

DETERMINING JOINT PERIODICITIES IN MULTI-TIME DATA WITH SAMPLING UNCERTAINTIES

David Svedberg[•], Filip Elvander^{*}, and Andreas Jakobsson[†]

[•] Uppsala University, Dept. of Electrical Engineering

^{*} KU Leuven, Dept. of Electrical Engineering, ESAT-STADIUS

[†] Lund University, Centre for Mathematical Sciences

ABSTRACT

In this work, we introduce a novel approach for determining a joint sparse spectrum from several non-uniformly sampled data sets, where each data set is assumed to have its own, and only partially known, sampling times. The problem originates in paleoclimatology, where each data point derives from a separate ice core measurement, resulting in that even though all measurements reflect the same periodicities, the sampling times and phases differ among the data sets, with the sampling times being only approximately known. The proposed estimator exploits all available data using a sparse reconstruction framework allowing for a reliable and robust estimation of the underlying periodicities. The performance of the method is illustrated using both simulated and measured ice core data sets.

Index Terms— Irregular Sampling, Multi-time, Misspecified Modelling, Paleoclimatology

1. INTRODUCTION

Ice and sea sediment cores provide an excellent record of long-term historic climate variability as well as solar activity via the molecular content trapped in the core [1–5]. By studying periodicities in the time-varying composition of the chemical content of the samples, information that allows for distinguishing between natural and anthropogenic climate change can be extracted [6]. One such molecular measure is the ratio $\delta^{18}O$, which is linearly linked to the local temperature [2]. The estimation of the frequency content for such data records is complicated by two factors. Firstly, measurements from ice and sea sediment cores constitute an irregularly sampled sequence, as, due to uneven accumulation and sedimentation processes, the samples are not uniformly distributed in time [7, 8]. Secondly, the actual sampling times of the measurements are only approximately known. Specifically, the sample time of a particular measurement is determined by using age-depth curves, i.e., an approximate function mapping the depth within the ice/sediment core at which the sample is extracted to its age or sampling time. However, due to the

global effect of astronomical influences, such as, e.g., eccentricity and solar activity, on the long-term climate variability, it is expected that measurements collected at different geographic locations should share similar spectral content in the low-frequency bands, thus opening for the possibility of exploiting multiple data sets to decrease the variability of the obtained spectral estimates [2, 9]. Taken together, these factors complicate inferring the spectral content using standard methods. In this work, we formulate the spectral estimation for ice/sediment core data as an inverse problem. Specifically, as it has been indicated that the low-frequency content of such measurements is concentrated to a small set of narrowband components [10, 11], we propose to model the time series as a sinusoidal mixture, utilizing a sparse reconstruction framework to estimate the spectral content [12, 13]. In order to allow for several data sets with similar but not identical spectral content, the spectral estimate is formed as the one closest to a set of spectra consistent with the different measurements, with a sparsity promoting penalty introduced to exploit the *a priori* knowledge that the expected number of signal components is small. The resulting estimator can be formulated as a convex program, allowing the spectral estimate to be reliably found using standard iterative solvers [14, 15]. Using numerical examples, we demonstrate that the proposed method allows for finding statistically efficient estimates of the frequency content for paleoclimatology data in the ideal case of known sample times, as well as displaying robustness to uncertainty in the sampling. Finally, we examine the estimation performance using both simulated and measured data.

2. PROBLEM FORMULATION

Inspired by ice/sediment core data studied in paleoclimatology, we consider data available from M different data sets, for this application typically being attained by analyzing the molecular content of gas trapped in the ice. In this case, the data consists of $\delta^{18}O$ measurements, formed as the ratio of the heavier oxygen isotope ^{18}O and ^{16}O as compared to a reference, which will vary with the volume of water in solid vs. liquid form [2]. From this data, one wants to estimate a

common, global, spectrum Φ , describing the frequency content of the climate variation. This spectrum can typically be well-modelled as a collection of point masses, corresponding to representing the signal by a finite sum of sinusoids. Specifically, the data from core m may be modeled as

$$s_t^{(m)} = \sum_{k=1}^K \rho_k^{(m)} \cos\{\omega_k(t + \Delta_t^{(m)}) + \varphi_k^{(m)}\}, \quad (1)$$

where $\Delta_t^{(m)}$ are the unknown missampling terms, i.e., sampling time errors, at the assumed sampling times $t = t_1^{(m)}, \dots, t_{N_m}^{(m)}$, generally being different for each core. The K frequencies, ω_k , are here assumed to be the same for each data set, whereas the amplitudes, $\rho_k^{(m)}$ and phases, $\varphi_k^{(m)}$ are, in general, different for each data set, implying that the data may be detailed using $K + 2KM$ unknown parameters. The measured signal is modelled as being embedded in additive noise, $v_t^{(m)}$, such that the observed data is $y_t^{(m)} = s_t^{(m)} + v_t^{(m)}$. Here, we assume that $v_t^{(m)}$ is normally distributed zero-mean white noise with variance σ_v^2 , with all observations being independent, such that $y_t^{(m)} \sim \mathcal{N}(s_t^{(m)}, \sigma_v^2)$. Each data set is assumed to be sampled at different sampling instants, with N_m available samples in the m -th data set. Let

$$\mathbf{T}^{(m)} \triangleq \begin{bmatrix} t_1^{(m)} & \dots & t_{N_m}^{(m)} \end{bmatrix}^T, \\ \Delta_{\mathbf{T}}^{(m)} \triangleq \begin{bmatrix} \Delta_{t_1^{(m)}}^{(m)} & \dots & \Delta_{t_{N_m}^{(m)}}^{(m)} \end{bmatrix}^T$$

denote the sampling instants and the corresponding missampling terms, respectively, allowing the noise-free and observed signals to be expressed as

$$\mathbf{s}^{(m)} \triangleq \begin{bmatrix} s_{t_1^{(m)}}^{(m)} & \dots & s_{t_{N_m}^{(m)}}^{(m)} \end{bmatrix}^T, \\ \mathbf{y}^{(m)} \triangleq \begin{bmatrix} y_{t_1^{(m)}}^{(m)} & \dots & y_{t_{N_m}^{(m)}}^{(m)} \end{bmatrix}^T,$$

where the subscript indicate the sampling instants. For notational simplicity, we will, without loss of generality, here assume that the sampling times of all data sets are unique and let N denote the total number of samples available from the M data sets, i.e. $N \triangleq \sum_{m=1}^M N_m$. Defining

$$\mathbf{s} \triangleq [(\mathbf{s}^{(1)})^T \dots (\mathbf{s}^{(M)})^T]^T, \\ \mathbf{y} \triangleq [(\mathbf{y}^{(1)})^T \dots (\mathbf{y}^{(M)})^T]^T,$$

we have that the observed $(N \times 1)$ -dimensional data vector may be modeled as the multivariate normal distributed vector $\mathbf{y} \sim \mathcal{N}(\mathbf{s}, \sigma_v^2 \mathbf{I}_{(N \times N)})$. Thus, the problem of interest is here to determine the $K + 2KM$ unknown parameters describing the observations \mathbf{y} while allowing for the sampling uncertainties.

3. PROPOSED ESTIMATION APPROACH

The proposed estimator is formed using a sparse reconstruction framework employing a separate wideband dictionary for each of the data sets, as described next. The resulting penalized regression problem exploits that the different data sets share spectral content in some frequency bands, thereby forming an underlying global spectrum, Φ . This spectrum is here represented using the non-negative vector $\Phi \triangleq [\Phi(-\omega_C) \dots \Phi(\omega_C)]^T \in \mathbb{R}^{2C+1}$, where $2C + 1$ denotes the number of considered spectral components (which may be different from the number of actual signal components). The proposed estimator is then formed as the extended penalized regression problem

$$\begin{aligned} & \underset{\{\beta^{(m)}, \alpha^{(m)} \geq 0\}_{m=1}^M, \Phi \geq 0}{\text{minimize}} \sum_{m=1}^M \|\mathbf{y}^{(m)} - \mathbf{D}^{(m)} \beta^{(m)}\|_2^2 \\ & + \sum_{m=1}^M \zeta^{(m)} \|\Phi - \alpha^{(m)}\|_1 + \lambda \|\Phi\|_1 \quad (2) \\ & \text{subject to } \alpha^{(m)} \geq |\beta^{(m)}|, \end{aligned}$$

where $\mathbf{D}^{(m)}$ denotes the m :th (wideband) dictionary, as discussed further below. The user-defined constants $\zeta^{(m)} \geq 0$ reflect the expected correlation between data set m and the rest of the records. Furthermore,

$$\beta^{(m)} \triangleq [\beta^{(m)}(-\omega_C) \dots \beta^{(m)}(\omega_C)]^T \in \mathbb{C}^{2C+1}$$

for $m = 1, \dots, M$, and

$$\alpha^{(m)} \triangleq [\alpha^{(m)}(-\omega_C) \dots \alpha^{(m)}(\omega_C)]^T \in \mathbb{R}^{2C+1}$$

representing the individual spectral estimates for the M data sets. The spectral estimate is thus formed as the spectral vector, Φ , that allows the M data sets to be best represented using the M different dictionaries, $\mathbf{D}^{(m)}$, at the assumed sampling times. The second term in the optimization enforces that the estimated global spectrum coincides, at least approximately, with the individual spectral estimates, $\alpha^{(m)}$. The third term is introduced to enforce sparsity on the resulting spectral estimate, whereas the constraint that $\alpha^{(m)} \geq |\beta^{(m)}|$ is introduced as the magnitude $|\beta^{(m)}|$ is, as illustrated in [16], expected to be lower than the global spectral components due to the sampling uncertainties. Thus, the constraint compensate for the loss of power from the missampling, while also ensuring that only differences in magnitude are penalized and thus ignoring differences in phase. By properly weighting these term using the weights $\zeta^{(m)}, m = 1, \dots, M$ and λ , one can find a trade-off between these properties. Herein, the weights $\zeta^{(m)}$ where, for simplicity, all chosen to be the same, whereas the value of λ was chosen such that each of the three terms of the cost function are of similar magnitude.

The M dictionaries $\mathbf{D}^{(m)}$ represent the range of potential frequencies. Here, in order to reduce the resulting computational complexity, we employ the use of iteratively refined

wideband dictionaries [17]. These are formed with the integrated dictionary elements¹

$$[\mathbf{D}]_{n,c}^{(m)} \triangleq \int_{\omega_c^s}^{\omega_c^e} e^{i\omega t_n^{(m)}} d\omega, \quad (3)$$

where ω_c^s and ω_c^e denotes the start and end frequencies of the c :th dictionary element, for $c = 1, \dots, C$, and for time $t_n^{(m)}$. By initially solving (2) using a coarse dictionary, the relevant frequency regions may be determined, whereafter the used dictionary may be iteratively refined in order to yield increasing resolution (see also [17]). As a sum of norms, the problem in (2) is convex and thus can be solved numerically using standard convex solvers, such as CVX [14, 15] or SeDuMi [18].

A detailed implementation of the proposed estimator is summarized in [16], and consists of two parts, firstly the wideband zooming procedure described above is employed to find the support of the model. In its final step, a grid-less narrowband search is employed in order to choose a single frequency in each of the active regions obtained from the earlier zooming step (see also [17]). The frequency estimates are found as

$$\hat{\omega} = \arg \min_{\Omega^s \leq \omega \leq \Omega^e} \sum_{m=1}^M \left\| \mathbf{y}^{(m)} - \left(\mathbf{A}_{\omega}^{(m)} \right)^{\dagger} \mathbf{y}^{(m)} \right\|_2^2, \quad (4)$$

where $\mathbf{A}^{\dagger} = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$, with where $\mathbf{A}_{\omega}^{(m)}$ being the real-valued narrowband dictionary for data set m , i.e.,

$$\mathbf{A}_{\omega}^{(m)} = \begin{bmatrix} \cos(\omega_1 \mathbf{T}^{(m)}) & \sin(\omega_1 \mathbf{T}^{(m)}) & \dots \\ \dots & \cos(\omega_C \mathbf{T}^{(m)}) & \sin(\omega_C \mathbf{T}^{(m)}) \end{bmatrix}.$$

This gridless search only results in frequency estimates; if one also wishes to find the amplitudes of the components at these frequencies, these may be found using least squares. A global amplitude estimate can then be formed by taking the mean of the amplitudes of the different data sets. It should be noted that, in general, the amplitudes may differ substantially between cores [2]. This simplistic approach may be problematic for certain applications where accuracy of the amplitude estimates is of great importance, but here it is only used to provide a visual comparison between different methods.

4. SIMULATION RESULTS

We proceed to evaluate the proposed method on both real and simulated data. To ensure that the simulated data displays the correct characteristics regarding, e.g., sampling irregularities, we employ a randomized sampling pattern construction. The sampling pattern used for the simulated data is here based on the Vostok ice core [1] and on the Taylor Dome

¹It is worth noting that although the signals of interest here are real-valued, it is still beneficial to make use of a complex-valued dictionary. This is as such a dictionary allows for different phases in each core (see also [16]).

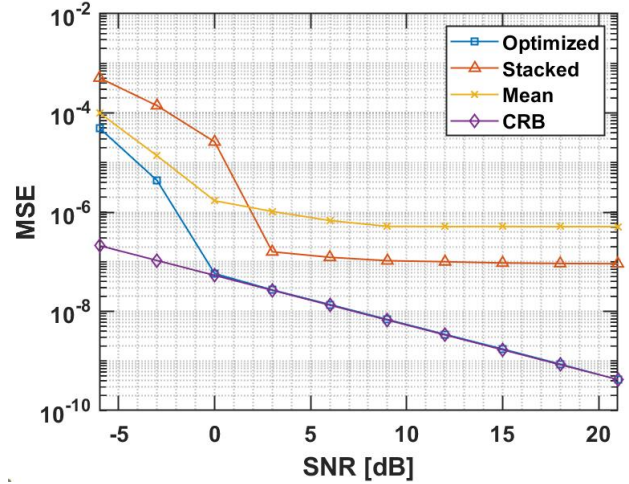


Fig. 1. The sum of MSE_k calculated from 1000 Monte-Carlo runs per SNR, as compared to the sum of averaged frequency CRB of the data assuming no missampling.

ice core [19, 20], as detailed in [16]. The global periodicities are simulated to mimic those that may be expected from the orbital theory of climate, having periods of 100, 41, 23, and 19 kyr [1, 2, 21]. The base amplitudes ρ_k for the frequencies are here set to be 1, 0.8, 0.6, and 0.6, with each data set having a deviation from this amplitude which is normal distributed with standard deviation $\rho_k/10$. The phase is drawn uniformly in $[0, \pi/5]$, in order to reflect the growth- and ablation rates of large ice sheets, which has a mean response time of roughly 10 kyr² [22]. For each simulation, 3 data sets are simulated, and the spectra for each are estimated using three different methods: the proposed estimator described above, the mean Lomb-Scargle periodogram [23], and the stacked periodogram [9, 24], both the latter being typical examples of state-of-the-art techniques used in the field.

Here, the mean periodogram is constructed by normalizing the data by N_m , such that each core contributes equally, and the mean of all periodograms is computed and considered to be the common spectrum. Since amplitude estimation accuracy is not considered in this application, the resulting estimate is scaled for ease of comparison in subsequent figures. The stacked periodogram is constructed by normalizing all data, stacking them into a single time-series and calculating the Lomb-Scargle periodogram of the resulting, stacked, time series. This stacking method has been used frequently to recreate the global signal [9, 11, 24]. To compare the methods, the sum of the mean squared error (MSE) of the frequency estimates are presented as a function of SNR (in dB), here, for white noise with variance σ_v^2 , defined as

$$\text{SNR} \triangleq 10 \log_{10} \frac{\sum_{k=1}^K (\rho_k^{(m)})^2}{2\sigma_v^2}.$$

²1 kyr = 1000 years.

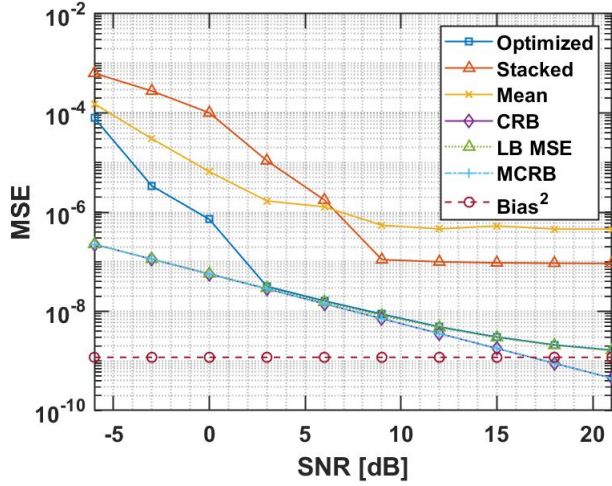


Fig. 2. The sum of MSE_k calculated from 2000 Monte-Carlo runs per SNR, as compared to the sum of averaged lower bound on the MSE on the frequency estimates, the MCRB, squared bias of the pseudo-true frequencies, and the CRB assuming no missampling.

With a total of N_R simulations of the 3 data sets, let $\hat{\omega}_k^n$ denote the estimate of ω_k in the n :th simulation. The frequency MSE is thus defined as

$$\text{MSE}_k \triangleq \frac{1}{N_R} \sum_{n=1}^{N_R} (\hat{\omega}_k^n - \omega_k)^2.$$

The frequency estimates for the periodogram-based methods are chosen as the K largest maxima of the spectral estimate.

We initially examine the case when the sampling patterns for each data set are known. Fig. 1 shows the sum of the frequency MSE_k for the three methods when there is no missampling, as compared to the corresponding CRB. The results for the proposed method were obtained using 4 zooming steps, starting with 16 initial bins with active threshold $\tau = 10^{-5}$, $\lambda = 15$, and $\zeta^{(m)} = 10$ for all m . As is clear from the figure, the MSE does not decrease for the periodogram-based methods. This effect is not due to a lack of decrease in variance as such, but rather due to bias of the frequency estimates (see also [16]). The proposed estimator on the other hand yields a statistically efficient estimate for higher SNR.

Fig. 2 shows the results using the same setup as above, but incorporating the expected missampling as described above. The frequency estimates using the proposed method are $\omega_1 = 9.014 \cdot 10^{-3}$, $\omega_2 = 2.471 \cdot 10^{-2}$, $\omega_3 = 4.331 \cdot 10^{-2}$, and $\omega_4 = 5.173 \cdot 10^{-2}$ [kyr^{-1}] with periods of 110.9, 40.47, 23.09, and 19.33 kyr, respectively. Here, $\lambda = 15$, $\zeta = 10$, with 4 zooming steps. It is worth noting that the misspecified CRB (MCRB), as presented in [16], and the CRB yields almost the same bound, except for the higher SNR cases, where the bias of the pseudo-true frequencies dominates such that

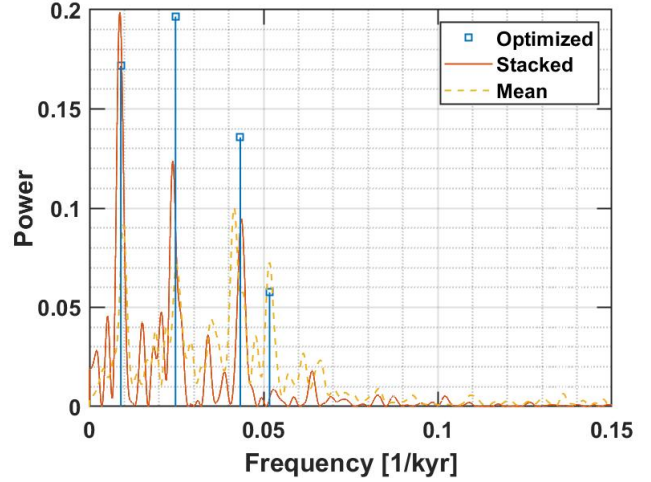


Fig. 3. The estimated global spectrum using measured ice-core data from three different ice-cores.

the lower bound decreases more slowly. The same bias problems for the periodogram-based methods at higher SNRs remain, implying that the MSE does not continue to decrease for higher SNRs. Again, the proposed estimator yields a statistically efficient estimate from SNR = 3 dB.

Using data from three different ice cores, the Vostok ice core [1], Taylor Dome ice core [19, 20, 25], and the Dome F ice core [7], the results of the method were computed; the results are shown in Fig. 3 and the estimated frequencies can be seen to be fairly close to those expected from orbital theory with periods of 100, 41, 23 and 19 kyr [2]. Interestingly, the stacked approach only shows three of the expected peaks, while all four are visible in the resulting estimates using the other two approaches. Also worth noting is that the relative amplitudes for each estimate is different, such that when using the proposed estimation approach, the second peak with a period of 40.47 kyr is the highest, whereas for the stacked method the first peak with a period of 114.61 kyr is the highest, and for the last approach the third peak with a period of 24.07 kyr is the highest. This shows that the choice of estimation approach significantly impacts the resulting estimate.

5. CONCLUSIONS

In this work, we have introduced a sparse reconstruction technique to estimate the global spectrum corresponding to the shared spectral structures of a set of non-uniformly sampled data sets, each sampled at different times and with significant uncertainties in the sampling times. The method is shown to yield statistically efficient estimates, as well as reaching the corresponding misspecified performance bound. The performance of the method is shown using realistic simulations of ice core data, including the growing uncertainty in the sampling times as the age of the samples grow, as well as with measured ice core data.

6. REFERENCES

- [1] J. R. Petit, J. Jouzel, D. Raynaud, N. Barkov, J. M. Barnola, I. Basile-Doelsch, et al., "Climate and Atmospheric History of the Past 420,000 Years from the Vostok Ice Core, Antarctica," *Nature*, vol. 399, pp. 429–436, 06 1999.
- [2] R. A. Muller and G. J. MacDonald, *Ice Ages and Astronomical Causes: Data, Spectral analysis and Mechanisms*, Springer-Verlag Berlin Heidelberg, 1 edition, 2000.
- [3] J. R. Petit, D. Raynaud, C. Lorius, J. Jouzel, G. Delaygue, N. Barkov, et al., "Historical Isotopic Temperature Record from the Vostok Ice Core (420,000 years BP-present)," Tech. Rep., CDIAC, 1 2000.
- [4] P. M. Grootes and M. Stuiver, "Oxygen 18/16 variability in Greenland snow and ice with 10-3 to 10-5 year time resolution," *Journal of Geophysical Research: Oceans*, vol. 102, no. C12, pp. 26455–26470, 1997.
- [5] S. J. Johnsen, D. Dahl-Jensen, N. Gundestrup, J. P. Steffensen, H.B. Clausen, H. Miller, et al., "Oxygen isotope and palaeotemperature records from six Greenland ice-core stations: Camp Century, Dye-3, GRIP, GISP2, Renland and NorthGRIP," *Journal of Quaternary Science*, vol. 16, no. 4, pp. 299–307, 2001.
- [6] J. Hansen and M. Sato, "Paleoclimate implications for human-made climate change," *Climate Change: Inferences from Paleoclimate and Regional Aspects*, 05 2011.
- [7] K. Kawamura, F. Parrenin, L. Lisiecki, R. Uemura, F. Vimeux, J. Severinghaus, et al., "Northern hemisphere forcing of climatic cycles in antarctica over the past 360,000 years," *Nature*, vol. 448, pp. 912–6, 09 2007.
- [8] C. Waelbroeck, B. C. Loughheed, N. Vazquez Riveiros, L. Missiaen, J. Pedro, T. Dokken, et al., "Consistently dated atlantic sediment cores over the last 40 thousand years," *Scientific Data*, vol. 6, no. 1, pp. 165, Sep 2019.
- [9] F. C. Bassinot, L. D. Labeyrie, E. Vincent, X. Quidelleur, N. J. Shackleton, and Y. Lancelot, "The astronomical theory of climate and the age of the Brunhes-Matuyama magnetic reversal," *Earth and Planetary Science Letters*, vol. 126, no. 1, pp. 91 – 108, 1994.
- [10] M. K. Milankovitch, "Kanon der Erdbestrahlung und seine Anwendung auf das Eiszeitenproblem," *Royal Serbian Academy Special Publication*, vol. 133, pp. 1–633, 1941.
- [11] J. Imbrie and J. Z. Imbrie, "Modeling the climatic response to orbital variations," *Science*, vol. 207, no. 4434, pp. 943–953, 1980.
- [12] 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," *Signal Processing, IEEE Transactions on*, vol. 59, pp. 35 – 47, 02 2011.
- [13] A. Maleki and D. L. Donoho, "Optimally tuned iterative reconstruction algorithms for compressed sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 330–341, 2010.
- [14] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. 2014.
- [15] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, V. Blondel, S. Boyd, and H. Kimura, Eds., Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008.
- [16] D. Svedberg, F. Elvander, and A. Jakobsson, "Determining joint periodicities in multi-time data with sampling uncertainties," *IEEE Transactions on Signal Processing*, 2021, Submitted Manuscript.
- [17] M. Butsenko, J. Swärd, and A. Jakobsson, "Estimating sparse signals using integrated wideband dictionaries," *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 4170–4181, 2018.
- [18] Jos F Sturm, "Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones," *Optimization methods and software*, vol. 11, no. 1-4, pp. 625–653, 1999.
- [19] P. M. Grootes, E. J. Steig, and M. Stuiver, "The oxygen isotope record from Taylor Dome, Antarctica," *EOS, Transactions of the American Geophysical Union*, , no. 76, pp. 176, 1994.
- [20] E. J. Steig, *Beryllium-10 in the Taylor Dome Ice Core: Applications to Antarctic Glaciology and Paleoclimatology*, Ph.D. thesis, University of Washington, 1996.
- [21] R. A. Muller and G. J. MacDonald, "Spectrum of 100-kyr glacial cycle: Orbital inclination, not eccentricity," *Proceedings of the National Academy of Sciences*, vol. 94, no. 16, pp. 8329–8334, 1997.
- [22] J. Weertman, "Rate of Growth or Shrinkage of Nonequilibrium Ice Sheets," *Journal of Glaciology*, vol. 5, no. 38, pp. 145–158, 1964.
- [23] J. D. Scargle, "Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data.," *Astrophysical Journal*, vol. 263, pp. 835–853, Dec. 1982.
- [24] J Imbrie, J D Hays, D G Martinson, A McIntyre, A C Mix, J J Morley, N G Pisias, W L Prell, , and N J Shackleton, "The orbital theory of pleistocene climate: support from a revised chronology of the marine $\delta^{18}\text{O}$ record. in "milankovitch and climate.,"" *Milankovitch and Climate*, 1984.
- [25] P. M. Grootes and E. J. Steig, "Low resolution stable isotopes of ice core Taylor Dome," <https://doi.org/10.1594/PANGAEA.473389>, 1994.