ESTIMATING SPARSE SIGNALS USING INTEGRATED WIDE-BAND DICTIONARIES

Maksim Butsenko[*], *Johan Swärd*[**], *and Andreas Jakobsson*[**]

[*]Dept. of Radio and Comm. Eng., Tallinn University of Technology, Estonia
[**]Dept. of Mathematical Statistics, Lund University, Sweden

ABSTRACT

In this paper, we present a technique for reducing the size of the dictionary in sparse signal reconstruction by formulating an initial dictionary containing elements that spans bands of the considered parameter space. We allow for the use of this banded dictionary in a first-stage estimation procedure, in which large parts of the parameter space is discarded for further analysis, thereby reducing the overall computationally complexity required to allow for a reliable signal reconstruction. We illustrate the presented principle on the problem of estimating sinusoidal components corrupted by white noise.

Index Terms— Sparse signal reconstruction, dictionary learning, convex optimization

1. INTRODUCTION

A wide range of applications yields signals that may be well approximated using a sparse reconstruction framework, and the area has attracted dramatic interest in the recent literature (see, e.g., [1–3] and the references therein). Much of this work has focused on formulating convex algorithms that exploit different sparsity inducing penalties, thereby encouraging solutions that are well represented using just a few elements from some known dictionary matrix, $\mathbf{D}$. If the dictionary is appropriately chosen, even very limited measurements can be shown to allow for an accurate signal reconstruction [4, 5]. Recently, increasing attention has been given to signals that are best represented using a continuous parameter space. In such cases, the discretization of the parameter space that is typically used to approximate the true parameters will not represent the noise-free signal exactly, resulting in solutions that are less sparse than desired. This problem has been examined in, e.g., [6–8], wherein discretization recommendations and new bounds of the reconstruction guarantees were presented, taking the grid mismatch into consideration. Typically, this results in the use of large and over-complete dictionaries, which, although quite efficient, often violate the assumptions required to allow for a perfect recovery guaranty.

This work was supported in part by the Swedish Research Council and Crafoord's foundation, and in part by the Estonian national scholarship program Kristjan Jaak, which is funded and managed by Archimedes Foundation in collaboration with the Ministry of Education and Research.

As an alternative, one may formulate the reconstruction problem using a continuous dictionary, such as in, e.g., [9–11]. Such formulations typically use an atomic norm penalty, as introduced in [12], which allows for a way to determine the most suitable convex penalty to recover the signal, even over a continuous parameter space. Such a solution often offers an accurate signal reconstruction, but typically requires one to solve large and rather complicated optimization problems, thereby limiting the size of the considered problem.

In this work, we examine an alternative way of approaching the problem, proposing the use of wide-band dictionary elements, such that the dictionary is formed over B subsets of the continuous parameter space. In the estimation procedure, the activated subsets are retained and refined, whereas non-activated sets are discarded from the further optimization. Without loss of generality, the proposed principle is here illustrated on the problem of estimating the frequencies of K complex-valued sinusoid corrupted by white circularly symmetric Gaussian noise. This is a classical estimation problem, originally expressed using a sparse reconstruction framework in [13], and having since attracting notable attention (see, e.g., [14–17]). Here, using the classical formulation, the resulting sinusoidal dictionary will allow for a K-sparse representation of frequencies on the grid, whereas the grid mismatch of any off-grid components will typically yield solutions with more than K components. Extending the dictionary to use a finely spaced dictionary, as suggested in, e.g., [8], will yield the desired solution, although at the cost of an increased complexity. In this work, we instead proceed to divide the spectrum into B (continuous) frequency bands, each band possibly containing multiple spectral lines. This allows for an initial coarse estimation of the signal frequencies, without (significantly) increasing the risk of missing any off-grid components.

The proposed principle may also be used when solving the reconstruction problem using gridless methods, such as the methods in [9–11]. It has been shown that if the reconstruction problem allows for any prior knowledge about the location of the frequencies, e.g., the frequencies are located within a certain region of the spectrum, one may use this information to improve the estimates [18]. The proposed method may then be used for attaining such prior information, and thus improving the overall estimates as a result.

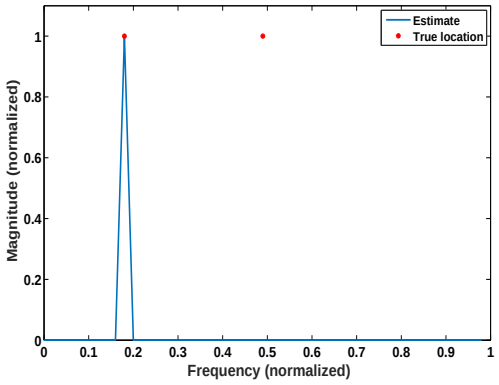

Fig. 1. The inner-product of a dictionary containing $L = 50$ (narrowband) candidate frequency elements and the noise-free signal, with $N = 100$.

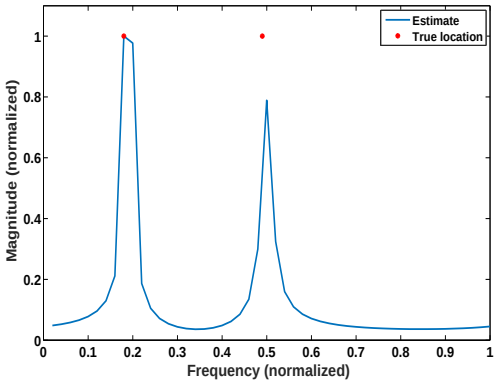

Fig. 2. The inner-product of a dictionary containing $B = 50$ (wide-band) candidate frequency elements and the noise-free signal, with $N = 100$.

2. PROBLEM STATEMENT

Consider the problem of estimating the frequencies f_k, for $k = 1, \ldots, K$, of a measured signal y_n, with

$$y_n = \sum_{k=1}^{K} \beta_k e^{2i\pi f_k t_n} + \epsilon_n \tag{1}$$

for $n = 1, \ldots, N$, and where K denotes the (unknown) number of sinusoids in the signal. Furthermore, let β_k and f_k denote the complex amplitude and frequency of the kth frequency, respectively, t_n the nth sample time, and ϵ_n the additive noise at time t_n. The classical sparse formulation of this estimation problem, as presented in [13], considers the LASSO minimization (see also [19])

$$\min_{\boldsymbol{x}} \frac{1}{2}||\mathbf{y} - \mathbf{D}\boldsymbol{x}||_2^2 + \lambda||\boldsymbol{x}||_1 \tag{2}$$

with

$$\mathbf{y} = \begin{bmatrix} y_1 & \ldots & y_N \end{bmatrix}^T \tag{3}$$

$$\mathbf{D} = \begin{bmatrix} \boldsymbol{d}_1 & \ldots & \boldsymbol{d}_L \end{bmatrix} \tag{4}$$

$$\boldsymbol{d}_\ell = \begin{bmatrix} e^{2i\pi \hat{f}_\ell t_1} & \ldots & e^{2i\pi \hat{f}_\ell t_N} \end{bmatrix}^T \tag{5}$$

where $\hat{f}_\ell$ for $\ell = 1, \ldots, L$ denotes the $L \gg K$ candidate frequencies in the dictionary, $\mathbf{D}$, typically selected to be closely spaced to allow for minimal grid mismatch, and $(\cdot)^T$ the transpose. The penalty on the 1-norm of $\boldsymbol{x}$ will ensure that the found solution, $\hat{\boldsymbol{x}}$, will be sparse, with λ denoting a user parameter governing the desired sparsity level of the solution. The desired frequencies, as well as their order, are then found as the non-zero elements in $\hat{\boldsymbol{x}}$. As shown in [8], the number of dictionary elements, L, typically has to be large to allow for reliable high-resolution frequency estimates.

As an alternative, one may use a zooming procedure, where one first employ an initial coarse frequency dictionary, $\mathbf{D}_1$, and then employ a fine dictionary, $\mathbf{D}_2$, centered around the initially found frequency estimates (see, e.g., [20,21] for similar approaches). This allows for computationally efficient solution of the optimization problem in (2), but suffers from the problem of possibly missing off-grid components far from the initial coarse frequency grid. This is illustrated in Figure 1, where the inner-product between the dictionary and the signal is depicted together with the location of the true peaks. In this noise-free example, we used $N = 100$ samples and $L = 50$ dictionary elements, with one of the frequencies being situated in between two adjacent grid points in the dictionary. As seen in the figure, the coarse initial estimate fails to detect the presence of the second sinusoid, which is thereby discarded as a possibility in the following refined estimate. Increasing the number of candidate frequencies will result in that the side-lobes of the more finely spaced frequencies will lessen the gap between the frequency grid points, making the inner-product between the dictionary and the signal larger for sinusoidal components that lies between two candidate frequencies. However, doing so will increase computational complexity correspondingly, begging the question if one may retain a low number of candidate frequencies, while reducing the likelihood of missing any off-grid components. This is the problem we examine in the following.

3. INTEGRATED WIDE-BAND DICTIONARIES

To allow for off-grid components, we here instead propose forming a wide-band dictionary over B frequency bands, with each integrated wide-band dictionary element being formed as

$$\mathbf{a}_b = \int_{f_b}^{f_{b+1}} e^{2i\pi f t} df \tag{6}$$

where f_b and f_{b+1} are the two frequencies bounding the frequency band, for $b = 1, \ldots, B$. The resulting elements are

then gathered into the dictionary, $\mathbf{A}$, formed as

$$\mathbf{A} = \left[\begin{array}{ccc} \mathbf{a}_1 & \dots & \mathbf{a}_B \end{array} \right] \tag{7}$$

with the bth dictionary element at time t_n being formed as

$$a_{b,n} = \frac{e^{2i\pi f_{b+1} t_n} - e^{2i\pi f_b t_n}}{2i\pi t_n} \tag{8}$$

where $a_{b,n}$ denotes the nth element in column b of $\mathbf{A}$. The inner-product between the proposed dictionary, $\mathbf{A}$, and the earlier signal is shown in Figure 2, using the same number of dictionary elements as in that case, i.e., with $B = 50$, clearly indicating that the proposed dictionary is able to locate the off-grid frequency. This is due to the wide-band nature of the proposed dictionary, which thus has less power concentrated at the grid points, but covers a wider range of frequencies, not reducing to zero, or close to zero, anywhere within the band (as is the case for the narrowband dictionary elements). As a result, using the wide-band dictionary elements, it is possible to use a smaller initial dictionary, thereby reducing the computational complexity, without increasing the risk of missing components in the signal.

4. EFFICIENT IMPLEMENTATION

To form a computationally efficient solution of the problem and to showcase the complexity reduction provided by the method proposed in this paper, we proceed to solve (2) using the popular ADMM algorithm [22]. In order to do so, the variable $\boldsymbol{x}$ is split into two variables, here denoted $\boldsymbol{x}$ and $\boldsymbol{z}$, after which the (scaled) augmented Lagrangian may be formulated as

$$L_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{u}} = ||\mathbf{y} - \mathbf{A}\boldsymbol{x}||_2^2 + \lambda||\boldsymbol{z}||_1 + \rho||\boldsymbol{x} - \boldsymbol{z} + \boldsymbol{u}||_2^2 \tag{9}$$

where $\boldsymbol{u}$ is the scaled dual variable and ρ is the step length (see [22] for a detailed discussion). The minimization is thus formed by iteratively solving (9) for $\boldsymbol{x}$ and $\boldsymbol{z}$, as well as updating the scaled dual variable $\boldsymbol{u}$. This is done by finding the (sub-)gradient for $\boldsymbol{x}$ and $\boldsymbol{z}$ of the augmented Lagrangian, and setting it to zero, fixing the other variables to their latest values. The steps for the jth iteration are thus

$$\boldsymbol{x}^{(j+1)} = \left(\mathbf{A}^H\mathbf{A} + \rho\mathbf{I}\right)^{-1}\left(\mathbf{A}^H\mathbf{y} + \boldsymbol{z}^{(j)} - \boldsymbol{u}^{(j)}\right) \tag{10}$$

$$\boldsymbol{z}^{(j+1)} = S(\boldsymbol{x}^{(j+1)} + \boldsymbol{u}^{(j)}, \lambda/\rho) \tag{11}$$

$$\boldsymbol{u}^{(j+1)} = \boldsymbol{u}^{(j)} + \boldsymbol{x}^{(j+1)} - \boldsymbol{z}^{(j+1)} \tag{12}$$

where $(\cdot)^H$ denotes the Hermitian transpose, $(\cdot)^{(j)}$ the jth iteration, and $S(\boldsymbol{x}, \kappa)$ the soft threshold operator, defined as

$$S(\boldsymbol{v}, \kappa) = \frac{\max\left(|\boldsymbol{v}| - \kappa, 0\right)}{\max\left(|\boldsymbol{v}| - \kappa, 0\right) + \kappa} \odot \boldsymbol{v} \tag{13}$$

where $\kappa \odot \boldsymbol{v}$ denotes the element-wise multiplication for any vector $\boldsymbol{v}$ and scalar κ.

Settings	Complexity ratio	Grid distance (10^{-3})
D1000	1	0.50
B20 Q25	31	1.0
B20 Q40	7	0.63
B40 Q25	26	0.50
B40 Q40	7	0.31
B75 Q25	16	0.27
B75 Q40	6	0.17
B75 Q323	1	0.02

Table 1. Complexity reduction compared to using the full dictionary and the distance between the final grid for different settings. Here, D1000 indicates the one-stage narrowband dictionary using a dictionary with $L = 1000$ elements, whereas $B20$ $Q25$ indicates the two-stage dictionary using $B = 20$ wide-band elements, followed by $Q = 25$ narrowband elements in the second-stage dictionary.

The computationally most demanding part of the resulting ADMM implementation is to form the inverse in (10) and to calculate $\mathbf{A}^H\mathbf{y}$. These steps are often done by QR factorizing the inverse prior to the iteration, so that this part is only calculated once, and then using the factors when forming the inner product. The total computational cost for the step in (10) depends on the size of the matrix $\mathbf{A}$ (or, correspondingly, $\mathbf{D}$, if using the narrowband dictionary). If $\mathbf{A}$ is an $N \times L$ matrix, and if $L < N$, computing the inverse will cost approximately L^3 operations, plus an additional L^2N operations to form the Gram-matrix $\mathbf{A}^H\mathbf{A}$. Furthermore, to compute $\mathbf{A}^H\mathbf{y}$ requires LN operations, and the final step to compute $\boldsymbol{x}$ costs L^2 operations. If instead $L > N$, one may make use of the Woodbury matrix identity [23], allowing the inverse to be formed using $N^3 + 3LN^2$ operations, whereafter one has to compute $\mathbf{A}^H\mathbf{y}$ and the final matrix-vector multiplication, together costing $LN + L^2$ operations. In total, the x-step will have the cost of roughly $L^3 + (N+1)L^2 + NL$, if $L < N$, or $N^3 + 3LN^2 + LN + L^2$, if $N < L$.

Since using the banded dictionary allows for a smaller dictionary, one may calculate the computational benefit of using the integrated dictionary as compared to just using an ordinary dictionary with large L. Consider using only a single-stage narrowband dictionary, $\mathbf{D}_1$, with $L > N$ dictionary elements. This requires $C_1 = N^3 + 3LN^2 + L^2 + LN$ operations if using the above ADMM solution, with the dictionary $\mathbf{D}_1$ in place of $\mathbf{A}$ in (10)-(12). If, on the other hand, one uses a two-stage wide-band dictionary with N dictionary elements in the initial coarse dictionary, $\mathbf{A}_1$ (which is more than required, but simplifies the calculations), the cost of forming the first stage (coarse) minimization is $C_2 = 2(N^3 + N^2)$. By taking the difference, i.e., forming $R = C_1 - C_2 = N^3 + 3LN^2 + L^2 + LN - 2(N^3 + N^2)$, one obtains the available computational resources, R, that are left for a second stage dictionary, $\mathbf{A}_2$, without increasing the overall com-

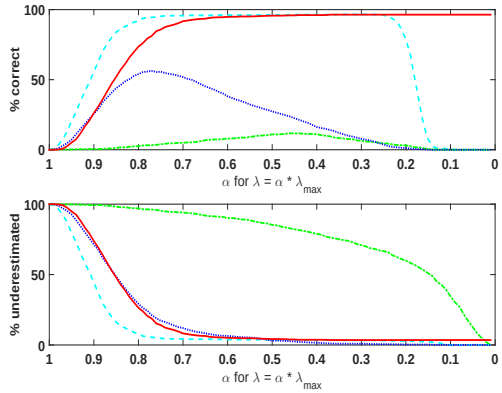

Fig. 3. The probability of (top) correctly estimating and (bottom) underestimating the number of spectral lines, for the (single-stage) narrowband dictionary, using $L = 1000$ elements (cyan, dashed) and $L = 75$ elements (green, dot-dashed), and for the wide-band dictionary, using $B = 75$ elements (blue, dotted), and the (two-stage) wide-band dictionary, using $B = 75$ elements, together with $Q = 25$ elements per activated bands in the refining dictionary (red, solid).

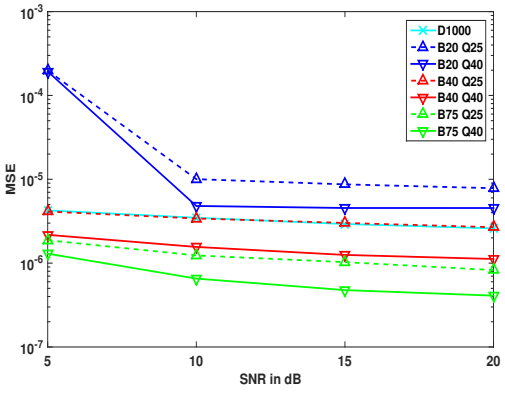

Fig. 4. Mean-square error curves for different SNR levels for the single-stage narrowband dictionary, using $L = 1000$, as compared to the two-stage dictionary, using B integrated wide-band elements in the first stage, followed by Q narrowband elements in the second stage.

putational cost above that of the narrowband dictionary solution. Assuming that the $\mathbf{A}_2$ dictionary has $Z > N$ grid points available, one may deduce the grid size by solving $R = N^3+3N^2Z+Z^2+ZN$, yielding that one is able to use a fine grid of $Z = (-3N^2+\sqrt{9N^4 + 2N^3 + N^2 + 4R}-N)/2$ candidates in a secondary refinement step, without increasing the total computational complexity, as compared to using the single stage narrowband dictionary. To illustrate the resulting difference, consider the following settings: $L = 1000$ and $N = 100$, yielding $Z \approx 936$ grid points to be distributed over the activated bands. If the number of activated bands are three in the settings above, that would yield a grid separation of $1.6 \cdot 10^{-5}$, which should be compared to the ordinary dictionary having a grid separation of $5 \cdot 10^{-4}$; a difference of roughly a factor 31.

5. NUMERICAL EXAMPLES

In this section, we proceed to examine the performance of proposed method, initially illustrating that the proposed (two-stage) wide-band estimator has the same estimation quality as when using the ordinary (one-stage) Lasso estimator. We considered a signal consisting of $N = 75$ samples containing $K = 3$ (complex-valued) sinusoids corrupted by a zero-mean white Gaussian noise with signal-to-noise ratio (SNR) of SNR$=$ 10dB. In each simulation, the sinusoidal frequencies are drawn from a uniform distribution, over $[0, 1)$, and all the amplitudes have magnitude 1 and phase drawn from a uniform distribution, over $[0, 2\pi)$. The performance is then computed using three different dictionaries, namely the (ordinary) narrowband dictionary, $\mathbf{D}$, with $L = 1000$ and $L = 75$ elements, respectively, and the proposed wide-band dictionary, $\mathbf{A}$, using $B = 75$ elements, followed by a second-stage narrowband dictionary using $Q = 25$ elements per active band. For each dictionary, we evaluate the performance for varying values of the user parameter α using $\lambda = \alpha\lambda_{max}$, where $\lambda_{max} = \max_i |\mathbf{x}_i^H \mathbf{y}_i|$ is the smallest tuning parameter value for which all coefficients in the solution are zero [24]. Each estimated result is then compared to the ground truth, counting the number of correct and underestimated model order estimates. The result is shown in Figure 3. As can be seen from the figure, the best results are achieved when $\alpha \leq 0.65$, in which case the proposed wide-band dictionary, using $B = 75$ bands, followed by a second stage narrowband dictionary, with $Q = 25$ per activated band, have similar performance to the narrowband dictionary using $L = 1000$ dictionary elements. Proceeding, we asses the mean-square error (MSE) for different settings of the two-stage dictionary, showing the MSE as a function of SNR for various sizes of the first-stage wide-band dictionary (B) and second-stage narrowband refining dictionary (Q). Figure 4 shows the resulting MSE, for the estimates with correctly estimated model order; Table 1 shows the corresponding complexity cost and the final grid distance of the second-stage dictionary. As can be seen from the figure, the two-stage dictionary using a wide-band dictionary, with $B = 40$ bands, followed by a refining dictionary using $Q = 25$ narrowband elements, achieves the same performance as the single-stage narrowband dictionary using $L = 1000$ elements, although the latter requires about 26 times fewer operations. Furthermore, it may be noted that using the same overall complexity, as resulting from using $B = 75$ and $Q = 323$, we achieve 25 times higher resolution as compared to the single-stage dictionary. All results are computed using 1000 Monte-Carlo simulations.

6. REFERENCES

[1] M. Unser and P. Tafti, *An introduction to sparse stochastic processes*, Cambridge University Press, 2013.

[2] M. Elad, *Sparse and Redundant Representations*, Springer, 2010.

[3] E. J. Candès and M. B. Wakin, "An Introduction To Compressive Sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, March 2008.

[4] E. J. Candès, J. Romberg, and T. Tao, "Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[5] D.L. Donoho, "Compressed Sensing," *IEEE Trans. Inf. Theory*, vol. 52, pp. 1289–1306, 2006.

[6] M. A. Herman and T. Strohmer, "Genral Deviants: An Analysis of Perturbations in Compressed Sensing," *IEEE J. Sel. Topics in Signal Processing*, vol. 4, no. 2, pp. 342–349, April 2010.

[7] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to Basis Mismatch in Compressed Sensing," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2182 – 2195, May 2011.

[8] P. Stoica and P. Babu, "Sparse Estimation of Spectral Lines: Grid Selection Problems and Their Solutions," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 962–967, Feb. 2012.

[9] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed Sensing Off the Grid," *IEEE Trans. Inform. Theory*, vol. 59, no. 11, pp. 7465–4790, Nov 2013.

[10] Y. Chi and Y. Chen, "Compressive Two-Dimensional Harmonic Retrieval via Atomic Norm Minimization," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1030–1042, Feb 2015.

[11] Z. Yang and L. Xie, "Enhancing Sparsity and Resolution via Reweighted Atomic Norm Minimization," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 995–1006, Feb 2016.

[12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The Convex Geometry of Linear Inverse Problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, Dec 2012.

[13] J. J. Fuchs, "On the Use of Sparse Representations in the Identification of Line Spectra," in *17th World Congress IFAC*, Seoul, jul 2008, pp. 10225–10229.

[14] P. Stoica, P. Babu, and J. Li, "New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 35–47, Jan 2011.

[15] P. Stoica and P. Babu, "SPICE and LIKES: Two hyperparameter-free methods for sparse-parameter estimation," *Signal Processing*, vol. 92, no. 7, pp. 1580–1590, July 2012.

[16] I. F. Gorodnitsky and B. D. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Reweighted Minimum Norm Algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, March 1997.

[17] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-Pitch Estimation Exploiting Block Sparsity," *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[18] Z. Yang and L. Xie, "Frequency-Selective Vandermonde Decomposition of Toeplitz Matrices With Applications," 2016, Publication: eprint arXiv:1605.02431.

[19] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[20] S. Sahnoun, E. H. Djermoune, and D. Brie, "Sparse Modal Estimation of 2-D NMR Signals," in *38th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 26-31 2013.

[21] J. Swärd, S. I. Adalbjörnsson, and A. Jakobsson, "High Resolution Sparse Estimation of Exponentially Decaying N-dimensional Signals," *IEEE Trans. Signal Process.*, submitted.

[22] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[23] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, 4^{th} edition, 2013.

[24] R. Tibshirani, J. Bienand, J. Friedman, T. Hastieand N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 2, pp. 245–266, 2012.