ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

On Feature Learning with State Space Models and Pulse Domain Signal Analysis

Hans-Andrea Loeliger

based on joint work with

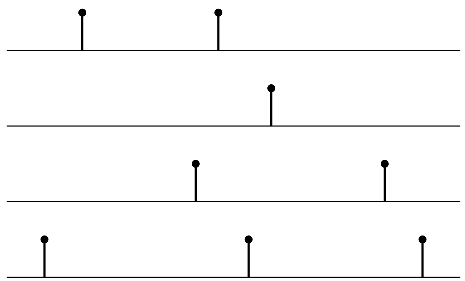
Sarah Neff and Nour Zalmai

ITA 2018

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

On Feature Learning with State Space Models and Pulse Domain Signal Analysis

Two vaguely related pieces of work and a view.



ITA 2018

Piece 1 (the solid piece):

A Single Trick and Algorithm for Many Problems in Signal Analysis

Combining

- linear state space models,
- normal priors with unknown variances (NUV) for sparsity,
- and expectation maximization (EM) for learning all parameters

can be used for sparse estimation, dictionary learning, unsupervised signal labeling, blind signal separation, and more,

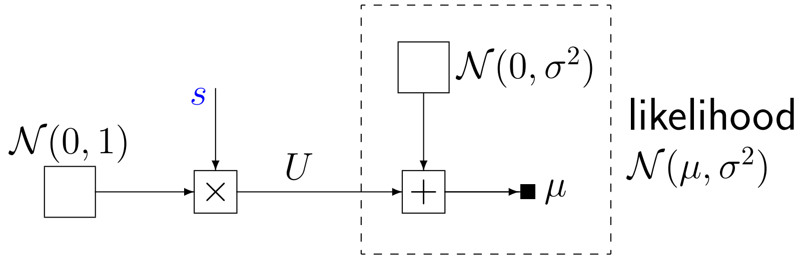
by variations of a single algorithm essentially consisting of repeated multivariate-Gaussian forward-backward message passing (i.e., recursions as in Kalman smoothing).

[ITA 2016], [EUSIPCO 2017], [PhD thesis Zalmai 2017]

Sparsity by NUV Priors (Normal with Unknown Variance)

- Originating from Bayesian inference [MacKay 1992, Neal 1996, ...]
- Basis of "automatic relevance determination" and sparse Bayesian learning [Neal, Tipping 2001, Wipf et al., ...]

Example: real $U \sim \mathcal{N}(0, s^2)$ with unknown variance s^2,
single observation $Y = U + Z = \mu \in \mathbb{R}$ with noise $Z \sim \mathcal{N}(0, \sigma^2)$:

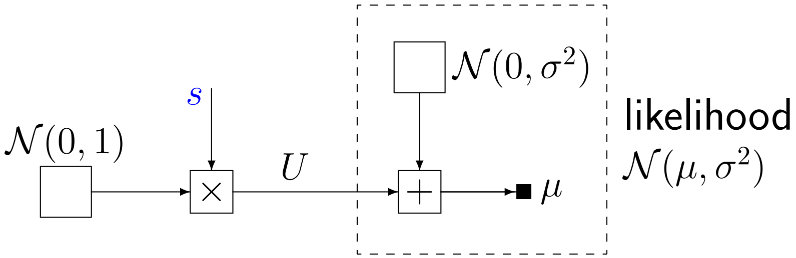


Maximum-likelihood estimate $\hat{s}^2_{\text{ML}} = \max\{0, \mu^2 - \sigma^2\}$

Sparsity by NUV Priors (Normal with Unknown Variance)

- Originating from Bayesian inference [MacKay 1992, Neal 1996, ...]
- Basis of "automatic relevance determination" and sparse Bayesian learning [Neal, Tipping 2001, Wipf et al., ...]

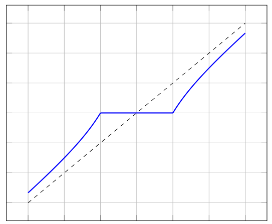
Example: real $U \sim \mathcal{N}(0, s^2)$ with unknown variance s^2,
single observation $Y = U + Z = \mu \in \mathbb{R}$ with noise $Z \sim \mathcal{N}(0, \sigma^2)$:



Maximum-likelihood estimate $\hat{s}^2_{\text{ML}} = \max\{0, \mu^2 - \sigma^2\}$

For fixed $s^2 = \hat{s}^2_{\text{ML}}$, U is Gaussian with posterior mean (MAP/MMSE/LMMSE estimate)

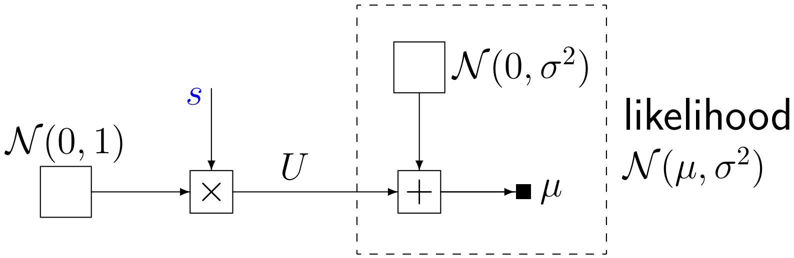
$$\hat{u} = \begin{cases} \mu \cdot \frac{\mu^2 - \sigma^2}{\mu^2} & \text{if } \mu^2 > \sigma^2 \\ 0, & \text{otherwise.} \end{cases}$$



Sparsity by NUV Priors (Normal with Unknown Variance)

- Originating from Bayesian inference [MacKay 1992, Neal 1996, ...]
- Basis of "automatic relevance determination" and sparse Bayesian learning [Neal, Tipping 2001, Wipf et al., ...]

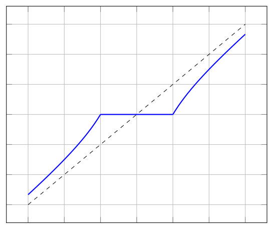
Example: real $U \sim \mathcal{N}(0, s^2)$ with unknown variance s^2,
single observation $Y = U + Z = \mu \in \mathbb{R}$ with noise $Z \sim \mathcal{N}(0, \sigma^2)$:



Maximum-likelihood estimate $\hat{s}^2_{\text{ML}} = \max\{0, \mu^2 - \sigma^2\}$

For fixed $s^2 = \hat{s}^2_{\text{ML}}$, U is Gaussian with posterior mean (MAP/MMSE/LMMSE estimate)

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



Still holds for $Y \in \mathbb{R}^N$ with likelihood $p(y|u) \propto e^{-(u-\mu(y))^2/2\sigma^2}$.

Sparsity by NUV Priors cont'd

General method:

- Model variables (or parameters) $U_1, \ldots, U_K$ of interest as independent zero-mean Gaussians, each with its own individual unknown variance $\sigma_1^2, \ldots, \sigma_K^2$.
- Determine $\sigma_1^2, \ldots, \sigma_K^2$ by ML (or some approximation thereof); e.g., by expectation maximization (EM).
  A *local* maximum of the likelihood suffices for sparsity.

  Specifically (for linear Gaussian models):

  1. Begin with an initial guess $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2$.
  2. Compute* the means m_{U_k} and the variances $\sigma_{U_k}^2$ of the (Gaussian) posterior distributions $p(u_k|y, \sigma_1^2, \ldots, \sigma_K^2)$ for $k = 1, \ldots, K$ with $\sigma_1^2, \ldots, \sigma_K^2$ fixed.
  3. Standard EM: update $\sigma_k^2 \leftarrow m_{U_k}^2 + \sigma_{U_k}^2$ for all k.
  4. Repeat 2 and 3 until convergence.

---

*by Gaussian message passing in the appropriate factor graph

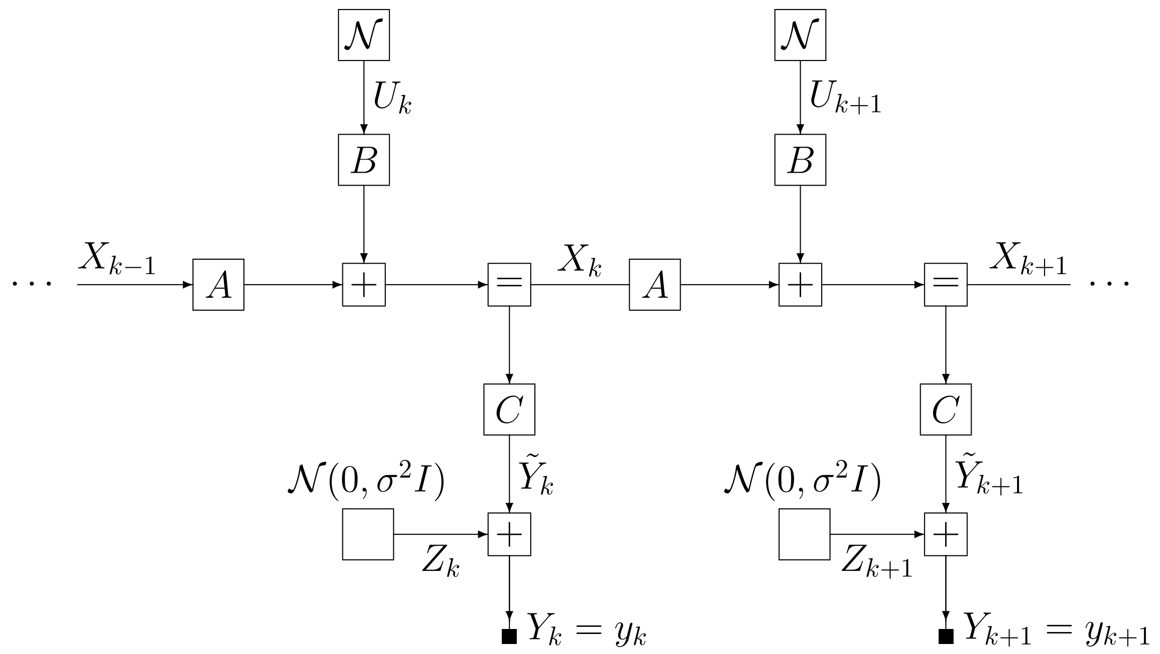
Linear State Space Models

State $X_k \in \mathbb{R}^n$ and observation $Y_k \in \mathbb{R}^L$ evolving according to

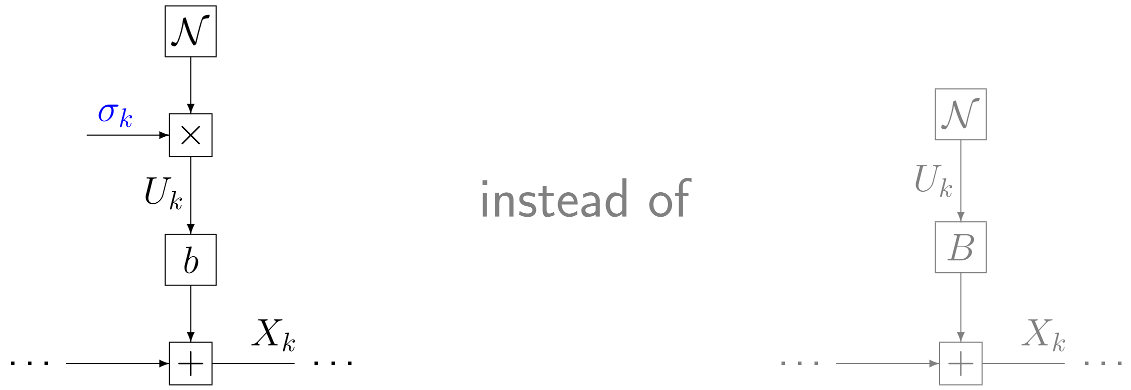
$$
\begin{aligned}
X_k &= AX_{k-1} + BU_k \\
Y_k &= CX_k + Z_k
\end{aligned}
$$

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.

Factor graph:

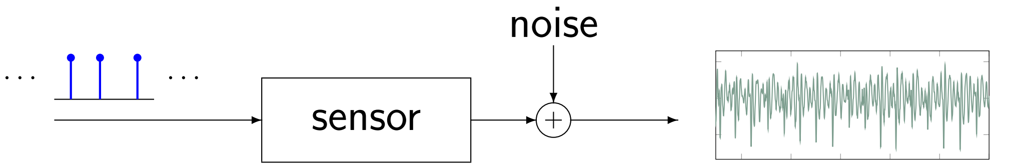


Sparse Scalar Input

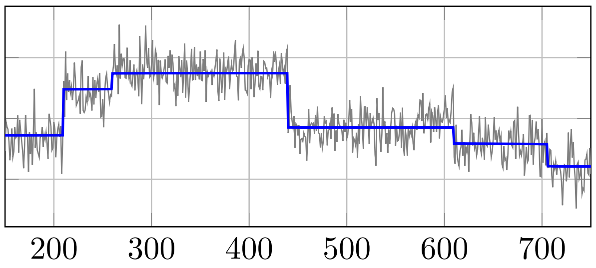


E.g.:

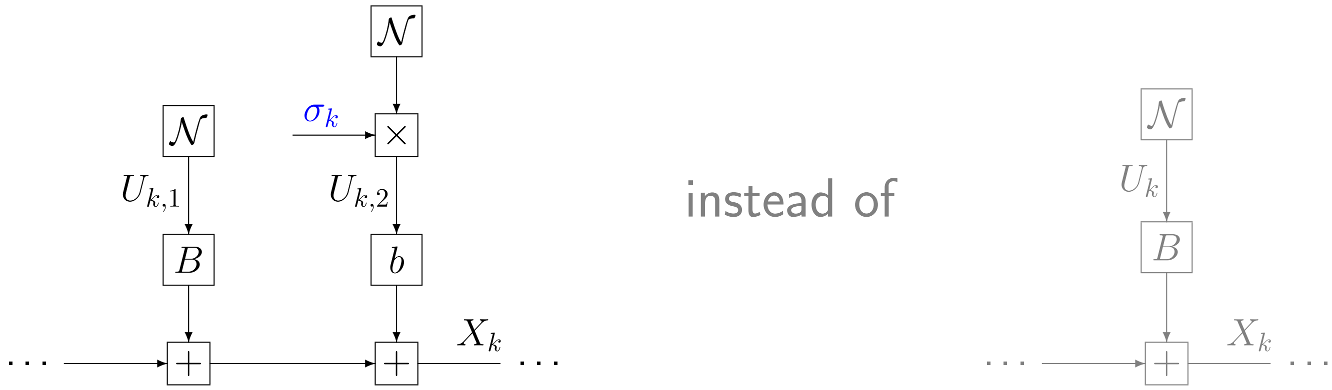
- Sparse input-signal estimation (e.g., heart beat [ISIT 2015]):



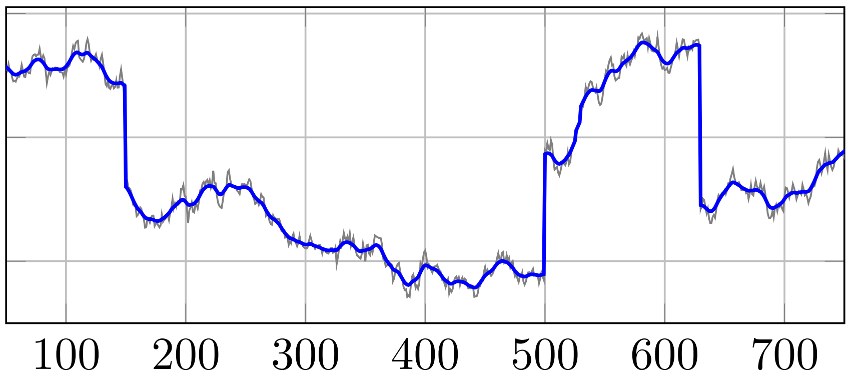
- Piecewise constant least-squares fit:



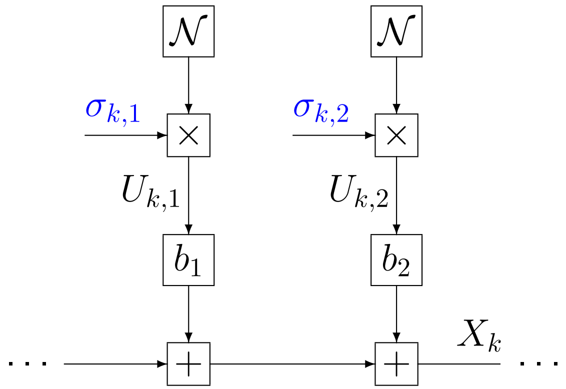
White Noise Input + Sparse Scalar Input



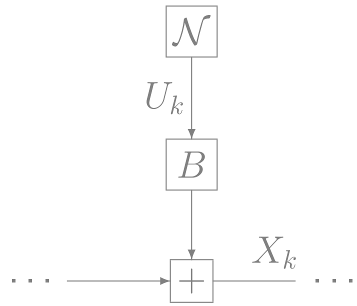
E.g., random walk with occasional jumps:



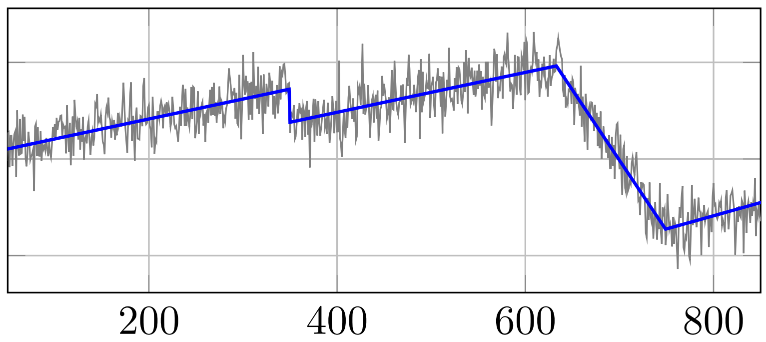
Multiple Sparse Scalar Inputs



instead of



E.g., least-squares fitting of straight-line segments:

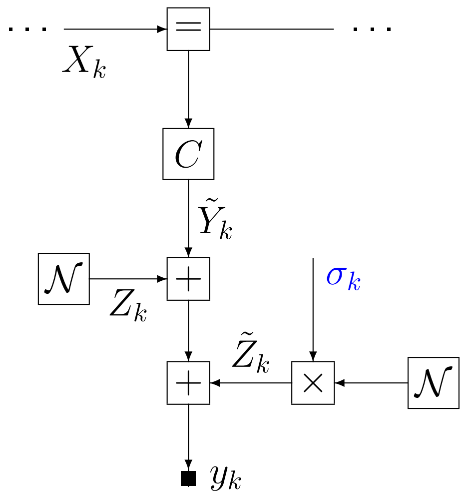


Obvious generalizations:

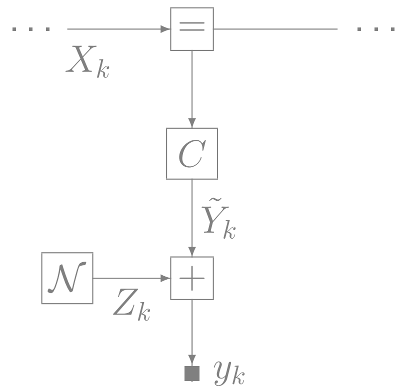
- polynomial segments
- enforcing continuity, or continuity of derivative(s)

Dealing with Outliers

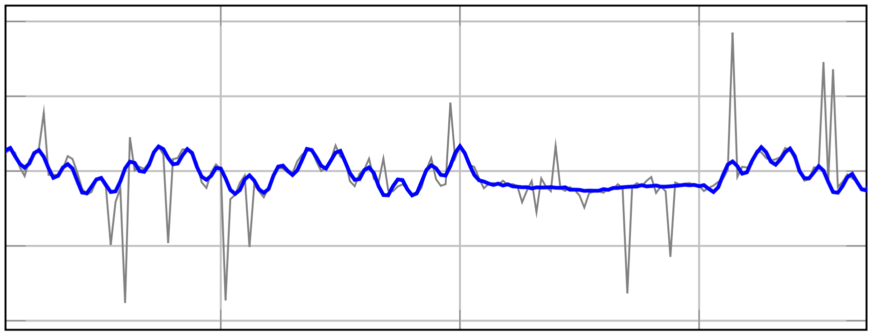
Simply replace $Y = CX_k + Z_k$
by $Y = CX_k + Z_k + \tilde{Z}_k$ with sparse $\tilde{Z}_k$, i.e.,



instead of

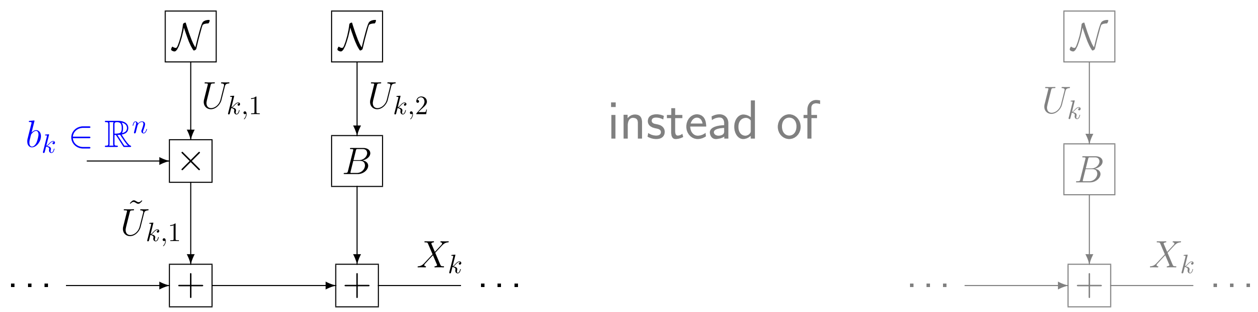


Example:



Linear state space models with sparse input: [ICASSP 2016]

Sparse Input Pulses with Individual Direction

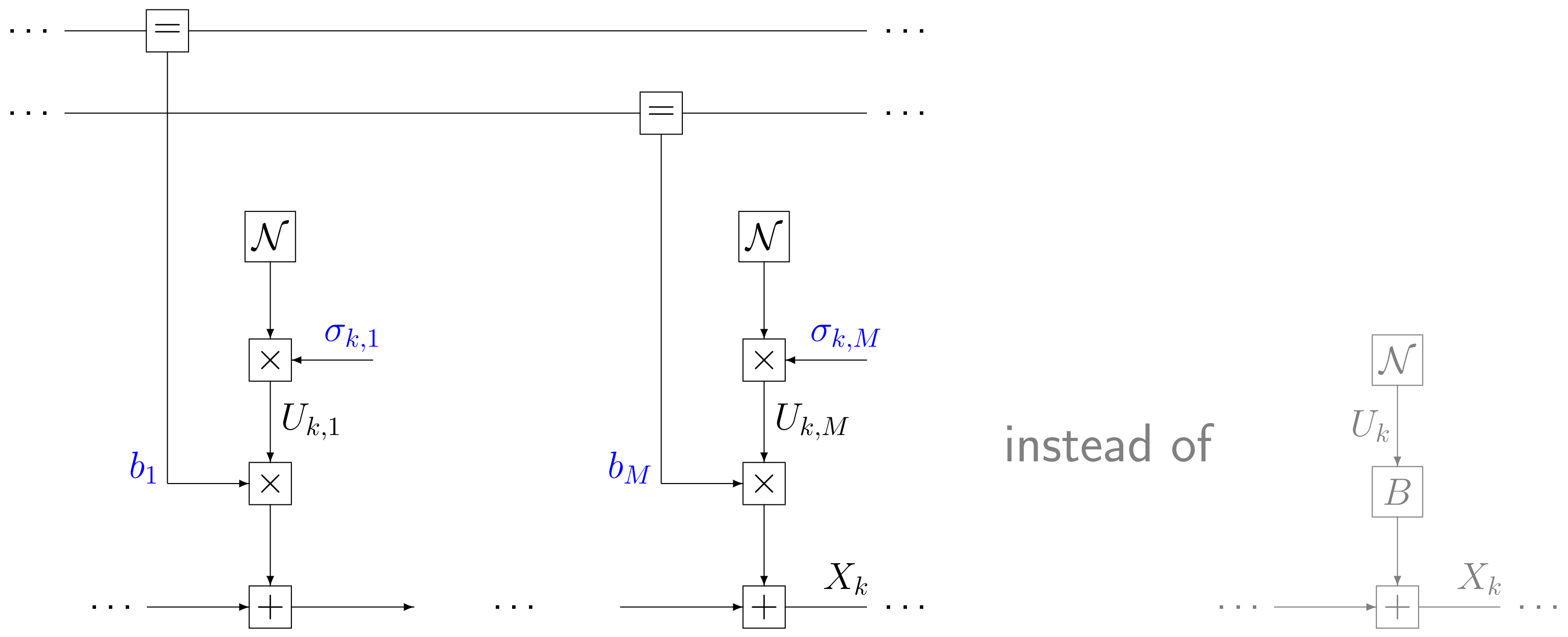


- Unknown scalar σ_k replaced by unknown vector $b_k \in \mathbb{R}^n$
- Still sparsifying, still learnable (e.g.) by EM

Applications:

- Occasional arbitrary jumps in the state space
- System identification from multiple unknown excitations
- ...

Recurring Unknown Sparse Input Pulses

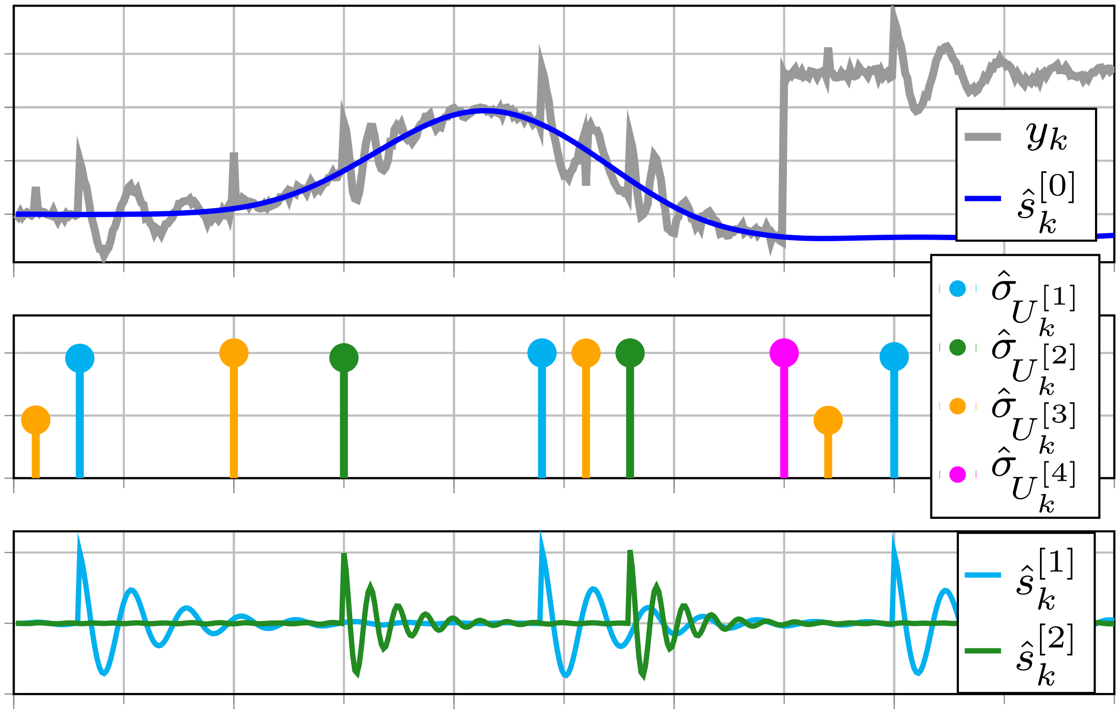


- Unknown input vectors $b_1, \ldots, b_M \in \mathbb{R}^n$, each with independent sparse input.
- Still learnable by EM. The state transition matrix A can also be learned.

Applications: unsupervised signal labeling, dictionary learning, blind signal separation, . . .

Unsupervised Feature Extraction, Signal Labeling, and Blind Signal Separation

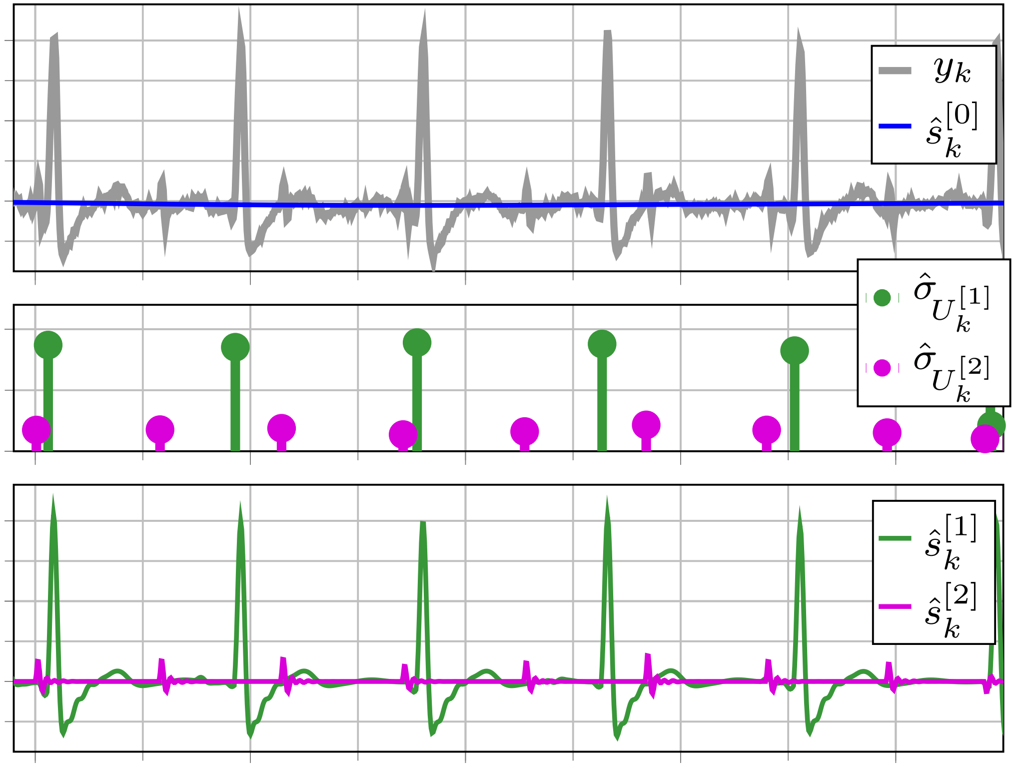
Artificial example: irregular occurances of localized signal shapes on top of a wandering baseline with jumps.



Everything (matrices A, B, input signals) is learned, unsupervised.

Unsupervised Feature Extraction, Signal Labeling, and Blind Signal Separation

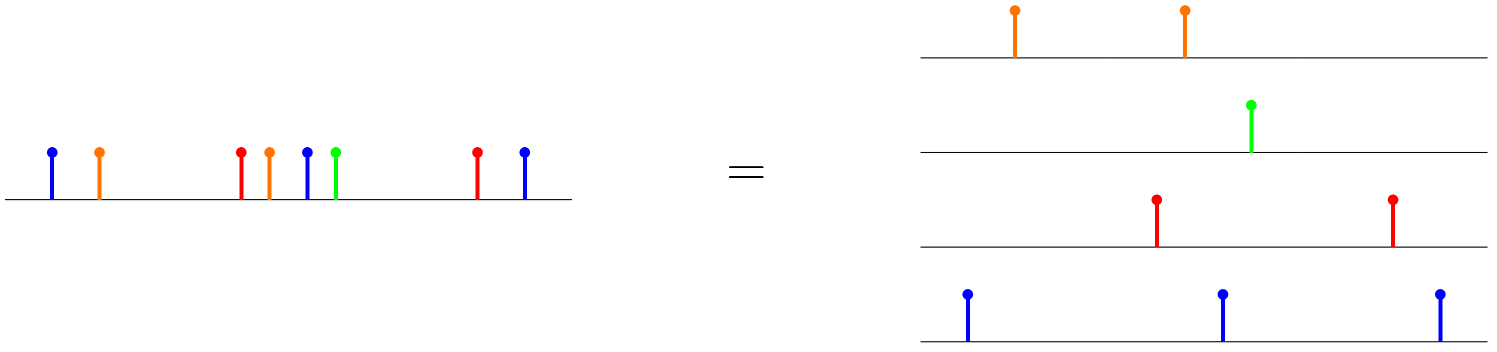
ECG recording of a pregnant woman:
decomposition into maternal and fetal heart beats.



Totel model order 24: 8 and 3 damped sinusoids, respectively, for the heart beats; local line model ($\approx$ cubic spline) for the baseline.

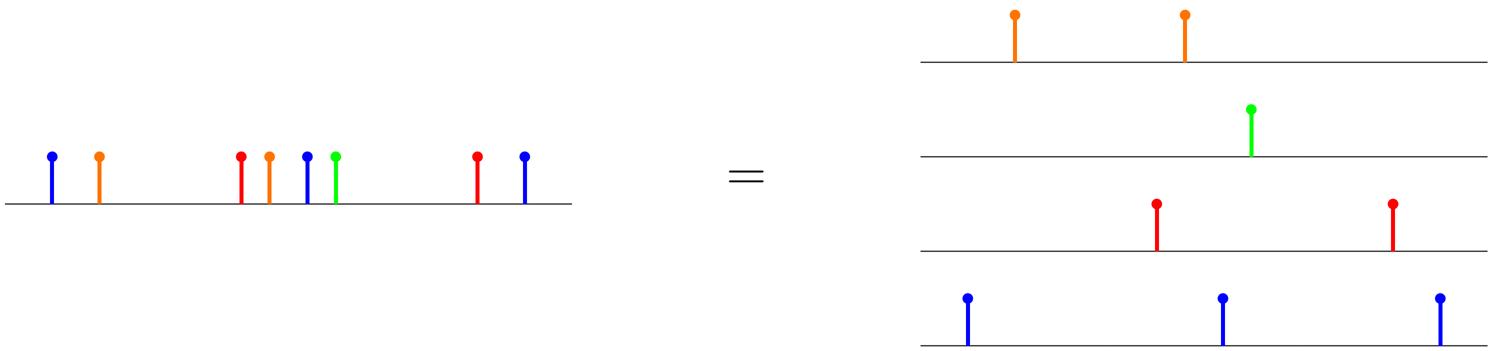
Multichannel Sparse Impulsive Signals as Data Type for Signal Analysis

The method just described yields sparse multichannel feature “signals”:



Multichannel Sparse Impulsive Signals as Data Type for Signal Analysis

The method just described yields sparse multichannel feature "signals":

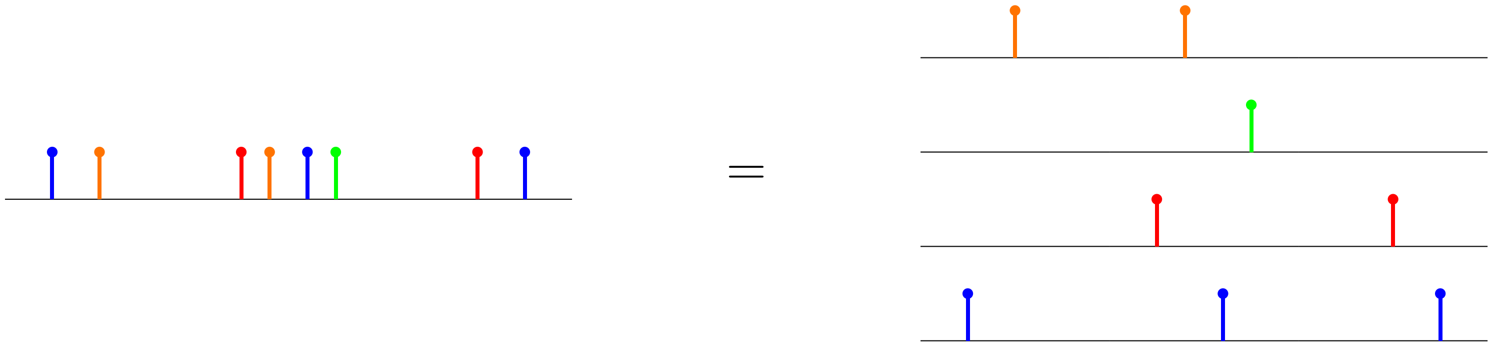


The same method can be applied again to such signals!
(Gaussian estimation $\approx$ least squares $\approx$ orthogonal projection.)

(Mentioned in [Zalmai thesis 2017], but no experience as yet.)

Multichannel Sparse Impulsive Signals as Data Type for Signal Analysis

The method just described yields sparse multichannel feature “signals”:



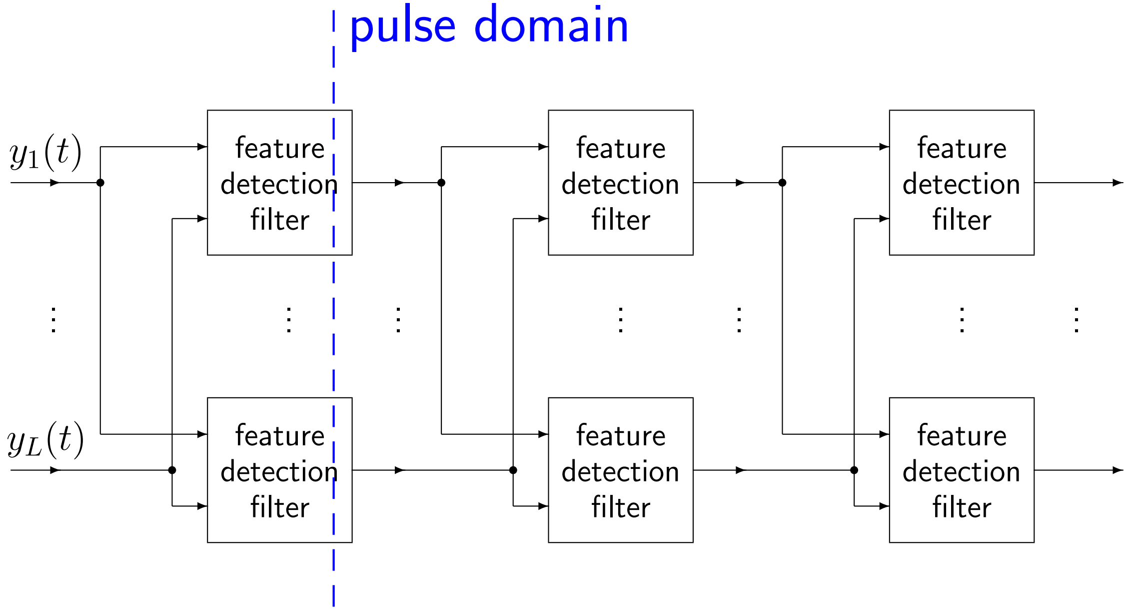
The same method can be applied again to such signals!
(Gaussian estimation $\approx$ least squares $\approx$ orthogonal projection.)
And again, and again . . . , to any depth (all unsupervised).

(Mentioned in [Zalmai thesis 2017], but no experience as yet.)

Piece 2 (more speculative): [ISIT 2015]

Layered Networks of Feature Detection Filters

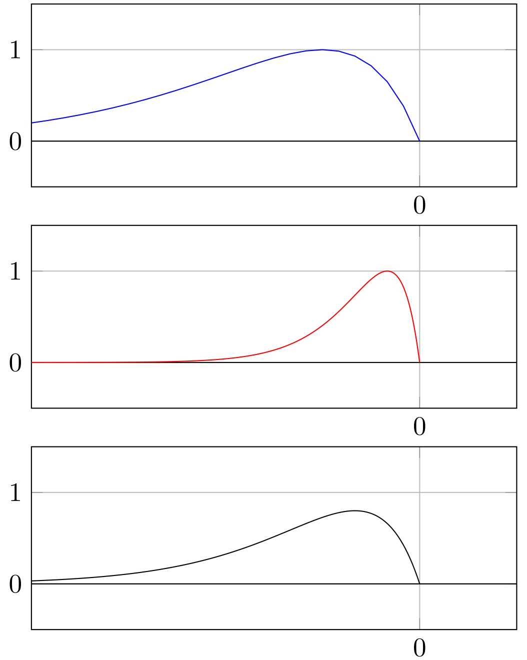
Such multichannel sparse feature signals have already been used in parallel work:



Feature detection filters ("neurons") work here as follows:

- A multi-input, single-output linear time-invariant filter (IIR) produces a score signal (= correlation with a smooth template).
- An isolated unit pulse is generated if the score signal exceeds some threshold. (Sparsity is essential: thresholding does not work.)

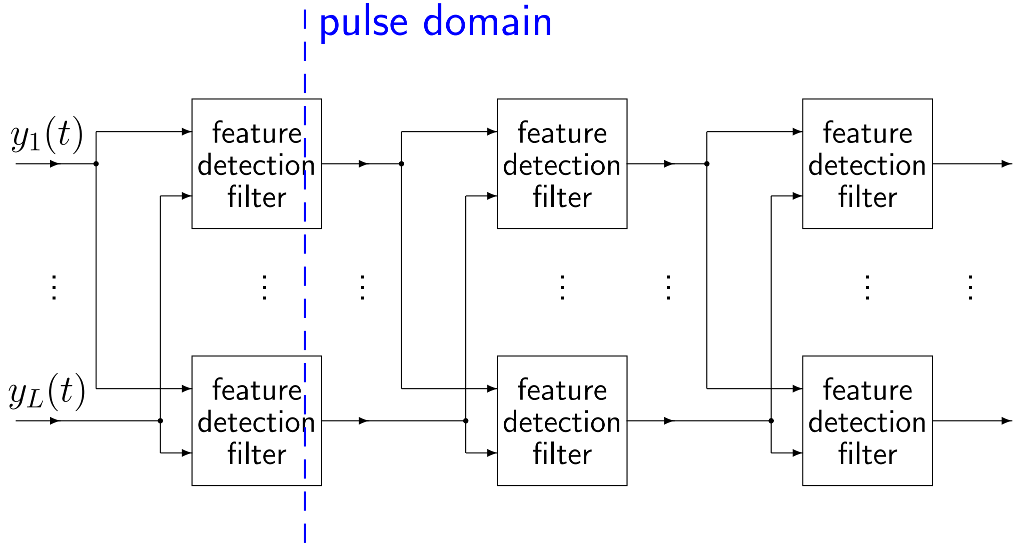
Toy Example of Three-Channel Template



- Time scale: at most one pulse in window
- Realizable with biological plausible neurons
- Realizable with simple analog circuits

Piece 2:

Layered Networks of Feature Detection Filters



Feature detection filters ("neurons"):

- Score signal (= correlation with smooth template) is computed by IIR filter.
- An isolated unit pulse is generated if the score signal exceeds some threshold.
- Allows biologically plausible neuron models.
- Supervised learning of deep network based on gradient back-propagation demonstrated (for toy example), apparently avoiding gradient degeneration [Neff thesis 2016].
- Promising for (non-digital) neuromorphic computation.

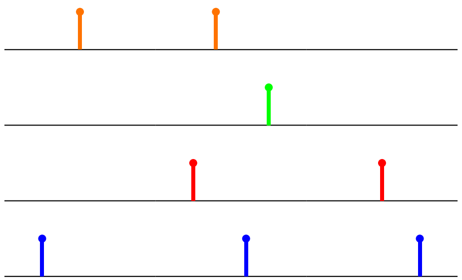
Conclusion

The solid piece:

- Linear state space models with NUV priors can be used for sparse estimation, dictionary learning, unsupervised signal labeling, blind signal separation, ...
- ... by variations of a single algorithm consisting essentially of repeated multivariate-Gaussian forward-backward message passing (i.e., recursions as in Kalman smoothing).

The view:
Sparse multichannel feature signals are an interesting data type for signal analysis. Features-of-features networks with such signals can be built as in Piece 1 or as in Piece 2.

(Did not discuss relations to convolutional neural networks, wavelets, ...)



Main References

- Loeliger, Bruderer, Malmberg, Wadehn, Zalmai, "On sparsity by NUV-EM, Gaussian message passing, and Kalman smoothing," ITA 2016
- Zalmai, Keusch, Malmberg, Loeliger, "Unsupervised feature extraction, signal labeling, and blind signal separation in a state space world," EUSIPCO 2017
- Zalmai, *A State Space World for Detecting and Estimating Events and Learning Sparse Signal Decompositions.* PhD thesis (24360), ETH, 2017

- Loeliger and Neff, "Pulse-domain signal parsing and neural computation," ISIT 2015
- Neff, *A New Approach to Information Processing with Filters and Pulses.* PhD thesis (23672), ETH, 2016