ACCEPTED FOR MLSP 2023

ON NUP PRIORS AND GAUSSIAN MESSAGE PASSING

Hans-Andrea Loeliger

ETH Zurich, Switzerland

ABSTRACT

Normals with unknown parameters (NUP) can represent many useful priors, and they allow to convert nontrivial model-based estimation problems into iterations of least-squares problems or linear-Gaussian estimation problems. Sparsity inducing NUP priors have been known for some time, and NUP priors for enforcing inequality constraints and discrete-level constraints have been proposed recently.

We review this approach, and we develop it further by proposing a NUP representation of certain non-Gaussian messages that occur in hierarchical models. For illustration, we use a state space model with piecewise constant observation noise variance.

Index Terms— Iteratively reweighted least squares, Gaussian message passing, NUV priors, factor graphs, jump Markov processes, variance estimation, outliers

1. INTRODUCTION

Normals with unknown variance (NUV) are a central idea of sparse Bayesian learning [1–4], and they are closely related to variational representations of sparsifying prior as in [5]. NUP representations (normal with unknown parameters) of binarizing priors [6] and of inequality constraints [7] have been proposed recently. The main attraction of such NUV and NUP priors is that they allow to convert non-Gaussian estimation problems into iteratively reweighted least-squares or iterations of linear-Gaussian estimation [8–10].

In consequence, NUP priors open a perspective of scalable model-based estimation and optimization far beyond variations of sparse recovery, cf. [10].

This paper is semi-tutorial in the sense that we first review the NUV/NUP approach (Sections 2 and 3). We then point out that NUP representations are not restricted to priors, but can also be used for messages in a factor graph [11]. Specifically, we will address a problem that arises in hierarchical modeling where variances of priors in a first-layer model are controlled by a second-layer model. The connection between these two models involves non-Gaussian messages, which we will show to admit a (exact) NUP representation. For illustration, we use a state space model with piecewise constant observation noise variance.

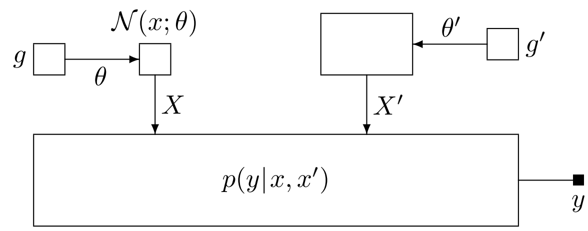

Fig. 1. Factor graph of system model (1) with fixed observation(s) $Y = y$, with a NUP prior on some variable (or parameter) X, and with X' subsuming other such variables.

The following notation will be used. $\|x\|$ denotes the Euclidean norm of $x \in \mathbb{R}^n$. $\mathcal{N}(x; m, V)$ and $\mathcal{N}(m, V)$ denote the normal probability density function with mean vector m and covariance matrix V, with or without argument x.

We will use Forney-style factor graphs as in [11, 12], cf. Fig. 1. For a directed edge X, the forward message and the backward messages along X will be denoted by $\overrightarrow{\mu}_X$ and $\overleftarrow{\mu}_X$, respectively.

2. BRIEF REVIEW OF NUP APPROACH—PART I: SYSTEM LEVEL

Consider a statistical system model of the form

$$f(y, x, x', \theta, \theta') \triangleq p(y|x, x')p(x; \theta)g(\theta)p(x'; \theta')g'(\theta') \quad (1)$$

with observation(s) Y and additional variables (random variables or parameters) X and X' such that, for fixed parameters θ and θ',

$$p(y, x, x'; \theta, \theta') \triangleq p(y|x, x')p(x; \theta)p(x'; \theta') \quad (2)$$

is a Gaussian probability density function in y, x, x'. The functions $g(\theta)$ and $g'(\theta')$ are chosen such that the factors $p(x; \theta)g(\theta)$ and $p(x'; \theta')g'(\theta')$ express some desired prior or constraint on X and X', respectively, as will be detailed in Section 3.

For fixed observation(s) $Y = y$, the variables X and X' and the parameters θ and θ' are estimated by the methods described below.

2.1. Joint MAP with Alternating Maximization

In this approach, the joint estimate is

$$(\hat{x}, \hat{x}', \hat{\theta}, \hat{\theta}') = \operatorname*{argmax}_{x,x',\theta,\theta'} f(y, x, x', \theta, \theta'). \tag{3}$$

It follows that

$$(\hat{x}, \hat{x}') = \operatorname*{argmax}_{x,x'} p(y|x, x')\rho(x)\rho'(x'), \tag{4}$$

where

$$\rho(x) \triangleq \max_{\theta} p(x;\theta)g(\theta) \tag{5}$$

and $\rho'(x') \triangleq \max_{\theta'} p(x';\theta')g'(\theta')$ are the effective priors on X and X', respectively.

The maximization in (3) is carried out by iterating the following two steps for $\ell = 1, 2, \ldots,$ until convergence:

Step 1: For fixed $\theta = \theta^{(\ell)}$ and $\theta' = \theta'^{(\ell)}$, compute

$$\begin{aligned} \big(x^{(\ell)}, x'^{(\ell)}\big) &= \operatorname*{argmax}_{x,x} f(y, x, x', \theta, \theta') && (6) \\ &= \operatorname*{argmax}_{x,x'} p(y|x, x')p(x;\theta)p(x';\theta') && (7) \\ &= \big(\mathrm{E}[X],\, \mathrm{E}[X']\big), && (8) \end{aligned}$$

where the expectation is conditioned on $Y = y$ and with fixed $\theta = \theta^{(\ell)}$ and $\theta' = \theta'^{(\ell)}$.

Step 2: For fixed $x = x^{(\ell)}$, compute

$$\theta^{(\ell+1)} = \operatorname*{argmax}_{\theta} p(x;\theta)g(\theta) \tag{9}$$

and likewise $\theta'^{(\ell+1)}$.

2.2. Type II MAP with EM [1, 2]

In this approach, we first determine the MAP estimate

$$(\hat{\theta}, \hat{\theta}') = \operatorname*{argmax}_{\theta,\theta'} f(y, \theta, \theta') \tag{10}$$

where

$$f(y, \theta, \theta') \triangleq \int\!\!\int f(y, x, x', \theta, \theta')\, dx\, dx'; \tag{11}$$

subsequently, we determine the estimate $(\hat{x}, \hat{x}')$ as in (7).

The maximization in (10) is carried out by expectation maximization (EM) with hidden variables X and X', which amounts to iterating

$$\big(\theta^{(\ell+1)}, \theta'^{(\ell+1)}\big) = \operatorname*{argmax}_{\theta,\theta'} \mathrm{E}\Big[\log f(y, X, X', \theta, \theta')\Big], \tag{12}$$

for $\ell = 1, 2, \ldots,$ where the expectation is conditioned on $Y = y$ and computed with $\theta = \theta^{(\ell)}$ and $\theta' = \theta'^{(\ell)}$. The point of using EM here is that the maximization in (12) splits into

$$\theta^{(\ell+1)} = \operatorname*{argmax}_{\theta} \mathrm{E}\big[\log p(X;\theta)g(\theta)\big] \tag{13}$$

and likewise for $\theta'^{(\ell+1)}$.

For the sake of clarity, we now specialize to

$$p(x;\theta) = \frac{1}{(2\pi\sigma^2)^{\nu/2}} \exp\!\left(\frac{-\|x\|^2}{2\sigma^2}\right) \tag{14}$$

for $x \in \mathbb{R}^\nu$ and $\theta = \sigma \geq 0$. Then (13) becomes

$$\theta^{(\ell+1)} = \operatorname*{argmin}_{\sigma\geq 0} \left(\frac{\mathrm{E}\big[\|X\|^2\big]}{2\sigma^2} + \nu \ln\sigma - \ln g(\sigma)\right), \tag{15}$$

and likewise for $\theta'^{(\ell+1)}$. Thus iterating (12) amounts to iterating the following two steps until convergence:

Step 1: For fixed $\theta = \theta^{(\ell)}$ and $\theta' = \theta'^{(\ell)}$, compute $\mathrm{E}\big[\|X\|^2\big]$ and $\mathrm{E}\big[\|X'\|^2\big]$.

Step 2: Compute (15) and likewise $\theta'^{(\ell+1)}$.

2.3. Remarks

1. Step 1 of Sec. 2.1 amounts to a least-squares problem. Step 1 of Sec. 2.2 amounts to computing posterior means and variances in a Gaussian model. Both cases can be handled by Gaussian message passing in a cycle-free factor graph of (2), cf. [11].
2. The methods of Sections 2.1 and 2.2 can get stuck in a local minimum or a saddle point.
3. The maximizations in (7) and (12) can be replaced by ascent steps.
4. The methods of Sections 2.1 and 2.2 can be mixed in various ways.
5. The maximizations in (5), (9), and (13) can sometimes be replaced by minimizations, cf. the comment below (17).

3. BRIEF REVIEW OF NUP APPROACH—PART II: SELECTED NUP "PRIORS" OLD AND RECENT

A selection of some useful NUP priors is given in Box 1, Box 2, and Table 1. The update rules for the parameter(s) θ of $p(x;\theta)$ are given in terms of the current estimate $\hat{x} = \mathrm{E}[X]$; for the EM update rules, we also need $\mathrm{E}[\|X\|^2]$ or $\mathrm{Var}[X]$, which are all computed with the current parameters (cf. Section 2).

Box 1 is about representing the "prior" (I.1) (with $p > 0$ and $\beta > 0$) in the form (5), with $p(x;\theta) = \mathcal{N}(x; 0, (s^2 + r^2)I)$ with fixed $r^2 \geq 0$. The scale factor

$$\gamma \triangleq \exp\!\left(-\frac{\beta(2-p)}{2}(\beta p r^2)^{\frac{p}{2-p}}\right) \tag{16}$$

$$\rho(x) = \begin{cases} \exp(-\beta\|x\|^p), & \text{if } \|x\|^{2-p} > \beta p r^2 \\ \gamma \exp\!\left(\frac{-\|x\|^2}{2r^2}\right), & \text{if } \|x\|^{2-p} \le \beta p r^2 \end{cases} \quad \text{(I.1)}$$

with γ as in (16) is obtained with the update rule

$$s^2 = \begin{cases} \dfrac{\|\hat{x}\|^{2-p}}{\beta p} - r^2, & \text{if } \|\hat{x}\|^{2-p} > \beta p r^2 \\ 0, & \text{if } \|\hat{x}\|^{2-p} \le \beta p r^2 \end{cases} \quad \text{(I.2)}$$

In particular, $\rho(x) = \exp(-\beta\|x\|)$ is obtained with

$$s^2 = \frac{\|\hat{x}\|}{\beta} \quad \text{(I.3)}$$

Box 1. NUV representation of $\exp(-\beta\|x\|^p)$ with $p > 0$ and $\beta > 0$, optionally with a Gaussian patch around the origin.

$$\rho(x) = \frac{1}{\|x\|^\beta} = \exp(-\beta \ln \|x\|) \quad \text{(II.1)}$$

is obtained with the update rule

$$s^2 = \frac{\|\hat{x}\|^2}{\beta} \quad \text{(II.2)}$$

EM update rule:

$$s^2 = \frac{\mathrm{E}\big[\|X\|^2\big]}{\beta} \quad \text{(II.3)}$$

Box 2. $p(x;\theta)g(\theta) = \mathcal{N}(x; 0, s^2 I) c s^{\nu-\beta}$.

makes (I.1) continuous with a continuous derivative. For $0 < p < 2$, (I.1) and (I.2) result from (5) with $\theta = s^2$ and

$$g(s^2) \triangleq \big(2\pi(s^2 + r^2)\big)^{\nu/2} \cdot \exp\!\left(-\frac{\beta(2-p)}{2}\big(\beta p(s^2+r^2)\big)^{\frac{p}{2-p}}\right), \quad (17)$$

where ν is the dimension of x.

For $0 < p < 2$, versions of (I.1) and (I.2) have been known for some time, cf. [5, 13]. But (I.1) and (I.2) work also for $p > 2$, which we have not seen in the prior literature. However, for $p > 2$, the derivation must be modified: the maximization in (5) must be replaced by minimization, and the friendly cooperation of minimization and maximization relies on the minimax theorem.

Box 2 is about $p(x;\theta)g(\theta)$ with $p(x;\theta) = \mathcal{N}(x; 0, s^2 I)$ and $g(\theta)$ of the form $cs^{\nu-\beta}$, where ν is the dimension of x. Using (5), the prior (II.1) is obtained with the update rule (II.2). With this same $p(x;\theta)g(\theta)$, the EM update rule (13) becomes (II.3).

For $\beta > 0$, both (II.2) and (II.3) are strongly sparsifying.

constraint	update rules: $\overrightarrow{\sigma}_X^2$	update rules: $\overrightarrow{m}_X$
$X \ge a$	$\dfrac{\lvert\hat{x} - a\rvert}{\beta}$	$\hat{x}$, if $\hat{x} \ge a$; $2a - \hat{x}$, if $\hat{x} < a$
$X \le b$	$\dfrac{\lvert\hat{x} - b\rvert}{\beta}$	$\hat{x}$, if $\hat{x} \le b$; $2b - \hat{x}$, if $\hat{x} > b$
$X \in \{a, b\}$	$\dfrac{1}{w_a + w_b}$	$\dfrac{w_a a + w_b b}{w_a + w_b}$
	with $w_a = \beta/\big(\mathrm{Var}[X] + (\hat{x} - a)^2\big)$ and $w_b = \beta/\big(\mathrm{Var}[X] + (\hat{x} - b)^2\big)$	

Table 1. NUP representations of constraints on $X \in \mathbb{R}$. The constraints can be enforced with sufficiently large β.

Generally speaking, (II.3) converges much slower, but is less likely to get trapped in a "bad" local maximum.

For $\beta > 0$, the idea of Box 2 is ancient, cf. [14]. Choosing $\beta < 0$ may seem nonsensical, but it is possible and will actually be used in Section 4; in this case, s^2 will be negative (and should be denoted by a different symbol), and the "normal prior" $p(x;\theta)$ with fixed $\theta = s^2$ is simply the function $\exp\big(\|x\|^2/(-2s^2)\big)$.

Table 1 summarizes recent NUP representations of constraints from [6, 7]. For fixed θ, $p(x;\theta) = \mathcal{N}(x; \overrightarrow{m}_X, \overrightarrow{\sigma}_X^2)$ is scalar Gaussian, with update rules for the mean $\overrightarrow{m}_X$ and the variance $\overrightarrow{\sigma}_X^2$ as in the table.

The derivation of the inequality constraints [7] begins with the prior

$$\rho(x) = \exp(-\beta|x-a|)\exp(-\beta|x-b|), \quad (18)$$

which is flat for $a \le x \le b$; for $a < b$ and sufficiently large β, this prior effectively restricts X to $[a, b]$. Using (I.3) for each of the two factors in (18) yields a NUP realization of this interval constraint. Finally, we form the limits $b \to \infty$ and $a \to -\infty$.

The binarizing constraint is derived analogously to the interval constraint, but with (II.3) instead of (I.3). For its further discussion, we refer to [6].

4. NUP MESSAGES IN A HIERARCHICAL MODEL

4.1. Motivation

Consider, for example, a model for piecewise constant data with level jumps at arbitrary unknown times, as illustrated in Fig. 2. We can write this as

$$X_k = X_{k-1} + U_k \quad (19)$$

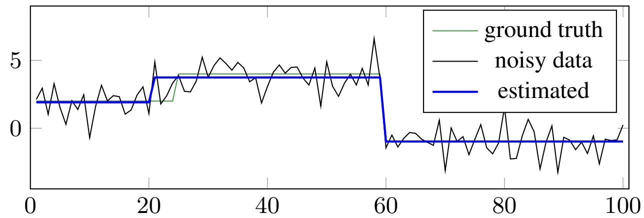

Fig. 2. Fitting a piecewise constant model to noisy data.

with scalar level X_k and sparse input U_k, $k = 1, \ldots, N$. We observe the noisy data $\breve{Y}_k = X_k + R_k$ with i.i.d. zero-mean Gaussian noise R_k, and we wish to recover X_k. Using the approach of Sections 2 and 3, we model U_k with a sparsifying prior from Box 1 or Box 2. In consequence, we can estimate X_k and U_k by iterations of least squares or Gaussian message passing as described in Section 2, with good results as illustrated in Fig. 2, cf. [12].

In a next step, we wish to use such a model as a second-layer model that controls the variance of "noise" in a first-layer model. For example, the first-layer model could also have the form (19), but with inputs U_k that are independent zero-mean Gaussians with (unknown) piecewise constant variance. For another example, the first-layer model is still as in (19), but the variance of the observation noise R_k is (unknown and) piecewise constant. Similar situations have, of course, been studied in the literature [15–20], without, however, fully achieving our present goal: we wish to treat the second-layer model conceptually and algorithmically like the first-layer model (i.e., with least squares or Gaussian message passing), and we wish to connect the two layers without approximations.

4.2. Problem Statement

For the sake of exposition, we now specialize to the setting of Fig. 3: a first model (Model 1) produces an output $Y_1, \ldots, Y_N \in \mathbb{R}$, of which we observe the noisy version

$$\breve{Y}_k = Y_k + R_k, \tag{20}$$

$k = 1, \ldots, N$. The observation noise variables $R_k \in \mathbb{R}$ are independent zero-mean Gaussians with unknown variances S_k^2. A second model (Model 2) produces $Z_1, \ldots, Z_N \in \mathbb{R}$, which determine $S_1^2, \ldots, S_N^2$ by

$$S_k^2 = g(Z_k) \tag{21}$$

for some given function g. Natural choices for g include $g(z_k) = z_k^\nu$ for $\nu \in \{1, 2, -1, -2\}$.

For given $\breve{Y}_k = \breve{y}_k$, $k = 1, \ldots, N$, we wish to estimate the other variables. We assume that both Model 1 and Model 2 are tractable by Gaussian message passing (perhaps with NUP priors as in Sections 2 and 3). Finally, we aim for an alternating maximization approach, i.e., we repeat the following two steps until convergence:

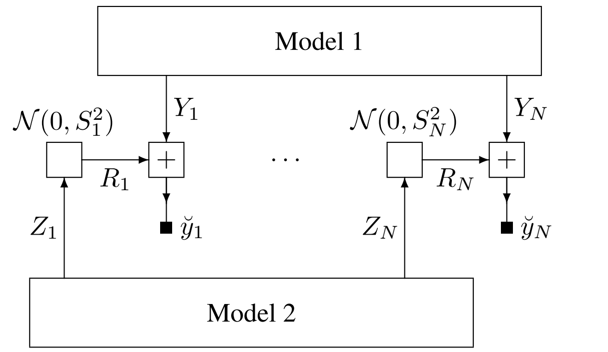

Fig. 3. Setting of Section 4.2: Model 2 is a joint prior on the observation noise variances $S_1^2, \ldots, S_N^2$ of Model 1.

Step 1: Estimate $Y_1, \ldots, Y_N$ for fixed $Z_1, \ldots, Z_N$ by Gaussian message passing in Model 1.

Step 2: Estimate $Z_1, \ldots, Z_N$ for fixed $Y_1, \ldots, Y_N$ by Gaussian message passing in Model 2.

Step 1 is conceptually trivial: in this step, $R_1, \ldots, R_N$ are independent zero-mean Gaussians with known variances $g(Z_k)$.

However, there is a problem with Step 2: even for fixed Y_k, the backward sum-product messages $\overleftarrow{\mu}_{Z_k}$ (cf. [11]) are not Gaussian (for any meaningful choice of g). Previous proposals to solve this problem resorted to approximations, cf. [18–20]. We are now going to propose an exact NUP representation of $\overleftarrow{\mu}_{Z_k}$ for

$$S_k = Z_k^{-1}. \tag{22}$$

4.3. Proposed Solution

For ease of notation, we here simplify R_k and S_k in Fig. 3 to R and S, respectively. We begin with the factor graph in Fig. 4 (left), which represents

$$\mathcal{N}(r; 0, s^2) = \int_{-\infty}^{\infty} \mathcal{N}(u; 0, 1)\delta(r - us)\, du, \tag{23}$$

where δ denotes the Dirac delta. With $s = 1/z$ and

$$\delta(r - us) = |z|\delta(u - zr), \tag{24}$$

(23) becomes

$$\mathcal{N}(r; 0, z^{-2}) = \gamma|z| \int_{-\infty}^{\infty} \mathcal{N}(u; 0, 1)\delta(u - zr)\, du \tag{25}$$

(with some immaterial constant γ) as in Fig. 4 (right).

But the function $|z|$ has a NUV representation as in Box 2 with $\beta = -1$ (which actually works, as mentioned in Section 3). In this way, for fixed $R = \hat{r}$, we obtain a Gaussian message $\overleftarrow{\mu}_Z$. However, this message has mean zero, which is unhelpful for the problem stated in Section 4.2.

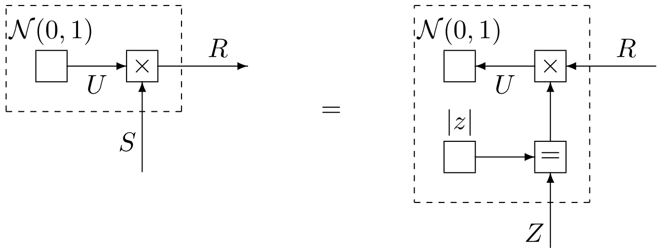

Fig. 4. Reversing the multiplier (with $Z = 1/S$).

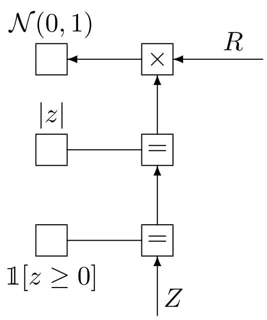

Fig. 5. Proposed solution: this factor graph with NUP representations of $|z|$ and of the constraint $Z \geq 0$.

Therefore, in a next step, we add the constraint $Z \geq 0$ (as shown in Fig. 5), a NUP representation of which was given in Table 1.

From Fig. 5 with fixed $R = \hat{r}$, and with the mentioned NUP representations, we easily obtain the Gaussian message $\overleftarrow{\mu}_Z$ with variance

$$\overleftarrow{\sigma}_Z^2 = \left(\hat{r}^2 - \hat{z}^{-2} + \beta|\hat{z}|^{-1}\right)^{-1} \tag{26}$$

and mean

$$\overleftarrow{m}_Z = \beta \overleftarrow{\sigma}_Z^2, \tag{27}$$

where $\hat{z}$ is the value of Z from which the NUP parameters were last updated. The parameter β stems from Table 1 and must be chosen large enough so that (26) is positive; otherwise, the choice of β is immaterial for correctness, but it may affect the speed of convergence.

4.4. Modifications for EM Setting

A natural modification of the problem statement of Section 4.2 is to estimate $Z_1, \ldots, Z_N$ by expectation maximization with hidden variables $Y_1, \ldots, Y_N$. There is no space here to develop this approach, but in result, the only change is that $\hat{r}^2$ in (26) is replaced by $\mathrm{E}[R^2]$, where the expectation is computed with $Z_1, \ldots, Z_N$ fixed to their present estimates $\hat{z}_1, \ldots, \hat{z}_N$.

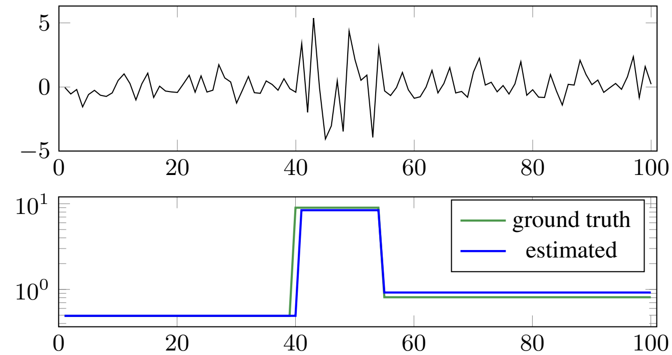

Fig. 6. Top: zero-mean Gaussian noise with piecewise constant variance. Bottom: actual and estimated variance S_k^2.

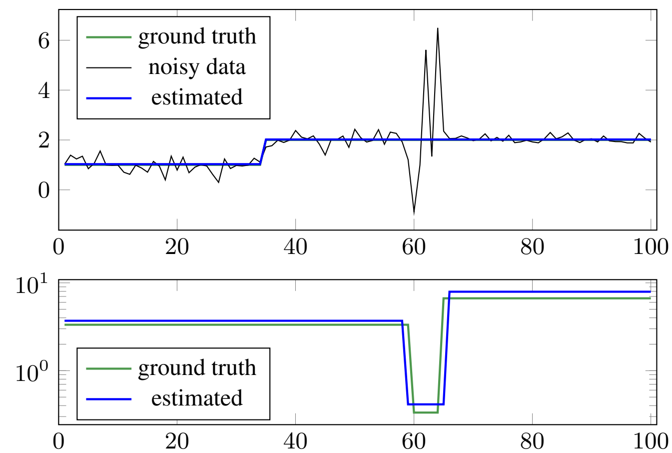

Fig. 7. Piecewise constant model with piecewise constant observation noise variance. Top: given noisy data, ground truth (covered by estimate), and estimate. Bottom: actual and estimated value of $Z_k = 1/\sqrt{S_k^2}$.

4.5. Examples

Two numerical examples with models as in Fig. 3 are shown in Figs. 6 and 7. In both examples, Model 2 is piecewise constant as in (19), with a sparsifying NUV prior on U_k of the form (II.1) (with $\beta > 0$) that is handled with (II.2). In Fig. 6, Model 1 is simply $Y_k = 0$ for all k, i.e., $\breve{Y}_1, \ldots, \breve{Y}_N$ is a sequence of zero-mean Gaussian random variables with piecewise constant variance.

In Fig. 7, Model 1 is itself piecewise constant as in (19). Note that the high-noise burst beginning at $k = 60$ does not derail the estimation of the actual data according to Model 1.

5. CONCLUSION

NUP represenions allow to convert non-Gaussian estimation problems into iteratively reweighted least squares or iterated linear-Gaussian estimation problems. We briefly reviewed this approach, including recently proposed NUP representations of binarizing priors and of inequality constraints. We then proposed a (exact) new NUP representation of non-Gaussian messages that occur when the variances of priors in a first model are controlled by a second model; in consequence, estimation in such models is reduced to iterated linear-Gaussian estimation, which we illustrated with a model of piecewise constant noise variance.

In conclusion, we venture to suggest that NUP representations of priors and messages open a new perspective of scalable multi-layer (or multi-component) models that are computationally tractable.

6. REFERENCES

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

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

[3] J. Palmer, D. P. Wipf, K. Kreutz-Delgado, and B. D. Rao, ”Variational EM algorithms for non-Gaussian latent variable models,” *Adv. Neural Inf. Proc. Systems (NIPS),* 2006.

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

[5] 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.

[6] R. Keusch and H.-A. Loeliger, “A binarizing NUV prior and its use for M-level control and digital-to-analog conversion,” arXiv:2105.02599, May 2021.

[7] R. Keusch and H.-A. Loeliger, “Half-space and box constraints as NUV priors: first results,” arXiv:2109.00036, Aug. 2021.

[8] D. Wipf and S. Nagarajan, “Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions,” *IEEE J. Select. Topics in Signal Proc.,* vol. 4, no. 2, April 2010, pp. 317–329.

[9] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, “Iteratively reweighted least squares minimization for sparse recovery,” *Comm. Pure & Appl. Math.,* Vol. 63, No. 1, pp. 1–38, 2010.

[10] R. Keusch and H.-A. Loeliger, “Model-predictive control with new NUV priors,” arXiv:2303.15806, Mar. 2023.

[11] 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.

[12] H.-A. Loeliger, L. Bruderer, H. Malmberg, F. Wadehn, and N. Zalmai “On sparsity by NUV-EM, Gaussian message passing, and Kalman smoothing,” *2016 Information Theory & Applications Workshop (ITA),* San Diego, CA, Feb. 2016.

[13] 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,” *Int. Symposium on Turbo Codes & Iterative Information Processing (ISTC) 2018*, Hongkong, China, Dec. 3–7, 2018.

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

[15] C.-J. Kim, C. R. Nelson et al., *State-space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications.* Cambridge, MA, USA: MIT Press, 1999.

[16] S. Särkkä and A. Nummenmaa, “Recursive noise adaptive Kalman filtering by variational Bayesian approximations,” *IEEE Trans. Automatic Control,* vol. 54, no. 3, March 2009.

[17] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods.* 2nd ed. Oxford, UK: Oxford University Press, 2012.

[18] 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, Sept. 2–6, 2019.

[19] A. Podusenko, W. M. Kouw, and B. de Vries, “Message passing-based inference for time-varying autoregressive models,” Entropy 2021, 23:683.

[20] A. Podusenko, B. van Erp, D. Bagaev, İ. Şenöz, and B. de Vries, “Message passing-based inference in switching autoregressive models, *30th European Signal Processing Conference (EUSIPCO),* Belgrade, Serbia, Aug. 29 – Sept. 2, 2022.