Received: 22 September 2015 | Revised: 19 August 2016 | Accepted: 19 December 2016

DOI 10.1002/acs.2756

WILEY

RESEARCH ARTICLE

Recursive Bayesian estimation of autoregressive model with uniform noise using approximation by parallelotopes

Lenka Pavelková | **Ladislav Jirsa**

Institute of Information Theory and Automation, Pod Vodárenskou Věží 4, Prague, Czech Republic

Correspondence
Lenka Pavelková, Institute of Information Theory and Automation, Pod Vodárenskou Věží 4, Prague, Czech Republic.
Email: pavelkov@utia.cas.cz

Funding information
MŠMT 7D12004 ; E!7262 ProDiSMon

Summary

This paper proposes a recursive algorithm for the estimation of a stochastic autoregressive model with an external input. The noise of the involved model is described by a uniform distribution. The model parameters are estimated using the Bayesian approach. Without an approximation, the support of the posterior distribution is a complex multidimensional polytope whose number of faces increases with time. We propose an approximation of this polytope in each time step by a parallelotope with a constant number of faces. The behaviour of the proposed algorithm is illustrated by simulations and compared with other methods.

KEYWORDS

approximate parameter estimation, ARX model, Bayesian estimation, bounded noise, Kullback-Leibler divergence, parallelotope

1 | INTRODUCTION

A linear regression model is often used in a range of adaptive decision-making tasks that include predictors, advising, and fault-detecting systems as well as adaptive controllers, see eg, in the previous study by Kárný et al[1] (Chapter 14), in the previous study by Bobál et al[2] (Chapter 8), and in the previous study by Young[3] (Part I). All these tasks need recursive estimation of unknown model parameters.

Bayesian estimators form a powerful tool for solving the parameter estimation. If the random disturbances entering the model are assumed to be Gaussian, then, fast and efficient estimation algorithms are based on least squares (LS).[4]

In practice, the involved noise is often bounded. Because the normal distribution has light tails, it can be usually accepted as a reasonable approximation of a bounded noise range. However, the unbounded support of the Gaussian distribution can cause difficulties if the estimated quantity is physically bounded because, for instance, it may give unreasonable negative estimates of naturally nonnegative variable. Then, the Monte Carlo methods[5] can be used for parameter estimation. These methods can handle models with arbitrarily distributed noise. Nevertheless, they require a large amount of samples to obtain acceptable results.

Alternatively, a nonprobabilistic "unknown-but-bounded errors" approach can be used for the parameter estimation. Then, the estimates are inside a bounded set.[6] The complexity of this set increases with time. Therefore, an approximation of this set is to be used, eg, by an orthotope,[7] by an ellipsoid[8] or by a parallelotope.[9] However, the nonprobabilistic interpretation makes solutions of related decision-making tasks unnecessarily difficult as a rich set of statistical tools is omitted. Moreover, a comparison of stochastic estimation theory and unknown-but-bounded approach in the previous study by Ninness and Goodwin[10] reveals that the latter provides results having exact stochastic equivalents.

A method, connecting advantages of a stochastic approach with the straightforwardness of nonprobabilistic unknown-but-bounded errors method, was presented in the previous study by Kárný and Pavelková.[11] There, a probabilistic ARX (auto-regressive with external input) model with uniform noise was introduced and its approximate Bayesian estimation was performed on a moving window. The maximum a posteriori probability estimation problem was converted into the problem of linear programming, yielding point estimates with no information about their precision. The ongoing paper[12] enriches the previous results in the sense that it provides the posterior probability distribution on

parameters, too. There, the approximate Bayesian estimation of the mentioned uniform ARX model uses the gradual orthogonal rotations. The complex support of an original posterior parameter distribution is approximated by a polytope whose number of faces does not increase with time.

This paper revises the results obtained in the previous study by Kárný and Pavelková.[12] It uses a similar estimation scheme but proposes a more efficient circumscription. The original complex support of the posterior parameter distribution is circumscribed by a parallelotope in each time step.

The paper is organised as follows. Section 2 describes the addressed problem and defines a linear autoregressive model with an external input (ARX model) with uniform noise. In Section 3, a theoretical solution of the estimation problem is proposed and the implementation aspects of the algorithm are discussed. Section 4 summarises algorithmic details. In Section 5, simulation experiments are presented and proposed algorithm is compared with other methods. Section 6 concludes the paper. Appendix provides some computational details.

Throughout: $'$ denotes transposition; $\ell_{\mathbf{x}}$ means the length of a vector x; vectors are always column; $\mathbf{I}_{(n)}$ is the square identity matrix of the order n; $\mathbf{1}_{(n)}$ means a unit vector of the length n; $\mathbf{0}_{(n)}$ means a zero vector of the length n; $\equiv$ is equality by definition; $x^\star$ denotes a set of all values of the variable x; $\chi_x(\bullet)$ is value of the indicator of a set defined by the conditions $\bullet$ at the point x; $\underline{x}$ and $\overline{x}$ are the lower and upper bound on $x \in x^\star$, respectively, ie, $\underline{x} \leqslant x \leqslant \overline{x}, \forall x \in x^\star$, the inequalities are meant entry-wise; t labels discrete-time moments, $t \in t^\star \equiv \{1, 2, \dots T\}$; $d_t = (y_t, u_t)$ is the data record at time t consisting of an observed system output y_t and of an optional system input u_t; $x(t)$ denotes the sequence $(x_1, \dots, x_t)$, $x \in \{d, y, u\}$. Θ_i denotes the i-th entry of the vector Θ. $f(\cdot|\cdot)$ denotes a conditional probability density function (pdf); names of arguments distinguish respective pdfs; no formal distinction is made between a random variable, its realisation and an argument of the pdf. Used integrals are always definite and multivariate; the integration domain coincides with the support of the pdf in its argument. $\mathcal{U}_x(\mu, r)$ is a uniform pdf of a scalar x given by the expectation μ and the half-width r.

2 | ADDRESSED PROBLEM

We model a system with a scalar output y_t at a discrete time t. Note that a modelling of a scalar output is sufficient because of the chain rule for pdfs,[1] $f(\alpha, \beta|\gamma) = f(\alpha|\beta, \gamma)f(\beta|\gamma)$. This rule treats dependence of the variables by their conditioning. Hence, the multivariate pdf is decomposed into a product of univariate pdfs (factors)[13] and parameters are estimated by factors.

We define the linear ARX model with a uniform noise as follows:

$$y_t = \psi_t'\vartheta + \varepsilon_t, \tag{1}$$

$$f(\varepsilon_t|r) = \mathcal{U}_\varepsilon(0, r), \tag{2}$$

where

- ψ_t is a finite-dimensional regression vector composed of past observed data and a current input in a known (recursively implementable) way,
- ϑ is a vector of unknown regression coefficients,
- ε_t is a uniformly distributed white noise at the time t (zero mean and uncorrelated with older observations),
- $r > 0$ is an unknown positive scalar half-width of a noise range.

The pair of formulae (1), (2) can be equivalently written as follows:

$$\begin{aligned} f(y_t|u_t, d(t-1), \vartheta, r) &\equiv f(y_t|\psi_t, \vartheta, r) = \mathcal{U}_{y_t}(\psi_t'\vartheta, r) \\ &= \frac{1}{2r} \quad \chi_{y_t}(-r \leqslant y_t - \psi_t'\vartheta \leqslant r) \\ &= \frac{1}{2} \quad \Theta_{\ell_\Psi}\chi_{y_t}(-1 \leqslant \Psi_t'\Theta \leqslant 1), \end{aligned} \tag{3}$$

where

$$\Psi_t = \left[\psi_t', y_t\right]', \tag{4}$$

$$\Theta = \left[-\frac{\vartheta'}{r}, \frac{1}{r}\right]'. \tag{5}$$

Note that the output y_t is scalar but the space of Θ is ℓ_Ψ-dimensional, where $\ell_\Psi \geqslant 2$. Therefore, the form that uses characteristic function χ will be preferred throughout the paper to express the conditions for values of Θ and to indicate the parameter set in the space of Θ.

We assume that the input generator meets natural conditions of control,[1] ie, the input is conditionally independent of the unknown parameters, $f(u_t|d(t-1),\Theta) = f(u_t|d(t-1))$. Under this assumption, the Bayes rule[14] provides the posterior pdf $f(\Theta|d(t))$ that has the following form:

$$\begin{aligned} f(\Theta|d(t)) &= f(\Theta|\mathbb{V}_t, \mathbb{L}_t, \mathbb{U}_t, \nu_t) \\ &= \frac{\Theta_{\ell_\Psi}^{\nu_t}\chi_{\Theta_{\ell_\Psi}}(0 < \Theta_{\ell_\Psi})\chi_\Theta(\mathbb{L}_t \leqslant \mathbb{V}_t\Theta \leqslant \mathbb{U}_t)}{\mathrm{J}(\mathbb{V}_t, \mathbb{L}_t, \mathbb{U}_t, \nu_t)}, \end{aligned} \tag{6}$$

with the normalising integral

$$\begin{aligned} \mathrm{J}(\mathbb{V}_t, \mathbb{L}_t, \mathbb{U}_t, \nu_t) = \int_{\Theta^\star} &\Theta_{\ell_\Psi}^{\nu_t}\chi_{\Theta_{\ell_\Psi}}(0 < \Theta_{\ell_\Psi}) \\ &\times \chi_\Theta(\mathbb{L}_t \leqslant \mathbb{V}_t\Theta \leqslant \mathbb{U}_t)\mathrm{d}\Theta. \end{aligned} \tag{7}$$

The support of $f(\Theta|\mathbb{V}_t, \mathbb{L}_t, \mathbb{U}_t, \nu_t)$ is a polytope in the parameter space. The values of the sufficient statistics $\mathbb{V}_t, \mathbb{L}_t, \mathbb{U}_t$, and ν_t update as follows:

$$\begin{aligned} &\mathbb{V}_t = \begin{bmatrix} \mathbb{V}_{t-1} \\ \Psi_t' \end{bmatrix}, \nu_t = \nu_{t-1} + 1. \\ &\mathbb{L}_t = \begin{bmatrix} \mathbb{L}_{t-1} \\ -1 \end{bmatrix} = \begin{bmatrix} \mathbb{L}_0 \\ -\mathbf{1}_{(\nu_t - \nu_0)} \end{bmatrix}, \mathbb{U}_t = \begin{bmatrix} \mathbb{U}_{t-1} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbb{U}_0 \\ \mathbf{1}_{(\nu_t - \nu_0)} \end{bmatrix}. \end{aligned} \tag{8}$$

Let us introduce a symbolic update operator $\mathcal{T}_t$ to simplify the notation in (8) as follows

$$(\mathbb{V}_t, \mathbb{L}_t, \mathbb{U}_t, \nu_t) = \mathcal{T}_t(\mathbb{V}_{t-1}, \mathbb{L}_{t-1}, \mathbb{U}_{t-1}, \nu_{t-1}). \tag{9}$$

The rules given above hold for the prior pdf

$$f(\Theta|\mathbb{V}_0, \mathbb{L}_0, \mathbb{U}_0, \nu_0) \tag{10}$$

determined by a (ℓ_Ψ,ℓ_Ψ)-matrix $\mathbb{V}_0$, ℓ_Ψ-vectors $\mathbb{L}_0$ and $\mathbb{U}_0$, and scalar ν_0. These prior statistics have to guarantee finiteness of the normalising integral (7). By appropriate choice of $\mathbb{V}_0$, ν_0, $\mathbb{L}_0$, and $\mathbb{U}_0$, a priori known, hard bounds on parameter estimates are respected and, moreover, the respective parameter set, ie, polytope, is bounded. Regular prior $\mathbb{V}_0$ will guarantee all parameter sets, given by the subsequent (regular) $\mathbb{V}$s, bounded. The value of statistic ν corresponds to the number of processed data.

The formula (6) and the update (8) cannot be used permanently in recursive estimation because the size of the matrix $\mathbb{V}_t$ and the vectors $\mathbb{L}_t$ and $\mathbb{U}_t$ increase with the number of processed data. Memory requirements permanently grow as well as the support of the posterior pdf becomes complex. As a consequence, implementation is data-limited and evaluation of posterior moments becomes hard.

Thus, an approximate recursive estimation is needed. Its design is the topic of this paper. We search for an approximation of the original statistics $\mathbb{V}_t$, $\mathbb{L}_t$, $\mathbb{U}_t$, and ν_t by suitable statistics V_t, L_t, U_t and ν_t, respectively, the sizes of which are bounded and independent of the value t.

3 | APPROXIMATE ESTIMATION OF UNIFORM ARX MODEL

For a recursively feasible estimation of the model (3) with posterior pdf (6) and statistics (8), we apply the projection-based approximation of the Bayes rule.[15,16] The projection-based approximation finds the best projection of the correct Bayesian pdf into the selected class of pdfs. The projection is optimal in the sense of Kullback-Leibler divergence (KLD).[17]

3.1 | Approximate posterior distribution

We aim to approximate the exact posterior pdf (6) by a pdf determined by a statistics of finite and bounded size. Respecting the form of (6), we choose the approximate pdf

$$\begin{aligned} f(\Theta|d(t)) &\approx f(\Theta|V_t, L_t, U_t, \nu_t) \\ &= \frac{\Theta_{\ell_\Psi}^{\nu_t}\chi_\Theta(L_t \leqslant V_t\Theta \leqslant U_t)}{\mathrm{J}(V_t, L_t, U_t, \nu_t)}, \end{aligned} \tag{11}$$

with the normalising integral $\mathrm{J}(V_t, L_t, U_t, \nu_t)$, cf (7),

$$\mathrm{J}(V_t, L_t, U_t, \nu_t) = \int_{\Theta^\star} \Theta_{\ell_\Psi}^{\nu_t}\chi_\Theta(L_t \leqslant V_t\Theta \leqslant U_t)\mathrm{d}\Theta \tag{12}$$

where V_t is a square matrix of size (ℓ_Ψ,ℓ_Ψ) with the last row $[\mathbf{0}'_{(\ell_\Psi-1)}, \upsilon]$, $\upsilon > 0$; L_t and U_t are vectors of length ℓ_Ψ, $L_{\ell_\Psi;t} > 0$. The form of the last row of V_t together with the last entries of L_t and U_t cover the condition $\chi_{\Theta_{\ell_\Psi}}(0 < \Theta_{\ell_\Psi})$ in (6).

Note that the inequalities in (11) describe a parallelotope of dimension ℓ_Ψ (intersection of ℓ_Ψ strips with linearly independent normal vectors corresponding to the rows of V_t and boundaries represented by L_t and U_t) with constant number of faces while the inequalities in characteristic function in original pdf (6) describe the polytope whose number of faces grows with time.

One estimation step consists of 2 stages: data update and projection.

Actually, the statistics $\mathbb{V}_t$, $\mathbb{L}_t$, $\mathbb{U}_t$, and ν_t (8) are not approximated. Projection uses the approximate posterior pdf (11) obtained in step $t-1$ as the true "old" posterior pdf in step t.

3.1.1 | Data update

The approximate statistics V_{t-1}, L_{t-1} and U_{t-1} from the previous time step are updated (see (9)) by the data vector Ψ_t (4)

$$(\tilde{V}_t, \tilde{L}_t, \tilde{U}_t, \tilde{\nu}_t) = \mathcal{T}_t(V_{t-1}, L_{t-1}, U_{t-1}, \nu_{t-1}) \tag{13}$$

to the pdf

$$f(\Theta|\tilde{V}_t, \tilde{L}_t, \tilde{U}_t, \tilde{\nu}_t) = \frac{\Theta_{\ell_\Psi}^{\nu_{t-1}+1}\chi_\Theta(\tilde{L}_t \leqslant \tilde{V}_t\Theta \leqslant \tilde{U}_t)}{\mathrm{J}(\tilde{V}_t, \tilde{L}_t, \tilde{U}_t, \tilde{\nu}_t)}. \tag{14}$$

This update damages the form of (11) by appending one row to the statistics.

3.1.2 | Projection

To close the recursion, we search for a posterior pdf $f(\Theta|V_t, L_t, U_t, \nu_t)$ of the form (11) that approximates $f(\Theta|\tilde{V}_t, \tilde{L}_t, \tilde{U}_t, \tilde{\nu}_t)$ (14). For the purpose of simplicity, we use the following notation in this subsection: $\tilde{f}_\Theta \equiv f(\Theta|\tilde{V}_t, \tilde{L}_t, \tilde{U}_t, \tilde{\nu}_t)$ (14), $f_\Theta \equiv f(\Theta|V_t, L_t, U_t, \nu_t)$ (11).

Projection-based approximation uses approximation by a pdf that minimises the KLD of the approximated pdf to the approximate pdf which is optimal within a Bayesian context,[18] ie, the pdf $f_\Theta \in f_\Theta{}^\star$ approximating the pdf $\tilde{f}_\Theta$ is to be a minimiser of KLD $\mathrm{D}(\tilde{f}_\Theta||f_\Theta)$ as follows:[17]

$${}^{O}f_\Theta \equiv \mathrm{Arg}\min_{f_\Theta \in f_\Theta^\star} \mathrm{D}(\tilde{f}_\Theta||f_\Theta) = \mathrm{Arg}\min_{f_\Theta \in f_\Theta^\star} \int_{\Theta^\star} \tilde{f}_\Theta \ln\frac{\tilde{f}_\Theta}{f_\Theta}\,\mathrm{d}\Theta. \tag{15}$$

3.2 | Minimisation of Kullback-Leibler divergence

The evaluation of (15) gives

$$\mathrm{D}(\tilde{f}_\Theta \| f_\Theta) = \ln \mathrm{J}(V_t, L_t, U_t, \nu_t) - \ln \mathrm{J}(\tilde{V}_t, \tilde{L}_t, \tilde{U}_t, \tilde{\nu}_t) + \\ + \int_{\Theta^\star} f_\Theta \ \ln \frac{\chi(\mathrm{supp}(\tilde{f}_\Theta))}{\chi(\mathrm{supp}(f_\Theta))} \, \mathrm{d}\Theta. \tag{16}$$

To guarantee finiteness of the KLD (16), the support of f_Θ has to include the support of $\tilde{f}_\Theta$, ie,

$$\mathrm{supp}(\tilde{f}_\Theta) \subset \ \mathrm{supp}\,(f_\Theta)\,. \tag{17}$$

Then, the last term in (16) is equal to 0. The second term is independent of f_Θ. Thus, KLD reaches its minimum by minimising $\ln \mathrm{J}(V_t, L_t, U_t, \nu_t)$ in (16) satisfying the condition (17). It means that we search for the optimal reduced statistics, ie, a square matrix V_t and vectors L_t and U_t of corresponding lengths that minimise $\ln \mathrm{J}(V_t, L_t, U_t, \nu_t)$ in (16). The value of $\nu_t = \min(\tilde{\nu}_t, \ell_\Psi)$.

It can be shown (see Appendix A.1) that, assuming a special form of the matrix V_t, this task is equivalent to the minimisation of the parallelotope volume, described by V_t, L_t, and U_t.

3.3 | Minimization of parallelotope volume

The characteristic function in (3) is geometrically a strip in the Θ-space bounded by 2 parallel hyperplanes with the normal vector Ψ_t (4) and zero point on the central hyperplane. The set of inequalities in (11) with a regular square V_t represents a parallelotope as an intersection of ℓ_Ψ strips whose normal vectors correspond to the rows of V_t. The set of inequalities in (14) with nonsquare $\tilde{V}_t$ represents a general polytope as an intersection of the parallelotope from the previous step and a strip coming from the newest observation.

We need to circumscribe this polytope by the smallest possible paralelotope. In the previous study by Vicino and Zappa,[9] a fast and simple method is proposed, which we adopt. The noise bound is supposed to be known there. Our method, because of a suitable parametrization (4) and (5), is more versatile as it enables estimation of noise bound r, too.

For the purpose of the method, the inequalities in (11) are equivalently rewritten as follows:

$$-\mathbf{1}_{(\ell_\Psi)} \leqslant \underbrace{D_t V_t}_{W_t} \Theta - \underbrace{D_t \frac{U_t + L_t}{2}}_{C_t} \leqslant \mathbf{1}_{(\ell_\Psi)}, \tag{18}$$

where D_t is a diagonal matrix with entries $D_{ii;t} = \frac{2}{U_{i;t} - L_{i;t}}$, $i = 1, \dots, \ell_\Psi$. We call this form a **[− 1, 1]** form while the previous form is called a **[L, U]** form.

Analogically, we can introduce $\tilde{W}_t$ and $\tilde{C}_t$ by row-by-row transformation of $\tilde{V}_t$, $\tilde{L}_t$ and $\tilde{U}_t$ using (18).

The algorithm[9] circumscribes a polytope, defined by $\tilde{V}_t$, $\tilde{L}_t$, and $\tilde{U}_t$ (13) with $\ell_\Psi + 1$ rows, by a minimum volume parallelotope. It uses the expression (18) and works with inversion of W_t. If new data are observed, a new strip (3) is generated and the intersection of all the strips changes into a general polytope. First, to remove redundancy, all the strips are tightened, ie, narrowed or shifted to minimise their widths with the same intersection. Second, it is shown that a minimum volume parallelotope circumscribing the polytope can be obtained by selecting such ℓ_Ψ rows from $\ell_\Psi + 1$ candidates into the matrix W_t that minimise $|\det W_t^{-1}|$.

We modified the algorithm[9] to keep the matrix V_t (equivalently W_t) in a special form (see (11)). After the tightening, the last row is always selected for the new parallelotope, together with those from remaining rows to minimise the volume. Because of the mentioned requirement on V_t, the circumscription never reaches absolute volume minimum as proposed in the previous study by Vicino and Zappa,[9] therefore the approximation is suboptimal.

The choice of ℓ_Ψ rows is actually done by replacing 1 row in the matrix W_{t-1} (except the last one) by the data strip, if it decreases the volume, which is computationally more practical.

3.4 | Choice of prior statistics

The prior information on the original parameters ϑ and r of the ARX model in (1) is supposed to be given (eg, by expert or user) in the form of lower and upper bounds as follows:

$$-\infty < \underline{\vartheta} \leqslant \vartheta \leqslant \overline{\vartheta} < \infty, \ \ 0 < \underline{r} \leqslant r \leqslant \overline{r} < \infty. \tag{19}$$

To obtain the prior pdf in the form (10), the given bounds are transformed using the (5) into

$$\underline{\Theta}_i = \min\left(-\frac{\overline{\vartheta}_i}{\underline{r}}, -\frac{\overline{\vartheta}_i}{\overline{r}}\right) \leqslant \Theta_i \leqslant \max\left(-\frac{\underline{\vartheta}_i}{\underline{r}}, -\frac{\underline{\vartheta}_i}{\overline{r}}\right) \\ = \overline{\Theta}_i, \quad i \in \{1, \dots, \ell_\Psi - 1\},$$

and

$$\underline{\Theta}_{\ell_\Psi} = \frac{1}{\overline{r}} \leqslant \Theta_{\ell_\Psi} \leqslant \frac{1}{\underline{r}} = \overline{\Theta}_{\ell_\Psi}.$$

Then,

$$f(\Theta | V_0, L_0, U_0, \nu_0) = \prod_{i=1}^{\ell_\Psi} \mathcal{U}_\Theta\left(\frac{L_{0;i} + U_{0;i}}{2}, \frac{U_{0;i} - L_{0;i}}{2}\right), \tag{20}$$

where $L_0 = \underline{\Theta}$, $U_0 = \bar{\Theta}$, $V_0 = \mathbf{I}_{(\ell_\Psi)}$ and $\nu_0 = 0$.

To be formally compatible with (11), the prior pdf can be expressed by the characteristic function

$$f(\Theta | V_0, L_0, U_0, \nu_0) = \frac{\Theta_{\ell_\Psi}^{\nu_0} \chi_\Theta(L_0 \leqslant V_0 \Theta \leqslant U_0)}{\mathrm{J}(V_0, L_0, U_0, \nu_0)}.$$

3.5 | Point parameter estimates and output prediction

For technical reasons, we split V (11) into the following blocks:

$$V = \begin{bmatrix} V_\psi & V_y \\ \mathbf{0}'_{(\ell_\Psi - 1)} & v \end{bmatrix}. \tag{21}$$

The normalising factor (12) then has the form

$$\mathrm{J}(V, L, U, \nu) = \frac{1}{|\det(V_\psi)|} \frac{1}{v^{\nu+1}} \frac{U_{\ell_\Psi}^{\nu+1} - L_{\ell_\Psi}^{\nu+1}}{\nu + 1} \prod_{i=1}^{\ell_\Psi - 1} (U_i - L_i), \tag{22}$$

where L_i and U_i are the i-th entries of L and U, respectively.

The expected values of ϑ and r in (1), (2) are, with respect to (21),

$$\begin{aligned} \begin{bmatrix} \hat{\vartheta} \\ \hat{r} \end{bmatrix} &= \mathrm{E}\left(\begin{bmatrix} \vartheta \\ r \end{bmatrix} \middle| V, L, U, \nu \right), \\ \hat{r} &= v \quad \frac{\nu + 1}{\nu} \frac{1 - \gamma^\nu}{1 - \gamma^{\nu+1}} \quad U_{\ell_\Psi}^{-1}, \qquad \gamma = \frac{L_{\ell_\Psi}}{U_{\ell_\Psi}} < 1, \\ \hat{\vartheta} &= V_\psi^{-1} V_y - \hat{r}\, V_\psi^{-1} [\mathbf{I}_{(\ell_\Psi - 1)}, \mathbf{0}] \frac{U + L}{2}, \end{aligned} \tag{23}$$

where $[\mathbf{I}_{(\ell_\Psi - 1)}, \mathbf{0}]$ denotes $\ell_\Psi - 1$ rows of unit ℓ_Ψ matrix. For computational details, see Appendix A.2.

The result provides immediately point output prediction as $\hat{\vartheta}'\psi$. Higher moments of parameters can be evaluated similarly and serve for evaluation of moments of the predicted output. Note that very similar formulas are obtained in the previous study by Kárný and Pavelková.[12] The substantial difference consists in the construction of V_t and in the way of its updating.

3.6 | Time-varying parameters—outline of adaptivity

The proposed estimation algorithm assumes the constant parameters. If the parameters vary slowly, we can apply in the estimation process a linear forgetting[19] which models a time evolution of the parameters. Using a forgetting factor $0 \ll \lambda < 1$, the update of statistics (13) changes to

$$(\tilde{V}_t, \tilde{L}_t, \tilde{U}_t, \tilde{\nu}_t) = \mathcal{T}_t(\lambda V_{t-1}, L_{t-1}, U_{t-1}, \nu_{t-1}). \tag{24}$$

Geometrically, it means that the respective polytope expands linearly by $1/\lambda$ to adjust its shape for intersection with new strips of different direction or different width, given by newly observed data.

The topic of adaptivity has a potential of treating eg outliers, however, it requires analysis more accurate than this outline.

4 | ALGORITHMIC SUMMARY

Initialisation

- Set $t = 0$, $T > 0$.
- Choose bounds (19) and compute V_0, L_0, U_0, and ν_0 according to (20).

On-line phase

1. Set $t \leftarrow t + 1$.
2. Construct the data vector Ψ_t (4) from the currently processed data.
3. Update statistics V_{t-1}, L_{t-1}, U_{t-1}, and ν_{t-1} to $\tilde{V}_t$, $\tilde{L}_t$, $\tilde{U}_t$, and $\tilde{\nu}_t$ according to (13) (or (24) when forgetting is applied).
4. Convert the statistics to the **[− 1, 1]** form (18).
5. Approximate $\tilde{W}_t$, $\tilde{C}_t$, and $\tilde{\nu}_t$ by W_t, C_t, and ν_t as follows:
   - tighten the respective strips, see Sec. 3.3,
   - remove 1 strip (except of the strip corresponding to the last row of $\tilde{W}_{t-1}$) so that volume of the respective parallelotope is minimal,
   - assign $\nu_t = \min(\tilde{\nu}_t, \ell_\Psi)$.
6. Convert the statistics to the **[L, U]** form (11) and compute $\hat{\vartheta}$, $\hat{r}$ (23).
7. If $t < T$, go to 1.

5 | EXPERIMENTS

In this section, we present an example with simulated data to compare the proposed algorithm with the uniform estimator presented in the previous study by Kárný and Pavelková[12] and with the least squares (LS) estimator that assumes normal noise. For the purpose of this paper, the proposed algorithm will be called LUP (linear uniform with parallelotopes). We want to demonstrate the performance of LUP estimator with the noise distribution it was designed for, and LS estimator, designed for normal noise.

5.1 | Simulation setup

The simulated system is represented by a second order ARX model (1), which is described by the following equation with $\psi = [y_{t-1},\ y_{t-2},\ u_t,\ u_{t-1}]'$, $\vartheta = [1.81, -0.83, -0.3,\ 0.1]'$, $t \in t^\star = \{1, 2, \dots, 200\}$:

$$y_t = 1.81 y_{t-1} - 0.83 y_{t-2} - 0.3 u_t + 0.1 u_{t-1} + \varepsilon_t,$$

where noise term ε_t is uniformly distributed with half-width $r = 10^{-3}$, ie, $\varepsilon_t \sim \mathcal{U}_\varepsilon(0, 10^{-3})$. The system is stimulated by a Gaussian random walk, $(u_t - u_{t-1}) \sim \mathcal{N}(0, 0.1)$.

The prior pdf for all the estimators was chosen flat to express ignorance of the estimators on the parameters' values.

The experiments are performed using Matlab*.

5.2 | Performance criteria

The evaluation of the estimation quality and the comparison of the LUP estimator with other estimators exploits the following criteria:

*Matlab is a numerical computing environment and fourth-generation programming language, see www.mathworks.com/products/matlab.

- convergence speed, ie, number of estimation (time) steps until the estimated value approaches the true value by less than 3 %,
- relative estimation error $\delta(\vartheta_{i;t})$ is defined as follows

$$\delta(\vartheta_{i;t}) = \frac{\hat{\vartheta}_{i;t} - \vartheta_{i;t}}{\vartheta_{i;t}}, \vartheta_{i;t} \neq 0,$$

where $\vartheta_{i;t}$, $i \in \{1, 2, 3, 4\}$, is the simulated value of the i-th regression coefficient, at time t, having its estimate $\hat{\vartheta}_{i;t}$.

5.3 | Results

The estimation started at $t = 3$ because of the model order. The estimates are stored for $t > 10$, ie, after the values stabilised (see below). The plots start at $t = 8$.

5.3.1 | Constant regression coefficients

Here, the parameters of the simulated system were estimated. No time evolution of the parameters was considered, ie, no forgetting applied.

First, we focused on convergence properties of the proposed algorithm. Table 1 shows number of time steps (cycles) to approach the estimates to the true values closer than 3 %. The cycles are counted from the beginning of the task, ie, from $t = 3$. Results of the LUP algorithm are in the first column.

Figure 1 shows the estimates of ϑ_3 and ϑ_4 obtained with the LUP algorithm (left) and an LS estimator (right) for $t \leqslant 100$.

Table 2 summarises mean value, median, and standard deviation of the relative estimation error $\delta(\vartheta_{i;t})$, $i = \{1, 2, 3, 4\}$, $t = \{3, \dots, 200\}$. Only the LUP algorithm (left part) and the LS estimator (right part) are included,

5.3.2 | Time-varying regression coefficients

Data for this task were simulated with $T = 200$. A time step increment magnitude was $\pm 10^{-2}$ for ϑ_1 and ϑ_2 and $\pm 5 \cdot 10^{-3}$ for ϑ_3 and ϑ_4, as shown in Figure 2. The forgetting factor for both the proposed algorithm (left) and the LS estimator (right) was set to $\lambda = 0.9$.

TABLE 1 Number of estimation steps to approach the true parameter value by less than 3 %. Results of the LUP algorithm are in the first column

Parameter	Algorithm assuming uniform noise (LUP)	Algorithm assuming normal noise (LS)	Estimator (see previous study by Kárný and Pavelková)[12]
ϑ_1	7	42	1 308
ϑ_2	9	68	3 570
ϑ_3	5	152	773
ϑ_4	9	> 200	9 954

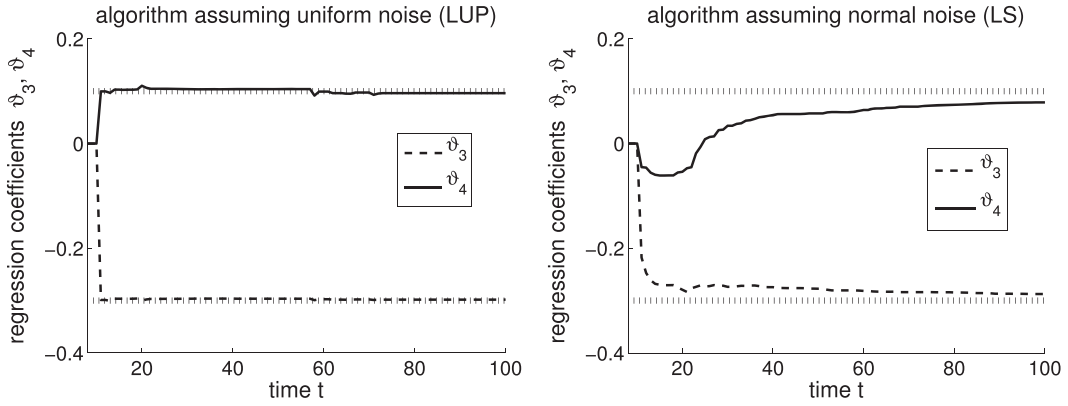

FIGURE 1 Estimates of constant parameters ϑ_3 and ϑ_4 obtained by the LUP algorithm (left) and the LS algorithm (right); the true values are represented by dotted lines. LUP, linear uniform with parallelotopes. LS, least squares

TABLE 2 Relative estimation errors of the parameters for linear uniform with parallelotopes and least squares algorithms. Results of the linear uniform with parallelotopes algorithm are in the first three columns

	algorithm assuming uniform noise			algorithm assuming normal noise		
	mean	median	std.dev.	mean	median	std.dev.
ϑ_1	0.0015	0.0007	0.0030	0.0155	0.0041	0.0351
ϑ_2	− 0.0031	− 0.0014	0.0064	0.0400	0.0081	0.1059
ϑ_3	0.0066	0.0062	0.0050	0.0503	0.0424	0.0257
ϑ_4	0.0060	0.0072	0.0037	0.2470	0.2350	1.8317

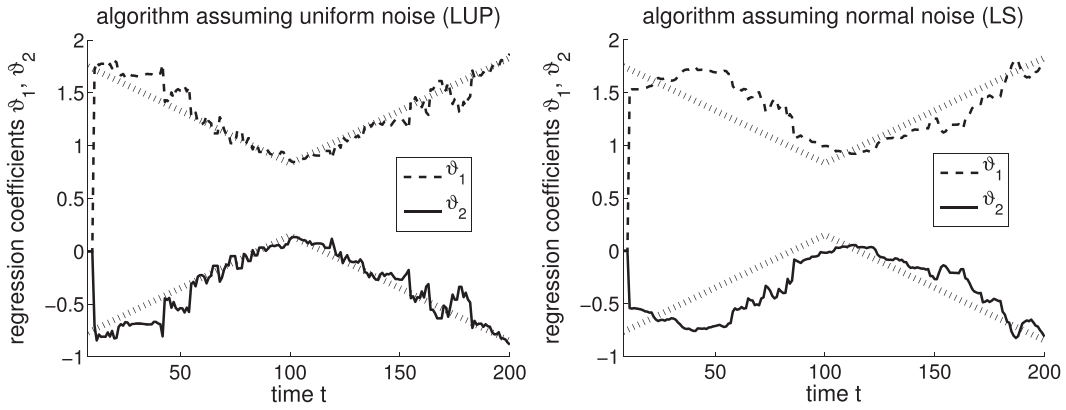

FIGURE 2 Estimates of time-varying parameters ϑ_1 and ϑ_2 obtained by the LUP algorithm (left) and the LS algorithm (right); the true values are represented by dotted lines, $\lambda = 0.9$. LUP, linear uniform with parallelotopes. LS, least squares

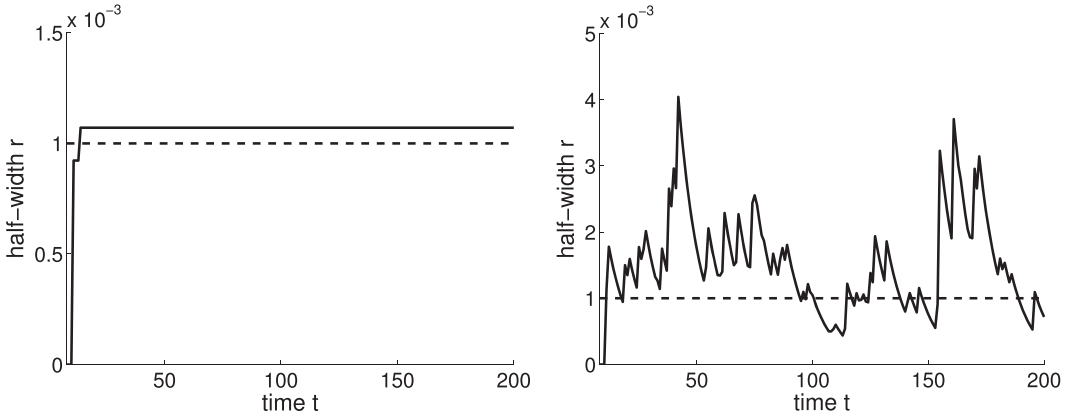

FIGURE 3 Estimates of noise half-width r by the proposed algorithm in the case of constant (left) and variable (right, $\lambda = 0.9$) regression coefficients; the true values are represented by dashed lines

5.3.3 | Noise

Further, the noise term r was estimated by the proposed algorithm, as presented in Figure 3. The left part shows estimates of r with regression coefficients constant (see Section 5.3.1, without forgetting), the right part shows estimates of r in the case of variable regression coefficients (see Section 5.3.2, with forgetting), while the value of r was kept constant.

5.4 | Discussion

Because of a flat prior, the transient period before stabilisation of the estimates is not shown. An informative prior pdf improves the estimation significantly, but we assumed ignorance of the parameters' value for the estimators.

Convergence of the estimates by the proposed (LUP) algorithm, as seen in Table 1, is very fast. The algorithm based on LS converges much slower, even if forgetting is employed (not shown). Finally, convergence of the algorithm with orthogonal rotation[12] is more than 100 × slower. This was actually the main motivation for this work.

In the case of constant regression coefficients, the LUP algorithm outperforms the LS (only a very informative prior for the LS algorithm improves its estimates to the level of the LUP algorithm). Because of the nature of the LUP algorithm (accepting/rejection of the data vector to modify the existing parallelotope), the estimates may exhibit slight jumps, although the true value has been already reached with a high precision, see eg Figure 1 left, ϑ_4. However, the estimation errors are by an order of magnitude lower than with the LS algorithm, see Table 2.

Time-varying regression coefficients are, again, estimated better with the LUP algorithm, see Figure 2. The estimation errors are not quantified because the difference is obvious. As in the constant case, even forgetting does not improve the estimation quality of the LS algorithm.

The noise half-width r, Figure 3, with regression coefficients either constant (left part, no forgetting) or varying (right, forgetting), is systematically overestimated. The reason may be in the special structure of the matrix V last row, see (21), and nonuniform marginal pdf of the half-width r (or Θ_{ℓ_ψ}, see, eg, (11)). For other systems, it is possible that r will be underestimated, but this property was not methodically tested. Anyway, the estimates are close to the true values. Variable noise estimation gives similar results.

The LS estimator can be used as an acceptable tool for data with uniform noise, but the LUP estimator performs better in the example shown here.

6 | CONCLUSIONS

We present a recursive LUP algorithm for the parameter estimation of a linear stochastic autoregressive model with an external input and the noise described by a uniform distribution.

The model parameters are estimated using Bayesian methodology. Because of nonexistence of finite sufficient statistic, we introduce such a statistic by a suitable approximation of the posterior pdf.† Geometrically, we search for a parallelotope of a minimum volume that circumscribes the given polytope in parameter space. The approximation is, however, suboptimal because we are technically limited by a special form of the statistic.

Our main contribution is that we include unknown noise parameter into estimation and we provide posterior pdfs, including evaluation of first moments of parameters. In comparison to the previous solution,[12] the convergence of the estimates was rapidly accelerated while keeping simplicity and speed of the algorithm. We outlined a capability of tracking slow changes of parameters' values by the linear forgetting.

The predecessors of the LUP, mentioned in the Introduction, are reported in both the previous studies by Kárný and Pavelková.[11,12] The LUP algorithm and the previous study by Kárný and Pavelková[12] were tested on simulation data only, and LUP outperforms the latter one. The maximum a posteriori probability approach[11] yielded reasonable results using both simulated and real traffic data. Other experiments with real data from the area of a cold rolling mill using model with uniform noise were made by the authors in both the previous studies by Pavelková and Jirsa.[20,21] Our experience with application of uniform estimators on real industrial data make us assume reasonable behaviour of the LUP, too, but this is a topic for further studies.

Analysis of the errors caused by the approximation is not a part of this paper, inspection of this type would be worth considering. As for a practical point, the LUP algorithm is vulnerable to outliers, which is another important topic for future work. This may correspond to systematic choice of the forgetting factor λ,[16] which, in our experiments, was determined by experience. Also, on-line resolving of potential conflicts between user-given parameter bounds ($\underline{\vartheta}$, $\overline{\vartheta}$, $\underline{r}$, $\overline{r}$) and the parallelotope bounds (particularly for time-varying parameters and/or forgetting) is a related and desirable task (we tested only $\underline{r}$ for positivity). Another potentially perspective topic would be introducing a directional forgetting on parallelotopes.

ACKNOWLEDGMENT

This research was partially supported by the grant MŠMT 7D12004 (E!7262 ProDiSMon). The authors also give their thanks to Dr. Miroslav Kárný for his valuable consultations.

†The statistics are sufficient for the chosen class of approximating functions (11), (21).

REFERENCES

1. Kárný M, Böhm J, Guy TV, et al. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. London: Springer, 2005.
2. Bobál V, Böhm J, Fessl J, Machácek J. *Digital Self-tuning Controllers: Algorithms, Implementation and Applications*. London: Springer Science & Business Media; 2006.
3. Young PC. *Recursive Estimation and Time-Series Analysis: An Introduction for the Student and Practitioner*. Berlin Heidelberg: Springer, 2011.
4. Haykin SS. *Adaptive Filter Theory*. New Delhi: Pearson Education India; 2008.
5. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis* (3rd edn.) Boca Raton: Chapman & Hall/CRC; 2014.
6. Milanese M, Belforte G. Estimation theory and uncertainty intervals evaluation in presence of unknown but bounded errors - Linear families of models and estimators. *IEEE T Automat Contr.* 1982;27(2):408–414.
7. Cerone V. Feasible parameter set for linear models with bounded errors in all variables. *Automatica.* 1993;29(6):1551–1555.
8. Polyak BT, Nazin SA, Durieu C, Walter E. Ellipsoidal parameter or state estimation under model uncertainty. *Automatica.* 2004;40(7):1171–1179.
9. Vicino A, Zappa G. Sequential approximation of feasible parameter sets for identification with set membership uncertainty. *IEEE T Automat Contr.* 1996;41(6):774–785.
10. Ninness BM, Goodwin GC. Rapprochement between bounded-error and stochastic estimation theory. *Int J Adapt Control.* 1995;9:107–132.
11. Kárný M, Pavelková L. Projection-based Bayesian recursive estimation of ARX model with uniform innovations. *Syst Control Lett.* 2007;56(9/10):646–655.
12. Kárný M, Pavelková L. Approximate Bayesian recursive estimation of linear model with uniform noise. *Preprints of the 16th IFAC Symposium on System Identification Sysid 2012*, Brussels, Belgium, July 11–13,2012; 1803–1807.
13. Kárný M. Parametrization of multi-output multi-input autoregressive-regressive models for self-tuning control. *Kybernetika* 1992;28(5):402–412.
14. Bernardo JM, Smith AFM. *Bayesian Theory* (2nd edn.) Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons; 1997.
15. Andrýsek J. Approximate recursive Bayesian estimation of dynamic probabilistic mixtures. *Multiple Participant Decision Making*. Workshop on Computer-Intensive Methods in Control and Data Processing, Prague, Czech Rep 2004;39–54.
16. Kárný M. Approximate Bayesian recursive estimation. *Inform Sciences.* 2014;285(1):100–111.
17. Kullback S, Leibler R. On information and sufficiency. *Ann Math Stat.* 1951;22:79–87.
18. Bernardo JM. Expected information as expected utility. *Ann Stat.* 1979;7(3):686–690.
19. Ljung L. *System Identification: Theory for the User*. London: Prentice-Hall, 1987.
20. Pavelková L, Jirsa L. Evaluation of sensor signal health using model with uniform noise. *Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, Vienna, Austria, 2014;671–677.
21. Jirsa L, Pavelková L. Estimation of uniform static regression model with abruptly varying parameters. *Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*. SCITEPRESS Science and Technology Publications, Colmar, France, 2015;603–607.

How to cite this article: Pavelková L, Jirsa L. Recursive Bayesian estimation of autoregressive model with uniform noise using approximation by parallelotopes. *Int J Adapt Control Signal Process*. 2017. https://doi.org/10.1002/acs.2756

APPENDIX

The results given below hold for the regular V separated to the blocks as in (21), with the last row zero except of the diagonal element.

A.1 Minimisation of $\ln \mathrm{J}$

Minimum of $\ln \mathrm{J}(V_t, L_t, U_t, \nu_t)$ (16) and $\mathrm{J}(V_t, L_t, U_t, \nu_t)$ (22) is achieved at the same point. The way to express the given admissible set of the parameter Θ in (11) in the **[L, U]** form is not unique. If the inequality within the characteristic function is left-multiplied by a diagonal matrix D_t with positive entries, the described set of Θ (ie, parallelotope) is unchanged. If $D_{ii;t} = \frac{2}{U_{i;t} - L_{i;t}}$ (see (18)) and if we denote $W = DV$, $\mathcal{L} = DL = C - \mathbf{1}_{(\ell_\Psi)}$ and $\mathcal{U} = DU = C + \mathbf{1}_{(\ell_\Psi)}$, then the difference of the multiplied bounds $\mathcal{U} - \mathcal{L} = 2 \times \mathbf{1}_{(\ell_\Psi)}$. Volume of the parallelotope then equals $2^{\ell_\Psi} / |\det W_t|$.

It can be easily shown that $\mathrm{J}(\mathrm{V}, \mathrm{L}, \mathrm{U}, \nu) = \mathrm{J}(\mathrm{W}, \mathcal{L}, \mathcal{U}, \nu)$. This term can be separated in several parts: $1/|\det W|$ (using Laplace expansion), which is proportional to the parallelotope volume, $\prod_{i=1}^{\ell_\Psi - 1}(\mathcal{U}_i - \mathcal{L}_i)$, which equals a constant term $2^{\ell_\Psi - 1}$, and $(\mathcal{U}_{\ell_\Psi}^{\nu+1} - \mathcal{L}_{\ell_\Psi}^{\nu+1})/(W_{\ell_\Psi,\ell_\Psi}^{\nu}(\nu + 1))$, concerning the last row of W. The posterior pdf must be kept in the class (11). Therefore, when new data come (13), the last row of W is only tightened around the new polytope and then kept untouched. In other words, the scaling factor $W_{\ell_\Psi \ell_\Psi}$ is maximised to minimise the range of Θ_{ℓ_Ψ}. The bounds and their powers $\mathcal{U}_{\ell_\Psi}^{\nu+1} - \mathcal{L}_{\ell_\Psi}^{\nu+1}$ are given by the data history.

Hence, minimisation of the Kullback-Leibler divergence represented by the term $\ln \mathrm{J}(V_t, L_t, U_t, \nu_t)$ is equivalent to minimisation of the volume of a circumscribing parallelotope, with respect to the class of pdfs defined in (11), for the special form of the matrix V (21) and the data update algorithm. For a general matrix and the circumscription algorithm as described in the previous study by Vicino and Zappa,[9] minima of Kullback-Leibler divergence and the parallelotope volume would differ because of the powers ν in the approximate posterior pdf (11).

A.2 Computation of $\mathrm{J}(V, L, U, \nu)$ and point estimates ϑ, r

The integration uses Fubini theorem, Laplace expansion and the substitution $x = V\Theta$, which has Jacobian equal to $|\det(V)| = |\det(V_\psi)v|$ and it holds $x_{\ell_\Psi} = v\Theta_{\ell_\Psi}$.

The normalizing factor J is evaluated in (22).

The expectation of $r = 1/\Theta_{\ell_\Psi}$ is obtained as follows

$$\begin{aligned}
\hat{r} &= \frac{\int_{\Theta^\star} \Theta_{\ell_\Psi}^{\nu-1} \chi(L \leqslant V\Theta \leqslant U)\mathrm{d}\Theta}{\mathrm{J}(V, L, U, \nu)} \\
&= \frac{\int_{x^\star} \left(\frac{x_{\ell_\Psi}}{v}\right)^{\nu-1} \chi(L \leqslant x \leqslant U)|\det(V)|^{-1}\mathrm{d}x}{\mathrm{J}(V, L, U, \nu)} \\
&= \frac{\mathrm{J}(V, L, U, \nu - 1)}{\mathrm{J}(V, L, U, \nu)} = v\frac{\nu + 1}{\nu}\frac{U_{\ell_\Psi}^{\nu} - L_{\ell_\Psi}^{\nu}}{U_{\ell_\Psi}^{\nu+1} - L_{\ell_\Psi}^{\nu+1}} \\
&= v\frac{\nu + 1}{\nu}\frac{1 - \gamma^{\nu}}{1 - \gamma^{\nu+1}}U_{\ell_\Psi}^{-1}
\end{aligned}$$

with $\gamma = \frac{L_{\ell_\Psi}}{U_{\ell_\Psi}} < 1$.

The inversion $V^{-1} = \begin{bmatrix} V_\psi^{-1} & -\frac{V_\psi^{-1}V_y}{v} \\ \mathbf{0}'_{(\ell_\Psi - 1)} & 1/v \end{bmatrix}$, see (21). Let $[\mathbf{I}_{(\ell_\Psi - 1)}, \mathbf{0}]$ denote $\ell_\Psi - 1$ rows of unit ℓ_Ψ matrix. According to (5), the expectation of $\vartheta = -[\mathbf{I}_{(\ell_\Psi - 1)}, \mathbf{0}]\Theta/\Theta_{\ell_\Psi}$ equals

$$\begin{aligned}
\hat{\vartheta} &= -\frac{\int_{\Theta^\star} [\mathbf{I}_{(\ell_\Psi - 1)}, \mathbf{0}]\Theta\Theta_{\ell_\Psi}^{\nu-1} \chi(L \leqslant V\Theta \leqslant U)\mathrm{d}\Theta}{\mathrm{J}(V, L, U, \nu)} \\
&= -\frac{\int_{x^\star} \left[V_\psi^{-1}, -\frac{V_\psi^{-1}V_y}{v} \right] x x_{\ell_\Psi}^{\nu-1} \chi(L \leqslant x \leqslant U)\mathrm{d}x}{\mathrm{J}(V, L, U, \nu)} \\
&= V_\psi^{-1}V_y - \hat{r}V_\psi^{-1}[\mathbf{I}_{(\ell_\Psi - 1)}, \mathbf{0}]\frac{U + L}{2}.
\end{aligned}$$