x}$ gives

$$\mathcal{I}(A,B,C) \;=\; \frac{\int_0^1 (x^B - x^{A-1})(1-x)^{C-2}\,\mathrm{d}x}{\Gamma(C)} = \frac{\mathcal{B}(C-1,A) - \mathcal{B}(C-1,B+1)}{\Gamma(C)}.$$

Finally $\quad \mathcal{I}(A,B,C) = \dfrac{\Gamma(A)}{(C-1)\Gamma(A+C-1)} - \dfrac{\Gamma(B+1)}{(C-1)\Gamma(B+C)}.$

For $B \to \infty$, $\quad \mathcal{I}(A,\infty,C) = \dfrac{\Gamma(A)}{(C-1)\Gamma(A+C-1)}.$

This formula does not work for $C = 1$. Proof of this case would be much more complicated, so we just give the result

$$\mathcal{I}(A,B,1) = \sum_{i=A}^{B} \frac{1}{i} = \Phi(0,B+1) - \Phi(0,A)$$

where $\Phi(\cdot,\cdot)$ is polygamma, specifically $\Phi(0,\cdot)$ is digamma function and $\gamma = -\Phi(0,1)$ is Euler's constant.

6. APPENDIX: LOCAL SEARCH FOR MAXIMUM A POSTERIORI STRUCTURE ESTIMATE

This appendix recalls basic idea of the local search for the best model structure.

Considered linear-in-parameters single-output y_τ Gaussian regression and auto-regression ($'$ denotes transposition)

$$f(y_\tau|\psi_\tau,\theta,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y_\tau-\theta'\psi_\tau)^2}{2\sigma^2}\right] \tag{18}$$

has an extreme application width due the flexibility of the regression vector ψ_t. It can be an arbitrary real finite-dimensional image of observed data and time. Moreover, chain rule for pdfs implies that multi-output case can always be treated as a collection of single-output cases by appropriately extending ψ.

The model (18) belongs to exponential family and has Gauss Inverse-Wishart pdf [17] as the conjugate prior of unknown parameters (θ,σ). It is seen by rewriting (18) into the form

$$f(y_\tau|\psi_\tau,\theta,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\mathrm{tr}\left(\Psi_\tau\Psi_\tau'\frac{\begin{bmatrix}-1 & \theta'\end{bmatrix}'\begin{bmatrix}-1 & \theta'\end{bmatrix}}{2\sigma^2}\right)\right],\ \Psi_\tau = \begin{bmatrix} y_\tau \\ \psi_\tau \end{bmatrix}. \tag{19}$$

The likelihood function, which is the product of $f(y_\tau|\psi_\tau,\theta,\sigma)$ with observed data inserted, is obviously determined by sufficient statistics that evolve recursively with the growing number of processed data ($V_{\tau;y}$ is scalar)

$$V_\tau = V_{\tau-1} + \Psi_\tau\Psi_\tau' = \begin{bmatrix} V_{\tau;y} & V_{\tau;y\psi}' \\ V_{\tau;y\psi} & V_\psi \end{bmatrix},\quad \nu_\tau = \nu_{\tau-1}+1. \tag{20}$$

They determine the posterior pdf, of the same form as the likelihood function, when the recursion (20) is initiated by the statistics $V_0 > 0$ (symmetric positive definite matrix) and $\nu_0 > 0$. Structure s (of the length ℓ_s) of the considered model determines individual regressors entering. The likelihood for estimating $s \in \boldsymbol{s}$ is the ratio of normalising factors $\mathcal{J}$ of the posterior and the prior Gauss Inverse-Wishart pdf, [9],

$$\mathcal{L}(s) = \mathcal{L}_\tau(s) = \frac{\mathcal{J}(V_{\tau;s},\nu_{\tau;s})}{\mathcal{J}(V_{0;s},\nu_{\tau;s})},\quad \mathcal{J}(V,\nu) = \Gamma(0.5\nu)2^{0.5\nu}(2\pi)^{0.5\ell}\frac{|V_\psi|^{0.5(\nu-1)}}{|V|^{0.5(\nu-1)}}. \tag{21}$$

The evaluation of determinants $|V_s|$, $|V_{s\psi}|$ for the considered structure is numerically the most sensitive and demanding operation. Paper [8] copes with it by: a) collecting the sufficient statistics for the richest $\bar{s} \in \boldsymbol{s}$ inspected structure; b) evaluating Choleski square root of V (typically, during sequential processing of observed data vectors $\Psi_{\tau;\bar{s}}$); c) noticing that V_s for any structure embedded in $\bar{s}$ has the statistic embedded in $V_{\bar{s}}$; d) designing and using efficient algorithms evaluating Choleski square root of V_s directly from Choleski square root of $V_{\tilde{s}}$ corresponding the structure $\tilde{s} \in \boldsymbol{s}$ in which s is embedded.

With the outlined machinery, the an efficient local search is designed. Using the initial guess of s, the likelihood is evaluated for all structures, which arise from it either by cancelling one regressor from it or by adding one regressors from $\psi_{\bar{s}}$. This requires just $\ell_{\bar{s}}$ values of $\mathcal{L}$ and selecting maximising argument a_m among them.

Details can be found in [8] while [9] describes LD factorised version of this algorithm, which is used during computationally demanding structure estimation of multivariate normal mixtures [9].

ACKNOWLEDGEMENTS

This work was partially supported by GAČR, grant 16-09848S. A historical predecessor of this paper was prepared together with L. Tesař and J. Šindelář. The author of this work acknowledge contributions of his former colleagues while attributing all deficiencies to himself.

(Received June 26, 2018)

REFERENCES

[1] E. Artin: The Gamma Function. Holt, Rinehart, Winston, NY 1964.

[2] O. Barndorff-Nielsen: Information and Exponential Families in Statistical Theory. Wiley, NY 1978. DOI:10.1002/9781118857281

[3] J. O. Berger: Statistical Decision Theory and Bayesian Analysis. Springer, NY 1985. DOI:10.1007/978-1-4757-4286-2

[4] K. K. Bharti and P. K. Singh: Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. Expert Systems Appl. *42* (2015), 3105–3114. DOI:10.1016/j.eswa.2014.11.038

[5] T. S. Ferguson: Who solved the secretary problem? Statist. Sci. *4* (1989), 3, 282–289. DOI:10.1214/ss/1177012493

[6] S. Foss, D. Korshunov, and S. Zachary: An Introduction to Heavy-Tailed and Subexponential Distributions. Springer Science and Business Media, 2013. DOI:10.1007/978-1-4614-7101-1

[7] R. Horst and H. Tuy: Global Optimization. Springer, 1996. DOI:10.1007/978-3-662-02947-3

[8] M. Kárný: Algorithms for determining the model structure of a controlled system. Kybernetika *9* (1983), 2, 164–178.

[9] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař: Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer, 2006. DOI:10.1007/1-84628-254-3

[10] M. Kárný and R. Kulhavý: Structure determination of regression-type models for adaptive prediction and control. In: Bayesian Analysis of Time Series and Dynamic Models (J. C. Spall, ed.), Marcel Dekker, New York 1988.

[11] D. E. Knuth: The Art of Computer Programming, Sorting and Searching. Addison-Wesley, Reading 1973.

[12] D. J. Lizotte: Practical Bayesian Optimization. PhD Thesis, Edmonton, Alta 2008.

[13] J. Novovičová and A. Malík: Information-theoretic feature selection algorithms for text classification. In: Proc. of the IJCNN 2005, 16th International Joint Conference on Neural Networks, Montreal 2005, pp. 3272–3277. DOI:10.1109/ijcnn.2005.1556452

[14] V. Peterka: Bayesian system identification. In: Trends and Progress in System Identification (P. Eykhoff, ed.), Pergamon Press, Oxford 1981, pp. 239–304. DOI:10.1016/b978-0-08-025683-2.50013-2

[15] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas: Taking the human out of the loop: A review of Bayesian optimization. Proc. IEEE *104* (2016), 1, 148–175. DOI:10.1109/jproc.2015.2494218

[16] D. H. Wolpert and W. G. Macready: No free lunch theorems for optimization. IEEE Trans. Evolutionary Comput. *1* (1997), 1, 67–82. DOI:10.1109/4235.585893

[17] A. Zellner: An Introduction to Bayesian Inference in Econometrics. J. Wiley, NY 1976.

Miroslav Kárný, Institute of Information Theory and Automation, The Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: school@utia.cas.cz