Aalborg Universitet

Multi-Pitch Estimation and Tracking Using Bayesian Inference in Block Sparsity

Karimian-Azari, Sam; Jakobsson, Andreas; Jensen, Jesper Rindom; Christensen, Mads Græsbøll

Published in:
2015 Proceedings of the 23rd European Signal Processing Conference (EUSIPCO 2015)

DOI (link to publication from Publisher):
10.1109/EUSIPCO.2015.7362336

Publication date:
2015

Link to publication from Aalborg University

Citation for published version (APA):
Karimian-Azari, S., Jakobsson, A., Jensen, J. R., & Christensen, M. G. (2015). Multi-Pitch Estimation and Tracking Using Bayesian Inference in Block Sparsity. In *2015 Proceedings of the 23rd European Signal Processing Conference (EUSIPCO 2015)* (pp. 16-20). IEEE (Institute of Electrical and Electronics Engineers). https://doi.org/10.1109/EUSIPCO.2015.7362336

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: December 05, 2025

AALBORG UNIVERSITY
DENMARK

Multi-Pitch Estimation and Tracking Using Bayesian Inference in Block Sparsity

Sam Karimian-Azari†, Andreas Jakobsson⋆, Jesper Rindom Jensen†, and Mads Græsbøll Christensen†
email: †{ska, jrj, mgc}@create.aau.dk, and ⋆aj@math.lth.se

LUND UNIVERSITY

Introduction

Multi-pitch estimation methods often make *a priori* assumptions on

- the number of measured sources
- the model orders of these sources

The measured signals are defined as:

$$\mathbf{y}_n \triangleq \sum_{m=1}^{M} \mathbf{Z}_m \mathbf{b}_m + \mathbf{v} = \mathbf{Z}\mathbf{b} + \mathbf{v} \tag{1}$$

where

$$\begin{aligned}
\mathbf{Z} &= \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 & \dots & \mathbf{Z}_M \end{bmatrix}\\
\mathbf{Z}_m &= \begin{bmatrix} \mathbf{z}_m & \mathbf{z}_m^2 & \dots & \mathbf{z}_m^{L_m} \end{bmatrix}\\
\mathbf{z}_m^l &= \begin{bmatrix} 1 & e^{jl\omega_m} & \dots & e^{jl\omega_m(N-1)} \end{bmatrix}^T\\
\mathbf{b} &= \begin{bmatrix} \mathbf{b}_1^T & \mathbf{b}_2^T & \dots & \mathbf{b}_M^T \end{bmatrix}^T\\
\mathbf{b}_m &= \begin{bmatrix} b_{m,1} & b_{m,2} & \dots & b_{m,L_m} \end{bmatrix}^T
\end{aligned}$$

(I) Known basis matrix, Z: Assuming $L_{\text{tot}} = \sum_{m=1}^{M} L_m \ll N$, the maximum likelihood (ML) problem $\hat{\mathbf{b}}_{\text{ML}} = \arg\max_{\mathbf{b}} \log P(\mathbf{y}_n|\mathbf{b}, \sigma_v)$ has the least-squares (LS) solution: $\hat{\mathbf{b}}_{\text{LS}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{y}_n$.

(II) Unknown situation: Imposing the large dictionary $\mathbf{W} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 & \dots & \mathbf{Z}_S \end{bmatrix}$ on $S \gg M$ feasible pitches with the maximum harmonics L_r for $r = 1, 2, \dots, S$ avoids such assumptions, i.e.,

$$\mathbf{y}_n \triangleq \mathbf{W}\mathbf{a} + \mathbf{v} \tag{2}$$

The spectral amplitudes of $L_{\text{ext}} = \sum_{r=1}^{S} L_r$ sinusoids of the dictionary are exceedingly sparse, containing $L_{\text{tot}} < L_{\text{ext}}$ non-zero values, and $\hat{\mathbf{a}}_{\text{LS}} = \cancel{(\mathbf{W}^H\mathbf{W})^{-1}\mathbf{W}^H\mathbf{y}_n}$ because of over-fitting problem ($L_{\text{ext}} \gg N$).

PEBS

The pitch estimation using block sparsity (PEBS) technique [1] (Based on the Lasso technique [2]):

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \frac{1}{2}\|\mathbf{y}_n - \mathbf{W}\mathbf{a}\|_2^2 + J \tag{3}$$

$$J = \lambda_{\text{L}}\|\mathbf{a}\|_1 + \lambda_{\text{GL}} \sum_{r=1}^{S} \|\mathbf{a}_r\|_2 \tag{4}$$

where λ_{L} and λ_{GL} are the regularization coefficients of the penalties. For a given $\tilde{M}$,

$$\hat{\mathbf{\Omega}} = \begin{bmatrix} \hat{\omega}_1 & \hat{\omega}_2 & \dots & \hat{\omega}_{\tilde{M}} \end{bmatrix}^T = \arg\max_{\mathbf{\Omega}} P(\{\|\hat{\mathbf{a}}_r\|_2\}_{r=1}^{S}|\mathbf{\Omega}) \tag{5}$$

Hypothesis

The regularization coefficients should not be identical for all components of the dictionary!

Imposing a Laplace distribution on the amplitudes like $P(a_{r,l}|\tau_{r,l}, \sigma_v) = \frac{\tau_{r,l}}{2\sigma_v} \exp\left(-\frac{\tau_{r,l}}{\sigma_v}|a_{r,l}|\right)$, where $\tau_{r,l}$ is the shrinkage coefficient, we interpret the PEBS as a Bayesian posteriori estimator. Bayesian Lasso:

$$P(\mathbf{a}|\mathbf{y}_n, \mathbf{\Psi}, \sigma_v) \propto \exp\left(-\frac{1}{2\sigma_v^2}\|\mathbf{y}_n - \mathbf{W}\mathbf{a}\|_2^2\right) \prod_{r=1}^{S}\prod_{l=1}^{L_r} \exp\left(-\frac{\tau_{r,l}}{\sigma_v}|a_{r,l}|\right)$$

Bayesian Group-Lasso:

$$P(\mathbf{a}|\mathbf{y}_n, \mathbf{\Psi}, \sigma_v) \propto \exp\left(-\frac{1}{2\sigma_v^2}\|\mathbf{y}_n - \mathbf{W}\mathbf{a}\|_2^2\right) \prod_{r=1}^{S} \exp\left(-\frac{\|\mathbf{\Psi}_r\|_2}{\sigma_v}\|\mathbf{a}_r\|_2\right)$$

where $\mathbf{\Psi} = \left\{\bigcup_{r=1}^{S}\bigcup_{l=1}^{L_r} \tau_{r,l}\right\}$, and $\mathbf{\Psi}_r = \left\{\bigcup_{l=1}^{L_r} \tau_{r,l}\right\}$.

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \log P(\mathbf{a}|\mathbf{y}_n, \mathbf{\Psi}, \sigma_v) \tag{6}$$

$$J = \|\boldsymbol{\psi}_{\text{L}} \odot \mathbf{a}\|_1 + \sum_{r=1}^{S} \|\boldsymbol{\psi}_{\text{GL},r}\|_2 \|\mathbf{a}_r\|_2 \tag{7}$$

where $\boldsymbol{\psi}_{\text{L}} = \begin{bmatrix} \boldsymbol{\psi}_{\text{GL},1}^T & \dots & \boldsymbol{\psi}_{\text{GL},S}^T \end{bmatrix}^T$ and $\boldsymbol{\psi}_{\text{GL},r} = \begin{bmatrix} \psi_{\text{GL},r,1} & \dots & \psi_{\text{GL},r,L_r} \end{bmatrix}^T$ are the real-valued and non-negative regularization coefficients.

Data-Dependent Penalties

To allow the efficient tracking of smooth change in pitch values, we assign data-dependent regularization coefficients, e.g., less shrinkage to the important contents.

Adaptive regularization coefficients [3]:

$$\|\boldsymbol{\psi}_{\text{GL},r}\|_2 = \frac{\hat{\sigma}_v}{(\|\tilde{\mathbf{a}}_r\|_2)^k} \tag{8}$$

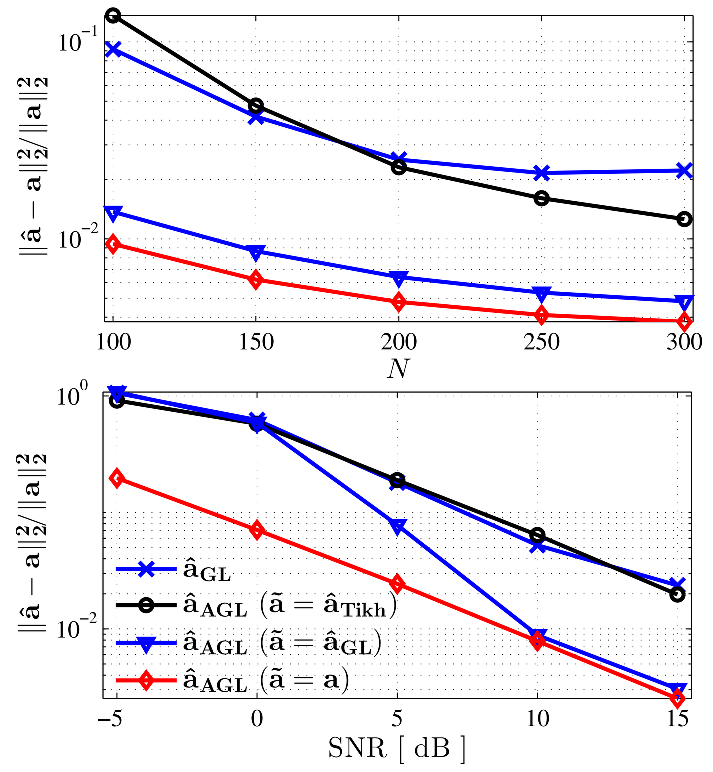
$$\psi_{\text{GL},r,l_r} = \frac{\hat{\sigma}_v}{(|\tilde{a}_{r,l_r}|)^k} \tag{9}$$

where $k > 0$, $\hat{\sigma}_v \approxeq \|\mathbf{y}_n - \mathbf{W}\tilde{\mathbf{a}}\|_2$, and $\tilde{\mathbf{a}} = E\{\mathbf{a}|\mathbf{\Psi}, \sigma_v\}$.

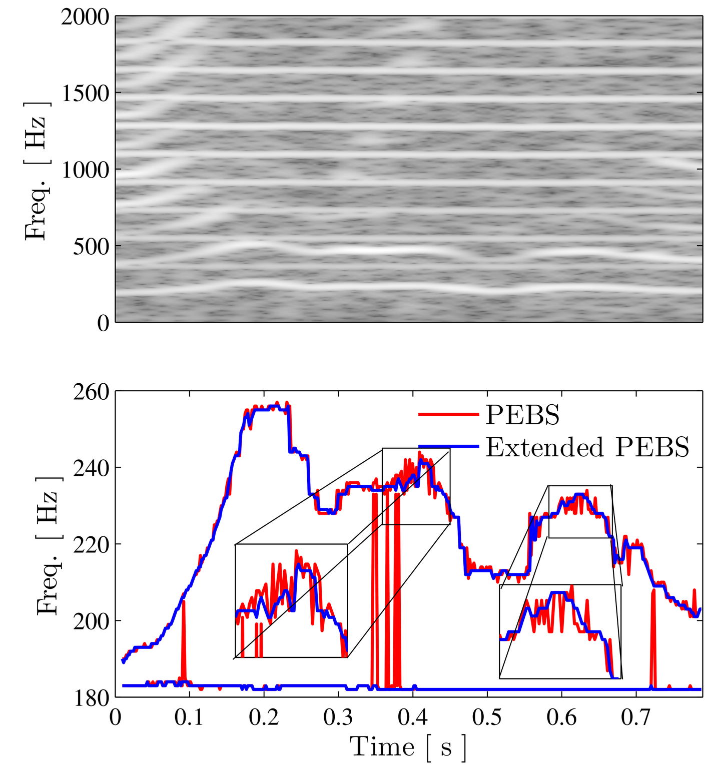
- The frequency contents of most audio signals are piecewise smooth, i.e., $E\{|\mathbf{a}(n)||\mathbf{\Psi}, \sigma_v\} \simeq E\{|\mathbf{a}(n+t)||\mathbf{\Psi}, \sigma_v\}$, so we can find $\tilde{\mathbf{a}}$ from the initial estimates $\hat{\mathbf{a}}(n)$ of the neighboring frames.
- For fast varying spectral content and poor initial estimates, we included a spectral smoothing, formed using kernel regression: $\bar{a}_{r,l} = \frac{\sum_{g=1}^{S}\sum_{l_g=1}^{L_g} K_{\mathbf{\Sigma}}(\mathbf{x}_g - \mathbf{x}_r)\, \tilde{a}_{g,l_g}}{\sum_{g=1}^{S}\sum_{l_g=1}^{L_g} K_{\mathbf{\Sigma}}(\mathbf{x}_g - \mathbf{x}_r)}$, with the kernel function $K_{\mathbf{\Sigma}}(\mathbf{x}_g - \mathbf{x}_r)$ that gives more weight at the data point $\mathbf{x}_g = [\omega_g, l_g\omega_g]^T$ that has a smaller Euclidean distance to $\mathbf{x}_r = [\omega_r, l\omega_r]^T$.

Experimental Results

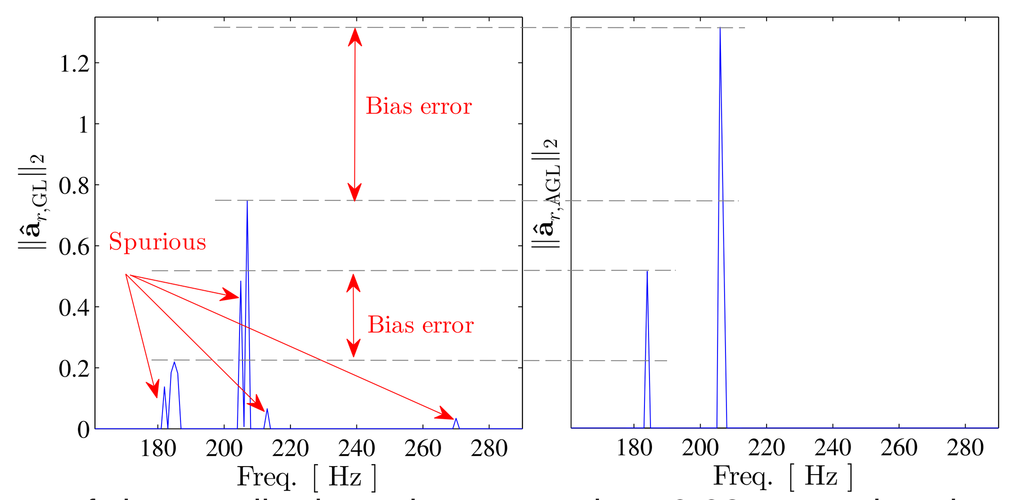
- A synthetic signal, uniformly drawn on $\omega_1 \in [160, 290] \times (2\pi/f_s)$, with a uniformly distributed number of harmonics $L_1 \in \mathcal{U}\{5, \lfloor \pi/\omega_1 \rfloor\}$, and unit amplitudes.
- Real signals, a mixture of a female voice and a trumpet signal, where $f_s = 8.0$ kHz, $S = 130$, $T = 1$, $k = 0.5$, $\lambda_{\text{L}} = 0.12$, $\lambda_{\text{GL},r} = 0.12\sqrt{L_r}$, and $\mathbf{\Sigma} = \text{diag}\{6.25, 0.01\} \times (2\pi/f_s)^2$.



Normalized MSE of the amplitude estimates at SNR = 10 dB (top), and using $N = 150$ (bottom).



Spectrogram of the examined real audio signals (top), and the resulting multi-pitch estimates (bottom).



ℓ_2-norm of the amplitude estimates at time 0.09 sec using the PEBS method [1] (left), and the extended PEBS (right).

Conclusion

- Data-dependent regularized LS, incorporated with an expectation on individual and grouped sinusoids
- Non-parametric smoothing
- Multi-pitch estimation and tracking without priori knowledge about sources

References

[1] S. I Adalbjörnsson, A. Jakobsson, and M. G Christensen. Multi-pitch estimation exploiting block sparsity. *Signal Processing*, 109:236–247, 2015.

[2] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[3] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Thanks to Villum Foundation, Carl Trygger's Foundation, and the Danish Council for Independent Research grant DFF-1337-00084.