Sparse Parameter Estimation for Linear Dynamical Systems

Behrad Soleimani
Department of Electrical and Computer Engineering, University of Maryland, College Park
Behrad@umd.edu

I. Introduction

Parameter estimation of linear systems is a well-known problem in the literature. The goal in such problems is that to find the underlying source dynamics given a set of indirect linear observations of a group of sources and assuming that the underlying dynamic model is known. However, in case that the observation vectors are low-dimensional in comparison with the source space dimension, the parameter estimation has been categorized as an ill-posed problem and is not straight forward to solve. On the other hand, assuming sparsity in the source space, meaning that the most of sources are inactive, enables to fully recover the underlying source dynamic. In this report, we utilize an instance of Expectation Maximization (EM) algorithm to dissolve the addressed problem. A MATLAB implementation of the proposed method is also accessible on Github [1].

II. Preliminaries and System Model

We consider a linear system whose measurement produces N_y dimensional observation vector $\mathbf{y}_t \in \mathbb{R}^{N_y}$ for time-points $t = 1, \cdots, T$. The measurement procedure that maps the source activities to the observations is considered as follows:

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{n}_t, \quad t = 1, \cdots, T, \tag{1}$$

where $\mathbf{C} \in \mathbb{R}^{N_y \times N_x}$ is the linear mapping (aka the sensing matrix), $\mathbf{x}_t \in \mathbb{R}^{N_x}$ is the source vector at time t, and $\mathbf{n}_t \in \mathbb{R}^{N_x}$ represents the observation noise at time t. The measurement noise is considered as a zero-mean Gaussian vector with covariance matrix $\mathbf{R}$, i.e., $\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ and i.i.d. across the time.
The underlying source dynamics is considered as a vector auto-regressive process with p lags, VAR(p), as follows

$$\mathbf{x}_t = \sum_{k=1}^{p} \mathbf{A}_k \mathbf{x}_{t-k} + \mathbf{B}\mathbf{e}_t + \mathbf{w}_t, \quad t = 1, 2, \cdots, T, \tag{2}$$

where $\{\mathbf{A}_k\}_{k=1}^{p} \in \mathbb{R}^{N_x \times N_x}$ represents the VAR coefficients, $\mathbf{B} \in \mathbb{R}^{N_x \times N_e}$ captures the contribution from external factors, $\mathbf{e}_t \in \mathbb{R}^{N_e}$ is the process corresponding to the external factor, and $\mathbf{w}_t \in \mathbb{R}^{N_x}$ is the source (state) noise vector which is also assumed to be a zero-mean Gaussian vector with covariance matrix $\mathbf{Q}$, i.e., $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ and i.i.d. across the time. Without loss of generality, we assume that $\mathbf{Q} = \operatorname{diag}\left(\sigma_1^2, \cdots, \sigma_{N_x}^2\right)$ is a diagonal matrix.
Given the observation and dynamical model in (1) and (2), the goal is to estimate source dynamic parameters $\{\mathbf{A}_k\}_{k=1}^{p}$, $\mathbf{B}$, and $\{\sigma_i^2\}_{i=1}^{N_x}$. It is worth mentioning that observations $\{\mathbf{y}_t\}_{t=1}^{T}$, linear mapping $\mathbf{C}$, and the measurement noise covariance matrix $\mathbf{R}$ are known and given. Moreover, number of lags in the underlying source dynamic VAR(p), i.e. p, can be obtained via standard variable selection techniques such as Akaike information criterion (AIC) [2].

III. Sparse Parameter Estimation

The unknown parameters is represented by $\boldsymbol{\theta} := (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{N_x})$ where $\boldsymbol{\theta}_i := \left(\sigma_i^2, \mathbf{a}_i, \mathbf{b}_i\right)$ is the corresponding parameters of the i^{th} source with $\mathbf{a}_i = \left[\left[\mathbf{A}_k\right]_{i,j}, \forall j, k\right]^\top$ and $\mathbf{b}_i = \left[\left[\mathbf{B}\right]_{i,j}, \forall j\right]^\top$.
Since, the log-likelihood function with respect to $\boldsymbol{\theta}$ has an intractable expression, we use an instance of the Expectation Maximization (EM) algorithm [3]. Starting from an initial guess $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(0)}$, at the l-th iteration, we compute the conditional expectation of $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(l)}) := \mathbb{E}\Big[\log p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}; \boldsymbol{\theta})\Big|\mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)}\Big]$ as a surrogate function and then maximize it with respect to $\boldsymbol{\theta}$ to obtain the updated estimate of $\widehat{\boldsymbol{\theta}}^{(l+1)}$. In the following, we summarize the two steps of EM procedure.

E-step: In the considered model, the surrogate function $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(l)})$ at the l-th iteration can be obtained as

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(l)}) = \mathcal{K}(\widehat{\boldsymbol{\theta}}^{(l)}) - \frac{T}{2}\sum_{i=1}^{N_x}\log(\sigma_i^2) - \sum_{i=1}^{N_x}\frac{1}{2\sigma_i^2}\left(\mathbf{a}_i{}^\top\mathbf{G}^{(l)}\mathbf{a}_i - 2\mathbf{h}_i^{(l)\top}\mathbf{a}_i + f_i^{(l)} + \mathbf{b}_i^\top\mathbf{S}\mathbf{b}_i - 2\mathbf{d}_i^{(l)\top}\mathbf{b}_i + 2\mathbf{b}_i^\top\mathbf{U}^{(l)}\mathbf{a}_i\right), \tag{3}$$

where $\mathcal{K}(\widehat{\boldsymbol{\theta}}^{(l)})$ denotes the constant terms with respect to $\boldsymbol{\theta}$, and $\mathbf{G}^{(l)}$, $\mathbf{S}$, $\mathbf{U}^{(l)}$, $\mathbf{h}_i^{(l)}$, $\mathbf{d}_i^{(l)}$, $f_i^{(l)}$ are functions of the first- and second-order moments of the conditional density $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)})$. The details of calculation are presented in Appendix A.

Algorithm 1 Sparse parameter estimation for linear dynamical systems

Input: $\{\mathbf{y}_t\}_{t=1}^T, \{\mathbf{e}_t\}_{t=1}^T, \mathbf{C}, \mathbf{R}, \boldsymbol{\lambda}, \boldsymbol{\gamma}$.

Output: $\{\mathbf{A}_k\}_{k=1}^p, \mathbf{B}, \{\sigma_i^2\}_{i=1}^{N_x}$.

Set the initial value ($l = 0$) $\widehat{\boldsymbol{\theta}}^{(l)}$: $\mathbf{A}_k = \mathbf{0}\ (\forall k),\ \mathbf{B} = \mathbf{0},\ \sigma_i^2 = 1\ (\forall i)$.

repeat

  Run the (Efficient) Forward Filtering and Backward Smoothing to obtain non-causal beliefs $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)})$ [5].

  Update $\mathbf{G}^{(l)}$, $\mathbf{U}^{(l)}$, $\mathbf{S}$, $\mathbf{h}_i^{(l)}$, $\mathbf{d}_i^{(l)}$, $f_i^{(l)}$ $(\forall i)$ from (12) and (13).

  Set initial value ($m = 0$): $\widehat{\mathbf{a}}_i^{(l),(m)} = \widehat{\mathbf{a}}_i^{(l)}$ and $\widehat{\mathbf{b}}_i^{(l),(m)} = \widehat{\mathbf{b}}_i^{(l)}$ $\forall i$.

  repeat

    Update the weight matrices as follows $(\forall i)$

$$\mathbf{W}^{a}{}_i^{(m)} = \mathrm{diag}\Big(\Big([\widehat{\mathbf{a}}_i^{(l),(m)}]_j^2 + \delta_a^2\Big)^{-\frac{1}{2}},\ \forall j\Big), \quad \mathbf{W}^{b}{}_i^{(m)} = \mathrm{diag}\Big(\Big([\widehat{\mathbf{b}}_i^{(l),(m)}]_j^2 + \delta_b^2\Big)^{-\frac{1}{2}},\ \forall j\Big).$$

    Update the coefficients as follows $(\forall i)$

$$\begin{bmatrix} \widehat{\mathbf{a}}_i^{(l),(m+1)} \\ \widehat{\mathbf{b}}_i^{(l),(m+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{G}^{(l)} + \lambda_i' \mathbf{W}^{a}{}_i^{(m)} & \mathbf{U}^{(l)\top} \\ \mathbf{U}^{(l)} & \mathbf{S} + \gamma_i' \mathbf{W}^{b}{}_i^{(m)} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{h}_i^{(l)} \\ \mathbf{d}_i^{(l)} \end{bmatrix},$$

    Set $m \leftarrow m + 1$.

  until Convergence $\left(\widehat{\mathbf{a}}_i^{(l),(m)}, \widehat{\mathbf{b}}_i^{(l),(m)},\ \forall i\right)$

  Update the variances as $(\forall i)$

$$\widehat{\sigma}_i^{2\,(l+1)} = \tfrac{1}{T}\left(\widehat{\mathbf{a}}_i^{(l+1)\top}\mathbf{G}^{(l)}\widehat{\mathbf{a}}_i^{(l+1)} - 2\mathbf{h}_i^{(l)\top}\widehat{\mathbf{a}}_i^{(l+1)} + f_i^{(l)} + \widehat{\mathbf{b}}_i^{(l+1)\top}\mathbf{S}\widehat{\mathbf{b}}_i^{(l+1)} - 2\mathbf{d}_i^{(l)\top}\widehat{\mathbf{b}}_i^{(l+1)} + 2\widehat{\mathbf{b}}_i^{(l+1)\top}\mathbf{U}^{(l)}\widehat{\mathbf{a}}_i^{(l+1)}\right).$$

  Set $l \leftarrow l + 1$.

until convergence of $Q(\widehat{\boldsymbol{\theta}}^{(l)}|\widehat{\boldsymbol{\theta}}^{(l)})$

One can show that due to the underlying Gaussian assumptions on $\mathbf{n}_t$ and $\mathbf{w}_t$, the conditional density of $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}; \boldsymbol{\theta})$ is Gaussian as well. Thus, the first- and second-order moments can be efficiently computed via the Fixed Interval Smoothing (FIS) algorithm [4]. The details and a MATLAB implementation are available on [5].

M-step: To avoid ill-posedness imposed by low-dimensionality of the observations, we add a regularization function in the update as follows

$$\widehat{\boldsymbol{\theta}}^{(l+1)} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\ \left\{Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(l)}) + R_q(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\theta})\right\}, \tag{4}$$

where $R_q(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = -\sum_{i=1}^{N_x}\left(\lambda_i \|\mathbf{a}_i\|_q^q + \gamma_i \|\mathbf{b}_i\|_q^q\right)$ is the regularization function and $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_{N_x}]^\top$ and $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_{N_x}]^\top$ are the $N_x \times 1$ regularization coefficients vectors. The closed-form solution for $q = 2$ can be obtained as

$$\begin{bmatrix} \widehat{\mathbf{a}}_i^{(l+1)} \\ \widehat{\mathbf{b}}_i^{(l+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{G}^{(l)} + \lambda_i'\mathbf{I} & \mathbf{U}^{(l)\top} \\ \mathbf{U}^{(l)} & \mathbf{S} + \gamma_i'\mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{h}_i^{(l)} \\ \mathbf{d}_i^{(l)} \end{bmatrix}, \tag{5}$$

$$\widehat{\sigma}_i^{2\,(l+1)} = \frac{1}{T}\left(\widehat{\mathbf{a}}_i^{(l+1)\top}\mathbf{G}^{(l)}\widehat{\mathbf{a}}_i^{(l+1)} - 2\mathbf{h}_i^{(l)\top}\widehat{\mathbf{a}}_i^{(l+1)} + f_i^{(l)} + \widehat{\mathbf{b}}_i^{(l+1)\top}\mathbf{S}\widehat{\mathbf{b}}_i^{(l+1)} - 2\mathbf{d}_i^{(l)\top}\widehat{\mathbf{b}}_i^{(l+1)} + 2\widehat{\mathbf{b}}_i^{(l+1)\top}\mathbf{U}^{(l)}\widehat{\mathbf{a}}_i^{(l+1)}\right). \tag{6}$$

To apply the sparsity condition, we may use $q = 1$, however, the closed-form solution does not exist. We use another EM algorithm known as Iteratively Re-weighted Least Squares (IRLS) to find the ℓ_1-norm regularized solution [6]. The detiled derivation is explained in Appendix B. The described EM procedure is presented in Algorithm 1.

APPENDIX A
E-STEP COMPUTATIONS

We start from the joint distribution of $\{\mathbf{x}_t\}_{t=1}^T$ and $\{\mathbf{y}_t\}_{t=1}^T$. From Baye's rule we have

$$\log p(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}; \boldsymbol{\theta}) = \log p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}; \boldsymbol{\theta}) + \log p(\mathbf{x}_{1:T}; \boldsymbol{\theta}). \tag{7}$$

The conditional distribution can be directly written from observation model in (1) as

$$\log p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}; \boldsymbol{\theta}) = \sum_{t=1}^{T} \log p(\mathbf{y}_t|\mathbf{x}_t; \boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi|\mathbf{R}|) - \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{C}\mathbf{x}_t\|_{\mathbf{R}^{-1}}, \tag{8}$$

where $\|\mathbf{a}\|_{\mathbf{B}} := \mathbf{a}^\top\mathbf{B}\mathbf{a}$ for notational convenience.

Using the fact that $\mathbf{Q} = \text{diag}(\sigma_1^2, \cdots, \sigma_{N_x}^2)$ along with the source dynamic model in (2), we have

$$\log p(\mathbf{x}_{1:T}; \boldsymbol{\theta}) = -\frac{T}{2} \log(2\pi \prod_{i=1}^{N_x} \sigma_i^2) - \sum_{i=1}^{N_x} \frac{1}{2\sigma_i^2} \left\| \underline{\mathbf{x}}_i - \mathcal{X}\mathbf{a}_i - \mathcal{S}\mathbf{b}_i \right\|_2^2, \tag{9}$$

where $\underline{\mathbf{x}}_i := [x_{i,q+1:T}]^\top$, $\mathbf{a}_i = \Big[[\mathbf{A}_k]_{i,j}, \forall j, k\Big]^\top$, $\mathbf{b}_i = \Big[[\mathbf{B}]_{i,j}, \forall j\Big]^\top$, and

$$\mathcal{X} := \Big[[x_{1,q:T-1}]^\top, \cdots, [x_{1,1:T-q}]^\top, \cdots, [x_{N_x,1:T-q}]^\top\Big],$$
$$\mathcal{S} := \Big[[e_{1,q+1:T}]^\top, [e_{2,q+1:T}]^\top, \cdots, [e_{N_e,1:T}]^\top\Big].$$

It is noteworthy to mention that $p(\mathbf{x}_{-q+1:0}) \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$ presenting the distribution of the sources initial condition is known. Now, by substituting (8) and (9) into (7) along with applying the expection, one can find $Q(.)$ function as follows

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(l)}) &= \mathbb{E}\big[\log p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}; \boldsymbol{\theta}) | \mathbf{y}_{1:T}, \widehat{\boldsymbol{\theta}}^{(l)}\big] \\ &= \mathcal{K}(\widehat{\boldsymbol{\theta}}^{(l)}) - \frac{T}{2} \sum_{i=1}^{N_x} \log(\sigma_i^2) - \sum_{i=1}^{N_x} \frac{1}{2\sigma_i^2} \left(\mathbf{a}_i^\top \mathbf{G}^{(l)} \mathbf{a}_i - 2{\mathbf{h}_i^{(l)}}^\top \mathbf{a}_i + f_i^{(l)} + \mathbf{b}_i^\top \mathbf{S} \mathbf{b}_i - 2{\mathbf{d}_i^{(l)}}^\top \mathbf{b}_i + 2\mathbf{b}_i^\top \mathbf{U}^{(l)} \mathbf{a}_i \right), \end{aligned} \tag{10}$$

where $\mathcal{K}(\widehat{\boldsymbol{\theta}}^{(l)})$ represents the constant terms with respect to $\boldsymbol{\theta}$

$$\mathcal{K}(\widehat{\boldsymbol{\theta}}^{(l)}) = -\frac{T}{2} \log(2\pi|\mathbf{R}|) - \frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}\big[\|\mathbf{y}_t - \mathbf{C}\mathbf{x}_t\|_{\mathbf{R}^{-1}} | \mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)}\big], \tag{11}$$

and

$$\mathbf{G}^{(l)} = \mathbb{E}\big[\mathcal{X}^\top \mathcal{X} | \mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)}\big], \quad \mathbf{S} = \mathcal{S}^\top \mathcal{S}, \quad \mathbf{U}^{(l)} = \mathcal{S}^\top \mathbb{E}\big[\mathcal{X} | \mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)}\big], \tag{12}$$

$$\mathbf{h}_i^{(l)} = \mathbb{E}\big[\mathcal{X}^\top \underline{\mathbf{x}}_i | \mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)}\big], \quad \mathbf{d}_i^{(l)} = \mathcal{S}^\top \mathbb{E}\big[\underline{\mathbf{x}}_i | \mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)}\big], \quad f_i^{(l)} = \mathbb{E}\big[{\underline{\mathbf{x}}_i}^\top \underline{\mathbf{x}}_i | \mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)}\big]. \tag{13}$$

The variables in (12) and (13) can be written as a function of first- and second- order moments of the conditional distribution $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}; \widehat{\boldsymbol{\theta}}^{(l)})$. For more details, check [5].

Appendix B
ℓ_1-norm regularization via IRLS

Since there is no closed-form ℓ_1-regularized solution for the optimization problem in (4), we use another instance of EM algorithm called IRLS to solve the problem in an iterative manner [6]. We denote the iteration index of the IRLS algorithm with m. At the l^{th} iteration of the main EM algorithm, we run another EM algortihm corresponding to the IRLS with the update step as follows

$$\begin{bmatrix} \widehat{\mathbf{a}}_i^{(l),(m+1)} \\ \widehat{\mathbf{b}}_i^{(l),(m+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{G}^{(l)} + \lambda_i' \mathbf{W}^{a_i^{(m)}} & {\mathbf{U}^{(l)}}^\top \\ \mathbf{U}^{(l)} & \mathbf{S} + \gamma_i' \mathbf{W}^{b_i^{(m)}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{h}_i^{(l)} \\ \mathbf{d}_i^{(l)} \end{bmatrix}, \tag{14}$$

where

$$\mathbf{W}^{a_i^{(m)}} = \text{diag}\Big(\Big([\widehat{\mathbf{a}}_i^{(l),(m)}]_j^2 + \delta_a^2\Big)^{-\frac{1}{2}}, \ \forall j\Big), \quad \mathbf{W}^{b_i^{(m)}} = \text{diag}\Big(\Big([\widehat{\mathbf{b}}_i^{(l),(m)}]_j^2 + \delta_b^2\Big)^{-\frac{1}{2}}, \ \forall j\Big). \tag{15}$$

$\mathbf{W}^{a_i^{(m)}}$ and $\mathbf{W}^{b_i^{(m)}}$ are $(pN_x) \times (pN_x)$ and $(N_e) \times (N_e)$ diagonal weight matrices corresponding to $\mathbf{a}_i$ and $\mathbf{b}_i$, respectively, and $\delta_a, \delta_b \in (0, 1)$ are two arbitrary constants.

References

[1] B. Soleimani, "SparsePELS," https://github.com/BehradSol/SparsePELS, 2020.
[2] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 6, pp. 16–34, 2018.
[3] 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: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
[4] B. D. Anderson and J. B. Moore, *Optimal Filtering*. Courier Corporation, 2012.
[5] B. Soleimani, "EfficientFFBS," https://github.com/BehradSol/EfficientFFBS, 2020.
[6] E. Pirondini, B. Babadi, G. Obregon-Henao, C. Lamus, W. Q. Malik, M. S. Hmlinen, and P. L. Purdon, "Computationally efficient algorithms for sparse, dynamic solutions to the eeg source localization problem," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 6, pp. 1359–1372, 2018.