Local Statistical Models from Deterministic State Space Models, Likelihood Filtering, and Local Typicality

Lukas Bruderer, Hans-Andrea Loeliger, and Nour Zalmai
ETH Zurich
Dept. of Information Technology & Electrical Engineering
{bruderer, loeliger, zalmai}@isi.ee.ethz.ch

***Abstract*—Surprisingly many signal processing problems can be approached by locally fitting autonomous deterministic linear state space models to the data. In this paper, we introduce local statistical models for such cases and discuss the computation both of the corresponding estimates and of local likelihoods for different models.**

I. Introduction

We consider signal processing problems involving deterministic state space models as follows. Let $y_1, \ldots, y_N \in \mathbb{R}$ (with $N \gg 1$) be a given signal that is to be analyzed. For $k = 0, 1, \ldots, N$, let $x_k \in \mathbb{R}^m$ be a vector that evolves according to

$$x_{k+1} = Ax_k \tag{1}$$

where $A \in \mathbb{R}^{m \times m}$ is a non-singular square matrix. Note that the state x_k at any time k completely determines the whole state trajectory $x_0, x_1, \ldots, x_N$. A corresponding output signal $\tilde{y}_1, \ldots, \tilde{y}_N \in \mathbb{R}$ is defined by

$$\tilde{y}_k = c^{\mathsf{T}} x_k \tag{2}$$

where c^{T} is a given row vector. At any time $n \in \{1, 2, \ldots, N\}$, we locally fit this model to the given signal $y_1, \ldots, y_N \in \mathbb{R}$ by forming an estimate $\hat{x}_n$ defined by

$$\hat{x}_n \triangleq \operatorname*{argmin}_{x_n \in \mathcal{S}} \sum_{k=1}^{N} \gamma^{|n-k|} \big(y_k - \tilde{y}_k(x_n)\big)^2 \tag{3}$$

where γ is a real parameter with $0 < \gamma < 1$, where $\tilde{y}_1(x_n), \ldots, \tilde{y}_N(x_n)$ is the output signal determined by x_n according to (1) and (2), and where $\mathcal{S} \subset \mathbb{R}^m$ is an admissible set for $\hat{x}_n$. We will be primarily interested in the case where $1 \ll n \ll N$ so that boundary effects can be neglected.

Note that, in general, these estimates $\hat{x}_n$ will *not* satisfy $\hat{x}_{n+1} = A\hat{x}_n$.

In a variation of (3), the estimate (3) is replaced by

$$\hat{x}_n \triangleq \operatorname*{argmin}_{x_n \in \mathcal{S}} \sum_{k=1}^{n} \gamma^{n-k} \big(y_k - \tilde{y}_k(x_n)\big)^2, \tag{4}$$

which amounts to online estimation using only the past values $y_1, \ldots, y_n$ of the signal.

What we may want to do with these estimates $\hat{x}_n$ depends on the application, cf. the examples in Section II. In any case, we will also be much interested in assessing the quality of the least-squares fit (3) or (4) in a way that allows a meaningful comparison of different models, even with different parameter γ.

Note also that the choice $\mathcal{S} = \{0\}$ turns any model of the form (1) and (2) into a noise-only model with clean signal $\tilde{y}_k = 0$ for all k, which may serve as a reference in detection problems.

Computing a single estimate (3) or (4) is a least-squares problem, and all estimates $\hat{x}_1, \ldots, \hat{x}_N$ can be computed simultaneously by variations of recursive least-squares algorithms.

In this paper, we convert these least-squares problems into equivalent statistical Gaussian estimation problems. We then show that all the following quantities can be both meaningfully defined and efficiently computed by recursions similar to those in Kalman filtering.

1) Local state estimates $\hat{x}_n$ as above.
2) A normalized local likelihood function
$$\breve{p}_n(y_1 \ldots, y_N; x_n) \propto p_n(y_1, \ldots, y_N | x_n) \tag{5}$$
that remains finite for $N \to \infty$.
3) Local estimates of the noise variance σ^2 and the corresponding normalized likelihood $\max_\sigma \breve{p}_n(y_1, \ldots, y_n; x_n)$.
4) A new measure of local typicality that allows meaningful comparisons of models with different damping γ.

While the first of these items is quite obvious, the others are new (to the best of our knowledge).

The paper is structured as follows. Some illustrative examples are given in Section II. The conceptual contributions of the paper are described in Sections III and IV. The actual algorithms (efficient recursions) and the corresponding estimates are given in Sections V and VI, respectively.

II. Examples

The broad scope of signal processing problems that are amenable to the approach of this paper is indicated by the following three examples.

Example 1 (Straight-Line Fitting) Let

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{6}$$

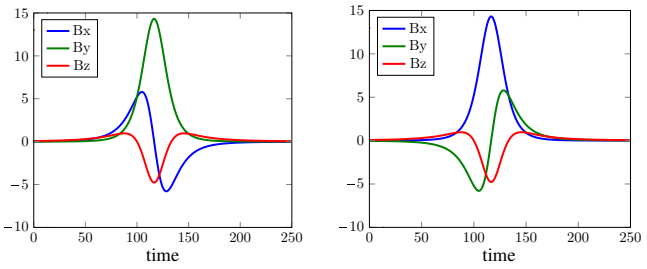

Fig. 1. Two examples of gestures (sweeps with an earphone magnet) as seen by the three channels of a 3-D magnetometer in a smartphone [5].

and $c^{\mathsf{T}} = (1, 0)$. The corresponding output signal (2) is a straight line, the slope of which is the second component of the state vector x_k. The estimate (3) with $\mathcal{S} = \mathbb{R}^2$ thus amounts to fitting a straight line to the signal $y_1, \dots, y_N$ around time n. □

The generalization of this example to polynomials is straightforward. Similar ideas have been described, e.g., in [1, Sec. 3.11], [2], [3].

The new results of this paper allow to evalute the goodness of such a straight-line fit (or of a polynomial fit) for different damping parameter γ (i.e., for different effective window size), for any time n independently, which can be used, e.g., for the detection of lines (or polynomials) or for adaptive smoothing.

Example 2 (PLL and Detection of a Sinusoid) Let

$$A = \begin{pmatrix} \cos(\Omega) & -\sin(\Omega) \\ \sin(\Omega) & \cos(\Omega) \end{pmatrix} \tag{7}$$

and $c^{\mathsf{T}} = (1, 0)$. The corresponding output signal (2) has the form

$$\tilde{y}_k = \beta \cos(\Omega k + \varphi), \tag{8}$$

where the amplitude $\beta > 0$ and the phase φ are determined by x_n.

From the local least-squares estimate (3) or (4) with $\mathcal{S} = \mathbb{R}^2$, we obtain a local estimate $\hat{\varphi}_n$ of the phase φ. If, around time n, the signal indeed contains a sinusoid with a frequency close to Ω, then the estimates $\hat{x}_n$ and $\hat{\varphi}_n$ will lock to this sinusoid. □

A very similar method was proposed in [4], except that localization in [4] is achieved with input noise rather than with a damping factor γ. The results of the present paper allow to use this PLL also for the local detection of a sinusoid at unknown signal-to-noise ratio.

Note that Examples 1 and 2 are meaningfull both with offline estimation as in (3) and with online estimation as in (4).

In the following example, the matrix A in (1) is not constant, but depends on the sign of $n - k$.

Example 3 (Gesture Detection with Magnetic Sensors) Contemporary smartphones contain a magnetometer that measures the magnetic field in three dimensions. Sweeping over the phone with a magnet (such as the magnet in typical earphones) results in 3-channel signals as shown in Figure 1. Such gestures can be used to give commands to the phone.

The detection of, and distinction between, such gestures can be based on local state space models of the form

$$A = \rho \begin{pmatrix} \cos(\Omega) & -\sin(\Omega) \\ \sin(\Omega) & \cos(\Omega) \end{pmatrix} \tag{9}$$

with $\rho > 1$ for $k < n$ and $\rho < 1$ for $k \geq n$ and with $c^{\mathsf{T}} = (1, 0)$. Moreover, the time-n state of the time-n model is restricted to be a scalar multiple of some vector $s \in \mathbb{R}^m$, i.e., $\mathcal{S} = \{\beta s : \beta \in \mathbb{R}\}$. The parameters ρ and s are chosen such that the signal (2) roughly approximates the clean sensor signal as in Figure 1. Multiple time scales (parameters ρ and γ) are necessary to detect gestures with different velocity and at different distance from the phone. □

The restriction of x_n to a set $\mathcal{S}$ as in Example 3 is an example of a glue factor as in [3], [6], [7], and the approach of this paper is easily adapted to more general glue factors.

III. Local Statistical Model

We now convert the least-squares problem (3) into an equivalent statistical estimation problem. (Problem (4) can be handled analogously.) To this end, we define, for each time $n \in \{1, \dots, N\}$, the Gaussian probability density

$$p_n(y_1, \dots, y_N | x_n) \triangleq \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\big(y_k - \tilde{y}_k(x_n)\big)^2}{2\sigma_k^2} \right) \tag{10}$$

with

$$\sigma_k^2 \triangleq \sigma^2 \gamma^{-|n-k|} \tag{11}$$

and where $\sigma > 0$ is a free parameter. For any fixed σ, we clearly have

$$\operatorname*{argmax}_{x_n \in \mathcal{S}} p_n(y_1, \dots, y_N | x_n) = \hat{x}_n \tag{12}$$

with $\hat{x}_n$ as in (3).

In the following, however, we prefer to work with the function (not a probability density)

$$\breve{p}_n(y_1, \dots, y_N; x_n)$$

$$\triangleq \prod_{k=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\big(y_k - \tilde{y}_k(x_n)\big)^2}{2\sigma^2} \right) \right)^{\gamma^{|n-k|}} \tag{13}$$

$$= \prod_{k=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^{\gamma^{|n-k|}} \exp\left(-\frac{\big(y_k - \tilde{y}_k(x_n)\big)^2}{2\sigma_k^2} \right), \tag{14}$$

which we will call the *local likelihood function.*

For fixed σ, we have

$$\breve{p}_n(y_1, \dots, y_N; x_n) \propto p_n(y_1, \dots, y_N | x_n) \tag{15}$$

where "$\propto$" denotes equality up to a scale factor, and thus

$$\operatorname*{argmax}_{x_n \in \mathcal{S}} \breve{p}_n(y_1, \dots, y_N; x_n) = \hat{x}_n \tag{16}$$

with $\hat{x}_n$ as in (3). However, while

$$\lim_{N \to \infty} p_n(y_1, \dots, y_N | x_n) = 0 \tag{17}$$

for every x_n and every $y_1, y_2, \dots$, the quantity $\lim_{N\to\infty} \breve{p}_n(y_1, \dots, y_N; x_n)$ is generically finite and nonzero, as will become obvious in Section V. Moreover, estimation of σ^2 from (10) yields absurd results for $N \to \infty$ (because of the factor $\prod_{k=1}^{N} \sigma^{-1}$). By contrast, σ^2 is properly localized in (13) and the estimate

$$\hat{\sigma}_n^2 \triangleq \operatorname*{argmax}_{\sigma^2} \max_{x_n \in \mathcal{S}} \breve{p}_n(y_1, \dots, y_N; x_n) \tag{18}$$

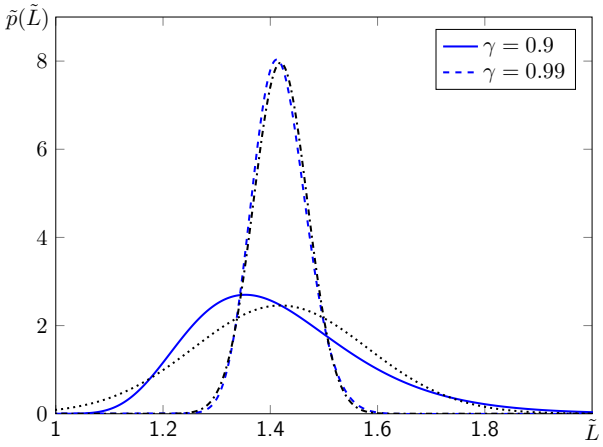

Fig. 2. The probability distribution $\tilde{p}$ from (23) and its Gaussian approximation (dotted) for two values of γ.

turns out to be the normalized squared error

$$\hat{\sigma}_n^2 = \frac{1}{\nu_n} \sum_{k=1}^{N} \gamma^{|n-k|} \big(y_k - \tilde{y}_k(\hat{x}_n)\big)^2 \tag{19}$$

where

$$\nu_n \triangleq \sum_{k=1}^{N} \gamma^{|n-k|} \tag{20}$$

is the effective window size.

A closely related quantity is the *local log-likelihood*

$$L_n \triangleq \log\Big(\max_{\sigma^2,\, x_n \in \mathcal{S}} \breve{p}_n(y_1, \ldots, y_N; x_n) \Big) \tag{21}$$

$$= -\nu_n \log\Big(\hat{\sigma}_n \sqrt{2\pi e} \Big). \tag{22}$$

For fixed $y_1, \ldots, y_N$, different local models with the same parameter γ can be compared, at each time n, using either $\hat{\sigma}_n^2$ or L_n. In particular, any local model can be compared with the noise-only model ($\mathcal{S} = \{0\}$). Efficient recursions for the computation of the sequences $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_N^2$ and $L_1, \ldots, L_N$ will be derived in Sections V to VI.

IV. Comparing Models over Different Windows by Means of Local Typicality

In order to compare models with different damping parameter γ (i.e., different effective window size), we propose the quantity

$$\tilde{p}(\tilde{L}_n) \tag{23}$$

where

$$\tilde{L}_n \triangleq -\frac{L_n}{\nu_n} \tag{24}$$

and where $\tilde{p}$ is a probability density with mean (34) and variance (35) that will be defined in Section IV-D. For γ close to 1, $\tilde{p}$ is well approximated by a Gaussian distribution, as illustrated in Figure 2.

We first explain the basic idea behind (23) in a setting outside the context of this paper.

A. The Basic Idea

Forget, for a moment, the context of this paper and consider the following hypothesis testing problem involving random variables $Y_1, \ldots, Y_N$ and two hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$. Under $\mathcal{H}_1$, the variables $Y_1, \ldots, Y_N$ are i.i.d. with known distribution $p(y)$; under $\mathcal{H}_2$, $Y_1, \ldots, Y_K$ (with $K < N$) are i.i.d. with the same distribution $p(y)$, but the distribution of $Y_{K+1}, \ldots, Y_N$ is unknown and arbitrary.

One approach[1] to this problem is to define

$$\mathcal{L}_N \triangleq -\frac{1}{N} \sum_{k=1}^{N} \log p(Y_k) \tag{25}$$

and to decide between $\mathcal{H}_1$ and $\mathcal{H}_2$ based on a comparison between $p_N(\mathcal{L}_N)$ and $p_K(\mathcal{L}_K)$, where $p_N(.)$ is the probability distribution of $\mathcal{L}_N$ under H_1. The probability distribution p_N has the following properties:

- Mean:
$$\mathrm{E}[\mathcal{L}_N] = -\mathrm{E}\big[\log p(Y_1)\big], \tag{26}$$
the entropy of $p(Y_1)$.
- Variance:
$$\mathrm{Var}[\mathcal{L}_N] = \frac{1}{N} \mathrm{Var}\big[\log p(Y_1)\big] \tag{27}$$
- Concentration: For $N \to \infty$, $p_N(\mathcal{L}_N)$ will concentrate around its mean and will converge to a Gaussian distribution.

In consequence, if the true hypothesis is $\mathcal{H}_1$ and both $N \gg 1$ and $K \gg 1$, then

$$\log \frac{p_N(\mathcal{L}_N)}{p_K(\mathcal{L}_K)} \approx \log \sqrt{\frac{N}{K}} \tag{28}$$

with probability close to one. By contrast, if the true hypothesis is $\mathcal{H}_2$ and $N - K \gg 1$, it is to be expected that $p_N(\mathcal{L}_N) < p_K(\mathcal{L}_K)$.

If $p(y)$ is Gaussian, then, under $\mathcal{H}_1$, p_N is a shifted and scaled version of a chi-squared distribution. However, this property will not carry over to the exponential-window setting of this paper.

Note that $p_N(\mathcal{L}_N(y_1, \ldots, y_N))$ may be viewed as a quantitative measure of typicality of the sequence $y_1, \ldots, y_N$: for sufficiently large N, $p_N(\mathcal{L}_N(y_1, \ldots, y_N))$ is large if $\log p(y_1, \ldots, y_n)$ is close to its mean, and small otherwise.

Finally, we note that this approach generalizes easily to the case where $\mathcal{H}_2$ uses i.i.d. variables $Y_1', \ldots, Y_K'$ with distribution $p(y')$ different from $p(y)$, and it generalizes even to non-i.i.d. variables provided that $\log p(Y_1, \ldots, Y_n)$ concentrates to its mean (i.e., satisfies an asymptotic equipartition property [8]).

[1] It is unlikely that this approach is new, but we have not yet spotted it in the literature.

B. Gaussian Case

Assume now that $p(y)$ is Gaussian with known mean and known variance σ^2. In this case (using the Gaussian approximation of the chi-squared distribution), a sign test of $\log p_N(\mathcal{L}_N) - \log p_K(\mathcal{L}_K)$ boils down to a sign test of

$$\left(K\left(\frac{q_K^2}{\sigma^2}-1\right)^2 - 2\log K\right) - \left(N\left(\frac{q_N^2}{\sigma^2}-1\right)^2 - 2\log N\right) \tag{29}$$

where q_K^2 and q_N^2 denote the empirical variance of $Y_1, \ldots, Y_K$ and $Y_1, \ldots, Y_N$, respectively. (The details are omitted).

As stated, the test (29) decides against a hypothesis not only if its likelihood is untypically small, but also if its likelihood is untypically large. If this is undesirable—and it is in our context— we may always decide in favor of $\mathcal{H}_1$ if $q_N \le q_K$ and use the test (29) only if $q_N^2 > q_K^2$.

C. What If σ^2 Needs to Be Estimated?

Assume now that $p(y)$ is Gaussian with known mean (as above), but σ^2 is not known and must be estimated from the data. It is not clear how this can be done in a principled way.

Estimating σ^2 separately for both hypotheses does not work: the natural estimate of σ^2 is the empirical variance (q_N^2 or q_K^2) of the data, which, when plugged into $p_N(\mathcal{L}_N)$ (or $p_K(\mathcal{L}_K)$, respectively) always indicates perfect typicality and $\mathcal{H}_1$ always wins.

A pragmatic proposal is to estimate σ^2 as

$$\hat{\sigma}^2 = \sqrt{q_N^2 q_K^2}, \tag{30}$$

which seems to work reasonably well.

D. Application to Local Statistical Models

We now adapt the idea of Sections IV-A–IV-C to the situation of this paper. (Concerning notation, we undefine all symbols defined in Sections IV-A–IV-C.) In particular, we now proceed to define the probability distribution $\tilde{p}$ in (23).

To this end, we need a distribution $\tilde{p}_n(y_1, \ldots, y_N)$ that reflects the idea that the given model is true (at least) throughout the effective window of the model. (The localized distribution (10) does not do this.) An arguable embodiment of this idea is the distribution

$$\tilde{p}_n(y_1, \ldots, y_N) \triangleq \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left(-\frac{\big(y_k - \tilde{y}_k(\hat{x}_n)\big)^2}{2\hat{\sigma}^2}\right) \tag{31}$$

where $\hat{x}_n$ is pragmatically chosen to be the maximizer in (21). For $\hat{\sigma}^2$, we propose the idea of (30), where the corresponding estimates of σ^2 for each model are obtained as described in Section VI.

Under the hypothesis that $Y_1, \ldots, Y_N$ are random variables with probability law (31), $\tilde{L}_n$ is a random variable with distribution $\tilde{p}(\tilde{L})$.

In order to understand the distribution $\tilde{p}(\tilde{L})$, consider

$$\tilde{L} = \frac{1}{\nu_n}\sum_{k=1}^{N} \gamma^{|n-k|}\left(\log\sqrt{2\pi}\hat{\sigma} + \frac{\big(Y_k - \tilde{y}_k(\hat{x}_n)\big)^2}{2\hat{\sigma}^2}\right). \tag{32}$$

Note that

$$\left(\frac{Y_k - \tilde{y}_k(\hat{x}_n)}{\hat{\sigma}}\right)^2 \tag{33}$$

is chi-squared with one degree of freedom and has mean 1 and variance 2. It follows that the mean of (32) is

$$\mathrm{E}[\tilde{L}] = \log(\sqrt{2\pi}\hat{\sigma}) + 1/2 \tag{34}$$

and the variance of (32) is

$$\mathrm{Var}[\tilde{L}] = \frac{1}{2\nu_n^2}\sum_{k=1}^{N}\gamma^{2|n-k|}; \tag{35}$$

for $1 \ll n \ll N$, (35) becomes

$$\lim_{1\ll n\ll N} \mathrm{Var}[\tilde{L}] = \frac{1}{2}(1-\gamma)\frac{1+\gamma^2}{(1+\gamma)^3} \tag{36}$$

It is then clear from (32) that $\tilde{p}(\tilde{L})$ concentrates around its mean (34) and becomes Gaussian for $\gamma \to 1$.

V. Recursive Computation of $\breve{p}_n(y_1, \ldots, y_N; x_n)$

For fixed $y_1, \ldots, y_N$, the function $\breve{p}_n(y_1, \ldots, y_N; x_n)$ can be computed for all $n \in \{1, \ldots, N\}$ with a total complexity that grows linearly with N. For this computation, we define the functions

$$\overrightarrow{\mu}_n(x_n) \triangleq \prod_{k=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\big(y_k - \tilde{y}_k(x_n)\big)^2}{2\sigma^2}\right)\right)^{\gamma^{n-k}} \tag{37}$$

and

$$\overleftarrow{\mu}_n(x_n) \triangleq \prod_{k=n+1}^{N}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\big(y_k - \tilde{y}_k(x_n)\big)^2}{2\sigma^2}\right)\right)^{\gamma^{k-n}} \tag{38}$$

which satisfy

$$\overrightarrow{\mu}_n(x_n)\overleftarrow{\mu}_n(x_n) = \breve{p}_n(y_1, \ldots, y_N; x_n). \tag{39}$$

(The analogous approach to the least-squares problem (4) requires only the forward recursion for $\overrightarrow{\mu}_n$; the quantity $\overleftarrow{\mu}_n$ is not needed.)

In order to cope with applications as in Example 3, we will allow that a different matrix A in (1) is used for the past of the time-n model than for its future, i.e.,

$$x_{k+1} = \begin{cases} A_{\mathrm{p}}x_k & \text{for } k < n \\ A_{\mathrm{f}}x_k & \text{for } k \ge n. \end{cases} \tag{40}$$

Beginning with $\overrightarrow{\mu}_0(x_0) = \overleftarrow{\mu}_N(x_N) = 1$ (for all x_0 and all x_N), we then have the recursions

$$\overrightarrow{\mu}_n(x_n) = \frac{\Big(\overrightarrow{\mu}_{n-1}(A_{\mathrm{p}}^{-1}x_n)\Big)^{\gamma}}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\big(y_n - c^{\mathsf{T}}x_n\big)^2}{2\sigma^2}\right) \tag{41}$$

and

$$\overleftarrow{\mu}_n(x_n) = \left(\frac{\overleftarrow{\mu}_{n+1}(A_{\mathrm{f}}x_n)}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\big(y_{n+1} - c^{\mathsf{T}}A_{\mathrm{f}}x_n\big)^2}{2\sigma^2}\right)\right)^{\gamma}. \tag{42}$$

The functions $\overrightarrow{\mu}_n$ and $\overleftarrow{\mu}_n$ can be parameterized as

$$\overrightarrow{\mu}_n(x) = \exp\left(-\frac{x^{\mathsf{T}}\overrightarrow{W}_n x - 2x^{\mathsf{T}}\overrightarrow{\xi}_n + \overrightarrow{\kappa}_n}{2\sigma^2} - \overrightarrow{\nu}_n \log\sqrt{2\pi\sigma^2}\right), \tag{43}$$

and

$$\overleftarrow{\mu}_n(x) = \exp\left(-\frac{x^{\mathsf{T}}\overleftarrow{W}_n x - 2x^{\mathsf{T}}\overleftarrow{\xi}_n + \overleftarrow{\kappa}_n}{2\sigma^2} - \overleftarrow{\nu}_n \log\sqrt{2\pi\sigma^2}\right), \tag{44}$$

respectively, where $\overrightarrow{W}_n$ and $\overleftarrow{W}_n$ are squares matrices, $\overrightarrow{\xi}_n$ and $\overleftarrow{\xi}_n$ are column vectors, and $\overrightarrow{\kappa}_n$, $\overleftarrow{\kappa}_n$, $\overrightarrow{\nu}_n$ and $\overleftarrow{\nu}_n$ are scalars. In terms of these parameters, the recursion (41) becomes

$$\overrightarrow{W}_n = \gamma\left(A_{\mathrm{p}}^{-1}\right)^{\mathsf{T}}\overrightarrow{W}_{n-1}A_{\mathrm{p}}^{-1} + cc^{\mathsf{T}} \tag{45}$$

$$\overrightarrow{\xi}_n = \gamma\left(A_{\mathrm{p}}^{-1}\right)^{\mathsf{T}}\overrightarrow{\xi}_{n-1} + y_n c \tag{46}$$

$$\overrightarrow{\kappa}_n = \gamma\overrightarrow{\kappa}_{n-1} + y_n^2 \tag{47}$$

$$\overrightarrow{\nu}_n = \gamma\overrightarrow{\nu}_{n-1} + 1 \tag{48}$$

with the initializations $\overrightarrow{W}_0 = 0$, $\overrightarrow{\xi}_0 = 0$, $\overrightarrow{\kappa}_0 = 0$, and $\overrightarrow{\nu}_0 = 0$. Similarly, the recursion (42) becomes

$$\overleftarrow{W}_n = \gamma\left(A_{\mathrm{f}}^{\mathsf{T}}\overleftarrow{W}_{n+1}A_{\mathrm{f}} + A_{\mathrm{f}}^{\mathsf{T}}cc^{\mathsf{T}}A_{\mathrm{f}}\right) \tag{49}$$

$$\overleftarrow{\xi}_n = \gamma\left(A_{\mathrm{f}}^{\mathsf{T}}\overleftarrow{\xi}_{n+1} + y_{n+1}A_{\mathrm{f}}^{\mathsf{T}}c\right) \tag{50}$$

$$\overleftarrow{\kappa}_n = \gamma\left(\overleftarrow{\kappa}_{n+1} + y_{n+1}^2\right) \tag{51}$$

$$\overleftarrow{\nu}_n = \gamma\left(\overleftarrow{\nu}_{n+1} + 1\right) \tag{52}$$

with the initializations $\overleftarrow{W}_N = 0$, $\overleftarrow{\xi}_N = 0$, $\overleftarrow{\kappa}_N = 0$, and $\overleftarrow{\nu}_N = 0$.

Finally, we obtain from (39)

$$\breve{p}(y_1, \ldots, y_N; x_n) = \exp\left(-\frac{x_n^{\mathsf{T}}W_n x_n - 2x_n^{\mathsf{T}}\xi_n + \kappa_n}{2\sigma^2} - \nu_n \log\sqrt{2\pi\sigma^2}\right) \tag{53}$$

with $W_n = \overrightarrow{W}_n + \overleftarrow{W}_n$, $\xi_n = \overrightarrow{\xi}_n + \overleftarrow{\xi}_n$, $\kappa_n = \overrightarrow{\kappa}_n + \overleftarrow{\kappa}_n$, and $\nu_n = \overrightarrow{\nu}_n + \overleftarrow{\nu}_n$.

We conclude this section with some remarks:

1) The parameters κ_n and ν_n are not required for state estimation, but they are needed for the computation of $\hat{\sigma}_n^2$, see Section VI.
2) The parameter $\nu_n = \overrightarrow{\nu}_n + \overleftarrow{\nu}_n$ agrees with (20), and it can easily be computed in closed form; in particular,
$$\lim_{1 \ll n \ll N} \nu_n = \frac{1+\gamma}{1-\gamma} \tag{54}$$
3) The computation of the parameters W_n and ξ_n amounts to a recursive least-squares algorithm with forgetting factor γ. However, standard recursive least-squares algorithms use only a single recursion while we here need both a forward recursion and a backward recursion.
4) The recursions for $\overrightarrow{W}_n$ and $\overleftarrow{W}_n$ do not depend on the data $y_1, \ldots, y_N$, as is usual in Kalman filtering and recursive least squares algorithms. In consequence, these recursions can be precomputed off-line.
For $1 \ll n \ll N$, the matrices $\overrightarrow{W}_n$ and $\overleftarrow{W}_n$ will not normally depend on n. (This applies, in particular, to all examples in Section II.) In many applications, only these steady-state solutions are of interest.

VI. Computation of Local State Estimate and Local Likelihood

The following quantities are easily derived from (53).

Estimation of σ_n^2 as in (18):

$$\hat{\sigma}_n^2 = \frac{1}{\nu_n}\left(\hat{x}_n^{\mathsf{T}}W_n\hat{x}_n - 2\hat{x}_n^{\mathsf{T}}\xi_n + \kappa_n\right). \tag{55}$$

Unconstrained state estimation:

$$\hat{x}_n = \operatorname*{argmax}_{x_n} \breve{p}_n(y_1, \ldots, y_N; x_n) \tag{56}$$

$$= W_n^{-1}\xi_n. \tag{57}$$

In this case, (55) becomes

$$\hat{\sigma}_n^2 = \frac{1}{\nu_n}\left(-\xi_n^{\mathsf{T}}W_n^{-1}\xi_n + \kappa_n\right). \tag{58}$$

Estimation of unknown amplitude, i.e., $x_n = \beta_n s$ for some given column vector s:

$$\hat{\beta}_n = \operatorname*{argmax}_{\beta \in \mathbb{R}} \breve{p}_n(y_1, \ldots, y_N; \beta s) \tag{59}$$

and

$$\hat{x}_n = \frac{(s^{\mathsf{T}}\xi_n)s}{s^{\mathsf{T}}W_n s} \tag{60}$$

In this case, (55) becomes

$$\hat{\sigma}_n^2 = \frac{1}{\nu_n}\left(-\frac{(s^{\mathsf{T}}\xi_n)^2}{s^{\mathsf{T}}W_n s} + \kappa_n\right). \tag{61}$$

In all these cases, the local likelihood (21) is easily obtained from $\hat{\sigma}_n^2$ by (22).

References

[1] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods.* Oxford Univ. Press, 2012.

[2] G. Wahba, *Spline Models for Observational Data.* SIAM, 1990.

[3] Christoph Reller, *State-Space Methods in Statistical Signal Processing: New Ideas and Applications.* PhD thesis at ETH Zurich No 20584, 2012.

[4] Yuan Qi, T. P. Minka, and R. W. Picara, "Bayesian spectrum estimation of unevenly sampled nonstationary data," *Proc. 2002 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP),* vol. 2, p. II–1473, 2002.

[5] C. Käslin, *Gesture Recognition on Smartphone with Likelihood Filtering.* Master thesis at ETH Zurich, 2014.

[6] H.-A. Loeliger, L. Bolliger, S. Korl, and Ch. Reller, "Localizing, forgetting, and likelihood filtering in state-space models," 2009 Information Theory & Applications Workshop, UCSD, La Jolla, CA, Feb. 8-13, 2009.

[7] Ch. Reller, M. V. R. S. Devarakonda, and H.-A. Loeliger, "Glue factors, likelihood computation, and filtering in state space models," *Proc. 50th Annual Allerton Conference on Communication, Control, and Computing,* Monticello, Illinois, USA, Oct. 1–5, 2012.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, 1991 and 2006.