Efficient Estimation of Decaying Sinusoids with Application in NMR Spectroscopy

Martin Jälmby

Master's thesis
2018:E20

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Abstract

In this thesis, the Wideband Sparse Exponential Mode Analysis (WSEMA) estimator is introduced. It combines two recently developed techniques, the wideband dictionary and the Sparse Exponential Mode Analysis (SEMA) to make for an efficient estimator. WSEMA estimates the parameters of decaying sinusoids, without a priori-information about the number of modes present in the signal. WSEMA works with arbitrary sampling schemes and is therefore compatible with sampling scheme optimization ideas presented recently. The suggested estimator is evaluated using both simulated data and real nuclear magnetic resonance (NMR) spectroscopy data. The results in this thesis suggests that WSEMA can be used to efficiently estimate the frequencies and dampings of multi-modal signals with minimum variance, although work remains concerning the handling of closely spaced peaks.

Parts of the content in this thesis have been published in the article *Computationally Efficient Estimation of Multi-dimensional Damped Modes Using Sparse Wideband Dictionaries*, accepted to the 26th European Signal Processing Conference (EUSIPCO 2018).

Keywords: decaying sinusoids, semi-parametric methods, wideband dictionary, NMR spectroscopy

Acknowledgements

I would like to extend my gratitude to my supervisor Prof. Andreas Jakobsson for his guidance and contagious enthusiasm. I would also like to thank Dr. Johan Swärd and Filip Elvander for their continuous encouragement and for always taking the time to answer my questions. Thanks to Jesper Jönsson for proofreading and providing helpful comments. Finally, thank you Alice, for always being on my team.

Contents

1 Background

1.1 Introduction

A sum of decaying sinusoids is a model used throughout a variety of applications, such as speech and audio signals [1], radar imaging [2], and nuclear magnetic resonance (NMR) spectroscopy [3]. The model traces its roots back to as early as 1795, when Gaspard Riche de Prony developed the so called Prony's method. Practical relevance was however delayed until the development of the modern computer [4]. On account of the number of possible applications, the topic has received a lot of attention in the recent literature and there exists several different types of algorithms to estimate the parameters of the model. The maximum likelihood (ML) estimator can attain optimal performance, but due to the need of a multidimensional search, the computational burden is huge, making it practically infeasible in more than two dimensions [5]. Amongst other methods are subspace based methods such as e.g. matrix pencil [6], M-D DMUSIC [7], ESPRIT [8] and PUMA [9]. These methods suffer from requiring a priori knowledge about the exact number of sinusoids present in the signal. Other approaches, like dCAPON and dAPES [10], have difficulty separating closely spaced components [11]. The recently introduced Sparse Exponential Mode Analysis (SEMA) [12] attempts to find a computationally efficient way to estimate the parameters by exploiting the Kronecker structure in the model. This, however, introduces restrictions on how the signal can be sampled. In this thesis, I propose the Wideband Sparse Exponential Mode Analysis (WSEMA) estimator for efficient estimation of decaying sinusoids. It is a general method although the application focus has been on NMR spectroscopy.

NMR is one of the most powerful, but still non-destructive, tools available when trying to determine the structure of chemical compounds. The mode of procedure is to alter the nuclear spins in a sample from equilibrium. After that, the time-dependant response is measured. This is called free induction decay (FID) and contains information from the entire sample. By spectral analysis, the goal is to identify the contributions made by the different parts of the sample and thereby determine it's structure [13, p. ix-x]. This alteration from equilibrium, accomplished by so called pulses, can be modified with respect to e.g. time between pulses. This way, the FID can become multidimensional, each dimension corresponding to a pulse setting. The multidimensional nature of applications like NMR fuels a need for efficient estimators. A desirable property of such an estimator is the ability to work with an arbitrary sampling scheme, since uniform sampling is practically impossible for multidimensional problems [11]. As an example, non-uniform sampling allowed for performing an NMR-experiment in 89 hours, instead of the 2.5 years it would have taken using conventional uniform sampling [14]. This motivates a section in this thesis about sampling scheme optimization. The section contains a discussion on the impact of the damping parameter and a conjecture about how it should be picked.

The proposed WSEMA-estimator combines two recently developed techniques. The first is the concept of wideband dictionary elements. In short, it utilizes integration over the frequency space to more efficiently find the frequency components present in the signal. The second technique is the aforementioned SEMA, that uses a relaxation-based procedure to estimate the parameters one mode a time to minimize the residual, rather than all at once.

This thesis is organized as follows. In the first chapter, the general framework regarding sparse spectral estimation, that is used throughout this thesis, is introduced. In chapter two, the concept of sampling scheme optimization is presented. The WSEMA-algorithm is introduced in chapter three, which also includes a discussion on spurious peaks. Results, using simulated as well as real NMR-data, is presented in chapter four. Finally, in chapter five, there is a discussion about the results as well as suggestions for further research.

1.2 Spectral estimation, parametric and non-parametric methods

Consider a discrete time, zero-mean, weak-sense stationary, random signal $y(t), t = 0, \pm 1, \pm 2, ...$ The autocovariance function for such a process is defined as

$$r(k) = \mathrm{E}\{y(t)y^*(t-k)\} \tag{1}$$

and assumed to depend only on the lag k. The power spectral density (PSD) is, under reasonable assumptions on $r(k)$, defined as

$$\phi(f) = \lim_{N\to\infty} \mathrm{E}\left\{ \frac{1}{N} \left| \sum_{t=1}^{N} y(t) e^{-2i\pi ft} \right|^2 \right\} \tag{2}$$

and represents the average power of the signal at the normalized frequency $f \in [0, 1]$. Estimating the PSD from a finite number of samples is the essence of spectral analysis [15].

Spectral estimation techniques are commonly divided into parametric and non-parametric methods. As the name suggest, non-parametric methods make no assumption about but the nature of the signal and instead estimates the spectrum from the definition of the PSD. An example of this is the periodogram,

$$\hat{\phi}_p(f) = \frac{1}{N} \left| \sum_{t=1}^{N} y(t) e^{-2i\pi ft} \right|^2. \tag{3}$$

In contrast to the non-parametric methods are the parametric methods. Here, the signal is assumed to satisfy a general model. The task at hand is then to estimate the parameters of that model. Non-parametric remain useful when little is known about the signal of interest and they are less sensitive to modelling errors, but assuming that the signal model is correct, parametric methods generally outperform non-parametric ones [15]. Along with parametric and non-parametric methods, there are semi-parametric methods. They contain elements from both the parametric and non-parametric realms and can be useful when some properties of the signal are known whereas others are not.

Consider a D-dimensional space with a K-mode decaying sinusoid. The number of dimensions,D, known, but K is here assumed to be unknown. The signal observed at $\boldsymbol{\tau}_n = \begin{bmatrix} t_n^{(1)} & t_n^{(2)} & \dots & t_n^{(D)} \end{bmatrix}^T$ can then be written as

$$x_{\boldsymbol{\tau}_n} = \sum_{k=1}^{K} \alpha_k \prod_{d=1}^{D} \xi(f_k^{(d)}, \beta_k^{(d)})^{t_n^{(d)}} + \epsilon_{\boldsymbol{\tau}_n}, \tag{4}$$

with

$$\xi(f_k^{(d)}, \beta_k^{(d)}) = e^{2\pi i f_k^{(d)} - \beta_k^{(d)}}. \tag{5}$$

Here, $\alpha_k \in \mathbb{C}$ denotes the amplitude of mode k, $f_k^{(d)} \in [0, 1]$, and $\beta_k^{(d)} \in \mathbb{R}^+$ the normalized frequency and damping respectively, in dimension d for the k:th mode, and $\epsilon_{\boldsymbol{\tau}_n}$ is an additive noise term. The noise is assumed to be circularly symmetric[1], white and Gaussian. Assuming that the signal is observed over N, possibly non-uniformly spaced, multidimensional sampling points $\Omega = \{\boldsymbol{\tau}_n \ , \ n = 1, 2, \dots, N\}$, the measurement vector $\mathbf{y}$ can be written as

$$\mathbf{y} = \sum_{k=1}^{K} \alpha_k \tilde{\mathbf{a}}_k^{\Omega} + \boldsymbol{\epsilon}^{\Omega}, \tag{6}$$

where

$$\tilde{\mathbf{a}}_k^{\Omega} = \begin{bmatrix} \prod_{d=1}^{D} \xi(f_k^{(d)}, \beta_k^{(d)})^{t_1^{(d)}} & \dots & \prod_{d=1}^{D} \xi(f_k^{(d)}, \beta_k^{(d)})^{t_N^{(d)}} \end{bmatrix}^T \tag{7}$$

$$\boldsymbol{\epsilon}^{\Omega} = \begin{bmatrix} \epsilon_{\boldsymbol{\tau}_1} & \dots & \epsilon_{\boldsymbol{\tau}_N} \end{bmatrix}^T, \tag{8}$$

[1] $e^{i\phi}\mathbf{z}$ has the same distribution as $\mathbf{z}$ for all real ϕ.

or in an equivalent matrix form

$$\mathbf{y} = \tilde{\mathbf{A}}^{\Omega}\boldsymbol{\alpha} + \boldsymbol{\epsilon}^{\Omega}, \tag{9}$$

with

$$\tilde{\mathbf{A}}^{\Omega} = \begin{bmatrix} \tilde{\mathbf{a}}_1^{\Omega} \dots & \tilde{\mathbf{a}}_K^{\Omega} \end{bmatrix} \tag{10}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \dots & \alpha_K \end{bmatrix}^T \tag{11}$$

1.3 Sparsity, regularized optimization, LASSO and ADMM

Estimating the order and parameters of (6) is an example of an often occurring situation, trying to fit a model to a set of measurements. This is commonly written on the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}. \tag{12}$$

Here, $\mathbf{y}$ is an vector with N elements, representing the measurements. The vector $\mathbf{x}$ is of length P and represent a set of coefficients. The linear map $\mathbf{A}$ is of size $N \times P$. Finally, $\boldsymbol{\epsilon}$ is a noise vector of length N. The goal is then to minimize the model residual, i.e. to find

$$\arg\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2. \tag{13}$$

In many applications, the assumption is made that $\mathbf{A}$ has full rank, that $N \geq P$ and that the noise terms are zero-mean, white, and independent. In this case, the least squares estimate can be found as

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 \\ &= \arg\min_{\mathbf{x}} \{[\mathbf{y} - \mathbf{A}\mathbf{x}]^H[\mathbf{y} - \mathbf{A}\mathbf{x}]\} \\ &= \arg\min_{\mathbf{x}} \{[\mathbf{x} - \mathbf{A}^{\dagger}\mathbf{y}]^H[\mathbf{A}^H\mathbf{A}][\mathbf{x} - \mathbf{A}^{\dagger}\mathbf{y}] + \mathbf{y}^H\mathbf{y} - \mathbf{y}^H\mathbf{A}\mathbf{A}^{\dagger}\mathbf{y}\} \\ &= \mathbf{A}^{\dagger}\mathbf{y}, \end{aligned} \tag{14}$$

where $\mathbf{A}^{\dagger}$ denotes the Moore-Penrose pseudoinverse, defined as $\mathbf{A}^{\dagger} \triangleq (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$.

In the common approach of the application considered in this thesis, however, $P \gg N$. The columns of $\mathbf{A}$ are closely spaced candidate frequencies, implying a discretization of the frequency space. This yields a vastly underdetermined system and (13) generally has infinitely many solutions. Another distinguishing feature of the application is that the system is assumed to be sparse. We don't necessarily know the exact number of components in the signal, but we know that it's a reasonably small number. This means that the majority of the entries in $\mathbf{x}$ should be equal to 0. In general, when solving (13), this will not be the case. This sparsity needs to be imposed on the solution.

To impose sparsity, equation (13) may be modified slightly. Instead of trying to minimize only the loss function, $\frac{1}{2}||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2$, a regularizing penalty term, $g(\mathbf{x})$, is added, to ensure that a less sparse solution significantly reduces the loss. This yields the problem

$$\underset{\mathbf{x}}{\text{minimize}} \ \frac{1}{2}||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 + \lambda g(\mathbf{x}), \tag{15}$$

where λ controls the degree of regularization. The choice of penalty function is not obvious. To effectively impose sparsity, one would ideally like to put a significant penalty on each non-zero term used in $\mathbf{x}$, no matter the size of the element, i.e.

$$g(\mathbf{x}) = ||\mathbf{x}||_0 = \sum_{m=1}^{M} |x_m|^0, \tag{16}$$

known as the ℓ_0-"norm". The quotation marks are highly deliberate, since ℓ_0 is not a proper norm, failing the homogeneity criterion[2]. The big drawback of using the ℓ_0-"norm" as regularizing function is that it would require a search over all possible combinations of zero and non-zero elements of $\mathbf{x}$, a task that is infeasible for almost all practically relevant problems. A popular remedy for this is to use a convex relaxation, see e.g. [16, p. 7], instead using the ℓ_1-norm, defined as

$$g(\mathbf{x}) = ||\mathbf{x}||_1 = \sum_{m=1}^{M} |\mathbf{x}_m|. \tag{17}$$

Equation (15) can now be written as

$$\underset{\mathbf{x}}{\text{minimize}} \ \frac{1}{2}||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 + \lambda||\mathbf{x}||_1, \tag{18}$$

known as the Least Optimal Shrinkage and Selection Operator, LASSO, introduced in [17]. The LASSO has the major advantage that it's convex, meaning that both the cost function and the set over which the function is minimized, are convex. Convex problems have the desirable property that a locally optimal point is also globally optimal. Much like the choice of penalty function, the choice of the parameter λ is not obvious. Commonly one prefers to connect the value of λ to the inner product between the signal and the dictionary,

$$\lambda = \gamma \max_{i=1,\dots,P} |\mathbf{a}_i^T \mathbf{y}|, \tag{19}$$

where a_i denotes the i:th column of $\mathbf{A}$, instead forcing a decision on γ. This is discussed in greater detail in [17] and [18]. In this thesis, γ has been selected to be in the range between 0.2 and 0.4. To further impose sparsity, a re-weighting scheme, as described in [19], will also be used. In such a scheme, instead of minimizing just the ℓ_1-norm of $\mathbf{x}$, one sets out to minimize $||\mathbf{W}\mathbf{x}||_1$, where $\mathbf{W}$ is a diagonal matrix consisting of weights $w_1, ..., w_P$, yielding

$$\underset{\mathbf{x}}{\text{minimize}} \ \frac{1}{2}||\mathbf{y} - \mathbf{A}\mathbf{x}||_2^2 + \lambda||\mathbf{W}\mathbf{x}||_1. \tag{20}$$

All weights are initialized as 1, then after solving (20), the weights for the q:th iteration are updated according to

$$w_p^{(q+1)} = \frac{1}{|x_p^{(q)}| + \delta} \tag{21}$$

where δ is a small number inserted to provide stability and to make sure that $x_p^{(q)} = 0$ not entirely excludes the possibility of $x_p^{(q+1)} \neq 0$. After that, (20) is solved again. This process is repeated a desired number of times. The reweighing scheme, that is utilized throughout the proposed algorithm whenever the LASSO is solved, also decreases the sensitivity of the choice of γ. To illustrate the difference between a sparse and a non-sparse solution, we simulate a simple, one-dimensional, undamped signal, uniformly sampled on $[0, N-1]$, where $N = 100$ denotes the number of samples. The signal consists of two modes with unit amplitude and with $f_1 = 0.25$ and $f_2 = 0.38$ respectively, disturbed by a noise with a variance $\sigma_e^2 = 4$. With a dictionary $\mathbf{A}$ consisting of 100 candidates, equations (13) and (18) are solved. The results are shown in figure 1. The least squares solution models the noise to large extent and violates the underlying sparsity assumption, whereas the LASSO solution consists of two components only. The LASSO solution at the bottom of figure 1 reveals an amplitude bias, common amongst sparse based methods. The estimated amplitudes are smaller than the ground truth. However, this is not a big problem, as the amplitude estimates may be refined later [16].

There exists several numerical solvers to convex optimization problems. Throughout this thesis, the Alternating Direction Method of Multipliers [20], is used. In our problem, the variable $\mathbf{x}$ in (18) is

[2] $g(a\mathbf{x}) \neq ag(\mathbf{x})$

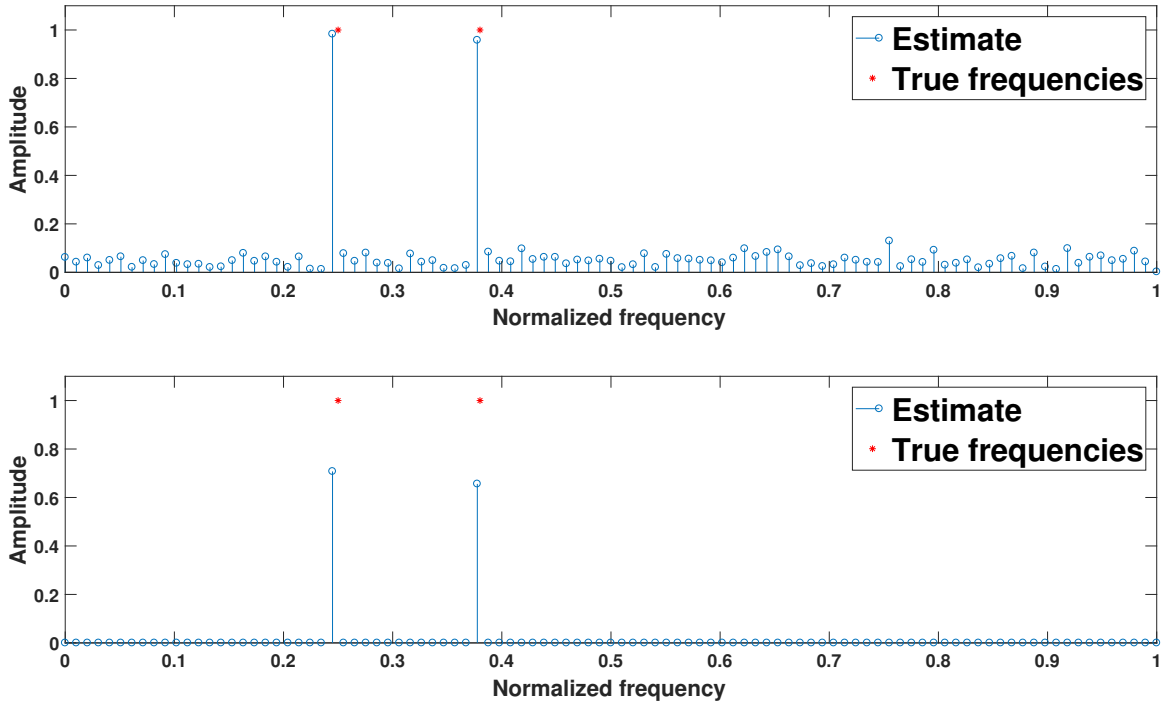

Figure 1: Non-sparse (least squares) solution (top), clearly modelling the noise to large extent, along with a sparse (LASSO) solution (bottom).

split into $\mathbf{x}$ and $\mathbf{z}$, and (18) is rewritten as

$$\underset{\mathbf{x},\,\mathbf{z}}{\text{minimize}}\ \frac{1}{2}||\mathbf{y}-\mathbf{A}\mathbf{x}||_2^2+\lambda||\mathbf{z}||_1,\quad \text{subject to}\quad \mathbf{x}=\mathbf{z}. \tag{22}$$

This is solved via the so called scaled augmented Lagrangian

$$\underset{\mathbf{x},\,\mathbf{z}}{\text{minimize}}\ \frac{1}{2}||\mathbf{y}-\mathbf{A}\mathbf{x}||_2^2+\lambda||\mathbf{z}||_1+\frac{\rho}{2}||\mathbf{x}-\mathbf{z}+\mathbf{u}||_2^2, \tag{23}$$

where $\mathbf{u}$ is the scaled dual variable and ρ is the step length. Equation (23) is solved iteratively for $\mathbf{x}$ and $\mathbf{z}$ and $\mathbf{u}$ is updated as well, according to the following set of equations,

$$\mathbf{x}^{(j+1)}=\left(\mathbf{A}^H\mathbf{A}+\rho\mathbf{I}\right)^{-1}\left(\mathbf{A}^H\mathbf{y}+\mathbf{z}^{(j)}-\mathbf{u}^{(j)}\right), \tag{24}$$

$$\mathbf{z}^{(j+1)}=S\left(\mathbf{x}^{(j+1)}+\mathbf{u}^{(j)},\lambda/\rho\right), \tag{25}$$

$$\mathbf{u}^{(j+1)}=\mathbf{u}^{(j)}+\mathbf{x}^{(j+1)}-\mathbf{z}^{(j+1)}, \tag{26}$$

where S is the soft threshold operator, defined as

$$S(\mathbf{x},\kappa)=\max\left(0,1-\kappa\odot\frac{1}{|\mathbf{x}|}\right)\odot\mathbf{x} \tag{27}$$

where $\odot$ denotes element-wise multiplication. This, of course, is defined only when $|\mathbf{x}|\neq 0$. If $|\mathbf{x}|=0$, $S(\mathbf{x},\kappa)=0$. More details on convex optimization in general can be found in [21] and on ADMM in particular in [20].

1.4 Dictionary considerations

As previously mentioned, $\mathbf{A}$ is often referred to as a dictionary in the sparse spectral estimation environment, since it can be seen as a dictionary of possible signal candidates. Each column of the

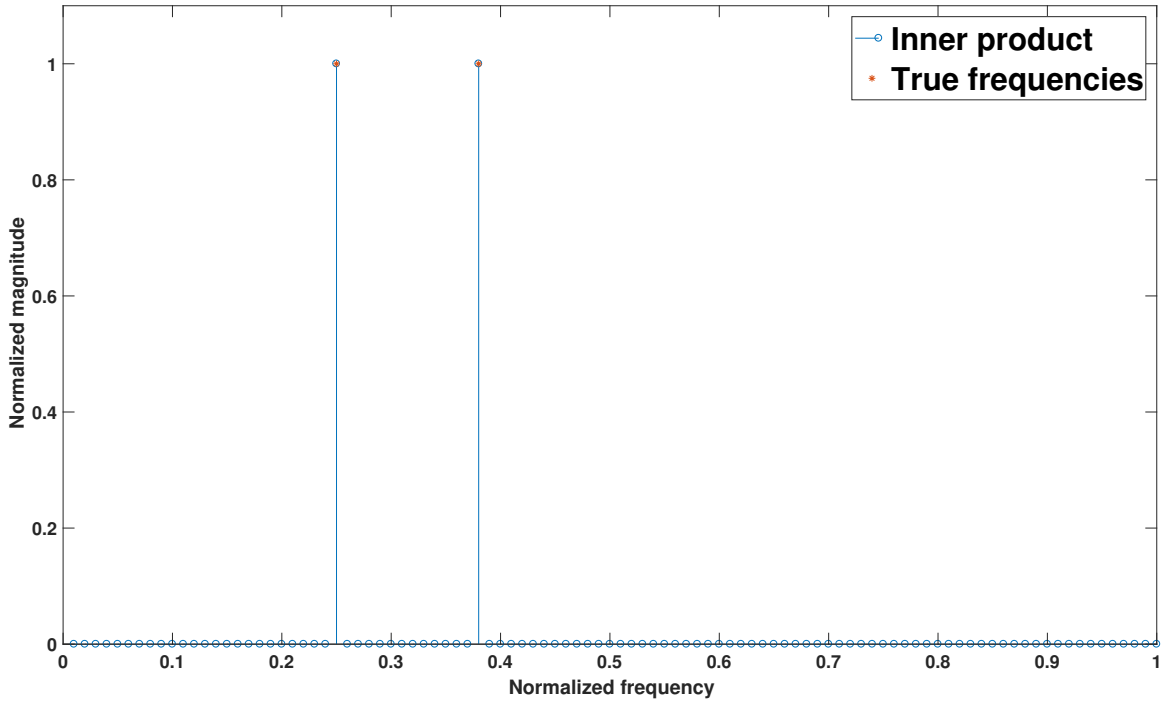

Figure 2: Inner product of the dictionary (100 elements) and the signal, N = 100, along with the true frequencies. The tight frequency grid ensures that both frequency components is close enough to at least one frequency candidate.

dictionary, called an atom, representing a unique combination of frequencies and dampings and their propagation over time

$$\bar{\mathbf{a}}_p^{\Omega} = \left[\prod_{d=1}^{D} \xi(f_p^{(d)}, \beta_p^{(d)})^{t_1^{(d)}} \quad \cdots \quad \prod_{d=1}^{D} \xi(f_p^{(d)}, \beta_p^{(d)})^{t_N^{(d)}} \right]^T \tag{28}$$

The dictionary $\bar{\mathbf{A}}^{\Omega}$ will then have the following form,

$$\bar{\mathbf{A}}^{\Omega} = \left[\bar{\mathbf{a}}_1 \ \cdots \ \bar{\mathbf{a}}_P \right]. \tag{29}$$

We here term this a narrowband dictionary. The assumption is that with a large enough number of candidate frequencies, the true frequencies will be close enough to a candidate frequency and it will be possible to recover the correct frequencies. This is illustrated in figure 2, where we see the inner product of a two-mode signal of length 100 and a dictionary containing 100 atoms. Both frequencies are successfully recovered. However if one tries to do this with a smaller dictionary, one might irretrievably lose components of the signal. This is illustrated in figure 3. Here, we have a one-dimensional, two mode signal without any damping. One of the frequency components, $f_1 = 0.38$ is located on the frequency grid, whereas the other, $f_2 = 0.25$ is off-grid, in between two dictionary elements, and in, in this case, in the null space of $\mathbf{A}$. To avoid losing any frequency components, the number of grid points in each dimension needs to be at least as many as the number of observations. In a multidimensional application such as NMR spectroscopy, this is devastating as the curse of dimensionality will balloon the size of the dictionary.

1.5 Integrated wideband dictionary elements

The idea of using integrated wideband dictionary elements (IWDE) is introduced in [18] as a possible remedy to using large dictionaries. In an integrated wideband framework, each dictionary element

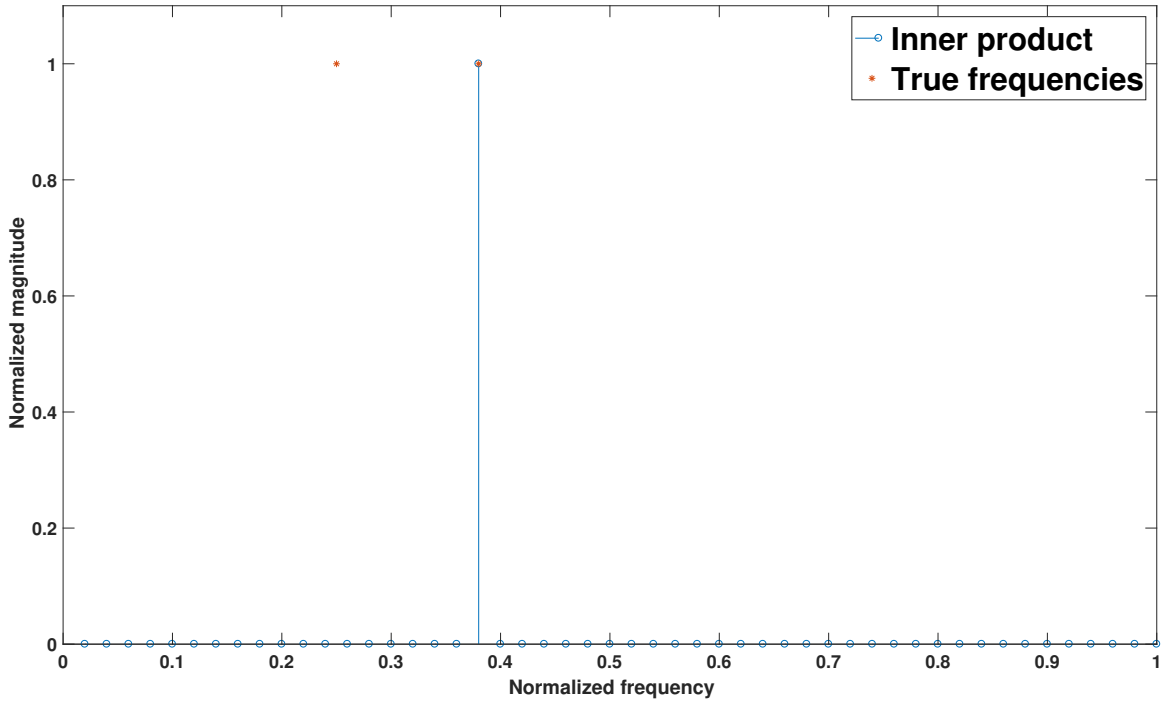

Figure 3: Inner product of the dictionary (50 elements) and the signal, N = 100, along with the true frequencies. The coarse grid causes the inability to recover one of the signal components.

is formed not as a single frequency but instead over an integrated range of the parameter space. In the one-dimensional undamped case, for two adjacent grid points f_a and f_b, this will be

$$\begin{cases} 1, & \text{for } f_a \leq f \leq f_b \\ 0, & \text{elsewhere} \end{cases} \quad \overset{\mathcal{F}^{-1}}{\rightarrow} \quad \frac{e^{2i\pi f_b t} - e^{2i\pi f_a t}}{2i\pi t}. \tag{30}$$

In a multidimensional setting, we can imagine dividing the parameter space into P D-dimensional hypercubes. Letting $\mathcal{H}_p$ denote one of these hypercubes, we then define the wideband dictionary as

$$\mathbf{a}_p^{\Omega} = \left[\begin{array}{ccc} \psi(\boldsymbol{\tau}_1, \boldsymbol{\beta}_p, \mathcal{H}_p) & \dots & \psi(\boldsymbol{\tau}_N, \boldsymbol{\beta}_p, \mathcal{H}_p) \end{array} \right]^T, \tag{31}$$

where

$$\psi(\boldsymbol{\tau}_n, \boldsymbol{\beta}, \mathcal{H}) = \int_{\mathcal{H}} \prod_{d=1}^{D} \xi(f^{(d)}, \beta^{(d)})^{t_n^{(d)}} df^{(1)} \dots df^{(D)}, \tag{32}$$

yielding the full dictionary

$$\mathbf{A} = \left[\begin{array}{ccc} \mathbf{a}_1^{\Omega} & \dots & \mathbf{a}_P^{\Omega} \end{array} \right]. \tag{33}$$

To demonstrate the IWDE, we use the same signal as in figure 3, using frequencies $f_1 = 0.38$, $f_2 = 0.25$. The inner product of the signal and a 25 element wideband dictionary is displayed in figure 4. Note that despite the dictionary being half the size compared to figure 3, the more evenly spread power of the elements allow for successful recovery of both the frequency components. The resolution of the estimate can then be improved by zooming in on and place a finer grid around the found peaks. The difference in distribution of power between the narrowband element and the wideband element is further illustrated in figure 5. This shows the larger, but significantly thinner peak of the narrowband dictionary, compared to the weaker, but more spread out wideband peak.

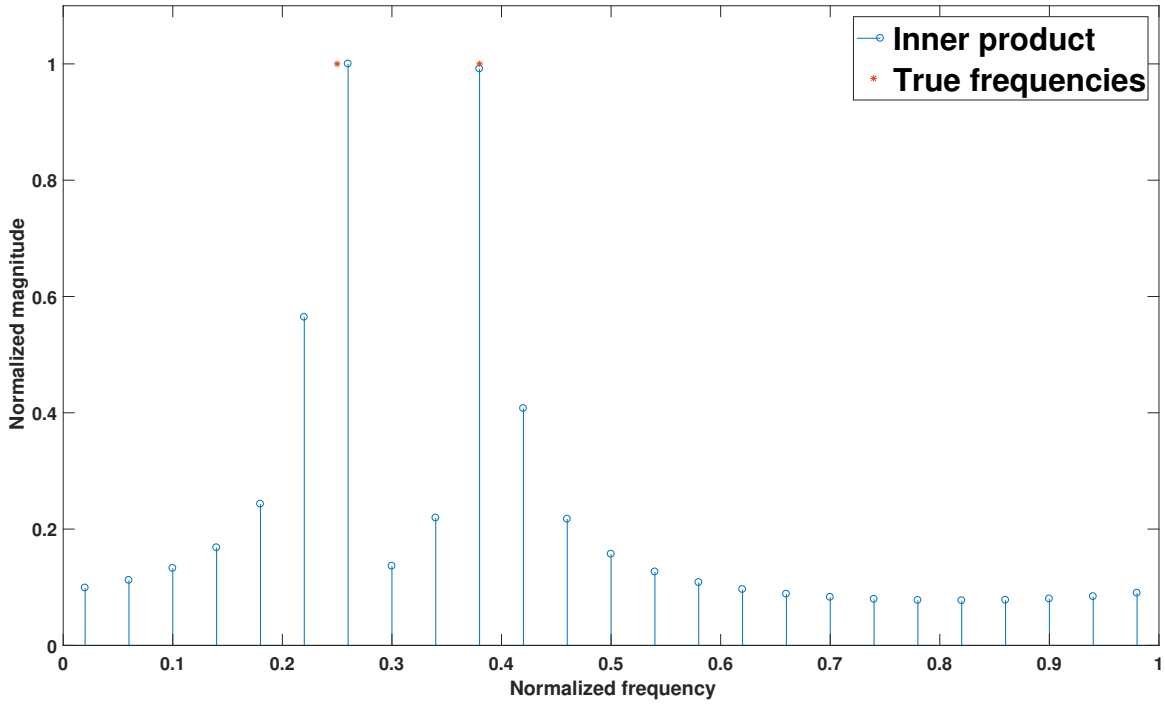

Figure 4: Inner product of the wideband dictionary (25 elements) and the signal, N = 100, along with the true frequencies. Despite the coarse grid, both signal components are recovered.

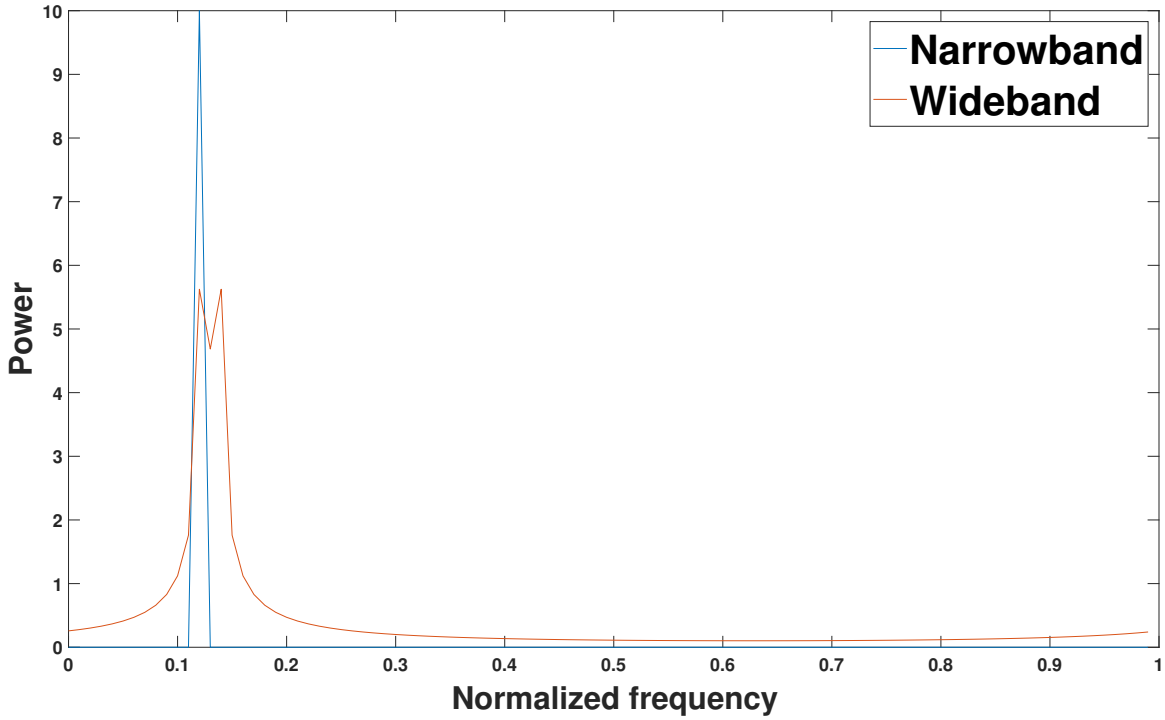

Figure 5: Absolute value of the Fourier transform of a narrowband and wideband dictionary element respectively of an $N = 100$ sample long signal and dictionary sizes of 50. The wideband element has less power at it's peak, but in return it covers more of the frequency space.

1.6 Computational complexity

The benefit of using an IWDE dictionary becomes evident when studying the computational complexity of solving the LASSO via ADMM, as done in [18]. The computationally most demanding steps is computing $\mathbf{A}^H\mathbf{A}$, creating the inverse and computing $\mathbf{A}^H\mathbf{y}$, all in (24). Fortunately, this can be done once before the iterations, commonly by doing a QR-factorization of $\mathbf{A}$. Assuming that $\mathbf{A}$ is of size $N \times P$, the cost of these operations depends on the relative size of N and P. If $P < N$, computing $\mathbf{A}^H\mathbf{A}$, forming the inverse and computing $\mathbf{A}^H\mathbf{y}$ is roughly $P^3+(N+1)P^2+NP$, whereas it's approximately $N^3+3PN^2+PN+P^2$ if $N < P$. Therefore, comparing a oversampled, one-stage, narrowband dictionary with $P > N$, with a wideband dictionary with $N = P$ (to simplify calculations) one can save $N^3+3PN^2+PN+P^2-2(N^3+N^2)$ computations in the initial step. Computations that can be used in the zooming procedure. Zooming can of course be performed using a narrowband dictionary, but to avoid missing off-grid components the size of the dictionary would have to be at least N in the initial step.

2 Sampling scheme optimization

2.1 Optimal sampling scheme

As previously mentioned, uniform sampling quickly becomes practically infeasible for multidimensional data. This yields the obvious question: how should non-uniform sampling be performed? This question is addressed in [11] where the aim is to minimize the Cramér-Rao Lower Bound (CRLB) of the parameters given the number of samples to be used. Among different bounds for the variance of an estimator, the CRLB is the bound most commonly used. The CRLB determines a lower bound on the optimal performance of an unbiased estimator. The CRLB is defined via the Fisher Information Matrix (FIM); which is, for a D-dimensional sampling time $\boldsymbol{\tau}_n$ and a set of parameters $\boldsymbol{\theta}$, defined as

$$\mathbf{F}(\boldsymbol{\tau}_n;\boldsymbol{\theta}) = \mathrm{E}\{\nabla_{\boldsymbol{\theta}}\ln(p(y(\boldsymbol{\tau}_n);\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}^H\ln(p(y(\boldsymbol{\tau}_n);\boldsymbol{\theta}))\}, \tag{34}$$

where $p(y(\boldsymbol{\tau};\boldsymbol{\theta}))$ denotes the probability density function. The CRLB of a parameter $\theta \in \boldsymbol{\theta}$ is then the corresponding element on the diagonal of the inverse of FIM. The sum of the CRLB of the parameters is consequently the trace of the inverse of FIM. Assuming that the samples are independent, the corresponding FIM for a set of samples indices $\boldsymbol{\Omega}$ is

$$\sum_{n\in\boldsymbol{\Omega}} \mathbf{F}(\boldsymbol{\tau}_n;\boldsymbol{\theta}). \tag{35}$$

Letting $\mathbf{w}$ denote the vector of possible sampling points and $w_n = 1$ meaning that the n:th sample is used and $w_n = 0$ that it is not used, the problem can be reformulated as

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathrm{tr}\left(\left(\sum_{n=1}^{N} w_n\mathbf{F}(\boldsymbol{\tau}_n;\boldsymbol{\theta})\right)^{-1}\right) \\ \text{subject to} \quad & ||\mathbf{w}||_1 \leq \gamma \\ & w_n \in \{0,1\}, n = 1,...,N, \end{aligned} \tag{36}$$

where $\gamma > 0$ is an upper bound on the norm of the number of samples chosen to be activated. Once again we are facing a combinatorial problem, why a convex relaxation, similar to the one used before, is used, allowing w_n to vary in the interval $[0,1]$. Further, since all entries in $\mathbf{w}$ are non-negative, we can replace $||\mathbf{w}||_1$ with $\mathbf{1}^T\mathbf{w}$ where $\mathbf{1}$ is a vector of ones of appropriate size, yielding

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathrm{tr}\left(\left(\sum_{n=1}^{N} w_n\mathbf{F}(\boldsymbol{\tau}_n;\boldsymbol{\theta})\right)^{-1}\right) \\ \text{subject to} \quad & \mathbf{1}^T\mathbf{w} \leq \gamma \\ & w_n \in [0,1], n = 1,...,N, \end{aligned} \tag{37}$$

Given a solution $\hat{\mathbf{w}}$ to (37) the Fisher information matrix then becomes

$$\mathcal{I}(\hat{\mathbf{w}};\boldsymbol{\theta}) = \sum_{j\in\boldsymbol{\Omega}} \mathbf{F}(\mathbf{t}_j;\boldsymbol{\Omega}), \quad \boldsymbol{\Omega} = \{j : \hat{w}_j > \xi\}, \tag{38}$$

where $\xi \geq 0$ is a threshold determining whether an element w_j should be rounded to 1 or 0. This allows for a minimization of the sum of the CRLB's of the parameters. However, since the parameters may be estimated with varying degree of precision, and the estimation of some parameters might be much more interesting than others, one would like to be able to weight the importance of the different parameters, to avoid a lopsided sum of the CRLB's. To allow for a parameter to be of arbitrary importance, including none, it is noted in [11] that

$$\mathrm{tr}\big(\mathbf{B}^{-1}\big) = \sum_{p=1}^{P} \mathbf{e}_p^T \mathbf{B}^{-1} \mathbf{e}_p, \tag{39}$$

with $\mathbf{e}_p$ denoting a vector with zeroes at every position except the p:th, where there is a one. Further, for a positive definite matrix[3] $\mathbf{B}$, scalar μ and vector $\mathbf{a}$

$$\mu - \mathbf{a}^T\mathbf{B}^{-1}\mathbf{a} \geq 0 \iff \begin{bmatrix} \mathbf{B} & \mathbf{a} \\ \mathbf{a}^T & \mu \end{bmatrix} \succeq 0 \tag{40}$$

where $\mathbf{C} \succeq \mathbf{0}$ means that $\mathbf{C}$ is positive semi-definite. From this it follows that

$$\underset{\mathbf{B}\succ\mathbf{0}}{\text{minimize}} \quad \mathbf{a}^T\mathbf{B}^{-1}\mathbf{a} \tag{41}$$

and

$$\begin{aligned} &\underset{\mu,\mathbf{B}\succ\mathbf{0}}{\text{minimize}} && \mu \\ &\text{Subject to} && \begin{bmatrix} \mathbf{B} & \mathbf{a} \\ \mathbf{a}^T & \mu \end{bmatrix} \succeq \mathbf{0} \end{aligned} \tag{42}$$

is minimized by the same matrix $\mathbf{B}$, $\mathbf{B} \succ \mathbf{0}$ meaning that $\mathbf{B}$ is positive definite. Thanks to this observation, and by letting ψ_p denote the weight of the p:th parameter, we may, as noted in [11], now reformulate (37) as a semi-definite program (SDP)

$$\begin{aligned} &\underset{\boldsymbol{\mu},\boldsymbol{w}}{\text{minimize}} && \sum_{p=1}^{P} \psi_p \mu_p \\ &\text{subject to} && \begin{bmatrix} \sum_{n=1}^{N} w_n \mathbf{F}(\mathbf{t}_n;\boldsymbol{\theta}) & \mathbf{e}_p \\ \mathbf{e}_p^T & \mu_p \end{bmatrix} \succeq \mathbf{0} \ \forall p \\ & && \sum_{n=1}^{N} w_n \mathbf{F}(\mathbf{t}_n;\boldsymbol{\theta}) \succ \mathbf{0} \\ & && \mathbf{1}\mathbf{w} \leq \gamma, \quad w_n \in [0,1], \ \forall\, n, \end{aligned} \tag{43}$$

which can be solved using for instance Matlab's CVX-toolbox. This framework also allow for incorporation of a priori-knowledge that might be available regarding some parameters. Another advantage is that it is possible to set a limit on the CRLB of a specific parameter. Once again using NMR spectroscopy as an example, the frequencies might be fairly well known, often it is the damping parameters that are of real interest. Letting Θ denote the set of possible parameters and

[3] $z^T B z > 0$ for any non-zero, real vector z.

λ_p the CRLB-bound on the p:th parameter,

$$\begin{aligned}
\underset{\boldsymbol{\mu},\boldsymbol{w}}{\text{minimize}} \quad & \sum_{p=1}^{P} \psi_p \mu_p \\
\text{subject to} \quad & \begin{bmatrix} \sum_{n=1}^{N} w_n \mathbf{F}(\mathbf{t}_n;\boldsymbol{\theta}) & \mathbf{e}_p \\ \mathbf{e}_p^T & \mu_p \end{bmatrix} \succeq \mathbf{0} \ \forall\, p, \ \forall\, \boldsymbol{\theta} \in \Theta \\
& \sum_{n=1}^{N} w_n \mathbf{F}(\mathbf{t}_n;\boldsymbol{\theta}) \succ \mathbf{0} \\
& \mathbf{1w} \leq \gamma, \quad w_n \in [0,1], \ \forall\, n \\
& \mu_p \leq \lambda_p, \ \forall\, p.
\end{aligned} \tag{44}$$

The solution μ_p is now the worst-case CRLB for the p:th parameter, assuming that all parameters are within the set Θ. The solution will therefore be

$$\mu_p = \arg \max_{\boldsymbol{\theta}\in\Theta} \ \mathbf{e}_p^T \mathcal{I}(\hat{\mathbf{w}};\boldsymbol{\theta})^{-1} \mathbf{e}_p. \tag{45}$$

2.2 Selection of β

Naturally, true values of the parameters is not known prior to the sampling scheme optimization. However, prior information in form of an interval regarding a parameter, might be available. The authors of [11] have proposed that if there's uncertainty regarding the damping parameter, it's better to overestimate the parameter than to underestimate it, choosing the largest value in the interval, when optimizing the sampling scheme. This way one ensures the best worst-case scenario CRLB. That this is a reasonable suggestion is demonstrated in figure 6. Here, we see two one-dimensional one-mode signals, identical with respect to frequency, amplitude, phase and noise, but with different damping. The top picture shows a signal with $\beta = 0.2$, whereas for the signal in the bottom picture $\beta = 0.05$. Out of the 50 possible samples, 13 have been chosen according to the method previously described. The optimal sampling schemes are shown in the pictures as well. Although the performance would be suboptimal, employing the sampling scheme from the top picture to the signal in the bottom picture, one would still get a descent estimate. Doing the opposite, using the sampling scheme from the bottom picture to the signal in the top picture, would be equivalent of throwing away the samples used between $t = 30$ and $t = 35$ since the signal in the top picture have decayed completely at that point.

To address this more rigorously, assume that the true value of the parameter β is the to us unknown $\tilde{\beta}$. We know that $\tilde{\beta} \in B$. in [11], the optimization was done by finely gridding B. Here, I formulate the proposition made by the authors of [11], that optimization of the sampling scheme should be done with respect to the largest value in B, β_{max}, to ensure the best worst-case scenario sampling scheme.

Conjecture 1 (Swärd-Elvander conjecture)**.** *Optimization of the sampling scheme should be done with the largest β in the uncertainty interval B.*

To be consistent with previous literature on the subject, a slightly different notation will be used throughout this section. The complex amplitude will be replaced by a real-valued α, equal to the absolute value of the previous α, and a phase shift $e^{i\phi}$, $\phi \in [0, 2\pi]$. Since any complex number $z = a + bi$ can be written as a $z = re^{i\phi}$, where $r = |z|$ and $\phi = \arg(z)$, theses forms of notation are equivalent. Further, we concentrate on the simplest thinkable case, considering only a single mode in one dimension, yielding

$$y_n = \alpha \ \exp\{2i\pi f t_n - \beta_k t_n + i\phi\} + \epsilon_n. \tag{46}$$

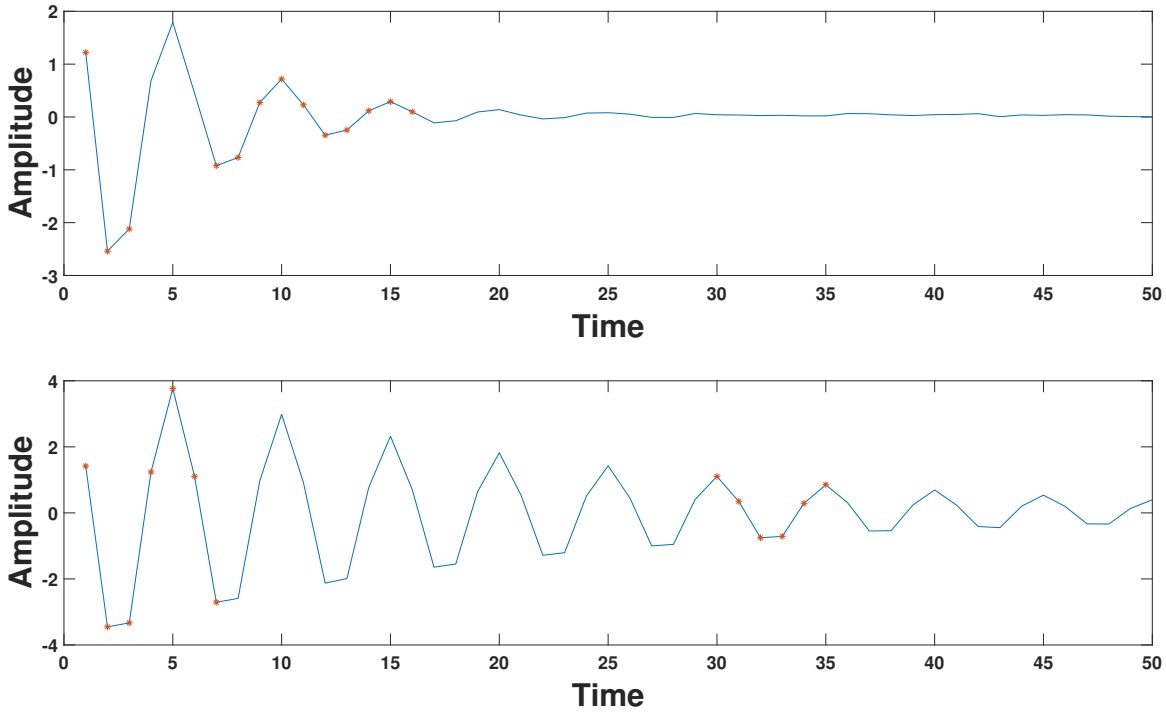

Figure 6: Comparison of the optimized sampling schemes (red dots) of two signals with different rate of decay. The faster decay of the signal in the top picture necessitates earlier sampling.

From [11], the Fisher Information Matrix (FIM) corresponding to the sampling time t_n, with a slight modification of the order of the parameters, is formulated as

$$\mathbf{F}(t_n;\boldsymbol{\theta}) = \frac{2}{\sigma^2}\begin{pmatrix} \mathbf{F}_{1,1}(t_n) & \cdots & \mathbf{F}_{1,K}(t_n) & 0 \\ \vdots & \cdots & \vdots & \vdots \\ \mathbf{F}_{K,1}(t_n) & \cdots & \mathbf{F}_{K,K}(t_n) & 0 \\ 0 & \cdots & 0 & 1/\sigma^2 \end{pmatrix} \tag{47}$$

where

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T \cdots \boldsymbol{\theta}_K^T], \tag{48}$$
$$\boldsymbol{\theta}_k = [\alpha_k \;\; \beta_k \;\; f_k \;\; \phi_k] \tag{49}$$

and

$$\mathbf{F}_{k,\ell}(t_n) = \begin{pmatrix} \frac{c_{k,\ell}(t_n)}{\alpha_k\alpha_\ell} & \frac{-t_n c_{k,\ell}(t_n)}{\alpha_k} & \frac{2\pi t_n s_{k,\ell}(t_n)}{\alpha_k} & \frac{s_{k,\ell}(t_n)}{\alpha_k} \\ \frac{-t_n c_{k,\ell}(t_n)}{\alpha_\ell} & t_n^2 c_{k,\ell}(t_n) & -2\pi t_n^2 s_{k,\ell}(t_n) & -t_n s_{k,\ell}(t_n) \\ \frac{-2\pi t_n s_{k,\ell}(t_n)}{\alpha_\ell} & 2\pi t_n^2 s_{k,\ell}(t_n) & (2\pi t_n)^2 c_{k,\ell}(t_n) & 2\pi t_n c_{k,\ell}(t_n) \\ \frac{-s_{k,\ell}(t_n)}{\alpha_\ell} & t_n s_{k,\ell}(t_n) & 2\pi c_{k,\ell}(t_n) & c_{k,\ell}(t_n) \end{pmatrix} \tag{50}$$

with

$$c_{k,\ell}(t_n) = \alpha_k\alpha_\ell e^{-(\beta_k+\beta_\ell)t_n}\cos(2\pi(f_k-f_\ell)t_n + (\phi_k-\phi_\ell)), \tag{51}$$
$$s_{k,\ell}(t_n) = \alpha_k\alpha_\ell e^{-(\beta_k+\beta_\ell)t_n}\sin(2\pi(f_k-f_\ell)t_n + (\phi_k-\phi_\ell)). \tag{52}$$

Given that there is only one mode, $k=\ell$, making the cosine-term in (51) equal to 1 and the sine-term

in (52) equal to 0. This simplifies the matrix in (50) quite a bit, yielding

$$\begin{pmatrix} e^{-2\beta t_n} & -t_n\alpha e^{-2\beta t_n} & 0 & 0 \\ -t_n\alpha e^{-2\beta t_n} & t_n^2\alpha^2 e^{-2\beta t_n} & 0 & 0 \\ 0 & 0 & (2\pi t_n\alpha)^2 e^{-2\beta t_n} & 2\pi t_n\alpha^2 e^{-2\beta t_n} \\ 0 & 0 & 2\pi t_n\alpha^2 e^{-2\beta t_n} & \alpha^2 e^{-2\beta t_n} \end{pmatrix}. \tag{53}$$

The FIM for the sum over all N sampling points will then be

$$\begin{pmatrix} \sum_n e^{-2\beta t_n} & -\alpha\sum_n t_n e^{-2\beta t_n} & 0 & 0 \\ -\alpha\sum_n t_n e^{-2\beta t_n} & \alpha^2\sum_n t_n^2 e^{-2\beta t_n} & 0 & 0 \\ 0 & 0 & (2\pi\alpha)^2\sum_n t_n^2 e^{-2\beta t_n} & 2\pi\alpha^2\sum_n t_n e^{-2\beta t_n} \\ 0 & 0 & 2\pi\alpha^2\sum_n t_n e^{-2\beta t_n} & \alpha^2\sum_n e^{-2\beta t_n} \end{pmatrix}. \tag{54}$$

The block nature of this matrix makes it fairly easy to invert and then compute the CRLB for the parameter β, being the second diagonal element of the inverse of (54),

$$\mathrm{CRLB}(\beta) = \frac{\sum_j e^{-2\beta t_j}}{\sigma^2\alpha^2\sum_n\sum_m (t_n^2 - t_n t_m)e^{-2\beta(t_n+t_n)}} \tag{55}$$

Proving that the β one should optimize with respect to, to yield the best worst-case CRLB, is the largest $\beta \in B$, is then equivalent to proving that CRLB(β_2) - CRLB(β_1) ≥ 0 under the assumption that $\beta_2 \geq \beta_1$. To prove this analytically has proven to be challenging. Using numerical simulations to provide support for the conjecture, 10^5 Monte Carlo-simulations was performed. For each simulation of CRLB(β_2) - CRLB(β_1), the number of samples was selected randomly on $[2, 1000]$, the time points $t_1 \leq ... \leq t_N$ was randomly selected on [0,100], and $\beta_2 \geq \beta_1$ was selected randomly on $[0, 5]$. The inequality held for every simulation, supporting that the conjecture may be plausible.

3 WSEMA

3.1 WSEMA algorithm

The initial step of the wideband SEMA (WSEMA) algorithm is to divide the frequency space into P hypercubes. Unless some a priori information regarding the location of the frequency peaks are available, it is advisable to let these hypercubes cover the entire frequency space, i.e., $\cup_{p=1}^{P}\mathcal{H}_p = [0,1)^D = \mathcal{D}$, $\mathcal{H}_p \cap \mathcal{H}_q = \emptyset$, $p \neq q$. The dictionary corresponding to these hypercubes is then formed according to (33). Because of computational reasons previously discussed, one would like to keep the size of the dictionary as small as possible. Therefore the dampings are set to zero to avoid having to grid in additional dimensions. Using this dictionary to solve (18) will yield a solution $\mathbf{x}$, likely containing both zero and non-zero elements. A zero-element of $\mathbf{x}$ corresponds to a hypercube where no power is found, why these areas can be removed from further consideration. Similarly, the non-zero elements corresponds to regions where some power has been detected. These hypercubes are subsequently divided into smaller hypercubes, the corresponding dictionary is created according to (33) and (18) is then solved anew. This way, regions of the frequency space containing no power is systematically disregarded while the resolution of the parts in fact containing power gets progressively better. This procedure can be iterated until desired resolution is attained. Worth noting is that, because of the dampings, one will get a less sparse solution, as the power of the damped modes will leak out to adjacent frequencies.

At this point, the frequency components should be fairly well located. Assume that $\hat{K}$ modes are found. To refine the frequency estimates and form estimates of the dampings of these $\hat{K}$ modes, non-linear least squares estimation is then performed in an iterative manner presented in [12]. There, the problem of minimizing the residual with respect to $\hat{K}$ modes at once is relaxed to instead forming frequency and damping estimates one mode at a time. This is done by first forming the model

residual

$$\mathbf{R} = \mathbf{y} - \sum_{k=1}^{\hat{K}} \hat{\alpha}_k \mathbf{a}_k \tag{56}$$

and then systematically add back one mode at a time

$$\mathbf{y} = \mathbf{R} + \hat{\alpha}_k \mathbf{a}_k, \tag{57}$$

and estimate its parameters using NLS

$$\min_{\theta} \sum_{n=1}^{N} (y_n - g(\theta_n))^2, \tag{58}$$

where g is a non-linear function of the parameter θ. In this case, $g(\theta) = \alpha e^{2i\pi f t_n - \beta t_n}$ is a function of several parameters and θ will by turns refer to f and β, while the other remain constant. Since, as previously established, the frequencies are fairly well known at this point and the damping parameters are reasonably small, the minimization in (58) is done approximately by finely gridding the vicinity of the parameter in question and minimizing the model residual. After the parameters of the mode is estimated, the mode is removed, the next mode is added to the residual, its parameters are then estimated, and so forth. This NLS-minimization procedure is repeated a desired number of times. The proposed WSEMA-algorithm is summarized in algorithm 1.

The WSEMA estimator possesses two important features. It requires no a priori-knowledge of the number of modes present in the signal, which is commonly not available [22]. This is advantageous since this is usually unknown in applications such as NMR spectroscopy [12]. Instead, one have to chose the hyper parameter λ, indirectly determining the number of peaks, a topic discussed in [23]. The second feature is that of being compatible with non-uniform sampling. As previously established, for multidimensional problems, uniform sampling is practically infeasible. Some previous estimators, like PUMA, requires uniform sampling. Others, such as SEMA, requires that the sampling scheme in one dimension is consistent throughout the other dimensions. In contrast, WSEMA works with an arbitrary sampling scheme, making it compatible with the ideas like the ones discussed in section 2.

3.2 Spurious peaks

A commonly occurring problem is that of spurious peaks. When a frequency peak ends on the grid, it triggers two non-zero elements in $\mathbf{x}$ in the solution of (18), i.e., it is interpreted as two different peaks. Being able to deal with this difficulty is an important component of an estimator since misjudgement of the number of modes present in the signal will impair the estimations of the dampings and amplitudes. What had been previously noted was that for such spurious peaks, the amplitude estimates stemming from the NLS-estimation of the frequencies and the dampings respectively, are not consistent. This can be used when trying to identify spurious peaks. For simulated data, the difference was clear enough to spot with a very crude criterion. If the maximum absolute difference between the amplitude estimate relating to the frequency estimate and the amplitude estimate relating to the damping estimate was larger than the quotient of the range of the absolute value of either amplitude estimate and the mean of difference, then it was concluded that there were spurious peaks present in the estimate. Subsequently, closely spaced peaks were merged into one and the NLS-procedure was repeated.

This was as a rule good enough for simulated data to generate satisfactory results, causing at most one extra iteration of the NLS-procedure. For real NMR data, being significantly more complicated than the signals previously simulated, this simple heuristic yielded sub par results, often prompting several time consuming iterations of the NLS-step. On account of this, other methods were tested. Amongst them the removal of closely located peaks before the NLS-estimation. The method showed some promise, but it has two obvious drawbacks. Firstly, it leaves the user with another hyper parameter choice, the criterion of two peaks being too close prior to the NLS-estimation.

Algorithm 1 WSEMA

1: Chose initial number of hypercubes P and create hypercubes $\mathcal{H}_p$, $p = 1, ..., P$
2: **repeat**
3: Create dictionary $\mathbf{A}$ according to (33)
4: Solve (18) using ADMM
5: Determine active components $\mathcal{I} = \{p : |\mathbf{x}_p| > 0 \}$
6: Construct new hypercubes subdividing $\mathcal{H}_p$, $p \in \mathcal{I}$ and discard hypercubes $\mathcal{H}_p$, $p \notin \mathcal{I}$
7: **until** Desired frequency resolution is attained
8: Determine active components $\mathcal{I} = \{p : |\mathbf{x}_p| > 0 \}$
9: Create dictionary $\mathbf{A}$ according to (33) with elements corresponding to $\mathbf{x}_p \in \mathcal{I}$
10: Solve (18) using ADMM
11: **for** $i = 1, ..., \text{iter}_{\max}$ **do**
12: Compute residual $\mathbf{R}$ according to (56)
13: **for** $k = 1, ..., \hat{K}$ **do**
14: Add mode to the residual according to (57).
15: Estimate parameters according to (58)
16: Remove the estimated mode from the residua
17: **end for**
18: **end for**
19: Determine active components $\mathcal{I} = \{p : |\mathbf{x}_p| > 0 \}$
20: Create dictionary $\mathbf{A}$ according to (33) with elements corresponding to $\mathbf{x}_p \in \mathcal{I}$
21: Solve (18) using ADMM

Secondly, closely spaced modes with different damping is characteristic in several spectroscopic applications [24] and precipitately merging them would prevent detecting closely spaced modes. An obvious alternative is ignoring the step of removing potentially spurious peaks and directly determine the amplitudes by taking the pseudo-inverse of the parts of the dictionary corresponding the the non-zero elements. This yields solutions very appealing to the naked eye, when comparing the WSEMA-estimate to the periodogram. When reviewing it closer however, the estimate was found to contain several closely spaced very large components with cancelling phases. Therefore, to further promote sparsity, additional steps was added to the process. After the zooming procedure is finished and desired frequency resolution is attained, the dictionary $\mathbf{A}$ was created according to (33), with the atoms of $\mathbf{A}$ corresponding to the non-zero elements of $\mathbf{x}$, and (18) is solved anew. This step is also performed after the NLS-estimation. These additions correspond to row 8-10 and 19-21 in algorithm 1 respectively. Together, these two late added steps helped provide a solution much sparser, but visibly just as good. The difficult balancing act of allowing closely spaced peaks and in the meanwhile provide a sparse solution will be discussed further in section 5. Since these extra steps were a late addition to the algorithm, they are employed only to the real data described in section 4.4.

4 Results

4.1 Multimodal recovery

To demonstrate the proposed algorithms ability to recover the correct frequencies of a multi-modal signal, a nine-mode signal was simulated. The frequencies for each dimension were randomly selected on the interval $[1/N_d, 1 - 1/N_d]$, where N_d denotes the number of samples in dimension $d = 1, 2$. The damping parameters $\beta_k^{(d)}$ where randomly picked on the interval $[0.015, 0.02]$, $k = 1, ..., 9$ and $d = 1, 2$. Further, checks were included to ensure that no two frequencies were closer than $1/N_d$. The periodogram estimate of a typical realisation can be found in figure 7. Then, for different settings of the number of samples per dimension and hypercubes per sample, the recovery rate was analysed. The frequencies were considered to be recovered if the estimated number of modes where correct and

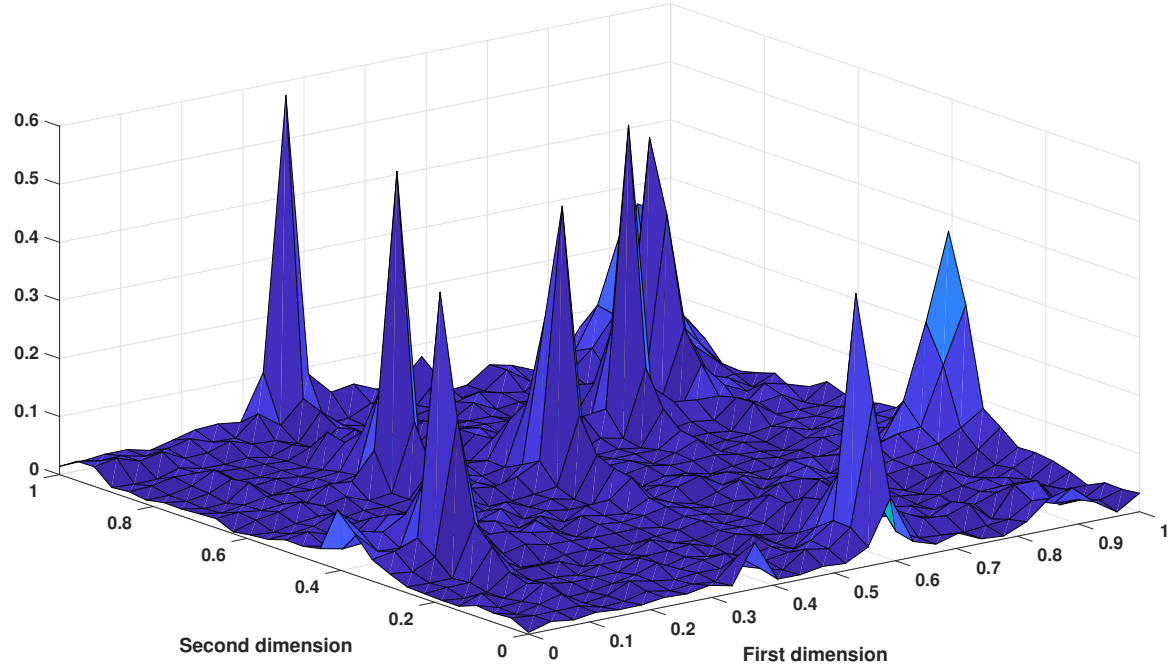

Figure 7: Periodogram estimate of a typical 9-mode signal used in the simulations in section 4.1.

each frequency estimate was within $1/(2N_d)$ of the ground truth. The results are computed using 150 Monte Carlo-simulations, each corrupted by circularly symmetric white Gaussian noise with a signal-to-noise ratio (SNR) of 15, defined as

$$\text{SNR} = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right), \tag{59}$$

where σ_x^2 denotes the variance of the signal, estimated as $\sigma_x^2 = \frac{1}{N-1}\sum_{n=1}^{N}|x_n - \mu|^2$ where N is the total number of samples in the signal and μ is the mean, and σ_e^2 denotes the variance of the noise. The results are shown in figure 8. As can be seen, the algorithm quickly achieves near perfect recovery of the frequency components.

4.2 Variance

To demonstrate the statistical performance of the proposed algorithm, a 2-dimensional signal was simulated. The signal consisted of two modes with unit amplitude and random phase. The frequencies of the components was $f_1^{(1)} = 0.2$, $f_1^{(2)} = 0.6$, $f_2^{(1)} = 0.7$, and $f_2^{(2)} = 0.3$. To make sure that the frequency components were off-grid, a perturbation, uniformly random selected on $[0, 0.04)$ was added to each frequency component. The damping parameter was randomly selected on the interval $[0.014, 0.022)$ for $\beta_k^{(d)}$, $d = 1, 2$, $k = 1, 2$. The signal was uniformly sampled on the 2-dimensional interval $[0, \sqrt{N}-1] \times [0, \sqrt{N}-1]$, where $N = 784$ is the total number of samples. Then, for a varying level of white circularly symmetric Gaussian noise, the proposed algorithms ability to estimate the frequency and damping was analysed. For each noise level, the root mean squared errors (RMSE),

$$\text{RMSE} = \sqrt{\sum_{m=1}^{M}\sum_{k=1}^{K}\left(\theta_{m,k} - \hat{\theta}_{m,k}\right)^2}, \tag{60}$$

where $\theta_{m,k}$ denotes the true parameter value, $\hat{\theta}_{m,k}$ the estimate, $M = 500$ the number of Monte Carlo-simulations and K denotes the number of modes, was calculated. For comparison, the RMSE

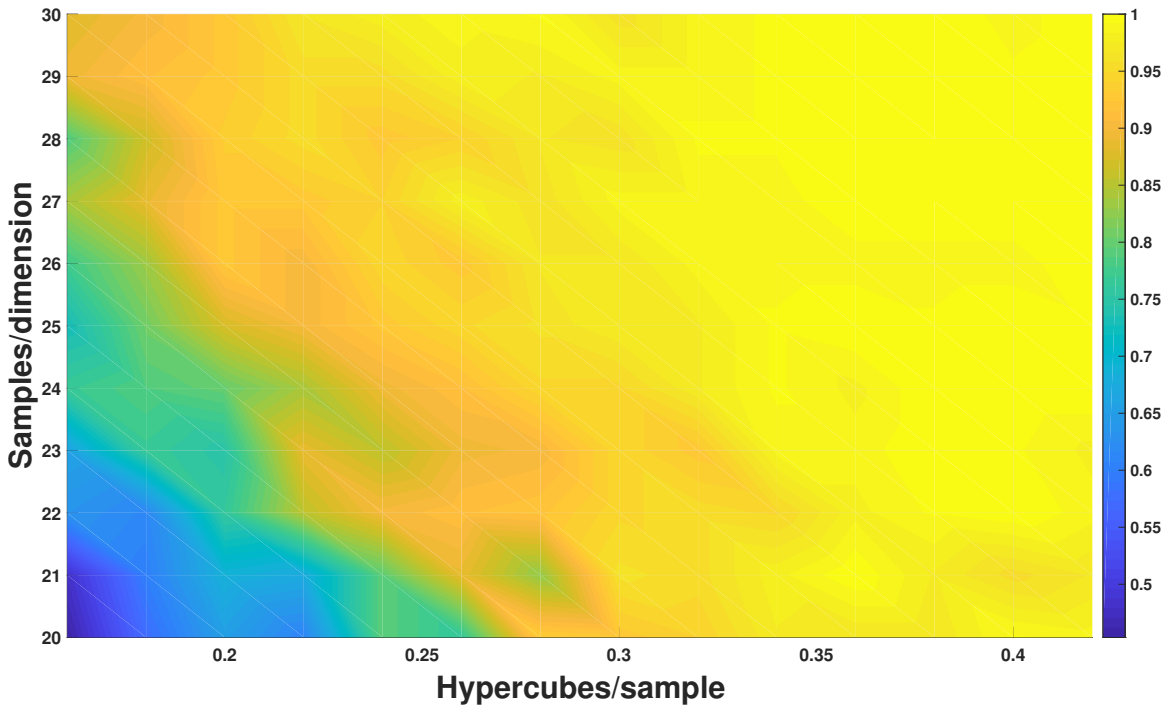

Figure 8: Recovery rate for the proposed algorithm as a function of number of samples and tightness of the grid. WSEMA achieves near-perfect recovery for a number as low as 0.3 hypercubes per sample.

was also computed for the PUMA estimator. Also in the plot is the CRLB. Because the frequencies and dampings are slightly altered for each simulation, the CRLB was computed for every Monte Carlo-simulation whereupon the average was computed,

$$\text{CRLB}_{avg} = \sqrt{\sum_{m=1}^{M}\sum_{k=1}^{K}(\text{CRLB}_{m,k})^2}. \tag{61}$$

To prevent outliers from spoiling the results, simulations where a method had failed to recover the correct frequencies was removed. A recovery was considered correct if the estimated frequency was within 0.05 of ground truth in each dimension. The number of samples kept is displayed in the figure 9. It should be stressed that only the PUMA estimator is given a priori-knowledge of the number of peaks. As shown in the figure, the WSEMA estimator reaches the CRLB.

4.3 Sampling scheme optimization

Up until this point all estimations by WSEMA have been done with a uniform sampling scheme. To demonstrate the algorithm's ability to work with a non-uniform sampling scheme we use simulated data to evaluate the performance. We use a simple two-mode signal with unit amplitude and random phase, $f_1^{(1)} = f_1^{(2)} = 0.6$, $f_2^{(1)} = 0.6$, $f_2^{(2)} = 0.4$ and $\beta_1^{(1)} = \beta_1^{(2)} = 0.04$, $\beta_2^{(1)} = \beta_2^{(2)} = 0.05$. The signal is disturbed by a white circular Gaussian noise with variance $\sigma^2 = 0.1$. The signal is first evaluated on a uniform 40×40-grid, shown in figure 10. Subsequently, the optimal sampling scheme was computed according to the procedure described in section 2. The optimization is done with perfect information, but solely with respect to the damping parameters. An example of such an optimized scheme, containing 400 samples, is shown in figure 11, where a 1 indicates that a possible sampling point is used, whereas a 0 indicates that it is not.

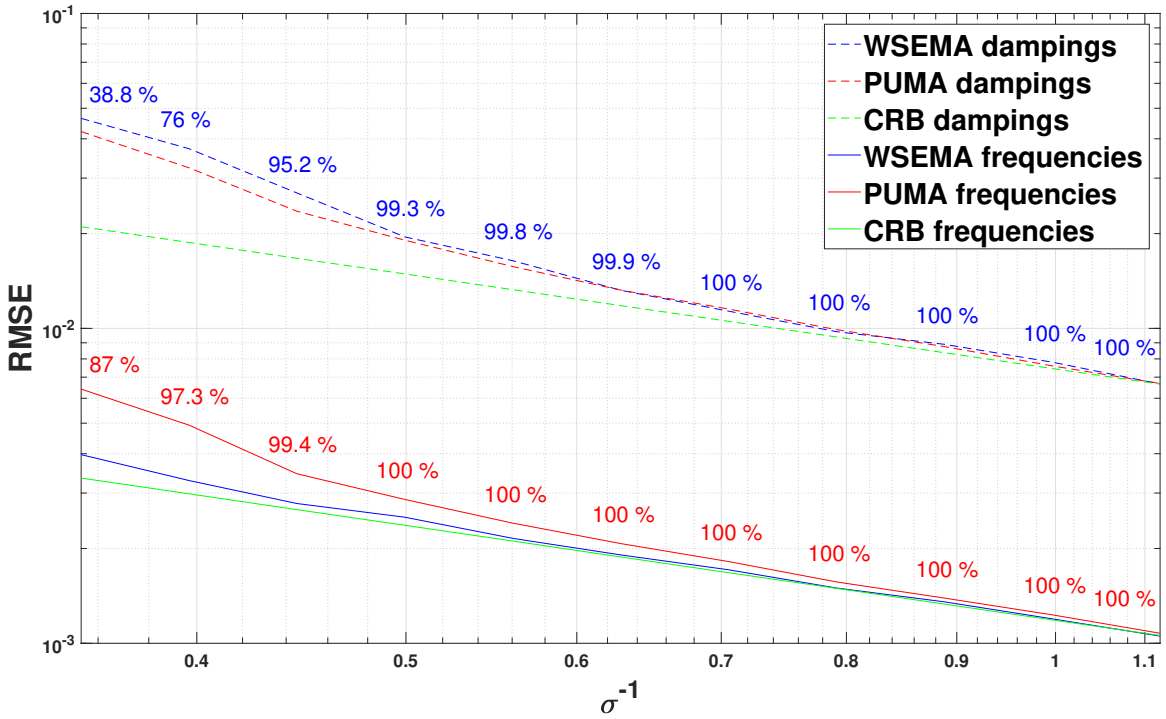

Figure 9: RMSE for WSEMA and PUMA respectively, as well as root CRB. The number of simulations kept for each noise level is displayed in blue for WSEMA and red for PUMA.

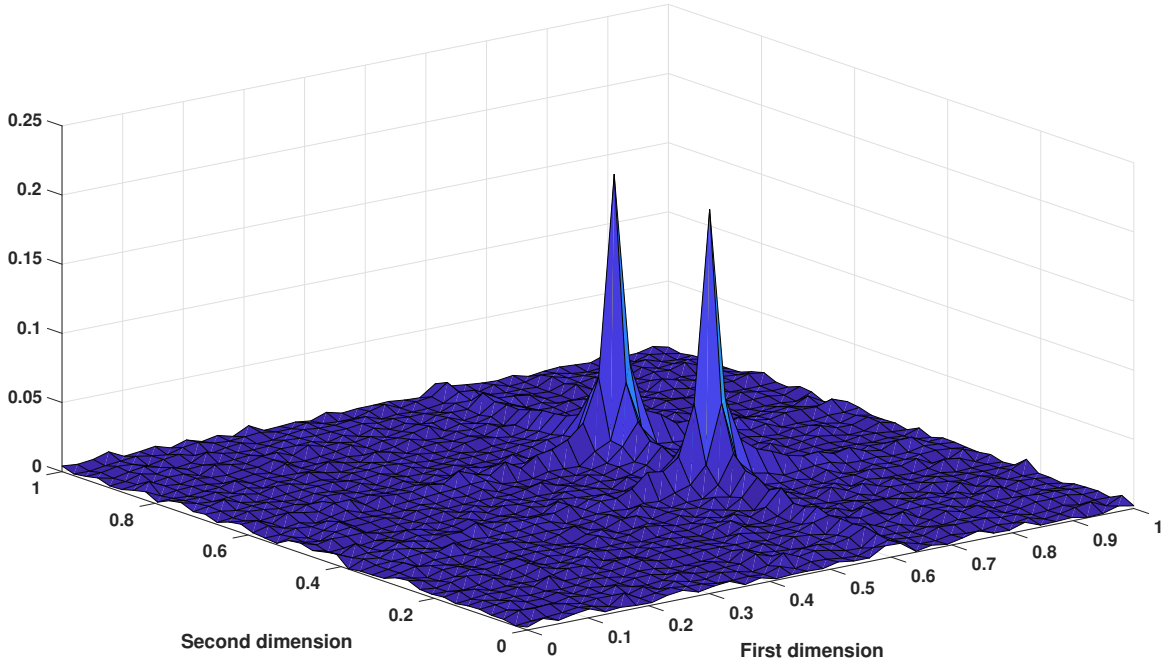

Figure 10: Periodogram of the signal described in section 4.3.

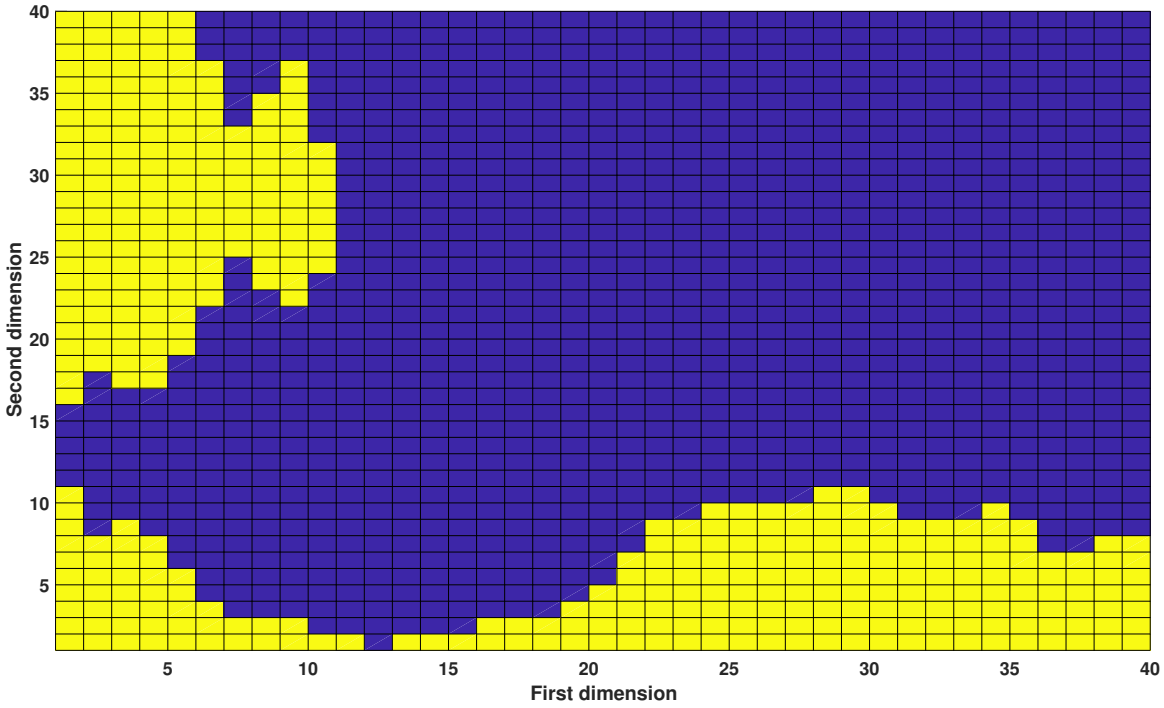

Figure 11: Optimal sampling scheme for the signal described in section 4.3. Samples used indicated by a one (yellow), unused samples by a zero (blue).

The WSEMA estimate of the reduced signal is done with 20 bands/dimension initially and 10 bands/dimension/mode in the one and only zoom step. The WSEMA estimate is displayed in figure 12. As can be seen, the recovery of the damped modes is successful. For comparison, in figure 13, we see the DFT formed for the same samples, where the interior have been padded with zeros to generate a regular sampling grid. This is the so called Lomb-Scargle periodogram and demonstrates the difficulty to handle a non-uniform grid.

To quantify the effect of the sampling scheme optimization, a simple two-mode signal was simulated. WSEMA estimated the parameters of the signal with optimal sampling scheme as well as with random sampling. The frequencies of the modes were randomly picked on $[1/(2N), 1 - 1/(2N)]$, however with the restriction that a peak was not allowed to be closer than $1/(2N)$, where $N = 30$ is the number of samples in each of the two dimensions. Then, out of the 900 possible sampling points, 512 were optimized, with perfect information, with respect to the damping parameters only. Subsequently, a white circularly symmetric Gaussian noise with variance of $\sigma^2 = 0.2$ was added and a random scheme was established. After that, WSEMA estimated the parameters of the model on the optimized and random sampling scheme respectively, and the relative error was determined, where the total relative RMSE (RRMSE) is defined as

$$\text{RRMSE} = \sqrt{\frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\frac{\theta_{m,k} - \hat{\theta}_{m,k}}{\theta_{m,k}} \right)^2}, \tag{62}$$

using the same notation as in equation (62). In total, 200 different signal parameters and optimal schemes were generated and for each of these, 10 different signals and random schemes was created. Similar to the procedure described in section 4.2, a simulation was disregarded if the recovery was considered unsuccessful, i.e. if any of the frequency parameters were not within $1/(2N)$ from the ground truth. The results are displayed in table 1.

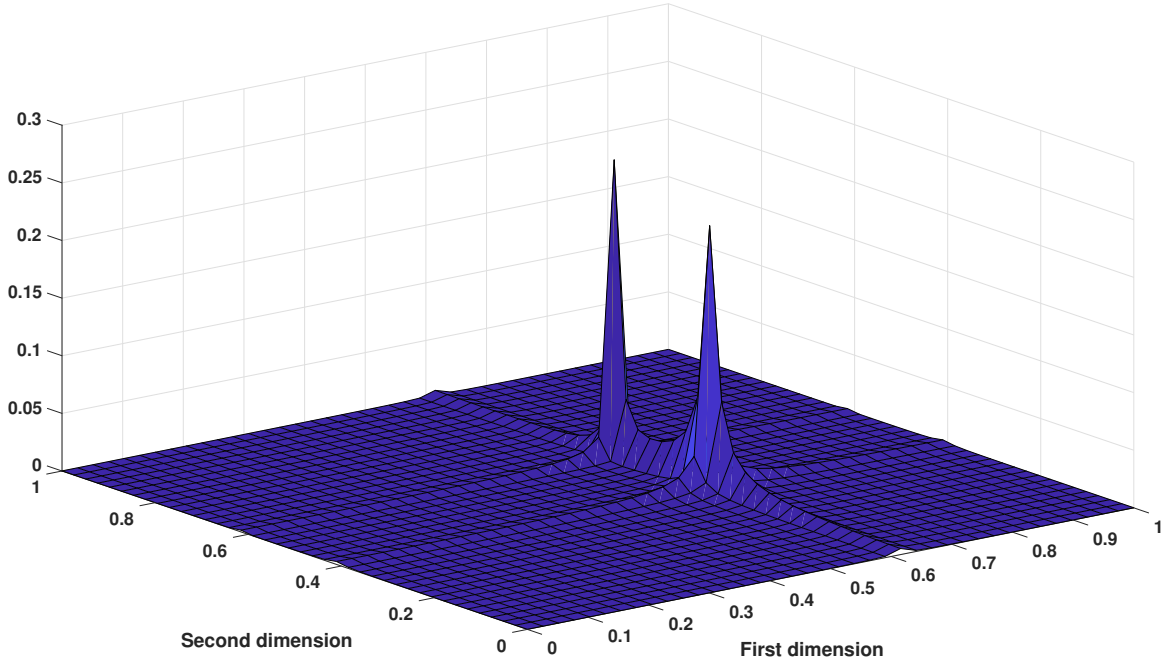

Figure 12: WSEMA estimate of the signal described in section 4.3.

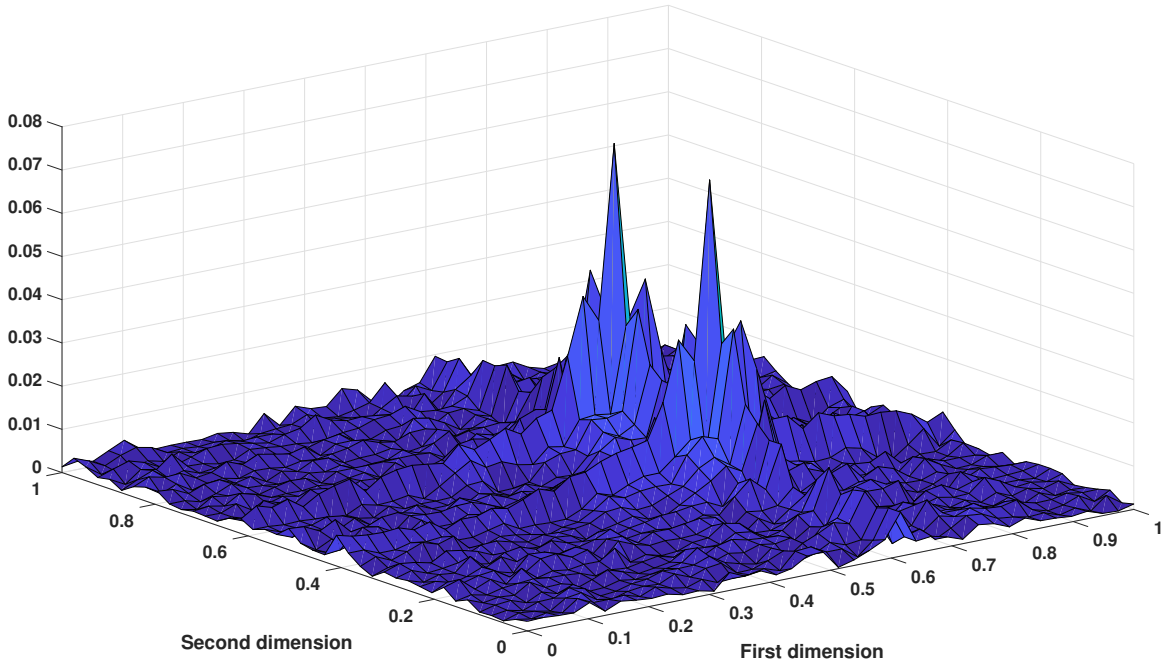

Figure 13: Lomb-Scargle periodogram of the signal described in section 4.3.

Table 1: RRMSE of WSEMA using optimal and random sampling scheme respectively

	Simulations used	RRMSE f	RRMSE β
Optimal scheme	99.25 %	0.0021	13.5518
Random scheme	99.45 %	0.0030	17.9065

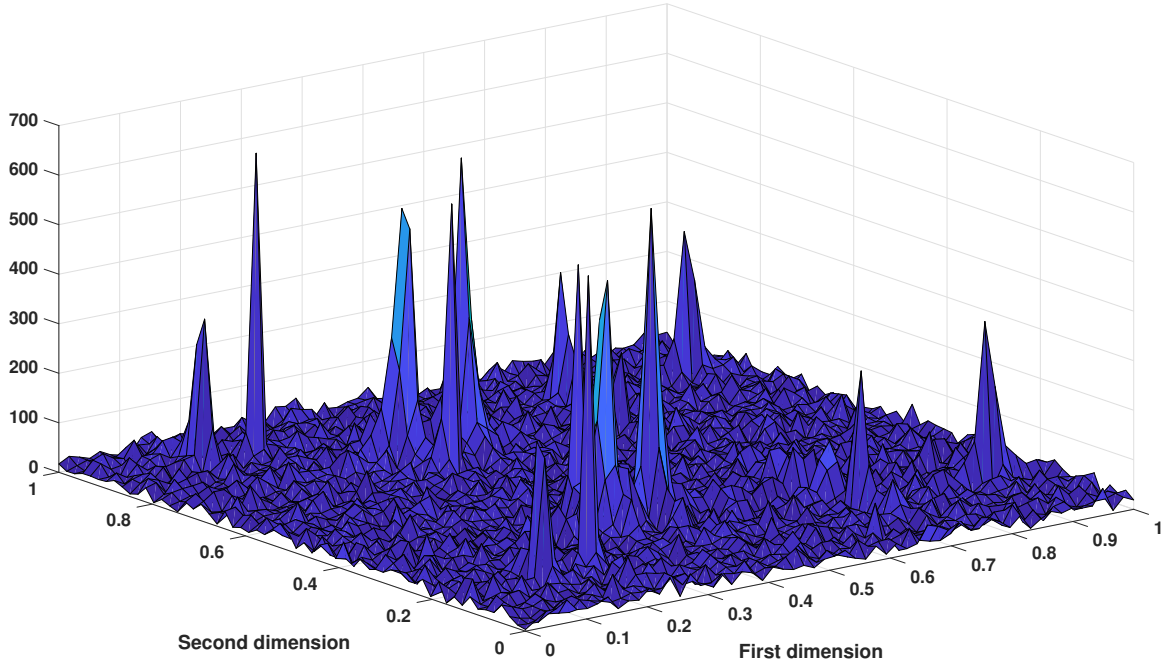

Figure 14: Periodogram of the FID-signal described in section 4.4.

The number of simulations disregarded is larger for the optimal sampling scheme than for the random one. A possible explanation for this is that the scheme is optimized with respect to the damping parameters, and not with respect to the frequency. Similar simulations, where the optimization instead is done with respect to frequency, indicates a lower rate of disregarded simulations for the optimized scheme compared to the random one.

4.4 Real data

It is of course of greatest importance that an estimator hoping to have any practical use is able to handle not only simulated data, but real data as well, with all difficulties that come with it. Therefore, WSEMA was given the task to estimate the parameters of real FID data. The FFT of the 60×60-signal can be seen in figure 14 and the corresponding WSEMA-estimate in figure 15. The estimation is done with 40 bands/dimension in the initial step and 10 bands/dimension/sample in the one and only zooming step. The hyper parameter γ is set to 0.2. Provided in figure 16 is the residual, i.e. the absolute difference between the real data and the estimated signal. With the exception of a couple of peaks, corresponding to the smaller peaks in the original signal, WSEMA's recovery appears successful.

5 Discussion

5.1 Evaluation of WSEMA

Estimation of the parameters of decaying sinusoids is a difficult task. In this thesis I introduce an efficient method to estimate the parameters of a signal with separated peaks whose damping coefficient are not too large, with $\beta \approx 25/N$, with N being the total number of samples. The introduced method is capable of estimating the parameters of the signal with the lowest possible variance for decent SNR-levels, stacking up well against an established estimator like PUMA, despite lacking the a priori-knowledge of the number of peaks required by the latter. Further, it is able to recover multi-modal signals using a dictionary as small as 0.3 hypercubes per sample.

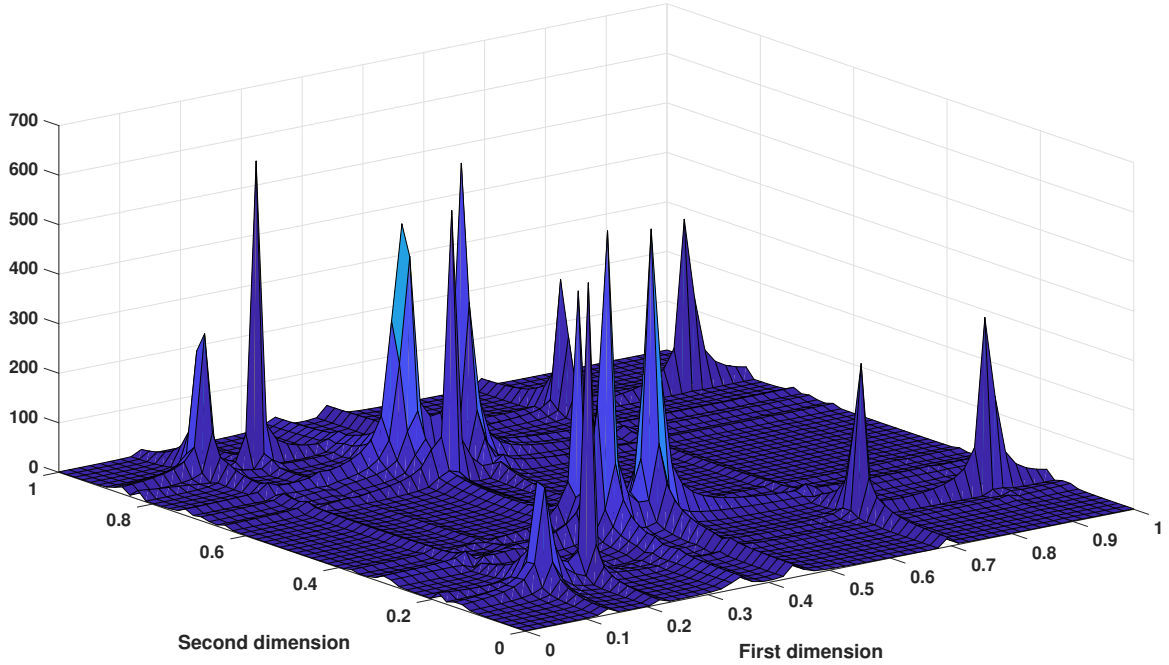

Figure 15: WSEMA estimate of the FID-signal described in section 4.4.

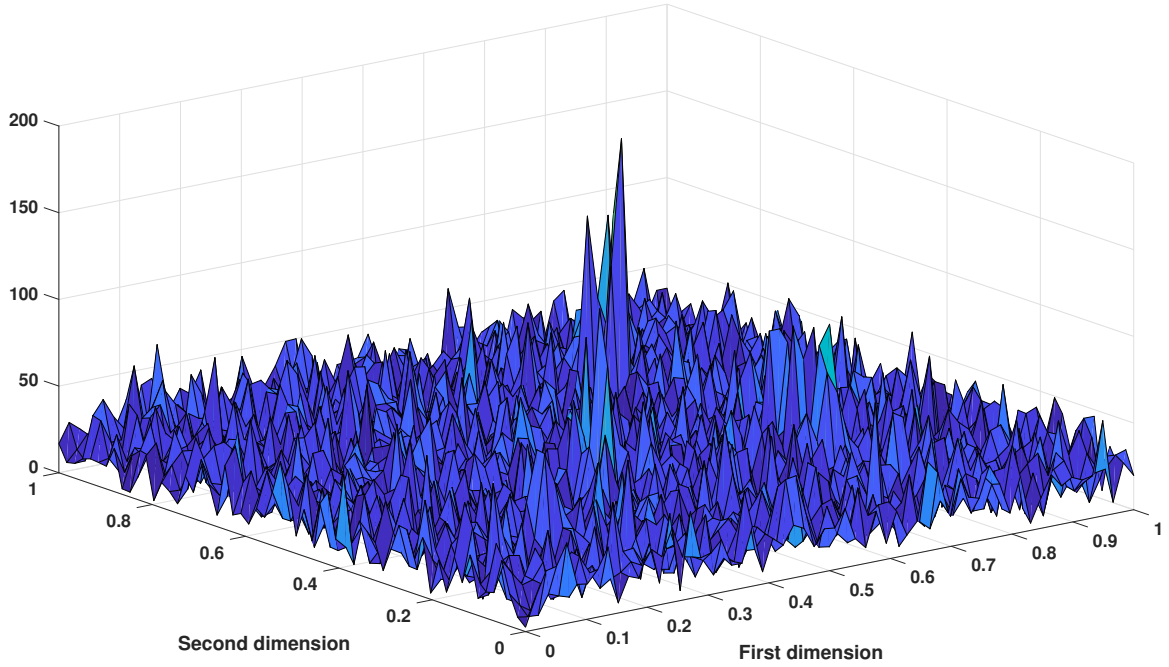

Figure 16: Residual between the FID-signal described in section 4.4 and the WSEMA estimate of the signal.

Regarding WSEMA:s ability to estimate real NMR spectroscopy data, the evaluation is a tougher task. No ground truth regarding the signal is available, so evaluation have to be made by comparing the WSEMA estimate to the FFT. The estimation and the residual, figure 15 and 16 respectively, shows that WSEMA successfully recovers a majority of the frequency peaks seen in figure 14. The peaks in the residual reveals the inherent difficulty of estimating a signal containing modes with widely varying amplitudes. However, comparing the scales of figure 14 and 16 shows that the size of the peaks that WSEMA missed is relatively small compared to the major components of the signal and the noise level. Here, the importance of proper management of spurious peaks is highlighted. Before the two additional LASSO-solving steps described at the end of section 3.2 was added, the solution contained 61 components. This is more than intuition would allow and certainly not as sparse as one would have liked. With the additional steps, however, the number of components was reduced to 14. With a smaller γ, rather than finding the ignored, smaller, peaks, the method adds peaks close to already existing ones.

5.2 Further research

To be able to fully reap the benefits of the WSEMA-estimator, further work has to be done. Firstly, a closer look at WSEMA's ability to detect closely spaced modes would be beneficial. Secondly, albeit being a general method, possible to use to estimate decaying sinusoids in different applications, it has been developed with NMR spectroscopy primarily in mind and it has with some success been used on real FID-signals. The evaluation have however been solely visual, comparing the signal estimated by WSEMA with the FFT. To more closely evaluate the quality of the algorithm, scrutiny by experts in the field is needed.

Further, an implementation of the algorithm in an arbitrary number of dimensions remains to be done. This would likely present some coding challenges and demand some afterthought to make sure that the computational time remains reasonable. The fundamental design however remains the same.

Another feature that would benefit the WSEMA is a more robust method to manage spurious peaks. Finally, proving the Swärd-Elvander conjecture eluded the author and this important task remains.

6 References

[1] J. Jensen, R. Heusdens, and S. H. Jensen. A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids. *IEEE Transactions on Speech and Audio Processing*, 12(2):121–132, March 2004.

[2] S. Rouquette, Y. Berthoumieu, and M. Najim. Subband decomposition based on the Hilbert transform applied to radar imaging. In *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, pages 1–4, Sept 1996.

[3] D. V. Rubtsov and J. L. Griffin. Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to nmr spectroscopy. *Journal of magnetic resonance*, 188 2:367–79, 2007.

[4] J. F. Hauer, C. J. Demeure, and L. L. Scharf. Initial results in Prony analysis of power system response signals. *IEEE Transactions on Power Systems*, 5(1):80–89, Feb 1990.

[5] W. Sun and H. C. So. Accurate and computationally efficient tensor-based subspace approach for multidimensional harmonic retrieval. *IEEE Transactions on Signal Processing*, 60(10):5077–5088, Oct 2012.

[6] Y. Hua and T. K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(5):814–824, May 1990.

[7] Y. Li, J. Razavilar, and K. J. R. Liu. A high-resolution technique for multidimensional nmr spectroscopy. *IEEE Transactions on Biomedical Engineering*, 45(1):78–86, Jan 1998.

[8] A. Paulraj, R. Roy, and T. Kailath. A subspace rotation approach to signal parameter estimation. *Proceedings of the IEEE*, 74(7):1044–1046, July 1986.

[9] H. C. So, F. K. W. Chan, W. H. Lau, and C. F. Chan. An efficient approach for two-dimensional parameter estimation of a single-tone. *IEEE Transactions on Signal Processing*, 58(4):1999–2009, April 2010.

[10] P. Stoica and T. Sundin. Nonparametric NMR Spectroscopy. *Journal of Magnetic Resonance*, 152(1):57–69, 2001.

[11] J. Swärd, F. Elvander, and A. Jakobsson. Designing sampling schemes for multi-dimensional data. *Signal Processing*, 150:1 – 10, 2018.

[12] J. Swärd, S. I. Adalbjörnsson, and A. Jakobsson. High resolution sparse estimation of exponentially decaying n-dimensional signals. *Signal Processing*, 128:309 – 317, 2016.

[13] J. C. Hoch and A. S. Stern. *NMR data processing*. Wiley-Liss, 1996.

[14] K. Kazimierczuk, A. Zawadzka-Kazimierczuk, and W. Koźmiński. Non-uniform frequency domain for optimal exploitation of non-uniform sampling. *Journal of Magnetic Resonance*, 205(2):286 – 292, 2010.

[15] P. Stoica and R. L. Moses. *Introduction to spectral analysis*. Prentice Hall, 1997.

[16] J. Swärd. *Parameter Estimation - in sparsity we trust*. PhD thesis, Lund University, 2017.

[17] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[18] M. Butsenko, J. Swärd, and A. Jakobsson. Estimating sparse signals using integrated wideband dictionaries. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4426–4430, March 2017.

[19] E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, dec 2008.

[20] 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.*, 3(1):1–122, January 2011.

[21] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[22] S. Sahnoun, E. H. Djermoune, and D. Brie. Sparse modal estimation of 2-d NMR signals. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8751–8755, May 2013.

[23] T. Kronvall and A. Jakobsson. Hyperparameter selection for group-sparse regression: a probabilistic approach. *Signal Processing*, 2018.

[24] G. O. Glentis and A. Jakobsson. Computationally efficient damped CAPON and APES spectral estimation. In *21st European Signal Processing Conference (EUSIPCO 2013)*, pages 1–5, Sept 2013.

Master's Theses in Mathematical Sciences 2018:E20
ISSN 1404-6342

LUTFMS-3343-2018

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/