Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Ladislav Jirsa, Lenka Pavelková

Normal and uniform noise — violation of the assumption on noise distribution in model identification

No. 2348 January 2015

ÚTIA AVČR, v.v.i., P.O.Box 18, 182 08 Prague, Czech Republic
Fax: (+420)286890378
http://www.utia.cas.cz
E-mail: utia@utia.cas.cz

Abstract

Mathematical modelling under uncertainty together with the field of applied statistics represent tools useful in many practical domains. Widely accepted assumption of normal (Gaussian) noise has created the basis for theoretical and algorithmic solutions of respective tasks. However, many continuous variables are strictly bounded and their uncertainty may have origin in various physical processes which causes a non-normal distribution of their noise. Furthermore, adaptation of algorithms based on normal model for identification of models with bounded noise can distort the estimates due to inconsistent handling of uncertainty. This report describes a study to compare results of estimation algorithms based on assumption of normal and uniform noise. Data sequences processed by the algorithms have normal noise bounded by a low limit with respect to standard deviation. We illustrate disparity between noise assumption and a true noise distribution and its influence on the quality of the estimates. It is a part of an effort to develop theory and fast algorithms for estimation with bounded noise, applicable in practice.

Keywords: uncertainty; bounded variable; uniform noise; model identification; assumption of normal noise; estimation comparison

Contents

1 Introduction

Many processes in practice generate data that are not normally distributed. For example, maximum and minimum value on a floating window are beta-distributed, radioactive decay is described by Poisson distribution [4], [6]. Other tasks involve strictly bounded variables, e. g. queue length in traffic [10], controlling of robots [2], etc. Model parameters in many practical tasks are hard constrained (bounded), because of either their physical nature (positive mass, length) or technological or other reasons (limited car velocity).

Probabilistic methodology, suitable estimation and prediction in such tasks must be chosen so that it respects nature of the data being processed. Among bounded pdfs, uniform pdf has been chosen for a research in estimation and control topics; theory and fast algorithms appear to be necessary for applications. Despite of a partial success in this effort, e. g. [8] or [9], the results, as mentioned below, are not satisfactory for practice.

In this report, an experiment is presented to demonstrate importance of corresponding and suitable theoretical and algorithmic tools with respect to the data nature.

This report aims to compare two estimation algorithms: (i) standard least square (LS) estimator based on Gaussian model [5] and (ii) estimator based on linear uniform (LU) model [8] under condition that data/noise is bounded.

2 Theory

From now on, pdf means probability density function, x^* is a set of possible values of x, $t \in t^* = \{1, \ldots, T\}$ is a discrete time label from a finite set. Observable outputs y_t and system inputs u_t form data $d_t = (y_t, u_t)$, data history is $d(t-1) \equiv (d_{t-1}, d_{t-2}, \ldots, d_1)$ and $d(0)$ represents prior information. Vector ψ_t is a finite-dimensional regression vector composed of past observed data and current input in a known (recursively implementable) way. ϑ is a vector of unknown regression coefficients, $\mathcal{U}_y(\mu, r)$ is a uniform pdf of y given by the expectation μ and the half-width $r > 0$, $\mathcal{N}_y(\mu, r)$ is a normal pdf of y with the expectation μ and variance r. Θ is used to denote all unknown parameters, e.g. $\Theta = (\vartheta, r)$. Symbol $\propto$ means proportionality up to a constant, $'$ is a transposition.

2.1 Calculus with pdfs

Let us consider the joint pdf $f(a, b, c)$. For any $(a, b, c) \in (a, b, c)^*$, the following relations between pdfs hold [7]:

Chain rule

$$f(a, b|c) = f(a|b, c)\, f(b|c) = f(b|a, c)\, f(a|c)$$

Marginalization

$$f(b|c) = \int f(a, b|c)\ \mathrm{d}a,$$

Bayes rule

$$\begin{aligned} f(b|a, c) &= \frac{f(a|b, c)\, f(b|c)}{f(a|c)} = \frac{f(a|b, c)\, f(b|c)}{\int f(a|b, c)\, f(b|c)\ \mathrm{d}b} \\ &\propto f(a|b, c)\, f(b|c). \end{aligned} \tag{1}$$

2.2 Basics of Bayesian learning

Using Bayesian approach, the system is described by probability density functions [7].

Generally, for $t \in t^*$, a system of interest can be characterized by data $d_t = (y_t, u_t)$ and internal quantities X_t that are never observed directly but influence the output and consist of system states x_t and/or model parameters Θ. We will focus on model parameters.

We assume that natural conditions of control [7] hold and that Θ contains entire information about the past evolution of the system. Then, based on the probabilistic description, the pdf $f(d_t, \Theta_t|d(t-1), \Theta(t-1))$ describing both the observed and internal quantities is constructed from the following elements:

observation model

$$f(y_t|u_t, d(t-1), \Theta_t) \tag{2}$$

time evolution model

$$f(\Theta_t|u_t, d(t-1), \Theta_{t-1}) \tag{3}$$

input generator / controller

$$f(u_t|d(t-1)). \tag{4}$$

To obtain estimates of Θ_t, pdf $f(\Theta_t|u_t, d(t-1))$ has to be computed. Bayesian filtering (estimation) solves the evolution of the pdf $f(\Theta_t|u_t, d(t-1))$, i.e. removes Θ from the condition in the time evolution model (3). Estimation requires knowledge of the models (2) and (3).

The evolution of the pdf $f(\Theta_t|u_t, d(t-1))$, called Bayesian estimation of unknown parameter Θ_t, is described by the following recursion that starts from the prior pdf $f(\Theta_0)$:

- data update

$$\begin{aligned} f(\Theta_t|d(t)) &= \frac{f(y_t|u_t, d(t-1), \Theta_t)\ f(\Theta_t|u_t, d(t-1))}{f(y_t|u_t, d(t-1))} \\ &\propto f(y_t|u_t, d(t-1), \Theta_t)\ f(\Theta_t|u_t, d(t-1)) \end{aligned} \tag{5}$$

  that incorporates information on the output y_t and the input u_t, and

- time update

$$f(\Theta_{t+1}|u_{t+1}, d(t)) = \int f(\Theta_{t+1}|u_{t+1}, d(t), \Theta_t)\ f(\Theta_t|d(t))\ \mathrm{d}\Theta_t \tag{6}$$

  that reflects the time evolution $\Theta_t \to \Theta_{t+1}$.

If the time evolution model (3) is unknown, it can be approximated e. g. by the exponential forgetting [11] in the time update step

$$f(\Theta_{t+1}|u_{t+1}, d(t)) \ \propto\ (f(\Theta_t|d(t)))^{\lambda}, \tag{7}$$

where λ is a forgetting factor, $0 \ll \lambda \leq 1$.

The (outer) model of the system $f(y_t|u_t, (t-1))$ obtained by filtering is called predictive pdf.

The described Bayesian estimation combines prior information in $f(\Theta_0)$, theoretical knowledge described by $f(y_t|u_t, d(t-1), \Theta_t)$, $f(\Theta_t|u_t, d(t-1), \Theta_{t-1})$ and observed data $d(t) = (y(t),\ u(t))$ by using deductive rules of the calculus with pdfs.

3 Models

We model the system with a scalar output y_t at discrete time t. The considered parametric model

$$f(y_t|u_t, d(t-1), \vartheta, r) \quad \equiv \quad f(y_t|\psi_t, \vartheta, r). \tag{8}$$

This pdf describes the ARX model with uniform white noise. The ARX model (8) can be expressed alternatively as

$$y_t - \psi_t'\vartheta \quad = \quad \varepsilon_t, \tag{9}$$

where ε_t is a white noise at the time t (zero mean, mutually uncorrelated and uncorrelated with older observation).

3.1 Gaussian model

Models assuming normal noise are highly elaborated with advanced theoretical and algorithmic results.

3.1.1 Definition

The considered parametric model (8) takes the form

$$f(y_t|\psi_t, \vartheta, r) \equiv \mathcal{N}_{y_t}(\psi_t'\vartheta, r) = (2\pi r)^{-\frac{1}{2}} \exp\left[-\frac{(y_t - \psi_t'\vartheta)^2}{2r}\right]. \tag{10}$$

3.1.2 Estimation

Estimation of the Gaussian model is performed using sufficient statistcs V_t and ν_t data-updated according to

$$V_t = V_{t-1} + \Psi_t \Psi_t', \tag{11}$$
$$\nu_t = \nu_{t-1} + 1, \tag{12}$$

where $\Psi_t = [y_t, \psi_t']'$ is the extended data vector. Time update can be represented by the stabilized exponential forgetting in the way

$$V_{t|t-1} = \lambda V_{t-1|t-1} + (1-\lambda) V_A, \tag{13}$$
$$\nu_{t|t-1} = \lambda \nu_{t-1|t-1} + (1-\lambda) \nu_A, \tag{14}$$

where V_A and ν_A are alternative (stabilizing) statistics, potentially zero. Symbol $V_{t|t-1}$ means statistic in time t conditioned by the data up to time $t-1$.

Matrix V_t is extended information matrix and it can be decomposed as a product $V = L'DL$, where L is a lower triangular matrix with unit diagonal and D is a diagonal matrix. This decomposition is used due to numerical stability and computational comfort. The matrices L and D can be updated directly without use of V. The matrices L and D can be further decomposed into blocks

$$L = \begin{bmatrix} 1 & 0 \\ L_{d\psi} & L_\psi \end{bmatrix}, \qquad D = \begin{bmatrix} D_d & 0 \\ 0 & D_\psi \end{bmatrix}, \tag{15}$$

where D_d is scalar.

Other details concerning algorithmic aspects, prior information etc. can be found e. g. in [7].

Standard estimation: The moments of the parameters' posterior pdfs are given by

$$\mathcal{E}[\vartheta | L, D, \nu] = L_\psi^{-1} L_{d\psi} \equiv \hat{\vartheta}, \tag{16}$$
$$\mathcal{E}[r | L, D, \nu)] = \frac{D_d}{\nu - 2} \equiv \hat{r}, \tag{17}$$
$$\mathrm{cov}[\vartheta | L, D, \nu] = \frac{D_d}{\nu - 2} \, L_\psi^{-1} D_\psi^{-1} \left(L_\psi'\right)^{-1} \equiv \hat{r} C. \tag{18}$$

Constrained estimation: Some methods describing estimation using least squares under constraints are described e. g. in [1], [3] or [12], however, they have been not used for experiments in this report. The comparison is intended after the improved algorithms for estimation with uniform noise are finished.

3.2 Uniform model

This model is a special and simplest case of bounded models. However, even for this model, the estimation is nontrivial.

3.2.1 Definition

The considered parametric model (8) takes the form

$$f(y_t | \psi_t, \vartheta, r) \equiv \mathcal{U}_{y_t}(\psi_t' \vartheta, r) = \frac{\chi_{y_t}(-r \le y_t - \psi_t' \vartheta \le r)}{2r}, \tag{19}$$

where $\chi_x(\bullet)$ is value of the indicator of a set defined by the conditions $\bullet$ at the point x.

This pdf describes the ARX model with uniform white noise.

In the alternative formula (9), it holds for ε_t

$$|\varepsilon_t| \le r. \tag{20}$$

3.2.2 Estimation

Because of non-existence of a finite-dimensional sufficient statistics in case of estimation with uniform noise, the solution of the estimation problem is based on a suitable approximation of dimension-increasing data matrix by the one with limited dimension. In this way, obsolete or insignificant information is lost.

Algorithms optimizing the posterior pdf of unknown parameters by linear programming [8] give satisfactory MAP estimate but they yield no information about the estimate precision. The approximation based on information matrix rotation [9] results in faster algorithms but because of suboptimal approximation, convergence of estimates is much slower.

Currently, the work on theory and algorithm approximating the support of the posterior pdf in parameter space by parallelotopes is being commenced.

If the unknown parameters are changing in time, time evolution of the parameters is modelled by (3), or a way of forgetting (e. g. exponential (7)) is applied in unbounded cases. Methods of approximating estimation of uniform parameters are based on storing the most relevant information carried by the data. Some information is always lost (e. g. removing the data vector from the sliding window or circumscription of a polytope by a parallelotope in the parameter space). In this way, these methods inherently contain a way to "forget" information and allow parameters to evolve, although the time evolution model (3) is unknown.

When the parameters are estimated on a sliding data window of length Δ time steps, it corresponds to a forgetting factor λ, see Section 3.1.2, by the relationship $\Delta = 1/(1-\lambda)$.

4 Experiments

4.1 Simulation setup

For the experiments, we simulated the 2^{nd} order ARX model (8) with $\psi = [y_{t-1},\, y_{t-2},\, u_t,\, u_{t-1}]'$, $\vartheta = [1.81,\, -0.83,\, -0.3,\, 0.1]'$, $t \in \{1, 2, \ldots, 500\}$

$$y_t \;=\; 1.81y_{t-1} - 0.83y_{t-2} - 0.3u_t + 0.1u_{t-1} + e_t, \tag{21}$$

where noise terms e_t are normally distributed and truncated as follows

$$e_t = \min(c\,\sigma, \max(-c\,\sigma, \overline{e}_t)),$$

where $\overline{e}_t \sim \mathcal{N}(0, \sigma)$. We chose $c = 0.1$, $\sigma = 5$.

The system is stimulated by deterministic bi-level signal. Figure 1 presents simulated data, input and output.

4.2 Estimation

LU model parameters were estimated using the mentioned trial and initial version of parallelotope circumscription (see Section 3.2.2) because it was the fastest, the most stable and had a good convergence properties. However, as it can be seen on the figures, the estimation is not optimal because of theoretical issues still unresolved, which is a challenge for the work in the near future. Furthermore, there is no direct analogy to forgetting, when the sliding window is used, i.e. the forgetting is unknown.

Unknown parameters of the normal model were estimated according to the formulae (16)–(18). The forgetting factor λ could not be fitted to the rate of information loss in case of LU model parameters, as mentioned above. Therefore, several values of λ were employed to observe the behaviour of the approaches.

4.3 Performance criteria

The estimation quality and comparison of the LU estimator with the LS exploits the following criteria.

Absolute prediction error $\mathcal{E}(y_t)$ and its mean $\overline{\mathcal{E}}(y)$ are defined as follows

$$\mathcal{E}_t(y) \equiv |y_t - \hat{y}_t|,\ t \in t^*, \qquad \overline{\mathcal{E}}(y) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{E}_t(y), \tag{22}$$

where y_t is the simulated output and $\hat{y}_t$ is its prediction based on estimates gained from data observed up to time $t-1$.

Absolute error of parameter estimates $\mathcal{E}_t(\vartheta)$ and its mean $\overline{\mathcal{E}}(\vartheta)$ defined as follows

$$\mathcal{E}_t(\vartheta) \equiv |\vartheta - \hat{\vartheta}_t|,\ t \in t^*, \qquad \overline{\mathcal{E}}(\vartheta) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{E}_t(\vartheta), \tag{23}$$

where ϑ comprises the true parameter values $\hat{\vartheta}_t$ is the current parameter estimate.

Note that ϑ, $\mathcal{E}_t(\vartheta)$ and $\overline{\mathcal{E}}(\vartheta)$ are vectors.

4.4 Results

The time course of parameter estimates are shown in Figures 2–4.

The mean absolute prediction error and the mean absolute error of parameter estimates are summarised in Table 1. LU estimate is only one, LS estimates vary according to the forgetting factor λ.

	$\overline{\mathcal{E}}(y)$	$\overline{\mathcal{E}}(\vartheta)$			
LU	0.473	0.022	0.020	0.033	0.060
LS, $\lambda = 1$	0.478	0.029	0.029	0.043	0.039
LS, $\lambda = 0.99$	0.477	0.045	0.046	0.084	0.069
LS, $\lambda = 0.98$	0.479	0.064	0.067	0.105	0.090
LS, $\lambda = 0.97$	0.480	0.081	0.084	0.121	0.113
LS, $\lambda = 0.96$	0.483	0.096	0.098	0.138	0.136
LS, $\lambda = 0.95$	0.485	0.110	0.112	0.154	0.156
LS, $\lambda = 0.90$	0.497	0.173	0.175	0.216	0.222

Table 1: Mean predictive error $\overline{\mathcal{E}}(y)$ and mean error of parameter estimates $\overline{\mathcal{E}}(\vartheta)$ for LU and LS

4.5 Discussion

Two approaches of parameter estimation were applied on data with truncated (non-Gaussian) noise: estimation assuming normally distributed noise and estimation assuming uniformly distributed noise. Data were simulated by a linear auto-regressive model of 2^{nd} order with input. Mean absolute error of output predictions and parameter estimates was computed. Plots are shown as well.

For LU estimation, unfinished (and unpublished) initial version of algorithm for approximation of posterior support by parallelotope in parameter space was applied. For LS estimation, algorithms assuming normal noise were applied in square root version.

Because LU estimation is specific in the sense of non-existence of finite-dimensional sufficient statistic, the data matrix is approximated to keep its dimension constant. In this way, forgetting is inherent to LU estimation. On the other hand, its rate is unknown (not solved) in the algorithm used. Furthermore, neither adaptability has been solved yet, therefore the model has constant parameters.

To match forgetting properties for both LU and LS estimation, several estimations were run for LS, each with different forgetting factor λ.

As seen in Table 1 and Figures 2–4, LU estimats demonstrate better results than LS, despite of forgetting rate λ. Even, the more forgetting, the worse results of LS, which may correspond to the fact that the parameters are constant. The quality decrease is visible even in case of slight decrease of λ (e. g.

$1 \longrightarrow 0.99$). This property may also be a consequence of the input part of the model and the bi-level periodic input data, In any case, LU estimation dominates over LS.

Comparison of LS model under constraints and LU has not been done because of unfinished LU methodology. These experiments are intended after the LU estimation theory and algorithms are completed.

5 Conclusion

We presented an illustrative experiment to demonstrate importance of assuption made on noise nature, i. e. the probabilistic model used to describe the system od data, and its violation. This experiment was motivated by the fact that many variables or parameters are strictly bounded and because of frequent use of estimation algorithms assuming Gaussian noise, which is unbounded. Therefore, an effort to develop corresponding theory and algorithms has been started in the past.

The study shows that algorithms assuming bounded (particularly uniform) noise, although imperfect and unfinished, give better results than algorithms assuming normal noise. Definitely, the range of cases and conditions should have been much wider but, despite this simplification, the tools for appropriate and correct treatment of uncertainty, satisfying the theoretical assumptions, are necessary for practical applications.

The outlined direction is promising and seems to be correct.

References

[1] R. Arablouei and K. Dogancay. Reduced-Complexity Constrained Recursive Least-Squares Adaptive Filtering Algorithm. *IEEE Transactions on Signal Processing*, 60(12):6687–6692, 2012.

[2] K. Belda and D. Vošmik. Explicit generalized predictive algorithms for speed control of PMSM drives — Fast explicit form with field weakening and current limitation. In *Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society*. AIT Austrian Institute of Technology GmbH, 2013.

[3] S. Bellavia, J. Gondzio, and B. Morini. Computational experience with numerical methods for nonnegative least-squares problems. *Numerical linear algebra with applications*, 18(3):363–385, 2011.

[4] C. M. Fonseca. Bayesian estimation of the intensity of low-level radiation sources. *Jaderná energie*, 37:83–97, 1991.

[5] S. S. Haykin. *Adaptive filter theory*. Pearson Education India, 2008.

[6] L. Jirsa. *Advanced Bayesian Processing of Clinical Data in Nuclear Medicine. Ph.D. Thesis*. PhD thesis, FJFI ČVUT, 1999.

[7] 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.

[8] M. Kárný and L. Pavelková. Projection-based Bayesian recursive estimation of ARX model with uniform innovations. *Systems & Control Letters*, 56(9/10):646–655, 2007.

[9] M. Kárný and L. Pavelková. Approximate Bayesian recursive estimation of linear model with uniform noise. In *Preprints of the 16th IFAC Symposium on System Identification Sysid 2012*, pages 1803–1807, Brussels, Belgium, July 11 – 13 2012.

[10] L. Pavelková. Nonlinear Bayesian state filtering with missing measurements and bounded noise: its application to vehicle position estimation. *Kybernetika*, 47(3):370–384, 2011.

[11] V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Pergamon Press, Oxford, 1981.

[12] Y. Zhu and X. R. Li. Recursive least squares with linear constraints. *Commun. Inf. Syst.*, 7(3):287–312, 2007.

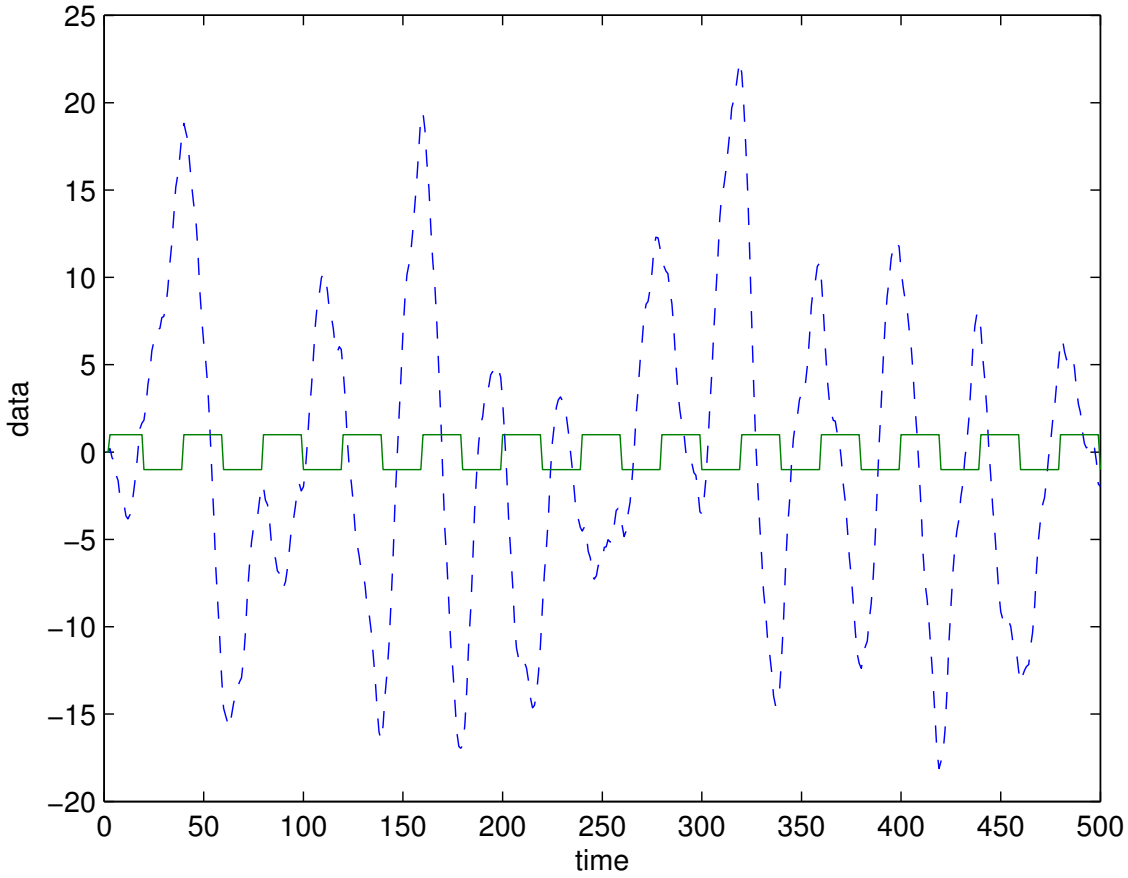

Figure 1: Simulated input (solid) and output (dashed).

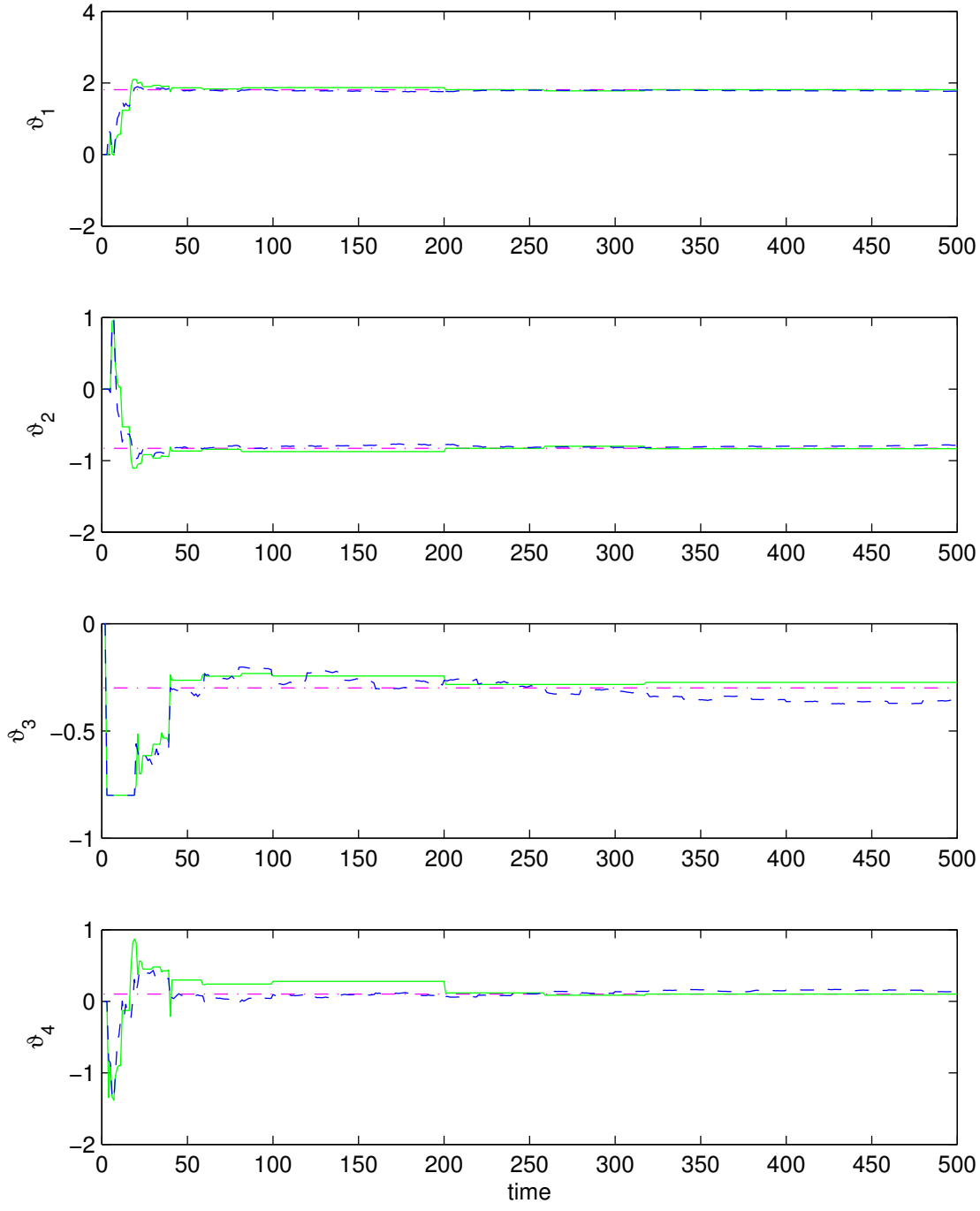

Figure 2: Time evolution of parameter estimates for both uniform (solid) and normal (dashed) estimators. The true parameter values are plotted in dash-dotted line, forgetting factor for the normal model $\lambda = 1$.

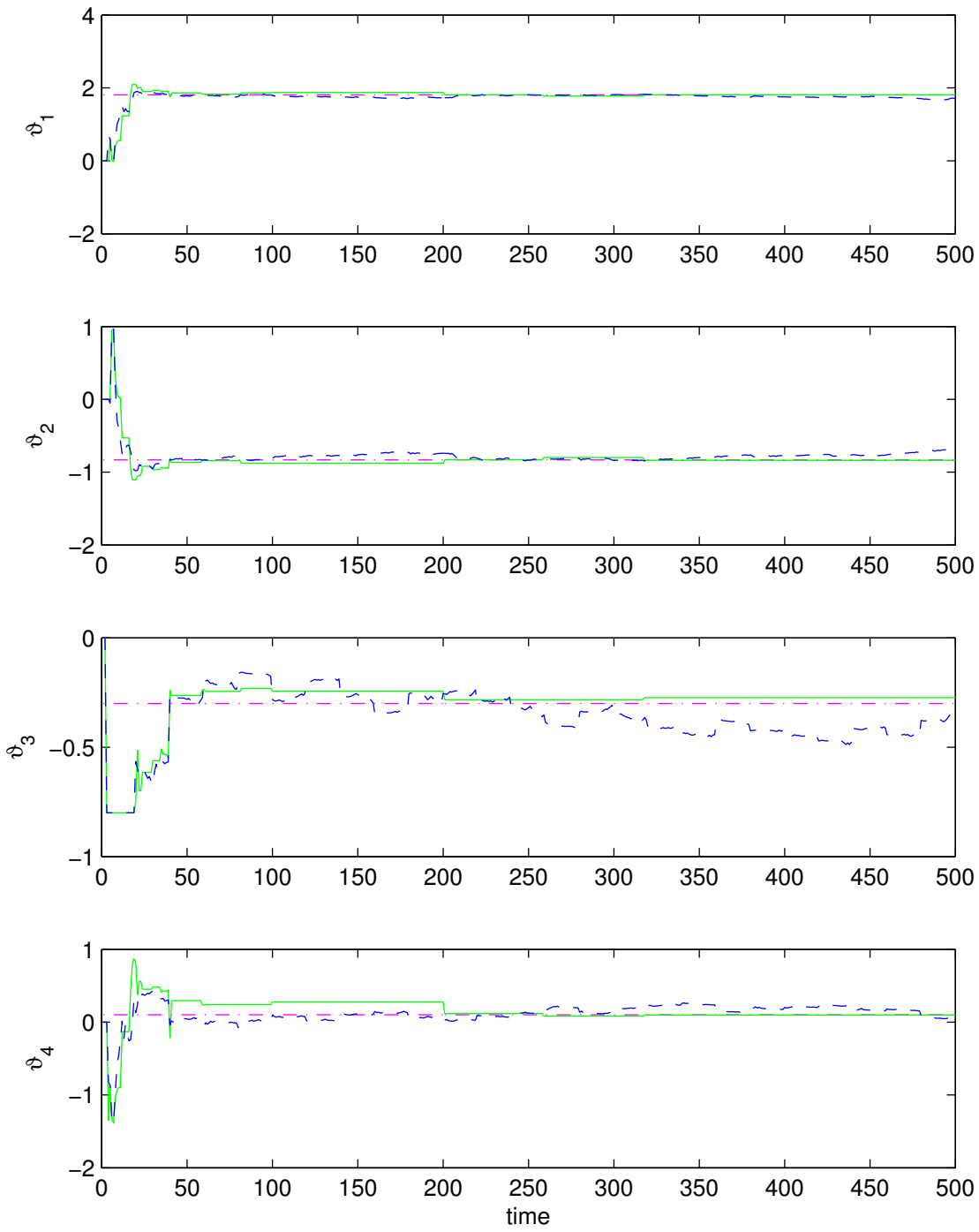

Figure 3: Time evolution of parameter estimates for both uniform (solid) and normal (dashed) estimators. The true parameter values are plotted in dash-dotted line, forgetting factor for the normal model $\lambda = 0.99$.

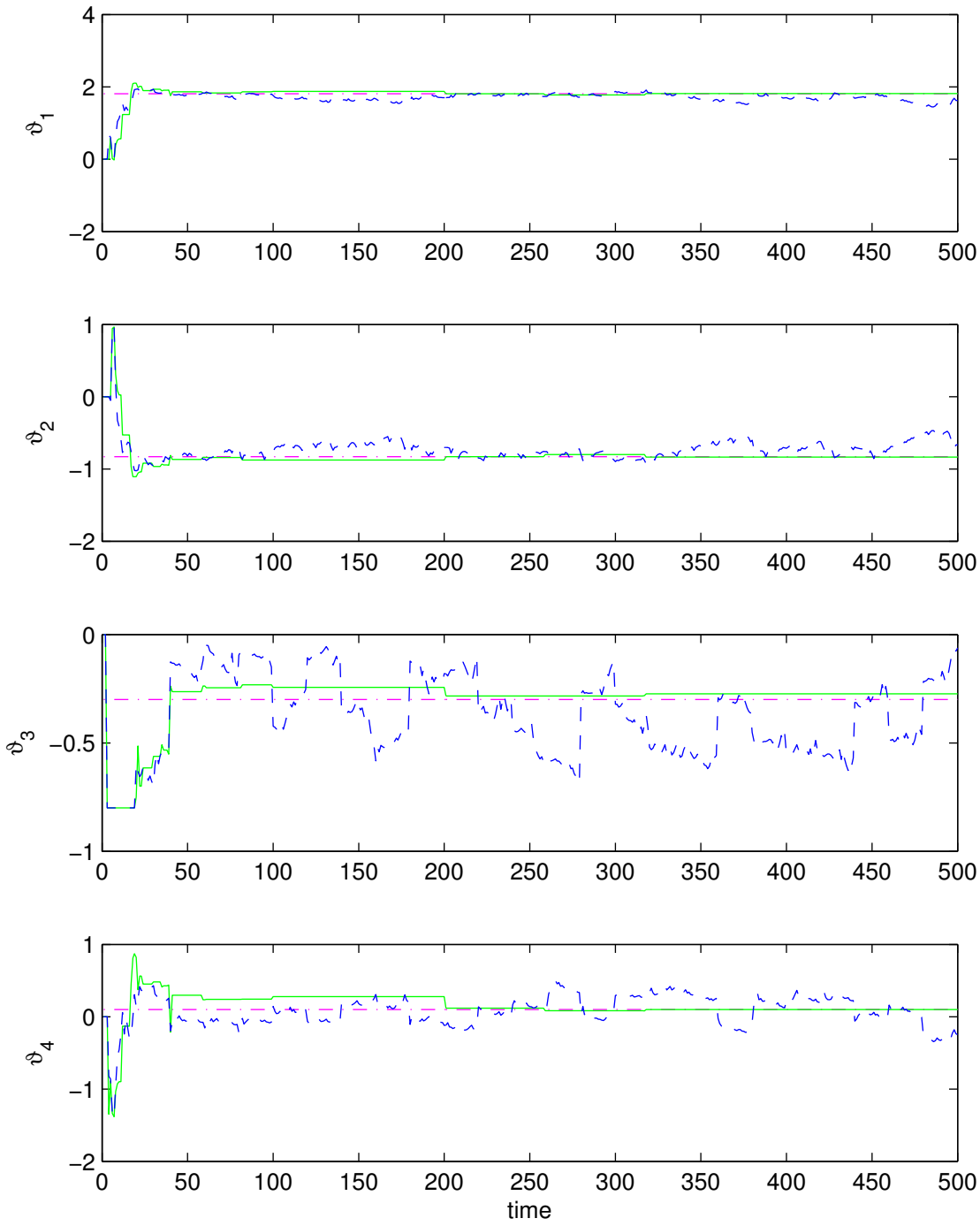

Figure 4: Time evolution of parameter estimates for both uniform (solid) and normal (dashed) estimators. The true parameter values are plotted in dash-dotted line, forgetting factor for the normal model $\lambda = 0.95$.