On Sparsity by NUV-EM, Gaussian Message Passing, and Kalman Smoothing

Hans-Andrea Loeliger, Lukas Bruderer, Hampus Malmberg, Federico Wadehn, and Nour Zalmai
ETH Zurich, Dept. of Information Technology & Electrical Engineering

***Abstract*—Normal priors with unknown variance (NUV) have long been known to promote sparsity and to blend well with parameter learning by expectation maximization (EM). In this paper, we advocate this approach for linear state space models for applications such as the estimation of impulsive signals, the detection of localized events, smoothing with occasional jumps in the state space, and the detection and removal of outliers.**

The actual computations boil down to multivariate-Gaussian message passing algorithms that are closely related to Kalman smoothing. We give improved tables of Gaussian-message computations from which such algorithms are easily synthesized, and we point out two preferred such algorithms.

I. Introduction

This paper is about two topics:

1) A particular approach to modeling and estimating sparse parameters based on zero-mean normal priors with unknown variance (NUV).
2) Multivariate-Gaussian message passing ($\approx$ variations of Kalman smoothing) in such models.

The main point of the paper is that these two things go very well together and combine to a versatile toolbox. This is not entirely new, of course, and the body of related literature is large. Nonetheless, the specific perspective of this paper has not, as far as known to these authors, been advocated before.

Concerning the second topic, linear state space models continue to be an essential tool for a broad variety of applications, cf. [1]–[4]. The primary algorithms for such models are variations and generalizations of Kalman filtering and smoothing, or, equivalently, multivariate-Gaussian message passing in the corresponding factor graph [5], [6] (or similar graphical model [1]). A variety of such algorithms can easily be synthesized from tables of message computations as in [6]. In this paper, we give a new version of these tables with many improvements over those in [6], and we point out two preferred such algorithms.

Concerning the first topic, NUV priors (zero-mean normal priors with unknown variance) originated in Bayesian inference [7]–[9]. The sparsity-promoting nature of such priors is the basis of automatic relevance determination (ARD) and sparse Bayesian learning developed by Neal [9], Tipping [10], [11], Wipf et al. [12], [13], and others.

The basic properties of NUV priors are illustrated by the following simple example. Let U be a variable or parameter of interest, which we model as a zero-mean real scalar Gaussian random variable with unknown variance s^2. Assume that we observe $Y = U + Z$, where the noise Z is zero-mean Gaussian with (known) variance σ^2 and independent of U. The maximum likelihood (ML) estimate of s^2 from a single sample $Y = \mu \in \mathbb{R}$ is easily determined:

$$\hat{s}^2 \triangleq \underset{s^2}{\operatorname{argmax}} \frac{1}{\sqrt{2\pi(s^2+\sigma^2)}} e^{-\mu^2/2(s^2+\sigma^2)} \tag{1}$$

$$= \max\{0,\, \mu^2 - \sigma^2\}. \tag{2}$$

In a second step, for s^2 fixed to $\hat{s}^2$ as in (2), the MAP/MMSE/LMMSE estimate of U is

$$\hat{u} = \mu \cdot \frac{\hat{s}^2}{\hat{s}^2 + \sigma^2} \tag{3}$$

$$= \begin{cases} \mu \cdot \frac{\mu^2 - \sigma^2}{\mu^2} & \text{if } \mu^2 > \sigma^2 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Equations (1)–(4) continue to hold if the scalar observation Y is generalized to an observation $Y \in \mathbb{R}^N$ such that, for fixed $Y = y$, the likelihood function $p(y|u)$ is Gaussian (up to a scale factor) with mean μ and variance σ^2. In fact, this is all we need to know in this paper about NUV priors per se.

The estimate (4) has some pleasing properties: first, it promotes sparsity and can thus be used to select features or relevant parameters; second, it has no *a priori* preference as to the scale of U, and large values of U are not scaled down. Note that the latter property is lost if ML estimation of s^2 is replaced by MAP estimation based on a proper prior on s^2.

In this paper, we will stick to basic NUV regularization as above, with no prior on the unknown variances: variables or parameters of interest are modeled as independent Gaussian random variables, each with its own unknown variance that is estimated (exactly or approximately) by maximum likelihood. We will advocate the use of NUV regularization in linear state space models, for applications such as the estimation of impulsive signals, the detection of localized events, smoothing with occasional jumps in the state space, and the detection and removal of outliers.

Concerning the actual computations, estimating the unknown variances is not substantially different from learning other parameters of state space models and can be carried out by expectation maximization (EM) [14]–[17] and other methods in such a way that the actual computations essentially amount to Gaussian message passing.

The paper is structured as follows. In Section II, we begin with a quick look at NUV regularization in a standard linear model. Estimation of the unknown variances is addressed in Section III. Factor graphs and state space models are reviewed in Sections IV and V, respectively, and NUV regularization in such models is addressed in Section VI. The new tables of Gaussian-message computations are given in Appendix A.

II. Sum of Gaussians and Least Squares with NUV Regularization

We begin with an elementary linear model (a special case of a relevance vector machine [10]) as follows. For $b_1, \ldots, b_K \in \mathbb{R}^n \setminus \{0\}$, let

$$Y = \sum_{k=1}^{K} b_k U_k + Z \tag{5}$$

where $U_1, \ldots, U_K$ are independent zero-mean real scalar Gaussian random variables with unknown variances $\sigma_1^2, \ldots, \sigma_K^2$, and where the "noise" Z is $\mathbb{R}^n$-valued zero-mean Gaussian with covariance matrix $\sigma^2 I$ and independent of $U_1, \ldots, U_K$. For a given observation $Y = y \in \mathbb{R}^n$, we wish to estimate, first, $\sigma_1^2, \ldots, \sigma_K^2$ by maximum likelihood, and second, $U_1, \ldots, U_K$ (with $\sigma_1^2, \ldots, \sigma_K^2$ fixed).

In the first step, we achieve sparsity: if the ML estimate of σ_k^2 is zero, then $U_k = 0$ is fixed in the second step.

The second step — the estimation of $U_1, \ldots, U_K$ for fixed $\sigma_1^2, \ldots, \sigma_K^2$ — is a standard Gaussian estimation problem where MAP estimation, MMSE estimation, and LMMSE estimation coincide and amount to minimizing

$$\frac{1}{\sigma^2}\Big\| y - \sum_{k\in\mathcal{K}^+} b_k u_k \Big\|^2 + \sum_{k\in\mathcal{K}^+} \frac{1}{\sigma_k^2}\|u_k\|^2, \tag{6}$$

where $\mathcal{K}^+$ denotes the set of those indices $k \in \{1, \ldots, K\}$ for which $\sigma_k^2 > 0$. A closed-form solution of this minimization is

$$\hat{u}_k = \sigma_k^2 b_k^\mathsf{T} \tilde{W} y \tag{7}$$

with

$$\tilde{W} \triangleq \left(\sum_{k=1}^{K} \sigma_k^2 b_k b_k^\mathsf{T} + \sigma^2 I \right)^{-1}, \tag{8}$$

as may be obtained from standard least-squares equations (see also [11]). An alternative proof will be given in Appendix B, where we also point out how $\tilde{W}$ can be computed without a matrix inversion.

In (2) and (4), the estimate is zero if and only if $y^2 \le \sigma^2$. Two different generalizations of this condition to the setting of this section are given in the following theorem. Let $p(y, \ldots)$ denote the probability density of Y and any other variables according to (5).

Theorem. Let $\sigma_1, \ldots, \sigma_K$ be fixed at a local maximum or at a saddle point of $p(y|\sigma_1^2, \ldots, \sigma_K^2)$. Then $\sigma_k^2 = 0$ if and only if

$$\big(b_k^\mathsf{T} W_k\, y\big)^2 \le b_k^\mathsf{T} W_k\, b_k \tag{9}$$

with

$$W_k \triangleq \left(\sum_{\ell=1}^{K} \sigma_\ell^2 b_\ell b_\ell^\mathsf{T} + \sigma^2 I - \sigma_k^2 b_k b_k^\mathsf{T} \right)^{-1}. \tag{10}$$

Moreover, with $\tilde{W}$ as in (8), we have

$$\big(b_k^\mathsf{T} \tilde{W} y\big)^2 \le b_k^\mathsf{T} \tilde{W} b_k^\mathsf{T}, \tag{11}$$

with equality if $\sigma_k^2 > 0$. □

(The proof will be given in Appendix B.) The matrices W_k and $\tilde{W}$ are both positive definite. The former depends on k, but not on σ_k^2; the latter depends also on σ_k^2, but not on k.

III. Variance Estimation

Following a standard approach, the unknown variances $\sigma_1^2, \ldots, \sigma_K^2$ in Section II (and analogous quantities in later sections) can be estimated by an EM algorithm as follows.

1) Begin with an initial guess of $\sigma_1^2, \ldots, \sigma_K^2$.
2) Compute the mean m_{U_k} and the variance $\sigma_{U_k}^2$ of the (Gaussian) posterior distribution $p(u_k|y, \sigma_1^2, \ldots, \sigma_K^2)$ with $\sigma_1^2, \ldots, \sigma_K^2$ fixed.
3) Update $\sigma_1^2, \ldots, \sigma_K^2$ according to (13) below.
4) Repeat steps 2 and 3 until convergence, or until some pragmatic stopping criterion is met.
5) Optionally update $\sigma_1^2, \ldots, \sigma_K^2$ according to (16) below.

The standard EM update for the variances is

$$\sigma_k^2 \leftarrow \mathrm{E}\big[U_k^2 | \sigma_1^2, \ldots, \sigma_K^2\big] \tag{12}$$

$$= m_{U_k}^2 + \sigma_{U_k}^2. \tag{13}$$

The required quantities $m_{U_k}^2$ and $\sigma_{U_k}^2$ are given by (85) and (88), respectively. With this update, basic EM theory guarantees that the likelihood $p(y|\sigma_1^2, \ldots, \sigma_K^2)$ cannot decrease (and will normally increase) in step 3 of the algorithm.

The stated EM algorithm is safe, but the convergence can be slow. The following alternative update rule, due to MacKay [10], often converges much faster:

$$\sigma_k^2 \leftarrow \frac{m_{U_k}^2}{1 - \sigma_{U_k}^2 / \sigma_k^2} \tag{14}$$

However, this alterative update rule comes without guarantees; sometimes, it is too agressive and the algorithm fails completely.

An individual variance σ_k^2 can also be estimated by a maximum-likelihood step as in (2):

$$\sigma_k^2 \leftarrow \operatorname*{argmax}_{\sigma_k^2} p(y|\sigma_1^2, \ldots, \sigma_K^2) \tag{15}$$

$$= \max\big\{0, (\overleftarrow{m}_{U_k})^2 - \overleftarrow{\sigma}_{U_k}^2\big\}, \tag{16}$$

The mean $\overleftarrow{m}_{U_k}$ is given by (104) and the variance $\overleftarrow{\sigma}_{U_k}^2$ is given by (95). However, for parallel updates (simultaneously for all $k \in \{1, \ldots, K\}$, as in step 3 of the algorithm above), the rule (16) is normally too agressive and fails.

Later on, the same algorithm will be used for estimating parameters or variables in linear state space models. In this case, we have no useful analytical expressions for (the analogs of) m_{U_k} and $\sigma_{U_k}^2$, but these quantities are easily computed by Gaussian message passing.

IV. On Factor Graphs and Gaussian Message Passing

From now on, we will heavily use factor graphs, both for reasoning and for describing algorithms. We will use factor graphs as in [5], [6], where nodes/boxes represent factors and

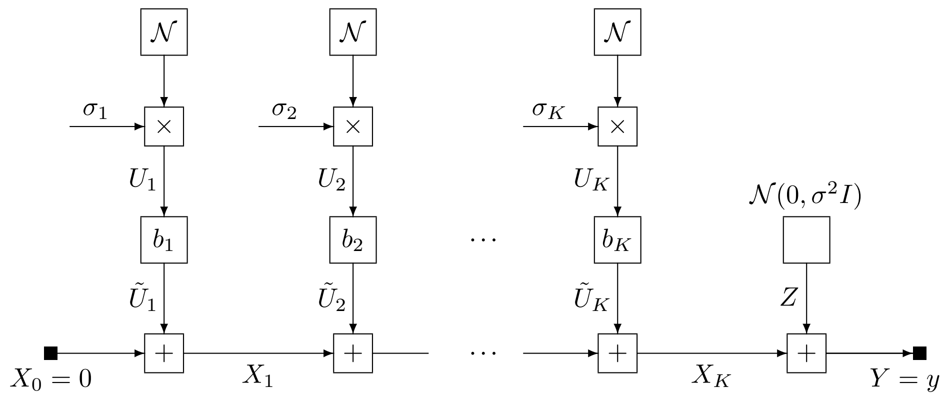


Fig. 1. Cycle-free factor graph of (5) with NUV regularization.

edges represent variables. (By contrast, factor graphs as in [18] have both variable nodes and factor nodes.)

Figure 1, for example, represents the probability density $p(y, z, u_1, \ldots, u_K | \sigma_1, \ldots, \sigma_K)$ of the model (5) with auxiliary variables $\tilde{U}_k \triangleq b_k U_k$ and $X_k \triangleq X_{k-1} + \tilde{U}_k$ with $X_0 \triangleq 0$. The nodes labeled "$\mathcal{N}$" represent zero-mean normal densities with variance 1; the node labeled "$\mathcal{N}(0, \sigma^2 I)$" represents a zero-mean multivariate normal density with covariance matrix $\sigma^2 I$. All other nodes in Figure 1 represent deterministic constraints.

For fixed $\sigma_1, \ldots, \sigma_K$, Figure 1 is a cycle-free linear Gaussian factor graph and MAP/MMSE/LMMSE estimation (of any variables) can be carried out by Gaussian message passing, as described in detail in [6]. Interestingly, in this particular example, most of the message passing can be carried out symbolically, i.e., as a technique to derive closed-form expressions for the estimates.

Every message in this paper is a (scalar or multivariate) Gaussian distribution, up to a scale factor. (Sometimes, we also allow a degenerate limit of a Gaussian, such as a "Gaussian" with variance zero or infinity, but we will not discuss this in detail.) Scale factors can be ignored in this paper. Messages can thus be parameterized by a mean vector and a covariance matrix. For example, $\overrightarrow{m}_{X_k}$ and $\overrightarrow{V}_{X_k}$ denote the mean vector and the covariance matrix, respectively, of the message traveling forward on the edge X_k in Figure 1, while $\overleftarrow{m}_{X_k}$ and $\overleftarrow{V}_{X_k}$ denote the mean vector and the covariance matrix, respectively, of the message traveling backward on the edge X_k. Alternatively, messages can be parameterized by the precision matrix $\overrightarrow{W}_{X_k}$ (= the inverse of the covariance matrix $\overrightarrow{V}_{X_k}$) and the precision-weighted mean vector

$$\overrightarrow{\xi}_{X_k} \triangleq \overrightarrow{W}_{X_k} \overrightarrow{m}_{X_k}. \tag{17}$$

Again, the backward message along the same edge will be denoted by reversed arrows.

In a directed graphical model without cycles as in Figure 1, forward messages represent priors while backward messages represent likelihood functions (up to a scale factor).

In addition, we also work with marginals of the posterior distribution (i.e., the product of forward message and backward message along the same edge [5], [6]). For example, m_{X_k} and V_{X_k} denote the posterior mean vector and the posterior covariance matrix, respectively, of X_k. An important role in this paper is played by the alternative parameterization with the dual precision matrix

$$\tilde{W}_{X_k} \triangleq \left(\overrightarrow{V}_{X_k} + \overleftarrow{V}_{X_k}\right)^{-1} \tag{18}$$

and the dual mean vector

$$\tilde{\xi}_{X_k} \triangleq \tilde{W}_{X_k} (\overrightarrow{m}_{X_k} - \overleftarrow{m}_{X_k}). \tag{19}$$

Message computations with all these parameterizations are given in Tables I–VI in Appendix A, which contain numerous improvements over the corresponding tables in [6].

V. Linear State Space Models

Consider a standard linear state space model with state $X_k \in \mathbb{R}^n$ and observation $Y_k \in \mathbb{R}^L$ evolving according to

$$X_k = AX_{k-1} + BU_k \tag{20}$$

$$Y_k = CX_k + Z_k \tag{21}$$

with $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{L \times n}$, and where U_k (with values in $\mathbb{R}^m$) and Z_k (with values in $\mathbb{R}^L$) are independent zero-mean white Gaussian noise processes. We will usually assume, first, that $L = 1$, and second, that the covariance matrix of U_k is an identity matrix, but these assumptions are not essential. A cycle-free factor graph of such a model is shown in Figure 2.

In Section VI, we will vary and augment such models with NUV priors on various quantities.

Inference in such a state space model amounts to Kalman filtering and smoothing [1], [2] or, equivalently, to Gaussian message passing in the factor graph of Figure 2 [5], [6]. (Estimating the input U_k is not usually considered in the Kalman filter literature, but it is essential for signal processing, cf. [19], [20].) With the tables in the appendix, it is easy to put together a large variety of such algorithms. The relative merits of different such algorithms depend on the particulars

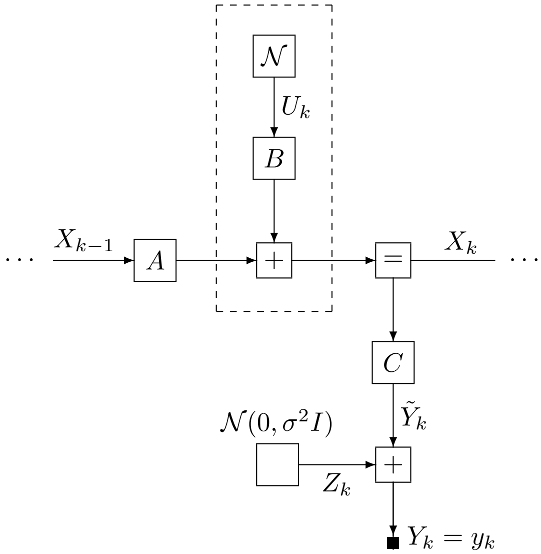


Fig. 2. One section of the factor graph of the linear state space model (20) and (21). The whole factor graph consists of many such sections and optional initial and/or terminal conditions. The dashed block will be varied in Section VI.

of the problem. However, we find the following two algorithms usually to be the most advantageous, both in terms of computational complexity and in terms of numerical stability. If both the input U_k and output Y_k are scalar (or can be decomposed into multiple scalar inputs and outputs), neither of these two algorithms requires a matrix inversion. The first of these algorithms is essentially the Modified Bryson–Frazier (MBF) smoother [21] augmented with input-signal estimation.

MBF Message Passing:

1) Perform forward message passing with $\overrightarrow{m}_{X_k}$ and $\overrightarrow{V}_{X_k}$ using (II.1), (II.2), (III.1), (III.2), (V.1), (V.2). (This is the standard Kalman filter.)
2) Perform backward message passing with $\tilde{\xi}_{X_k}$ and $\tilde{W}_{X_k}$, beginning with $\tilde{\xi}_{X_N} = 0$ and $\tilde{W}_{X_N} = 0$ at the end of the horizon, using (II.6), (II.7), (III.7), (III.8), and either (V.4), (V.6), (V.8) or (V.5), (V.7), (V.9).
3) Inputs U_k may then be estimated using (II.6), (II.7), (III.7), (III.8), (IV.9), (IV.13).
4) The posterior mean m_{X_k} and covariance matrix V_{X_k} of any state X_k (or of an individual component thereof) may be obtained from (IV.9) and (IV.13)
5) Outputs $\tilde{Y}_k \triangleq CX_k$ may then (very obviously) be estimated using (I.5), (I.6), (III.5), (III.6). □

In step 2, the initialization with $\tilde{W}_{X_N} = 0$ corresponds to the typical situation with no *a priori* information about the state X_N at the end of the horizon. MBF message passing is especially attractive for input signal estimation (as in step 3 above), without steps 4 and 5.

The second algorithm is an exact dual to MBF message passing and especially attractive for state estimation and output signal estimation (i.e., for standard Kalman smoothing), without steps 4 and 5 below. This algorithm—backward recursion with time-reversed information filter, forward recursion with marginals (BIFM)—does not seem to be widely known.

BIFM Message Passing:

1) Perform backward message passing with $\overleftarrow{\xi}_{X_k}$ and $\overleftarrow{W}_{X_k}$ using (I.3), (I.4), (III.3), (III.4), and (VI.1), (VI.2) with the changes "for the reverse direction" stated in Table VI. (This is a time-reversed version of the standard information filter.)
2) Perform forward message passing with m_{X_k} and V_{X_k} using (I.5), (I.6), (III.5), (III.6), and either (VI.4), (VI.6), (VI.8) or (VI.5), (VI.7), (VI.9).
3) Outputs $\tilde{Y}_k$ may then (very obviously) be estimated using (I.5), (I.6), (III.5), (III.6).
4) The dual means $\tilde{\xi}_{X_k}$ and the dual precision matrices $\tilde{W}_{X_k}$ may be obtained from (IV.3) and (IV.7).
5) Inputs U_k may then be estimated using (II.6), (II.7), (III.7), (III.8), (IV.9), (IV.13). □

VI. Sparsity by NUV in State Space Models

Sparse input signals are easily introduced: simply replace the normal prior on U_k in (20) and in Figure 2 by a NUV prior, as shown in Figure 3. This approach was used in [22] to estimate the input signal $U_1, U_2, \ldots$ itself.

However, we may also be interested in the clean output signal $\tilde{Y}_k = CX_k$. For example, consider the problem of approximating some given signal $y_1, y_2, \ldots \in \mathbb{R}$ by constant segments, as illustrated in Figure 8. The constant segments can be represented by the simplest possible state space model with $n = 1$, $A = C = (1)$, and no input. For the occasional jumps between the constant segments, we use a sparse input signal $U_1, U_2, \ldots$ with a NUV prior (and with $B = b = (1)$) as in Figure 3. The sparsity level—i.e., the number of constant segments—can be controlled by the assumed observation noise σ^2.

The sparse scalar input signal of Figure 3 can be generalized in several different directions. A first obvious generalization is to combine a primary white-noise input with a secondary sparse input as shown in Figure 4. For example, the constant segments in Figure 8 are thus generalized to random-walk segments as in Figure 9.

Another generalization of Figure 8 is shown in Figure 10, where the constant-level segments are replaced by straight-line segments, which can be represented by a state space model of order $n = 2$. The corresponding input block, with two separate sparse scalar input signals, is shown in Figure 5; the first input, $U_{k,1}$, affects the magnitude and the second input, $U_{k,2}$, affects the slope of the line model. The further generalization to polynomial segments is obvious. Continuity can be enforced by omitting the input $U_{k,1}$, and continuity of derivatives can be enforced likewise.

More generally, Figure 5 (generalized to an arbitrary number of sparse scalar input signals) can be used to allow occasional jumps in individual components of the state of arbitrary state space models.

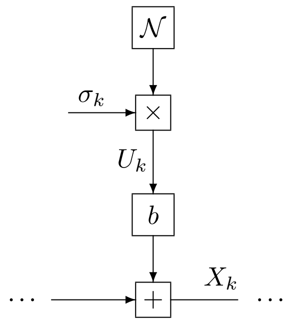


Fig. 3. Alternative input block (to replace the dashed box in Figure 2) for a sparse scalar input signal $U_1, U_2, \ldots$.

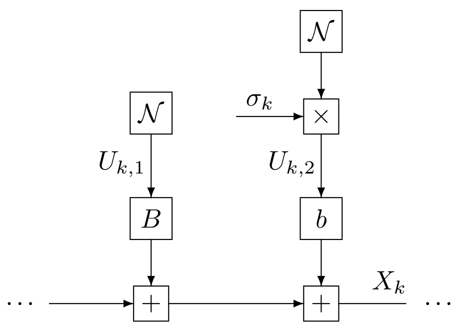


Fig. 4. Input block with both white noise and additional sparse scalar input.

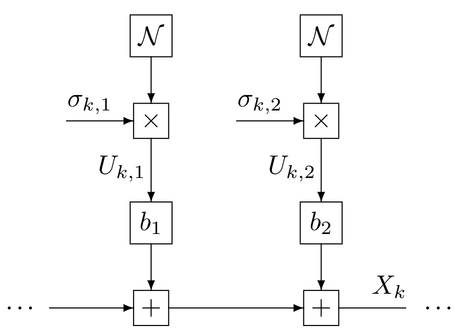


Fig. 5. Input block with two separate sparse scalar inputs for two degrees of freedom such as in Figure 10.

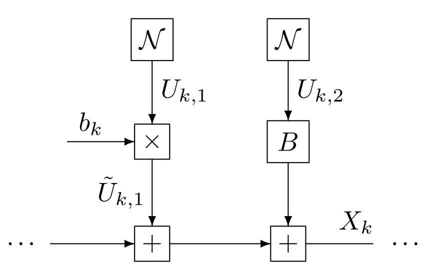


Fig. 6. Input block allowing general sparse pulses, each with its own signature b_k, in addition to full-rank white noise.

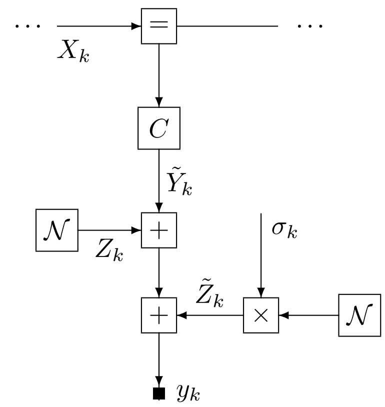


Fig. 7. Alternative output block for scalar signal with outliers.

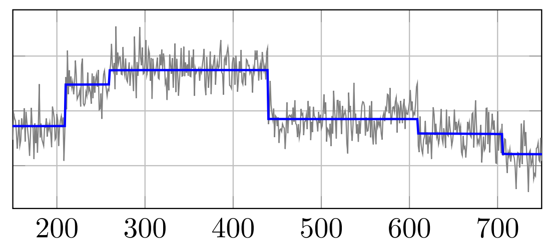


Fig. 8. Estimating (or fitting) a piecewise constant signal.

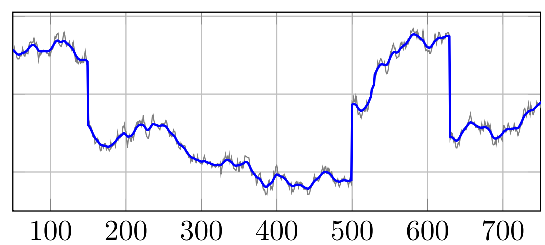


Fig. 9. Estimating a random walk with occasional jumps.

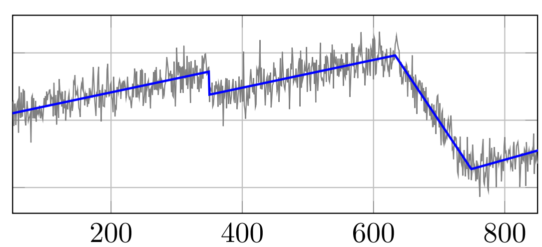


Fig. 10. Approximation with straight-line segments.

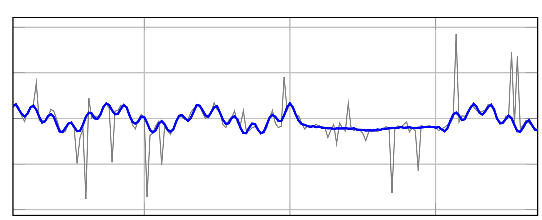


Fig. 11. Outlier removal according to Figure 7.

In all these examples, the parameters σ_k^2 (or $\sigma_{k,\ell}^2$) can be learned as described in Section III, and the required quantities m_{U_k} and $\sigma_{U_k}^2$ (or $m_{U_{k,\ell}}$ and $\sigma_{U_{k,\ell}}^2$, respectively) can be computed by message passing in the pertinent factor graph as described in Section V.

A more substantial generalization of Figure 3 is shown in Figure 6, with σ_k of Figure 3 generalized to $b_k \in \mathbb{R}^n$. We mention without proof that this generalized NUV prior on $\tilde{U}_{k,1} \triangleq b_k U_{k,1}$ still promotes sparsity and can be learned by EM (provided that BB^T has full rank) [23]. This input model allows quite general events to happen, each with its own signature b_k. The estimated nonzero vectors $\hat{b}_1, \hat{b}_2, \ldots$ may be viewed as features of the given signal $y_1, y_2, \ldots$ that can be used for further analysis.

Finally, we turn to the output block in Figure 2. A simple and effective method to detect and to remove outliers from the scalar output signal of a state space model is to replace (21) with

$$Y_k = CX_k + Z_k + \tilde{Z}_k \tag{22}$$

with sparse $\tilde{Z}_k$, as shown in Figure 7 [24]. Again, the parameters σ_k can be estimated by EM essentially as described in Section III, and the required quantities $m_{\tilde{Z}_k}$ and $\sigma_{\tilde{Z}_k}^2$ can be computed by message passing as described in Section V. An example of this method is shown in Figure 11 for some state space model of order $n = 4$ with details that are irrelevant for this paper.

VII. Conclusion

We have given improved tables of Gaussian-message computations for estimation in linear state space models, and we have pointed out two preferred message passing algorithms: the first algorithm is essentially the Modified Bryson-Frazier smoother, the second algorithm is a dual of it. In addition, we have advocated NUV priors (together with EM algorithms) from sparse Bayesian learning for introducing sparsity into linear state space models and outlined several applications.

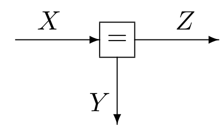
In this paper, all factor graphs were cycle-free so that Gaussian message passing yields exact marginals. The use of NUV regularization in factor graphs with cycles, and its relative merits in comparison with, e.g., AMP [25], remains to be investigated.

Appendix A
Tabulated Gaussian-Message Computations

Tables I–VI are improved versions of the corresponding tables in [6]. The notation for the different parameterizations of the messages was defined in Section IV. The main novelties of this new version are the following:

1) New notation $\overrightarrow{\xi} \triangleq \overrightarrow{W}\overrightarrow{m}$ and $\overleftarrow{\xi} \triangleq \overleftarrow{W}\overleftarrow{m}$.
2) Introduction of the dual marginal $\tilde{\xi}$ (IV.1) with pertinent new expressions in Tables I–V, and new expressions with the dual precision matrix, especially (V.4)–(V.9). These results (from [20]) are used both in Appendix B and in the two preferred algorithms in Section V.
3) New expressions (VI.4)–(VI.9) for the marginals, which are essential for the BIFM Kalman smoother in Section V.

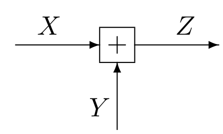
TABLE I
Gaussian message passing through an equality-constraint.



Constraint $X = Y = Z$,
expressed by factor $\delta(z - x)\,\delta(y - x)$

$\overrightarrow{\xi}_Z = \overrightarrow{\xi}_X + \overleftarrow{\xi}_Y$	(I.1)
$\overrightarrow{W}_Z = \overrightarrow{W}_X + \overleftarrow{W}_Y$	(I.2)
$\overleftarrow{\xi}_X = \overleftarrow{\xi}_Y + \overleftarrow{\xi}_Z$	(I.3)
$\overleftarrow{W}_X = \overleftarrow{W}_Y + \overleftarrow{W}_Z$	(I.4)
$m_X = m_Y = m_Z$	(I.5)
$V_X = V_Y = V_Z$	(I.6)
$\tilde{\xi}_X = \tilde{\xi}_Y + \tilde{\xi}_Z$	(I.7)

TABLE II
Gaussian message passing through an adder node.



Constraint $Z = X + Y$,
expressed by factor $\delta(z - (x + y))$

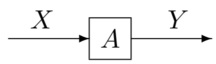
$\overrightarrow{m}_Z = \overrightarrow{m}_X + \overrightarrow{m}_Y$	(II.1)
$\overrightarrow{V}_Z = \overrightarrow{V}_X + \overrightarrow{V}_Y$	(II.2)
$\overleftarrow{m}_X = \overleftarrow{m}_Z - \overrightarrow{m}_Y$	(II.3)
$\overleftarrow{V}_X = \overleftarrow{V}_Z + \overrightarrow{V}_Y$	(II.4)
$m_Z = m_X + m_Y$	(II.5)
$\tilde{\xi}_X = \tilde{\xi}_Y = \tilde{\xi}_Z$	(II.6)
$\tilde{W}_X = \tilde{W}_Y = \tilde{W}_Z$	(II.7)

The proofs (below) are given only for the new expressions; for the other proofs, we refer to [6].

Proof of (I.7): Using (IV.3), (I.3), (I.4), and (I.5), we have

$$\tilde{\xi}_X = \overleftarrow{W}_X m_X - \overleftarrow{\xi}_X \tag{23}$$

TABLE III
GAUSSIAN MESSAGE PASSING THROUGH A MATRIX MULTIPLIER NODE WITH ARBITRARY REAL MATRIX A.



Constraint $Y = AX$, expressed by factor $\delta(y - Ax)$

$$\overrightarrow{m}_Y = A\overrightarrow{m}_X \qquad \text{(III.1)}$$
$$\overrightarrow{V}_Y = A\overrightarrow{V}_X A^{\mathsf{T}} \qquad \text{(III.2)}$$

$$\overleftarrow{\xi}_X = A^{\mathsf{T}}\overleftarrow{\xi}_Y \qquad \text{(III.3)}$$
$$\overleftarrow{W}_X = A^{\mathsf{T}}\overleftarrow{W}_Y A \qquad \text{(III.4)}$$

$$m_Y = Am_X \qquad \text{(III.5)}$$
$$V_Y = AV_X A^{\mathsf{T}} \qquad \text{(III.6)}$$

$$\tilde{\xi}_X = A^{\mathsf{T}}\tilde{\xi}_Y \qquad \text{(III.7)}$$
$$\tilde{W}_X = A^{\mathsf{T}}\tilde{W}_Y A \qquad \text{(III.8)}$$

TABLE IV
GAUSSIAN SINGLE-EDGE MARGINALS (m, V) AND THEIR DUALS $(\tilde{\xi}, \tilde{W})$.

$$
\begin{aligned}
\tilde{\xi}_X &\triangleq \tilde{W}_X(\overrightarrow{m}_X - \overleftarrow{m}_X) && \text{(IV.1)}\\
&= \overrightarrow{\xi}_X - \overrightarrow{W}_X m_X && \text{(IV.2)}\\
&= \overleftarrow{W}_X m_X - \overleftarrow{\xi}_X && \text{(IV.3)}\\
\tilde{W}_X &\triangleq (\overrightarrow{V}_X + \overleftarrow{V}_X)^{-1} && \text{(IV.4)}\\
&= \overrightarrow{W}_X V_X \overleftarrow{W}_X && \text{(IV.5)}\\
&= \overrightarrow{W}_X - \overrightarrow{W}_X V_X \overrightarrow{W}_X && \text{(IV.6)}\\
&= \overleftarrow{W}_X - \overleftarrow{W}_X V_X \overleftarrow{W}_X && \text{(IV.7)}
\end{aligned}
$$

$$
\begin{aligned}
m_X &= V_X(\overrightarrow{\xi}_X + \overleftarrow{\xi}_X) && \text{(IV.8)}\\
&= \overrightarrow{m}_X - \overrightarrow{V}_X\tilde{\xi}_X && \text{(IV.9)}\\
&= \overleftarrow{m}_X + \overleftarrow{V}_X\tilde{\xi}_X && \text{(IV.10)}\\
V_X &= (\overrightarrow{W}_X + \overleftarrow{W}_X)^{-1} && \text{(IV.11)}\\
&= \overrightarrow{V}_X\tilde{W}_X\overleftarrow{V}_X && \text{(IV.12)}\\
&= \overrightarrow{V}_X - \overrightarrow{V}_X\tilde{W}_X\overrightarrow{V}_X && \text{(IV.13)}\\
&= \overleftarrow{V}_X - \overleftarrow{V}_X\tilde{W}_X\overleftarrow{V}_X && \text{(IV.14)}
\end{aligned}
$$

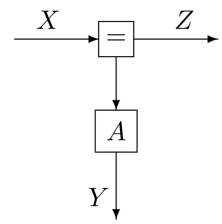
$$
\begin{aligned}
&= \big(\overleftarrow{W}_Y + \overleftarrow{W}_Z\big)m_X - \big(\overleftarrow{\xi}_Y + \overleftarrow{\xi}_Z\big) && (24)\\
&= \big(\overleftarrow{W}_Y m_Y - \overleftarrow{\xi}_Y\big) + \big(\overleftarrow{W}_Z m_Z - \overleftarrow{\xi}_Z\big) && (25)\\
&= \tilde{\xi}_Y + \tilde{\xi}_Z. && (26)
\end{aligned}
$$

□

Proof of (II.6): We first note

$$
\begin{aligned}
\overrightarrow{m}_X - \overleftarrow{m}_X &= \overrightarrow{m}_X + \overrightarrow{m}_Y - \overleftarrow{m}_Z && (27)\\
&= \overrightarrow{m}_Z - \overleftarrow{m}_Z, && (28)
\end{aligned}
$$

TABLE V
GAUSSIAN MESSAGE PASSING THROUGH AN OBSERVATION BLOCK.



$$
\begin{aligned}
\overrightarrow{m}_Z &= \overrightarrow{m}_X + \overrightarrow{V}_X A^{\mathsf{T}} G\,(\overleftarrow{m}_Y - A\overrightarrow{m}_X) && \text{(V.1)}\\
\overrightarrow{V}_Z &= \overrightarrow{V}_X - \overrightarrow{V}_X A^{\mathsf{T}} G A \overrightarrow{V}_X && \text{(V.2)}\\
\text{with } G &\triangleq \big(\overleftarrow{V}_Y + A\overrightarrow{V}_X A^{\mathsf{T}}\big)^{-1} && \text{(V.3)}
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\xi}_X &= F^{\mathsf{T}}\tilde{\xi}_Z + A^{\mathsf{T}}\overleftarrow{W}_Y\big(A\overrightarrow{m}_Z - \overleftarrow{m}_Y\big) && \text{(V.4)}\\
&= F^{\mathsf{T}}\tilde{\xi}_Z + A^{\mathsf{T}} G\,(A\overrightarrow{m}_X - \overleftarrow{m}_Y) && \text{(V.5)}\\
\tilde{W}_X &= F^{\mathsf{T}}\tilde{W}_Z F + A^{\mathsf{T}}\overleftarrow{W}_Y A F && \text{(V.6)}\\
&= F^{\mathsf{T}}\tilde{W}_Z F + A^{\mathsf{T}} G A && \text{(V.7)}\\
\text{with } F &\triangleq I - \overrightarrow{V}_Z A^{\mathsf{T}}\overleftarrow{W}_Y A && \text{(V.8)}\\
&= I - \overrightarrow{V}_X A^{\mathsf{T}} G A && \text{(V.9)}
\end{aligned}
$$

For the reverse direction, replace $\overrightarrow{m}_Z$ by $\overleftarrow{m}_X$, $\overrightarrow{V}_Z$ by $\overleftarrow{V}_X$, $\overrightarrow{m}_X$ by $\overleftarrow{m}_Z$, $\overrightarrow{V}_X$ by $\overleftarrow{V}_Z$, exchange $\tilde{\xi}_X$ and $\tilde{\xi}_Z$, exchange $\tilde{W}_X$ and $\tilde{W}_Z$, and change "+" to "−" in (V.4) and (V.5).

and (II.6) follows from (II.7). □

Proof of (III.7): Using [6, eq. (III.9)], we have

$$
\begin{aligned}
\tilde{\xi}_X &= \tilde{W}_X(\overrightarrow{m}_X - \overleftarrow{m}_X) && (29)\\
&= \tilde{W}_X\overrightarrow{m}_X - \tilde{W}_X\overleftarrow{m}_X && (30)\\
&= A^{\mathsf{T}}\tilde{W}_Y A\overrightarrow{m}_X - A^{\mathsf{T}}\tilde{W}_Y\overleftarrow{m}_Y && (31)\\
&= A^{\mathsf{T}}\tilde{W}_Y(\overrightarrow{m}_Y - \overleftarrow{m}_Y). && (32)
\end{aligned}
$$

□

Proof of (IV.9) and (IV.2): Using (IV.13) and (IV.12), we have

$$
\begin{aligned}
m_X &= V_X\overrightarrow{\xi}_X + V_X\overleftarrow{\xi}_X && (33)\\
&= \Big(\overrightarrow{V}_X - \overrightarrow{V}_X\tilde{W}_X\overrightarrow{V}_X\Big)\overrightarrow{\xi}_X + \overrightarrow{V}_X\tilde{W}_X\overleftarrow{V}_X\overleftarrow{\xi}_X && (34)\\
&= \overrightarrow{m}_X - \overrightarrow{V}_X\tilde{W}_X\left(\overrightarrow{m}_X - \overleftarrow{m}_X\right) && (35)\\
&= \overrightarrow{m}_X - \overrightarrow{V}_X\tilde{\xi}_X, && (36)
\end{aligned}
$$

and (IV.2) follows by multiplication with $\overrightarrow{W}_X$. □

Proof of (V.9): From (I.2) and (III.4), we have

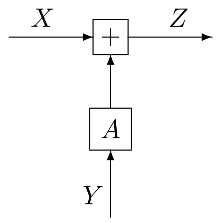
$$\overrightarrow{W}_Z = \overrightarrow{W}_X + A^{\mathsf{T}}\overleftarrow{W}_Y A, \qquad (37)$$

from which we obtain

$$
\begin{aligned}
\overrightarrow{W}_X &= \overrightarrow{W}_Z - A^{\mathsf{T}}\overleftarrow{W}_Y A && (38)\\
&= \overrightarrow{W}_Z F. && (39)
\end{aligned}
$$

TABLE VI

GAUSSIAN MESSAGE PASSING THROUGH AN INPUT BLOCK.



$$\overrightarrow{\xi}_Z = \overrightarrow{\xi}_X + \overrightarrow{W}_X A H (\overleftarrow{\xi}_Y - A^\mathsf{T} \overrightarrow{\xi}_X) \tag{VI.1}$$

$$\overrightarrow{W}_Z = \overrightarrow{W}_X - \overrightarrow{W}_X A H A^\mathsf{T} \overrightarrow{W}_X \tag{VI.2}$$

$$\text{with } H \triangleq \big(\overleftarrow{W}_Y + A^\mathsf{T} \overrightarrow{W}_X A\big)^{-1} \tag{VI.3}$$

$$m_X = \tilde{F}^\mathsf{T} m_Z + A \overrightarrow{V}_Y \big(A^\mathsf{T} \overleftarrow{\xi}_Z - \overrightarrow{\xi}_Y\big) \tag{VI.4}$$

$$= \tilde{F}^\mathsf{T} m_Z + A H \big(A^\mathsf{T} \overleftarrow{\xi}_X - \overrightarrow{\xi}_Y\big) \tag{VI.5}$$

$$V_X = \tilde{F}^\mathsf{T} V_Z \tilde{F} + A \overrightarrow{V}_Y A^\mathsf{T} \tilde{F} \tag{VI.6}$$

$$= \tilde{F}^\mathsf{T} V_Z \tilde{F} + A H A^\mathsf{T} \tag{VI.7}$$

$$\text{with } \tilde{F} \triangleq I - \overleftarrow{W}_Z A \overrightarrow{V}_Y A^\mathsf{T} \tag{VI.8}$$

$$= I - \overleftarrow{W}_X A H A^\mathsf{T} \tag{VI.9}$$

For the reverse direction, replace $\overrightarrow{\xi}_Z$ by $\overleftarrow{\xi}_X$, $\overrightarrow{W}_Z$ by $\overleftarrow{W}_X$, $\overrightarrow{\xi}_X$ by $\overleftarrow{\xi}_Z$, $\overrightarrow{W}_X$ by $\overleftarrow{W}_Z$, exchange m_X and m_Z, exchange V_X and V_Z, and replace $\overleftarrow{\xi}_Y$ by $-\overrightarrow{\xi}_Y$.

Thus $\overrightarrow{V}_Z \overleftarrow{W}_X = F$ and

$$\overrightarrow{V}_Z = F \overrightarrow{V}_X. \tag{40}$$

On the other hand, we have

$$\overrightarrow{V}_Z = \big(I - \overrightarrow{V}_X A^\mathsf{T} G A\big) \overrightarrow{V}_X \tag{41}$$

from (V.2), and $F = I - \overrightarrow{V}_X A^\mathsf{T} G A$ follows. □

Proof of (V.4): Using (I.7), (III.7), and (IV.3), we have

$$\tilde{\xi}_X = \tilde{\xi}_Z + A^\mathsf{T} \tilde{\xi}_Y \tag{42}$$

$$= \tilde{\xi}_Z + A^\mathsf{T} \big(\overleftarrow{W}_Y m_Y - \overleftarrow{W}_Y \overleftarrow{m}_Y\big). \tag{43}$$

Using (III.5) and (IV.9), we further have

$$m_Y = A m_Z \tag{44}$$

$$= A\big(\overrightarrow{m}_Z - \overrightarrow{V}_Z \tilde{\xi}_Z\big), \tag{45}$$

and inserting (45) into (43) yields (V.4). □

Proof of (V.5): We begin with $m_X = m_Z$. Using (IV.9), we have

$$\overrightarrow{m}_X - \overrightarrow{V}_X \tilde{\xi}_X = \overrightarrow{m}_Z - \overrightarrow{V}_Z \tilde{\xi}_Z \tag{46}$$

$$= \overrightarrow{m}_X + \overrightarrow{V}_X A^\mathsf{T} G (\overleftarrow{m}_Y - A \overrightarrow{m}_X) - \overrightarrow{V}_X F^\mathsf{T} \tilde{\xi}_Z, \tag{47}$$

where the second step uses (V.1) and $\overrightarrow{V}_Z = F \overrightarrow{V}_X = (F \overrightarrow{V}_X)^\mathsf{T}$ from (40). Subtracting $\overrightarrow{m}_X$ and multiplying by $\overrightarrow{V}_X^{-1}$ yields (V.5). □

Proof of (V.7): We begin with $V_X = V_Z$. Using (IV.13), we have

$$\overrightarrow{V}_X - \overrightarrow{V}_X \tilde{W}_X \overrightarrow{V}_X = \overrightarrow{V}_Z - \overrightarrow{V}_Z \tilde{W}_Z \overrightarrow{V}_Z \tag{48}$$

$$= \overrightarrow{V}_X - \overrightarrow{V}_X A^\mathsf{T} G A \overrightarrow{V}_X - \overrightarrow{V}_X F^\mathsf{T} \tilde{W}_Z F \overrightarrow{V}_X, \tag{49}$$

where the second step uses (V.2) and (40). Subtracting $\overrightarrow{V}_X$ and multiplying by $\overrightarrow{V}_X^{-1}$ yields (V.7). □

Proof of (V.6): As we have already established (V.7), we only need to prove

$$A^\mathsf{T} G A = A^\mathsf{T} \overleftarrow{W}_Y A F. \tag{50}$$

Using (V.9), we have

$$A^\mathsf{T} G A = \overleftarrow{W}_X (I - F) \tag{51}$$

$$= \overleftarrow{W}_X \overrightarrow{V}_Z A^\mathsf{T} \overleftarrow{W}_Y A \tag{52}$$

$$= A^\mathsf{T} \overleftarrow{W}_Y A \overrightarrow{V}_Z \overleftarrow{W}_X, \tag{53}$$

where the last step follows from $A^\mathsf{T} G A = (A^\mathsf{T} G A)^\mathsf{T}$. Inserting (39) then yields (50). □

Proof of (VI.9): From (II.2) and (III.2), we have

$$\overrightarrow{V}_Z = \overrightarrow{V}_X + A \overrightarrow{V}_Y A^\mathsf{T}, \tag{54}$$

from which we obtain

$$\overrightarrow{V}_X = \overrightarrow{V}_Z - A \overrightarrow{V}_Y A^\mathsf{T} \tag{55}$$

$$= \overrightarrow{V}_Z \tilde{F}. \tag{56}$$

Thus $\overleftarrow{W}_Z \overrightarrow{V}_X = \tilde{F}$ and

$$\overleftarrow{W}_Z = \tilde{F} \overleftarrow{W}_X. \tag{57}$$

On the other hand, we have

$$\overleftarrow{W}_Z = \big(I - \overleftarrow{W}_X A H A^\mathsf{T}\big) \overleftarrow{W}_X \tag{58}$$

from (VI.2), and $\tilde{F} = I - \overleftarrow{W}_X A H A^\mathsf{T}$ follows. □

Proof of (VI.4): Using (II.3), (III.5), and (IV.9), we have

$$m_X = m_Z - A m_Y \tag{59}$$

$$= m_Z - A\big(\overrightarrow{m}_Y - \overrightarrow{V}_Y \tilde{\xi}_Y\big). \tag{60}$$

Using (II.6), (III.7), and (IV.2), we further have

$$\tilde{\xi}_Y = A^\mathsf{T} \tilde{\xi}_Z \tag{61}$$

$$= A^\mathsf{T} \big(\overleftarrow{\xi}_Z - \overleftarrow{W}_Z m_Z\big), \tag{62}$$

and inserting (62) into (60) yields (VI.4). □

Proof of (VI.5): We begin with $\tilde{\xi}_X = \tilde{\xi}_Z$ from (II.6). Using (IV.2), we have

$$\overrightarrow{\xi}_X - \overrightarrow{W}_X m_X = \overrightarrow{\xi}_Z - \overrightarrow{W}_Z m_Z \tag{63}$$

$$= \overrightarrow{\xi}_X + \overrightarrow{W}_X A H (\overleftarrow{\xi}_Y - A^\mathsf{T} \overrightarrow{\xi}_X) - \overrightarrow{W}_X \tilde{F}^\mathsf{T} m_Z, \tag{64}$$

where the second step uses (VI.1) and $\overrightarrow{W}_Z = (\tilde{F}\overrightarrow{W}_X)^{\mathsf{T}}$ from (57). Subtracting $\overrightarrow{\xi}_X$ and multiplying by $\overrightarrow{V}_X$ yields (VI.5). □

Proof of (VI.7): We begin with $\tilde{W}_X = \tilde{W}_Z$ from (II.7). Using (IV.6), we have

$$\overrightarrow{W}_X - \overrightarrow{W}_X V_X \overrightarrow{W}_X = \overrightarrow{W}_Z - \overrightarrow{W}_Z V_Z \overrightarrow{W}_Z \tag{65}$$

$$= \overrightarrow{W}_X - \overrightarrow{W}_X A H A^{\mathsf{T}} \overrightarrow{W}_X - \overrightarrow{W}_X \tilde{F}^{\mathsf{T}} V_Z \tilde{F} \overrightarrow{W}_X, \tag{66}$$

where the second step uses (VI.2) and (57). Subtracting $\overrightarrow{W}_X$ and multiplying by $\overrightarrow{V}_X$ yields (VI.7). □

Proof of (VI.6): Since we have already established (VI.7), we only need to prove

$$AHA^{\mathsf{T}} = A\overrightarrow{V}_Y A^{\mathsf{T}} \tilde{F}. \tag{67}$$

Using (VI.9), we have

$$AHA^{\mathsf{T}} = \overrightarrow{V}_X (I - \tilde{F}) \tag{68}$$

$$= \overrightarrow{V}_X \overrightarrow{W}_Z A \overrightarrow{V}_Y A^{\mathsf{T}} \tag{69}$$

$$= A\overrightarrow{V}_Y A^{\mathsf{T}} \overrightarrow{W}_Z \overrightarrow{V}_X, \tag{70}$$

where the last step follows from $AHA^{\mathsf{T}} = (AHA^{\mathsf{T}})^{\mathsf{T}}$. Inserting (57) then yields (67). □

Appendix B
Message Passing in Figure 1 and Proofs

In this appendix, we demonstrate how all the quantities pertaining to computations mentioned in Sections II and III, as well as the proof of the theorem in Section II, are obtained by symbolic message passing using the tables in Appendix A. The key ideas of this section are from [6, Section V.C].

Throughout this section, $\sigma_1, \ldots, \sigma_K$ are fixed.

A. Key Quantities $\tilde{\xi}_{X_k}$ and $\tilde{W}_{X_k}$

The pivotal quantities of this section are the dual mean vector $\tilde{\xi}_{\tilde{U}_k}$ and the dual precision matrix $\tilde{W}_{\tilde{U}_k}$. Concerning the former, we have

$$\tilde{\xi}_{\tilde{U}_k} = \tilde{\xi}_{X_k} = \tilde{\xi}_{X_0} = \tilde{\xi}_Y \tag{71}$$

$$= -\tilde{W}_Y y, \tag{72}$$

for $k = 1, \ldots, K$, where (71) follows from (II.6), and (72) follows from

$$\tilde{\xi}_Y = \tilde{W}_Y (\overrightarrow{m}_Y - \overleftarrow{m}_Y) \tag{73}$$

$$= -\tilde{W}_Y y \tag{74}$$

since $\overrightarrow{m}_Y = 0$.

Concerning $\tilde{W}_{X_k}$, we have

$$\tilde{W}_{\tilde{U}_k} = \tilde{W}_{X_k} = \tilde{W}_{X_0} = \tilde{W}_Y \tag{75}$$

$$= \tilde{W} \text{ as defined in (8)} \tag{76}$$

for $k = 1, \ldots, K$, where (75) follows from (II.7), and (76) follows from

$$\tilde{W}_Y = \big(\overrightarrow{V}_Y + \overleftarrow{V}_Y\big)^{-1} \tag{77}$$

with $\overleftarrow{V}_Y = 0$ and

$$\overrightarrow{V}_Y = \sum_{k=1}^{K} \sigma_k^2 b_k b_k^{\mathsf{T}} + \sigma^2 I. \tag{78}$$

The matrix $\tilde{W}$ can be computed without matrix inversion as follows. First, we note that

$$\tilde{W}_{X_0} = \big(\overrightarrow{V}_{X_0} + \overleftarrow{V}_{X_0}\big)^{-1} \tag{79}$$

$$= \big(0 + \overleftarrow{V}_{X_0}\big)^{-1} \tag{80}$$

$$= \overleftarrow{W}_{X_0}. \tag{81}$$

Second, using (VI.2), the matrix $\overleftarrow{W}_{X_0}$ can be computed by the backward recursion

$$\overleftarrow{W}_{X_{k-1}} = \overleftarrow{W}_{X_k} - (\overleftarrow{W}_{X_k} b_k)(\sigma_k^{-2} + b_k^{\mathsf{T}} \overleftarrow{W}_{X_k} b_k)^{-1} (\overleftarrow{W}_{X_k} b_k)^{\mathsf{T}} \tag{82}$$

starting from $\overleftarrow{W}_{X_K} = \sigma^{-2} I$. The complexity of this alternative computation of $\tilde{W}$ is $O(n^2 K)$; by contrast, the direct computation of (8) (using Gauss-Jordan elimination for the matrix inversion) has complexity $O(n^2 K + n^3)$.

B. Posterior Distribution and MAP estimate of U_k

For fixed $\sigma_1, \ldots, \sigma_K$, the MAP estimate of U_k is the mean m_{U_k} of the (Gaussian) posterior of U_k. From (IV.9) and (III.7), we have

$$m_{U_k} = \overrightarrow{m}_{U_k} - \overrightarrow{V}_{U_k} \tilde{\xi}_{U_k} \tag{83}$$

$$= 0 - \sigma_k^2 b_k^{\mathsf{T}} \tilde{\xi}_{\tilde{U}_k}, \tag{84}$$

and (72) yields

$$m_{U_k} = \sigma_k^2 b_k^{\mathsf{T}} \tilde{W} y, \tag{85}$$

which proves (7).

For re-estimating the variance σ_k^2 as in Section III, we also need the variance $\sigma_{U_k}^2$ of the posterior distribution of U_k. From (IV.13) and (III.8), we have

$$\sigma_{U_k}^2 = \overrightarrow{V}_{U_k} - \overrightarrow{V}_{U_k} \tilde{W}_{U_k} \overrightarrow{V}_{U_k} \tag{86}$$

$$= \sigma_k^2 - \sigma_k^2 b_k^{\mathsf{T}} \tilde{W}_{\tilde{U}_k} b_k \sigma_k^2, \tag{87}$$

and (76) yields

$$\sigma_{U_k}^2 = \sigma_k^2 - \sigma_k^2 b_k^{\mathsf{T}} \tilde{W} b_k \sigma_k^2. \tag{88}$$

C. Likelihood Function and Backward Message of U_k

We now consider the backward message along the edge U_k, which is the likelihood function $p(y|u_k, \sigma_1, \ldots, \sigma_K)$, for fixed y and fixed $\sigma_1, \ldots, \sigma_K$, up to a scale factor. For use in Section B-D below, we give two different expressions both for the mean $\overleftarrow{m}_{U_k}$ and for the variance $\overleftarrow{\sigma}_{U_k}^2$ of this message.

As to the latter, we have

$$\overleftarrow{W}_{U_k} = b_k^{\mathsf{T}} \overleftarrow{W}_{\tilde{U}_k} b_k \tag{89}$$

from (III.4), and thus

$$\overleftarrow{\sigma}_{U_k}^2 = \big(b_k^{\mathsf{T}} \overleftarrow{W}_{\tilde{U}_k} b_k\big)^{-1}. \tag{90}$$

We also note (from (II.4)) that

$$\overleftarrow{W}_{\tilde{U}_k} = \big(\overrightarrow{V}_{X_{k-1}} + \overleftarrow{V}_{X_k}\big)^{-1} \tag{91}$$
$$= W_k \text{ as defined in (10).} \tag{92}$$

Alternatively, we have

$$\overleftarrow{\sigma}^2_{U_k} = \tilde{W}_{U_k}^{-1} - \overrightarrow{V}_{U_k} \tag{93}$$
$$= (b_k^\mathsf{T} \tilde{W}_{\tilde{U}_k} b_k)^{-1} - \sigma_k^2 \tag{94}$$
$$= (b_k^\mathsf{T} \tilde{W} b_k)^{-1} - \sigma_k^2, \tag{95}$$

where we used (IV.4), (III.8), and (76).

As to the mean $\overleftarrow{m}_{U_k}$, we have

$$\overleftarrow{\xi}_{U_k} = b_k^\mathsf{T} \overleftarrow{\xi}_{\tilde{U}_k} \tag{96}$$
$$= b_k^\mathsf{T} \overleftarrow{W}_{\tilde{U}_k} \overleftarrow{m}_{\tilde{U}_k} \tag{97}$$
$$= b_k^\mathsf{T} \overleftarrow{W}_{\tilde{U}_k} (\overleftarrow{m}_{X_k} - \overrightarrow{m}_{X_{k-1}}) \tag{98}$$
$$= b_k^\mathsf{T} \overleftarrow{W}_{\tilde{U}_k} y \tag{99}$$

from (III.3) and (II.3), and thus

$$\overleftarrow{m}_{U_k} = \overleftarrow{\sigma}^2_{U_k} \overleftarrow{\xi}_{U_k} \tag{100}$$
$$= \big(b_k^\mathsf{T} \overleftarrow{W}_{\tilde{U}_k} b_k\big)^{-1} b_k^\mathsf{T} \overleftarrow{W}_{\tilde{U}_k} y \tag{101}$$

from (90). Alternatively, we have

$$\overleftarrow{m}_{U_k} = \overrightarrow{m}_{U_k} - \tilde{W}_{U_k}^{-1} \tilde{\xi}_{U_k} \tag{102}$$
$$= 0 - (b_k^\mathsf{T} \tilde{W}_{\tilde{U}_k} b_k)^{-1} b_k^\mathsf{T} \tilde{\xi}_{\tilde{U}_k} \tag{103}$$
$$= (b_k^\mathsf{T} \tilde{W} b_k)^{-1} b_k^\mathsf{T} \tilde{W} y, \tag{104}$$

where we used (IV.1), (III.8), (III.7), (72), and (76).

D. Proof of the Theorem in Section II

Let $\sigma_1, \ldots, \sigma_K$ be fixed at a local maximum or at a saddle point of the likelihood $p(y|\sigma_1, \ldots, \sigma_K)$. Then

$$\sigma_k = \underset{\sigma_k}{\operatorname{argmax}}\, p(y|\sigma_1, \ldots, \sigma_K) \tag{105}$$

and

$$\sigma_k^2 = \max\{0, \overleftarrow{m}^2_{U_k} - \overleftarrow{\sigma}^2_{U_k}\} \tag{106}$$

from (2). From (101) and (90), we have

$$\overleftarrow{m}^2_{U_k} - \overleftarrow{\sigma}^2_{U_k} = \frac{\big(b_k^\mathsf{T} \overleftarrow{W}_{\tilde{U}_k} y\big)^2}{\big(b_k^\mathsf{T} \overleftarrow{W}_{\tilde{U}_k} b_k\big)^2} - \frac{1}{b_k^\mathsf{T} \overleftarrow{W}_{\tilde{U}_k} b_k} \tag{107}$$

With (92), it is obvious that $\overleftarrow{m}^2_{U_k} - \overleftarrow{\sigma}^2_{U_k} \leq 0$ if and only if (9) holds.

As to (11), we have

$$\overleftarrow{m}^2_{U_k} - \overleftarrow{\sigma}^2_{U_k} = \frac{\big(b_k^\mathsf{T} \tilde{W} y\big)^2}{\big(b_k^\mathsf{T} \tilde{W} b_k\big)^2} - \frac{1}{b_k^\mathsf{T} \tilde{W} b_k} + \sigma_k^2 \tag{108}$$

from (104) and (95). We now distinguish two cases. If $\sigma_k^2 > 0$, (106) and (108) together imply

$$\frac{\big(b_k^\mathsf{T} \tilde{W} y\big)^2}{\big(b_k^\mathsf{T} \tilde{W} b_k\big)^2} - \frac{1}{b_k^\mathsf{T} \tilde{W} b_k} = 0. \tag{109}$$

On the other hand, if $\sigma_k^2 = 0$, (106) and (108) imply

$$\frac{\big(b_k^\mathsf{T} \tilde{W} y\big)^2}{\big(b_k^\mathsf{T} \tilde{W} b_k\big)^2} - \frac{1}{b_k^\mathsf{T} \tilde{W} b_k} \leq 0. \tag{110}$$

Combining these two cases yields (11).

References

[1] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural Computation,* vol. 11, pp. 305–345, Feb. 1999.

[2] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation.* Prentice Hall, NJ, 2000.

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

[4] C. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[5] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Proc. Mag.,* Jan. 2004, pp. 28–41.

[6] H.-A. Loeliger, J. Dauwels, Junli Hu, S. Korl, Li Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proceedings of the IEEE,* vol. 95, no. 6, pp. 1295–1322, June 2007.

[7] D. J. C MacKay, "Bayesian interpolation," *Neural Comp.,* vol. 4, n. 3, pp. 415–447, 1992.

[8] S. Gull, "Bayesian inductive inference and maximum entropy," in *Maximum-entropy and Bayesian Methods in Science and Engineering,* G. J. Erickson and C. R. Smith, eds., Kluwer 1988, pp. 53–74.

[9] R. M. Neal, *Bayesian Learning for Neural Networks,* New York: Springer Verlag, 1996.

[10] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Research,* vol. 1, pp. 211–244, 2001.

[11] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," *Proc. 9th Int. Workshop on Artificial Intelligence and Statistics,* 2003.

[12] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," *Advances in Neural Information Processing Systems,* pp. 1625–1632, 2008.

[13] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Proc.,* vol. 52, no. 8, Aug. 2004, pp. 2153–2164.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society,* vol. 39, Series B, pp. 1–38, 1977.

[15] P. Stoica and Y. Selén, "Cyclic minimizers, majorization techniques, and the expectation-maximization algorithm: a refresher," *IEEE Signal Proc. Mag.,* January 2004, pp. 112–114.

[16] Z. Ghahramani and G. E. Hinton, *Parameter Estimation for Linear Dynamical Systems.* Techn. Report CRG-TR-96-2, Univ. of Toronto, 1996.

[17] J. Dauwels, A. Eckford, S. Korl, and H.-A. Loeliger, "Expectation maximization as message passing—Part I: principles and Gaussian messages," arXiv:0910.2832.

[18] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, *"Factor graphs and the sum-product algorithm," IEEE Trans. Information Theory,* vol. 47, pp. 498–519, Feb. 2001.

[19] L. Bruderer and H.-A. Loeliger, "Estimation of sensor input signals that are neither bandlimited nor sparse," *2014 Information Theory & Applications Workshop (ITA),* San Diego, CA, Feb. 9–14, 2014.

[20] L. Bruderer, *Input Estimation and Dynamical System Identification: New Algorithms and Results.* PhD thesis at ETH Zurich No 22575, 2015.

[21] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation.* New York: Academic Press, 1977.

[22] L. Bruderer, H. Malmberg, and H.-A. Loeliger, "Deconvolution of weakly-sparse signals and dynamical-system identification by Gaussian message passing," *2015 IEEE Int. Symp. on Information Theory (ISIT),* Hong Kong, June 14–19, 2015.

[23] N. Zalmai, H. Malmberg, and H. A. Loeliger, "Blind deconvolution of sparse but filtered pulses with linear state space models," *41th IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* Shanghai, China, March 20–25, 2016.

[24] F. Wadehn, L. Bruderer, V. Sahdeva, and H.-A. Loeliger, "Outlier-insensitive Kalman smoothing and marginal message passing," in preparation.

[25] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing", *Proc. National Academy of Sciences,* vol. 106, no. 45, pp. 18914–18919, 2009.