Automatic Regularization and Estimation of Multiple Scale Factors in Linear Gaussian and Non-Gaussian Models

Alessio Lukaj
Dept. of Information Technology and Electrical Engineering
ETH Zurich, Switzerland
Email: lukaj@isi.ee.ethz.ch

Federico Wadehn
ZHAW School of Engineering
Winterthur, Switzerland
Email: wadh@zhaw.ch

Hans-Andrea Loeliger
Dept. of Information Technology and Electrical Engineering
ETH Zurich, Switzerland
Email: loeliger@isi.ee.ethz.ch

***Abstract*—The paper considers the joint estimation of multiple constant or time-varying scale factors in linear Gaussian and non-Gaussian models. A recently proposed variational NUP (normal with unknown parameters) representation allows to address such problems by iterated Gaussian message passing in a pertinent factor graph. However, this method requires to adaptively tune a hyperparameter for each scale factor, which hinders its use for estimating multiple scale factors.**

In this paper, we improve this NUP representation and replace that hyperparameter by one that need not be tuned adaptively. We then show the effectiveness and versatility of the proposed approach by two applications: (i) elastic net regression with automatic estimation of the regularization parameters, and (ii) a state space model with joint estimation of piecewise constant driving noise and piecewise constant observation noise.

***Index Terms*—factor graphs, variance estimation, regularization, variational representations.**

I. Introduction

NUP representations (normal with unknown parameters) [1] are a generalization of NUV representations (normal with unknown variance), which originate from sparse Bayesian learning [2]–[5] and from variational representations of loss functions [6], [7]. Such representations can be used to convert nontrivial model fitting problems into iterated versions of least-squares or Gaussian message passing [7]–[12], and thereby offer a prospect of scalable hierarchical and multi-component models that are computationally tractable.

In this paper, we consider the estimation of scale factors of Gaussian or non-Gaussian priors using the NUP approach, for the case where such scale factors are themselves governed by a nontrivial model. A solution to this problem was proposed in [1]. In the present paper, we improve the proposal of [1] by replacing its problematic hyperparameter with one that need not be actively controlled. We then show the effectiveness and versatility of the proposed approach by its application to two problems: first, elastic net regression with automatic estimation of the regularization parameters; second, a state space model with joint estimation of piecewise constant driving noise and piecewise constant observation noise.

Throughout, we will use factor graphs and the message passing notation as in [13], [14].

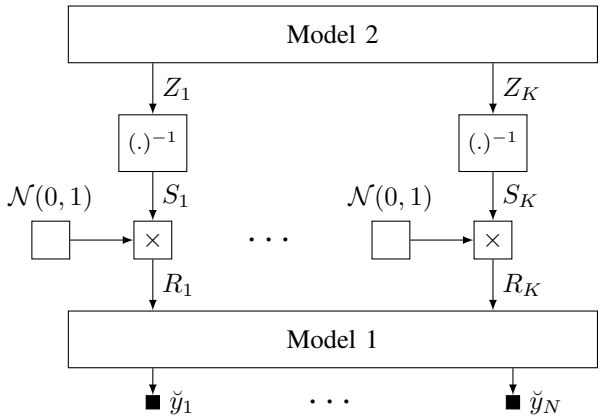

Fig. 1. Factor graph of the setting of Section II.

II. Improving the NUP Representation of a non-Gaussian Message Proposed in [1]

A. Setting and Summary of the Proposal in [1]

Consider a setting as in Fig. 1 where $R_1, \ldots, R_K$ are (scalar) Gaussian random variables that are inputs into some statistical model (Model 1) with given observations $\breve{y}_1, \ldots, \breve{y}_N$. Let $S_1^2, \ldots, S_K^2$ be the variances of $R_1, \ldots, R_K$, respectively; these variances are not fixed, but controlled by a second statistical model (Model 2). For reasons that will become clear below, Model 2 works with $Z_1 \triangleq S_1^{-1}, \ldots, Z_K \triangleq S_K^{-1}$ rather than directly with $S_1, \ldots, S_K$.

Our aim is to estimate all variables (i.e., all R_k, all Z_k, and all variables inside Model 1 and Model 2) by some algorithm that repeats the following two steps until convergence:

1) For fixed $Z_1 = \hat{z}_1, \ldots, Z_K = \hat{z}_K$, compute estimates $\hat{r}_1, \ldots, \hat{r}_K$ of $R_1, \ldots, R_K$. In this step, $R_1, \ldots, R_K$ are independent Gaussian random variables, and we assume that estimation in Model 1 is straightforward in this case.
2) For fixed $R_1 = \hat{r}_1, \ldots, R_K = \hat{r}_K$, compute new estimates $\hat{z}_1, \ldots, \hat{z}_K$ of $Z_1, \ldots, Z_K$ using Model 2.

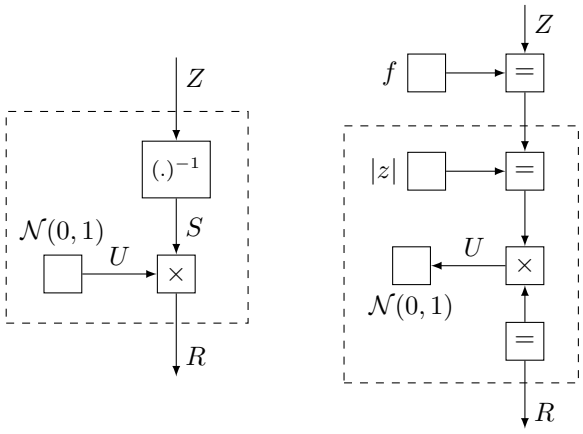

Fig. 2. About the trick proposed in [1]: the dashed boxes left and right represent the same function (2).

The problem here is Step 2, for which we wish to use some algorithm that is designed to work with Gaussian messages, but

$$\overleftarrow{\mu}_{Z_k}(z_k) \propto p(r_k|z_k) \tag{1}$$

is not Gaussian in z_k. Moreover, we do not wish to resort to approximations (such as moment matching or the Laplace approximation [15]).

In [1], a solution to this problem was proposed in form of a NUP representation of (1) as summarized below. In the following, we will focus on a single index k in (1) and drop that index (e.g., $Z = Z_k$).

The derivation in [1] begins with writing (1) as

$$\frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r^2}{2s^2}\right)\bigg|_{s\,=\,z^{-1}} = \frac{|z|}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)\bigg|_{u\,=\,rz} \tag{2}$$

(cf. Fig. 2) and uses the NUP representation of the factor $|z|$ in (2) as

$$|z| = \min_{\sigma \geq 0} g(z,\sigma) \quad \text{with} \quad g(z,\sigma) \triangleq \frac{1}{|\sigma|} \exp\left(\frac{z^2}{2\sigma^2}\right) \sigma^2 e^{-1/2} \tag{3}$$

with unique minimizer $\sigma = |z|$. For fixed $\sigma = |\hat{z}|$ according to some previous estimate $\hat{z}$, $g(z,\sigma)$ is an improper Gaussian factor in z with mean zero and "variance" $\sigma^2 = -\hat{z}^2$, and the resulting message $\overleftarrow{\mu}_Z$ is Gaussian with mean $\overleftarrow{m}_Z = 0$ and variance $\overleftarrow{\sigma}_Z^2 = \left(\hat{r}^2 - \hat{z}^{-2}\right)^{-1}$.

There are still two problems: (i) $\overleftarrow{\sigma}_Z^2$ can be negative and (ii) $\overleftarrow{m}_Z = 0$ independently of r. Both problems can be addressed by a suitable factor $f(z)$ in Fig. 2 that encourages $\hat{z} \geq 0$. In [1], it was proposed to use $f(z) = \exp(-\kappa(z))$ with

$$\kappa(z) = \begin{cases} 0, & \text{if } z \geq 0 \\ \beta|z|, & \text{if } z < 0 \end{cases} \tag{4}$$

for some $\beta > 0$. Using a NUP representation of $f(z)$, we then obtain a Gaussian message $\overleftarrow{\mu}_Z$ with variance and mean given by

$$\overleftarrow{\sigma}_Z^2 = \left(\hat{r}^2 - \hat{z}^{-2} + \beta|\hat{z}|^{-1}\right)^{-1} \quad \text{and} \quad \overleftarrow{m}_Z = \beta\overleftarrow{\sigma}_Z^2. \tag{5}$$

The parameter β needs to be controlled adaptively such that $\overleftarrow{\sigma}_Z^2 \geq 0$. But this fiddling with β affects the point to which the algorithm converges, and we find that controlling multiple such parameters can be challenging.

B. New Proposal

We now propose to choose the factor $f(z)$ in Fig. 2 to have the NUP representation $f(z) = \max_s h(z,s) = 1$ with

$$h(z,s) = \exp\left(-\frac{(z-s)^2}{2(\alpha s)^2}\right) \tag{6}$$

(with hyperparameter α) and update rule $s = \hat{z}$. The factor $f(z)$ does not bias the model, but (6) affects the dynamics of the algorithm of Section II-A by expressing a preference for z to be close to the previous estimate $\hat{z}$ (akin to proximal regularization [16], but with adaptive scaling of "close"). With (6), we obtain

$$\overleftarrow{\sigma}_Z^2 = \left(\hat{r}^2 + (\alpha^{-2}-1)\hat{z}^{-2}\right)^{-1} \quad \text{and} \quad \overleftarrow{m}_Z = \overleftarrow{\sigma}_Z^2 \alpha^{-2} \hat{z}^{-1}. \tag{7}$$

Note that $0 < \alpha \leq 1$ guarantees $\overleftarrow{\sigma}_Z^2 > 0$ for $\hat{r}^2 > 0$. If the algorithm is initialized with $\hat{z} > 0$, $\overleftarrow{m}_Z > 0$ can be maintained throughout.

Empirically, we find that any fixed $\alpha \in (0,1]$ works (but the choice affects the speed of convergence).

C. Modifications for EM Setting

As pointed out in [1], a natural alternative to the alternating maximization algorithm of Section II-A is to estimate $Z_1, \ldots, Z_K$ by expectation maximization (EM) [17], [18] (i.e., an empirical Bayes approach or type-II MAP estimation [2]), which amounts to repeating the following two steps until convergence:

1) For fixed $Z_1 = \hat{z}_1$, …, $Z_K = \hat{z}_K$, compute $\mathbb{E}[R_1^2]$, …, $\mathbb{E}[R_K^2]$ using Model 1.
2) Compute new estimates $\hat{z}_1, \ldots, \hat{z}_K$ of $Z_1, \ldots, Z_K$ using Model 2 with independent Gaussian messages $\overleftarrow{\mu}_{Z_1}$, …, $\overleftarrow{\mu}_{Z_K}$ with

$$\overleftarrow{\sigma}_Z^2 = \left(\mathbb{E}[R^2] + (\alpha^{-2}-1)\hat{z}^{-2}\right)^{-1} \tag{8}$$

and $\overleftarrow{m}_Z$ as in (7).

The details are omitted due to space constraints.

III. Estimating a Single Common Variance

The point of the approach of Section II is to enable the estimation of multiple variances or scale factors in nontrivial multi-component models that are challenging for, or beyond the reach of, classical methods such as cross-validation [1].

Nonetheless, as a sanity check, we now consider the estimation of a single common variance. Specifically, consider a statistical model where $R_1, \ldots, R_K$ are i.i.d. zero-mean Gaussian random variables with unknown variance S^2 and where observations $\breve{Y}_1 = \breve{y}_1, \ldots, \breve{Y}_N = \breve{y}_N$ are given. We first recall two ways to estimate this variance by standard methods.

Fig. 3. Factor graph of Model 2 in Fig. 1 for the setting in Section III.

A. Joint ML/MAP Estimation by Alternating Maximization

Repeat the following two steps until convergence:

1) For fixed $S = \hat{s}$, compute estimates $\hat{r}_1, \ldots, \hat{r}_K$ of $R_1, \ldots, R_K$.
2) For fixed $R_1 = \hat{r}_1, \ldots, R_K = \hat{r}_K$, compute a new estimate

$$\hat{s}^{(\text{new})} = \operatorname*{argmax}_s \prod_{k=1}^{K} p(\hat{r}_k|s) \tag{9}$$

$$= \operatorname*{argmax}_s \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{\hat{r}_k^2}{2s^2}\right) \tag{10}$$

$$= \sqrt{\frac{1}{K}\sum_{k=1}^{K} \hat{r}_k^2}. \tag{11}$$

B. ML Estimation by EM

Expectation maximization with hidden variables $R_1, \ldots, R_K$ (after some calculations) boils down to repeating the following two steps until convergence:

1) For fixed $S = \hat{s}$, compute $\mathbb{E}[R_1^2]$, …, $\mathbb{E}[R_K^2]$.
2) Compute new estimates

$$\hat{s}^{(\text{new})} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[R_k^2]}. \tag{12}$$

C. Using the Proposed Method

Note that we have the situation of Fig. 1 with a general (unspecified) Model 1 while Model 2 is simply a chain of equality constraints as shown in Fig. 3.

Step 2 of the alternating maximization algorithm of Section II amounts to scalar Gaussian message passing in Fig. 3, which can easily be carried out analytically. We thus obtain

$$\hat{s}^{(\text{new})} = \left(\hat{z}^{(\text{new})}\right)^{-1} \tag{13}$$

$$= \left(\hat{r}^2 + (\alpha^{-2} - 1)\hat{s}^2\right)\frac{\alpha^2}{\hat{s}} \tag{14}$$

$$= (1-\alpha^2)\hat{s} + \alpha^2 \frac{\hat{r}^2}{\hat{s}} \tag{15}$$

with $\hat{s} \triangleq \hat{s}^{(\text{old})}$ and

$$\hat{r}^2 \triangleq \frac{1}{K}\sum_{k=1}^{K} \hat{r}_k^2 \quad \text{or} \quad \hat{r}^2 \triangleq \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[R_k^2], \tag{16}$$

where the second version results from using (8) instead of (7). It is easily verified that (15) with the two versions of (16) has the same fixed points as (11) and (12), respectively.

We empirically observe that $\alpha = 1$ works, but (in this example) smaller values of α make the algorithm converge faster.

IV. Automatic Elastic-Net Regularization

In this section, we consider a model that is more complex, but still within the reach of specialized prior-art methods.

A. The Problem

Consider the elastic net problem [19], which is to compute the minimizer $u \in \mathbb{R}^K$ of

$$\frac{1}{2\sigma_0^2}\|\breve{y} - Au\|_2^2 + \frac{1}{s_1}\sum_{k=1}^{K}|u_k| + \frac{1}{2s_2^2}\sum_{k=1}^{K} u_k^2 \tag{17}$$

for given $A \in \mathbb{R}^{N\times K}, \breve{y} \in \mathbb{R}^N$, and $\sigma_0, s_1, s_2 > 0$. We are here interested in automatically determining also "good" parameters σ_0, s_1, s_2 in a Bayesian setting. (One of these parameters is obviously redundant, but we will find it helpful to use all three of them.)

B. Statistical Model and Factor Graph

For fixed σ_0, s_1, s_2, the minimizer u of (17) equals the maximizer u of

$$\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(c_n u - \breve{y}_n)^2}{2\sigma_0^2}\right) p_{S_1}(s_1) p_{S_2}(s_2) \cdot \prod_{k=1}^{K} \frac{1}{2s_1}\exp\left(-\frac{|u_k|}{s_1}\right) \cdot \frac{1}{\sqrt{2\pi s_2^2}}\exp\left(-\frac{u_k^2}{2s_2^2}\right), \tag{18}$$

where $c_n \in \mathbb{R}^{1\times K}$ denotes the nth row of A, and p_{S_1} and p_{S_2} are optional (and possibly improper) priors on S_1 and S_2, respectively. Note that (18) admits a meaningful joint MAP estimate of σ_0, s_1, s_2 and u.

Let $b_k = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathsf{T}} \in \mathbb{R}^K$ (with a single "1" at position k). Using the decomposition

$$X_k \triangleq X_{k-1} + b_k u_k \tag{19}$$

with $X_0 \triangleq 0$ (resulting in $X_K = U$), and using $Z_0 \triangleq \sigma_0^{-1}$, $Z_1 \triangleq S_1^{-1}$, and $Z_2 \triangleq S_2^{-1}$, the model (18) can be represented by the factor graph in Fig. 4, where p_{S_1} and p_{S_2} are replaced by equivalent priors p_{Z_1} and p_{Z_2}, respectively.

C. Required NUP Priors

The factor graph of Fig. 4 is amenable to maximization by Gaussian message passing, treating the inverter nodes as proposed in Section II. In addition, we need the well-known NUP representation of the Laplace distribution [6], [7],

$$\mathcal{L}(e_k; 0, 1) = \frac{1}{2}\exp(-|e_k|), \tag{20}$$

which yields

$$\vec{m}_{E_k} = 0 \quad \text{and} \quad \vec{\sigma}_{E_k}^2 = |\hat{e}_k|. \tag{21}$$

With this NUP representation, a Gaussian message $\overleftarrow{\mu}_{Z_k'}$ can be obtained using (7) or (8); from the latter, we obtain

$$\overleftarrow{\sigma}_{Z_k'}^2 = \left(\frac{\mathbb{E}[U_k^2]}{\vec{\sigma}_{E_k}^2} + (\alpha^{-2} - 1)\hat{s}_1^2\right)^{-1} \text{ and } \overleftarrow{m}_{Z_k'} = \overleftarrow{\sigma}_{Z_k'}^2 \alpha^{-2}\hat{s}_1. \tag{22}$$

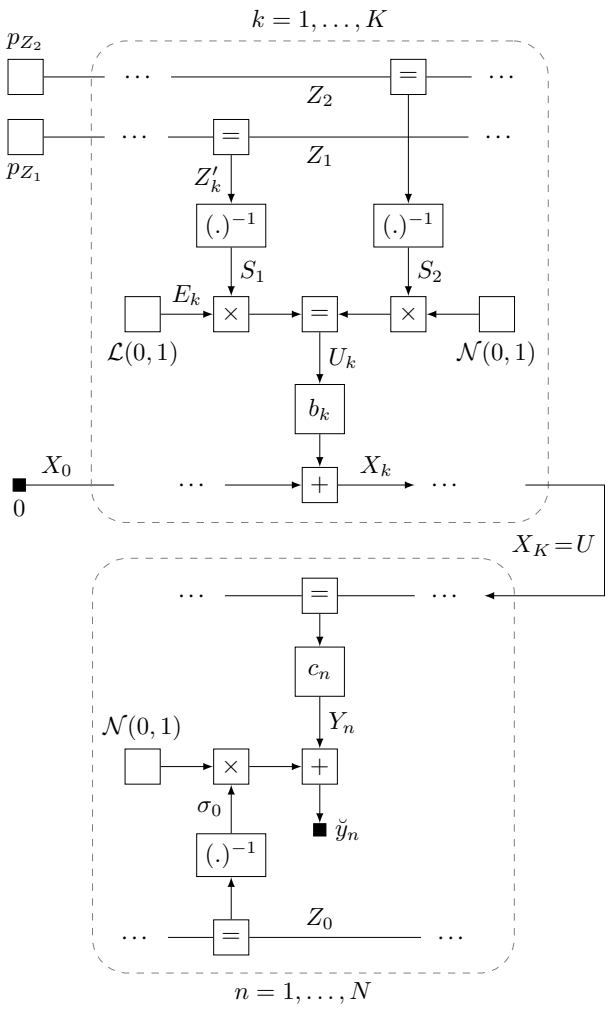

Fig. 4. Factor graph of the Elastic Net model (18) with unknown scale factors σ_0, S_1, S_2. The variables $X_0, \ldots, X_K = U$ take values in $\mathbb{R}^K$; all other variables are scalars. The factor $\mathcal{L}(0,1)$ denotes (20).

We further (heuristically) choose the (improper) priors p_{Z_1} and p_{Z_2} to have the NUV form

$$p_{Z_\nu}(z;\sigma_\nu) \propto \mathcal{N}(z;0,\sigma_\nu^2)\sigma_\nu^{1-\beta_\nu} \tag{23}$$

with $\beta_\nu \geq 1$, which discourages $\hat{z}_\nu \to \infty$ (i.e., $\hat{s}_\nu \to 0$). The EM update rule for σ_ν is well-known [1, eq. (II.3)] to be

$$\sigma_\nu^2 = \mathbb{E}[Z_\nu^2]/\beta_\nu. \tag{24}$$

D. Proposed Algorithm

The proposed algorithm iterates between Gaussian message passing for fixed parameters σ_0, S_1, S_2, and updating these parameters using (8). Specifically:

1) Initialize $\hat{\sigma}_0, \hat{s}_1, \hat{s}_2, \overrightarrow{\sigma}^2_{E_k}, \hat{\sigma}_1, \hat{\sigma}_2$. Set $\beta_1 = \beta_2 = 1$ in (23).
2) For fixed $\sigma_0 = \hat{\sigma}_0$, $S_1 = \hat{s}_1$, $S_2 = \hat{s}_2$, compute $\mathbb{E}[U_k]$, $\mathbb{E}[U_k^2]$, $\mathbb{E}[Y_n^2]$, and $\hat{e}_k = \mathbb{E}[U_k]/\hat{s}_1$ by Gaussian message passing (preferably by the BIFM algorithm or the augmented MBF algorithm [14]).
3) Only every 5th iteration: compute $\mathbb{E}[Z_0], \mathbb{E}[Z_1], \mathbb{E}[Z_2]$ as well as $\mathbb{E}[Z_1^2]$ and $\mathbb{E}[Z_2^2]$ using (8) and (22) and scalar Gaussian message passing. Then set $\hat{\sigma}_0 = \mathbb{E}[Z_0]^{-1}$, $\hat{s}_1 = \mathbb{E}[Z_1]^{-1}$, $\hat{s}_2 = \mathbb{E}[Z_2]^{-1}$. Then update $\hat{\sigma}_1$ and $\hat{\sigma}_2$ by (24).
4) Update $\overrightarrow{\sigma}^2_{E_k}$ using (21).
5) Repeat from Step 2 until convergence, or until $\hat{s}_1$ or $\hat{s}_2$ are smaller than 10^{-4}; in the latter case, increase both β_1 and β_2 by 5 and restart with Step 1.

The computational complexity is dominated by Step 2. Concerning the last step, increasing β_1 and β_2 more than once is rarely necessary.

Obviously, there are several heuristic choices in this algorithm. In particular, the priors (23) and the increase of its parameters β_1 and β_2 in Step 5 are heuristics to prevent the algorithm from diverging with $\hat{z}_1 \to \infty$ or $\hat{z}_2 \to \infty$.

E. Numerical Experiments

We demonstrate the viability of the NUP approach by comparing the proposed algorithm with two specialized prior-art methods for a linear regression problem with automatic feature selection.

In the experiments, each row of the matrix $A \in \mathbb{R}^{N\times K}$ is independently generated according to some probability law as specified below. The observation vector $\breve{y}$ is generated according to $\breve{y} = Au + W$, where $u \in \mathbb{R}^K$ is a fixed ground truth (with some zero components) and the noise W is a random vector with i.i.d. normal components with mean zero and variance σ_W^2. For testing, a second matrix $A' \in \mathbb{R}^{N'\times K}$ and a second observation vector $\breve{y}' \in \mathbb{R}^{N'}$ is generated in the same way (and with the same ground truth u) as A and $\breve{y}$, respectively.

We consider three different scenarios, which are similar to those in [20]:

Scenario 1: $u = (3, 1.5, 0, 0, 2, 0, 0, 0)^\mathsf{T}$, $N = 40$, $N' = 200$, and $\sigma_W^2 = 9$. Each row $(a_1, \ldots, a_8)$ of $A \in \mathbb{R}^{N\times 8}$ and $A' \in \mathbb{R}^{N'\times 8}$ is generated according to the multivariate normal distribution with mean 0, variance 1, and correlation $0.5^{|k-k'|}$ between a_k and $a_{k'}$.

Scenario 2: $u = (\underbrace{3, ..., 3}_{15}, \underbrace{0, ..., 0}_{15})^\mathsf{T}$, $N = 200$, $N' = 400$, and $\sigma_W^2 = 25$. Each row $(a_1, \ldots, a_{30})$ of $A \in \mathbb{R}^{N\times 30}$ and $A' \in \mathbb{R}^{N'\times 30}$ is generated as follows. First, b_1, b_2, and $b_3 \in \mathbb{R}$ are generated independently from $\mathcal{N}(0,1)$. Then $a_k = b_1 + \epsilon_k$ for $k = 1, \ldots, 5$, $a_k = b_2 + \epsilon_k$ for $k = 6, \ldots, 10$, $a_k = b_3 + \epsilon_k$ for $k = 11, \ldots, 15$, where $\epsilon_1, \ldots, \epsilon_{15}$ are generated i.i.d. from $\mathcal{N}(0, 0.01)$. Finally, $a_{16}, \ldots, a_{30}$ are generated i.i.d. from $\mathcal{N}(0,1)$.

Scenario 3: $u = (\underbrace{3, ..., 3}_{10}, \underbrace{0, ..., 0}_{10}, \underbrace{3, ..., 3}_{10})^\mathsf{T}$, $N = 25$, $N' = 30$, and everything else is as in Scenario 2. Note that $N < K$ in this case.

We report three different figures of merit:

Median MSE on test data: the median of $\frac{1}{N'}\|A'\hat{u} - \breve{y}'\|^2$ over 500 trials, each trial with new matrices A and A', and new $\breve{y}$ and $\breve{y}'$. We also give the bootstrapped standard deviation (std) obtained from 500 samples (with replacement) from these 500 MSE values.

TABLE I

EXPERIMENTAL RESULTS FOR THE ELASTIC NET: STANDARD ELASTIC NET WITH 10-FOLD CROSS-VALIDATION [19], BAYESIAN ELASTIC NET [20], AND PROPOSED METHOD (WITH $\alpha = 1$; OTHER VALUES OF α YIELD VIRTUALLY IDENTICAL SCORES). THE BEST NUMERICAL SCORES ARE SET IN **BOLD**.

	Scenario 1 $\left(A \in \mathbb{R}^{40\times 8}\right)$			Scenario 2 $\left(A \in \mathbb{R}^{200\times 30}\right)$			Scenario 3 $\left(A \in \mathbb{R}^{25\times 30}\right)$		
Algorithm	median MSE (boot. std.)	MCC	average $\|\hat{u}-u\|_2$	median MSE (boot. std.)	MCC	average $\|\hat{u}-u\|_2$	median MSE (boot. std.)	MCC	average $\|\hat{u}-u\|_2$
ElNet with CV [19]	10.61 (0.08)	**0.51**	1.46	**26.53 (0.11)**	**0.72**	4.28	69.88 (1.47)	0.29	9.62
Bayes. ElNet [20]	10.82 (0.06)	**0.51**	1.55	28.89 (0.20)	0.47	5.01	88.43 (2.72)	**0.55**	7.84
Proposed	**10.54 (0.07)**	0.44	**1.43**	27.00 (0.09)	0.43	**4.25**	**63.88 (1.22)**	0.52	**6.24**

MCC: The Matthews correlation coefficient [21] measures the success in identifying the nonzero components of u:

$$\text{MCC} = \frac{\text{TP}\cdot\text{TN} - \text{FP}\cdot\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{25}$$

where TP, TN, FP, FN are the entries of the confusion matrix, and MCC $= 1$ is the perfect score. We identify $\hat{u}_k$ as "zero" if $|\hat{u}_k| < 0.1$.

Estimation error: the average of $\|\hat{u}-u\|_2$ over the 500 trials.

Table I reports the results with the algorithm of Section IV-D and two prior-art methods: In [19], the (effectively two) scale factors in (17) are determined by 10-fold cross-validation; in [20], $\hat{u}$ is determined by Gibbs sampling on a probability density function that is more complicated than (18).

Note that Table I supports the claim that, in these simulations, the proposed method is competitive with specialized prior-art methods. Note also that the proposed method can easily be adapted to LAD-lasso [22], fused lasso [23], group lasso [24], etc.

V. ESTIMATING PIECEWISE CONSTANT OBSERVATION NOISE AND PIECEWISE CONSTANT DRIVING NOISE OF A RANDOM WALK

Estimating the variance of the driving noise in a state space problem is not an easy problem [25]–[28], and the joint estimation of time-varying driving noise variance and time-varying observation noise variance remains challenging. We now demonstrate by an example that the proposed approach can handle such problems. (Quantitative evaluations and comparisons are beyond the scope of the present paper.)

Consider a simple state space model with a scalar state X_k that develops according to

$$X_{k+1} = X_k + U_k \tag{26}$$

$$\breve{Y}_k = X_k + R_k, \tag{27}$$

with driving noise $U_k \sim \mathcal{N}(0, \sigma^2_{U_k})$ such that, conditioned on $\sigma^2_{U_k}$ (for all k), $U_k, U_{k+1} \ldots$ are independent. The variances $\sigma^2_{U_k}$ are not fixed, but piecewise constant with unknown level jumps at unknown change points. Likewise, the observation noise $R_k \sim \mathcal{N}(0, \sigma^2_{R_k})$ has piecewise constant variances $\sigma^2_{R_k}$ with unknown level jumps at unknown change points. We observe $\breve{Y}_k = \breve{y}_k$ (for all k) and wish to jointly estimate X_k, σ_{U_k}, and σ_{R_k}.

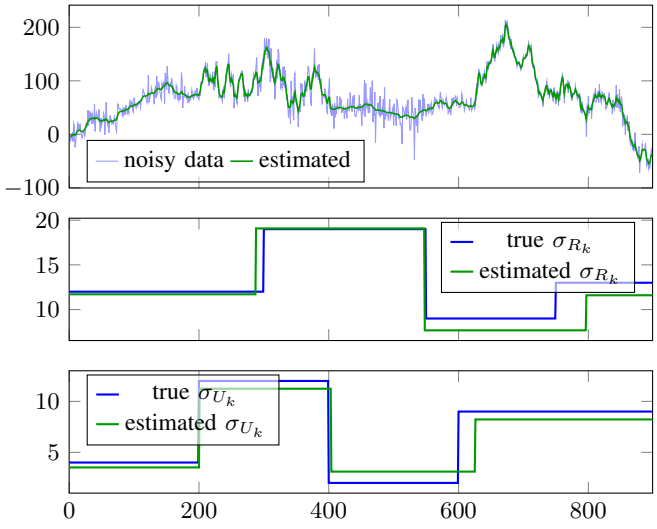

Fig. 5. Simulation example for Section V. Top: given noisy data $\breve{y}_k$ and final estimate of X_k. Middle: true and estimated observation noise std. dev. σ_{R_k}. Bottom: true and estimated driving noise std. dev. σ_{U_k}.

As suggested in [1], we use a state space model for $Z_{U_k} \triangleq \sigma^{-1}_{U_k}$ of the form

$$Z_{U_{k+1}} = Z_{U_k} + J_{Z_k} \tag{28}$$

with sparse i.i.d. $J_{Z_k}, J_{Z_{k+1}}, \ldots$ (with a sparsifying NUP prior), and likewise for $Z_{R_k} \triangleq \sigma^{-1}_{R_k}$. The estimation amounts to iterating the following steps until convergence:

1) For fixed $\sigma^2_{U_k}$ and $\sigma^2_{R_k}$, compute posterior means and variances of X_k, U_k, R_k by Gaussian message passing.
2) Using (8), estimate Z_{U_k} by Gaussian message passing, and likewise for Z_{R_k}.

An exemplary simulation is shown in Fig. 5. The improvement proposed in Section II is critical for the stability and reliability of the algorithm.

VI. CONCLUSION

We improved the NUP representation from [1] by replacing its annoying hyperparameter by one that need not be actively controlled. We then demonstrated the use of this approach by its application to automatic elastic net regularization and to joint estimation of observation noise and driving noise in a state space model. The point of these examples is not the specific application, but to demonstrate the viability, versatility, and scalability of the (now improved) proposal of [1].

REFERENCES

[1] H.-A. Loeliger, "On NUP priors and Gaussian message passing," *IEEE 33rd Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Rome, Italy, 2023.

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

[3] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, Jun. 2001.

[4] J. Palmer, D. P. Wipf, K. Kreutz–Delgado, and B. D. Rao, "Variational EM algorithms for non-Gaussian latent variable models," *Advances in Neural Information Processing Systems (NIPS)*, vol. 18, 2005.

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

[6] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[7] H.-A. Loeliger, Boxiao Ma, H. Malmberg and F. Wadehn, "Factor Graphs with NUV Priors and Iteratively Reweighted Descent for Sparse Least Squares and More," *IEEE 10th Int. Symposium on Turbo Codes & Iterative Information Processing (ISTC)*, Hong Kong, China, 2018.

[8] G. Marti, R. Keusch, and H.-A. Loeliger, "Multiuser MIMO detection with composite NUV priors," *11th Int. Symposium on Topics in Coding 2021 (ISTC)*, Montreal, Canada, Aug. 30 – Sept. 3, 2021.

[9] Boxiao Ma, N. Zalmai, and H.-A. Loeliger, "Smoothed-NUV priors for imaging," *IEEE Trans. on Image Processing*, vol. 31, pp. 4663-4678, July 2022.

[10] R. Keusch and H.-A. Loeliger, "Model-predictive control with NUP priors," arXiv:2303.15806.

[11] Yun Peng Li and H.-A. Loeliger, "Backward filtering forward deciding in linear non-Gaussian state space models," *7th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, May 2024.

[12] Yun Peng Li and H.-A. Loeliger, "Dual NUP representations and min-maximization in factor graphs," *IEEE Int. Symposium on Information Theory (ISIT)*, June 2025.

[13] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. 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, 2007.

[14] H.-A. Loeliger, L. Bruderer, H. Malmberg, F. Wadehn and N. Zalmai, "On sparsity by NUV-EM, Gaussian message passing, and Kalman smoothing," *Information Theory and Applications Workshop (ITA)*, La Jolla, CA, USA, 2016.

[15] F. Wadehn, T. Weber and H.-A. Loeliger, "State space models with dynamical and sparse variances, and inference by EM message passing," *27th European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, 2019.

[16] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

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

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

[19] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, April 2005.

[20] Q. Li and N. Lin, "The Bayesian elastic net," *Bayesian Analysis*, vol. 5, no. 1, pp. 151–170, March 2010.

[21] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 6, 2020.

[22] H. Wang, G. Li and G. Jiang, "Robust regression shrinkage and consistent variable selection through the LAD-lasso," *Journal of Business & Economic Statistics*, vol. 25, no. 3, pp. 347–355, July 2007.

[23] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society,* vol. 67, no. 1, pp. 91–108, Feb. 2005.

[24] M. Yuan, Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society*, vol. 84, no. 5, pp. 49–67, Dec. 2005.

[25] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Series Anal.*, vol. 3, no. 4, pp. 253–264, 1982.

[26] R. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Trans. on Automatic Control*, vol. 15, no. 2, pp. 175–184, April 1970.

[27] L. Zhang, D. Sidoti, A. Bienkowski, K. R. Pattipati, Y. Bar-Shalom and D. L. Kleinman, "On the identification of noise covariances and adaptive Kalman filtering: a new look at a 50 year-old problem," in *IEEE Access*, vol. 8, pp. 59362–59388, 2020.

[28] S. Sarkka and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *IEEE Trans. on Automatic Control*, vol. 54, no. 3, pp. 596–600, 2009.