2020 54th Asilomar Conference on Signals, Systems, and Computers | 978-0-7381-3126-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/IEEECONF51394.2020.9443532

Nuclear Norm Based Spectrum Estimation for Molecular Dynamic Simulations

Shuang Li, Stephen Becker, and Michael B. Wakin

***Abstract*—Molecular dynamic (MD) simulations are used to probe molecular systems in regimes not accessible to physical experiments. A common goal of these simulations is to compute the power spectral density (PSD) of some component of the system such as particle velocity. In certain MD simulations, only a few time locations are observed, which makes it difficult to estimate the autocorrelation and PSD. This work develops a novel nuclear norm minimization-based method for this type of sub-sampled data, based on a parametric representation of the PSD as the sum of Lorentzians. We show results on both synthetic data and a test system of methanol.**

I. Introduction

Molecular dynamic (MD) simulations are widespread computer simulation-based tools used to study the properties of a chemical system by analyzing the physical movements of its atoms or molecules [1]–[4]. MD simulations are used in a variety of applications including drug design [5], calculating vibrational or rotational modes [6], optical absorption spectra [7], titanium dioxide polymorphs [8], and circular dichroism spectra [9].

In this work, we consider the problem of estimating the power spectral density (PSD) of a stationary stochastic process $\{y_i\}$. Based on the Wiener–Khinchin theorem, we may equivalently estimate the autocorrelation function (ACF) of $\{y_i\}$, defined as

$$R_y(\tau) = \mathbb{E}(y_i y_{i+\tau}),$$

where τ is the time lag and the expectation is taken over i (and over all particles if y_i is a vector). The PSD is then obtained by taking the (discrete) Fourier transform of the ACF.

In some types of molecular dynamic (MD) simulations, such as time-dependent density functional theory [10], y_i is a quantity like polarizability that is slow to simulate, and therefore it is only calculated at a few possibly irregularly spaced time points. Other types of MD may have access to the full sequence $\{y_i\}$ but then purposely save only a subset of the sequence in order to achieve compression. Motivated by these types of MD, we seek a method to accurately estimate the spectrum of the stochastic process from a small amount of *subsampled* data from $\{y_i\}$.

The PSD of the particle velocity has a special structure that can be represented as a sparse superposition of Lorentzian functions, which, near their centers, each have a similar shape to a Gaussian function. A Lorentzian function (in the frequency domain) is defined as

$$\mathcal{L}(f) = \frac{A}{2\pi\left((f-c)^2 + \left(\frac{w}{2}\right)^2\right)},$$

where A, c, and w denote the amplitude, center, and width of the function. In the time (lag) domain, a Lorentzian is a sinusoid with exponential decay. In particular, we have

$$L(\tau) = \frac{A}{w} e^{j2\pi c\tau - w\pi|\tau|}.$$

Therefore, for particle velocity data, the autocorrelation function $R_y(\tau)$ will consist of a sparse superposition of decaying sinusoids.

II. The Proposed Method

To estimate the PSD based on subsampled velocity data $\{y_i\}$, we first construct an estimate $\widehat{R}_y(\tau)$ of the ACF using the sample mean of quantities $y_i y_{i+\tau}$ over all indices i where both y_i and $y_{i+\tau}$ are available. Such an estimate may be noisy and may even contain gaps at certain τ values where there are no data pairs $(y_i, y_{i+\tau})$ available. Our second step then involves formulating a matrix optimization problem to estimate the clean and complete ACF. In particular, we formulate a Hankel matrix $\mathbf{H}_y$ from the estimated ACF $\widehat{R}_y(\tau)$. Due to the small number of Lorentzian functions in the PSD, we can view $\mathbf{H}_y$ as a noisy and possibly incomplete matrix that has low rank.

S. Li was with the Department of Electrical Engineering, Colorado School of Mines, Golden, CO 80401, USA. She is now with the Department of Mathematics, University of California, Los Angeles, CA 90095, USA (email: shuangli@math.ucla.edu). M. B. Wakin is with the Department of Electrical Engineering, Colorado School of Mines, Golden, CO 80401, USA (e-mail: mwakin@mines.edu). S. Becker is with the Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, USA (e-mail: stephen.becker@colorado.edu).

Denote $\mathbf{W}_1$ and $\mathbf{W}_2$ as two Hankel matrices formulated from two weighting vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. In particular, $\boldsymbol{w}_1$ is a vector with entries being all 1's except those corresponding to missing entries of the estimated ACF $\widehat{R}_y(\tau)$, i.e., $\boldsymbol{w}_1(\tau) = 0$ if the estimated ACF $\widehat{R}_y(\tau)$ has a missing entry at index τ; and $\boldsymbol{w}_2$ is a vector with entries being the inverse of the number of diagonal entries in the Hankel matrix $\mathbf{H}_y$. We propose the following weighted nuclear norm minimization to update the estimated ACF:

$$\begin{aligned} &\min_{\mathbf{H}_x \in \mathcal{H}} \ \|\mathbf{H}_x\|_* \\ &\ \text{s.t.} \ \ \|\mathbf{W}_1 \odot \mathbf{W}_2 \odot (\mathbf{H}_x - \mathbf{H}_y)\|_F^2 \leq \epsilon, \end{aligned} \tag{II.1}$$

where $\odot$ denotes element-wise multiplication, $\mathcal{H}$ denotes a set that contains all of the Hankel matrices with proper size, ϵ is a parameter corresponding to the noise variance, $\mathbf{W}_1$ serves as a mask, as in matrix completion, and $\mathbf{W}_2$ standardizes the entries to have equal variance.

One can use any off-the-shelf SDP solver, such as CVX [11], to solve the above optimization (II.1). As CVX can return both a *primal solution* $\widehat{\mathbf{H}}_x$ and a *dual solution* $\mathbf{Q}$, we propose two ways to estimate the spectrum in this work. The first way is to use the primal solution $\widehat{\mathbf{H}}_x$, which is a Hankel matrix corresponding to the updated ACF, and transform the updated ACF to the frequency domain to get the spectrum.

The second way is to use the dual solution $\mathbf{Q}$, which is inspired from the dual analysis used in atomic norm minimization [12]–[15]. In particular, we can first compute a dual polynomial with the dual solution. Recalling that the estimated ACF consists of samples of a sum of a few decaying exponentials, the dual polynomial is then computed as

$$\mathcal{Q}(r, f) = \mathbf{Q}^H \boldsymbol{a}(r, f),$$

where $\boldsymbol{a}(r, f)$ is a vector with entries being uniform time samples of a decaying exponential $r^t e^{j2\pi f t}$. Here, $f \in [0, 1)$ and $r \in [0, 1]$ denote the normalized frequency and damping ratio, respectively. Then, we can estimate the frequency components and damping ratios contained in the estimated autocorrelation function by localizing the places where the ℓ_2 norm of the dual polynomial achieves 1, as in our previous work [16]. With the estimated (r, f) pairs, we can then reconstruct the ACF and compute its PSD by transforming the ACF to the frequency domain. Note that the atomic norm minimization based methods cannot address signals with damping since there is no semi-definite program for the corresponding atomic norm [16]. Therefore, we use nuclear norm minimization here as in [16].

III. Experiments

In this section, we test the proposed methods with both synthetic data and Methanol velocity data.

A. Synthetic data

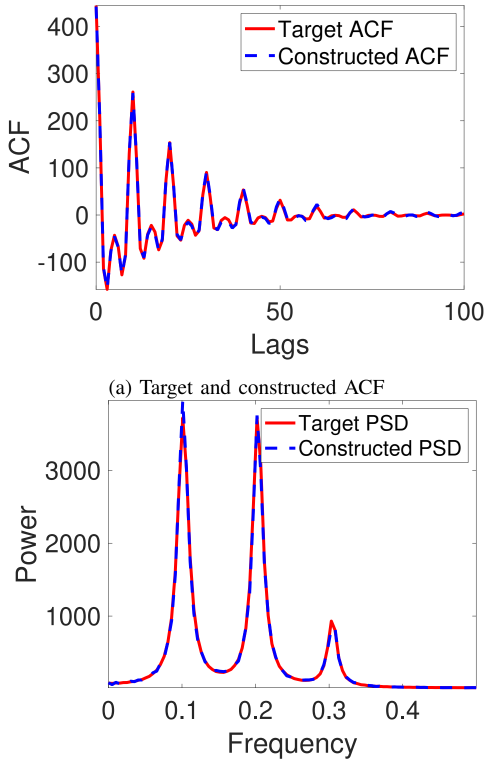
First, we generate a target ACF as a sum of three damped sinusoids (red line in Figure 1 (a)). The PSD of this target ACF is shown in Figure 1 (b) (see red line). Given the target PSD (denoted as $\mathcal{P}_Y(f)$), we create a system with frequency response $H(f)$ satisfying

$$|H(f)|^2 = \mathcal{P}_Y(f).$$

Then, we can get the frequency response $H(f)$ by taking the square root of $\mathcal{P}_Y(f)$, and we can get the impulse response $h(t)$ by taking inverse Fourier transform of $H(f)$. Once we have the impulse response, we then input a sequence of white noise $x(t)$ with length 10^5 into this system and measure the output signal $y(t)$. Then, $y(t)$ should have an ACF close to the target ACF. We present the ACF and PSD of the constructed $y(t)$ in Figure 1 (blue dashed lines). It can be seen that the constructed ACF and PSD are very close to the target ones. Note that the PSDs shown in Figure 1 (b) are computed via the Fast Fourier Transform (FFT).

Next, we estimate the PSD of the constructed time domain data $y(t)$ via two classical methods: Welch and Yule-Walker, in both the full data case and the missing data case. As is shown in Figures 2-4, the two classical methods work well in estimating the PSD when there are enough data samples. However, when there are too few data samples available (e.g., Figure 4), the two methods can miss the third frequency component.

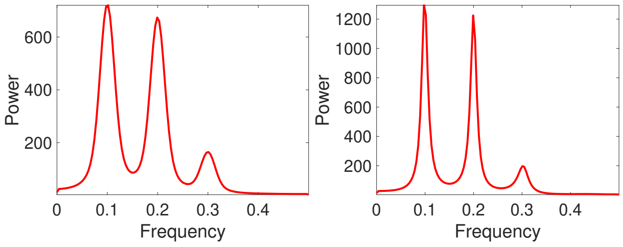
For the missing data case, we first keep 20% of the time domain data. In this case, there are no gaps (missing entries) in the estimated ACF. We build a 50×51 Hankel matrix $\mathbf{H}_y$ with these samples and use CVX to solve the optimization problem (II.1) with $\epsilon = 4.9497 \times 10^3$, which is obtained by computing the variance of the estimated ACF. The dual polynomial and estimated parameters are shown in Figure 5. Note that the dual polynomial is symmetric as the ACF signal is real. The estimated ACF and PSD are shown in Figure 6. It can be seen that both the primal method and the dual method work well in estimating the PSD. Then, we repeat the above experiment by keeping only 9% of the time domain data with $\epsilon = 2.4180 \times 10^4$. The results are shown in Figures 7 and 8. It can be seen that the primal method still works well in estimating the PSD. Since there are some spurious (r, f) pairs localized by the dual polynomial, the dual method can only roughly estimate the PSD in this case.



(a) Target and constructed ACF

(b) Target and constructed PSD

Figure 1: The ACF and PSD of a constructed signal $y(t)$ v.s. the target ACF and PSD.



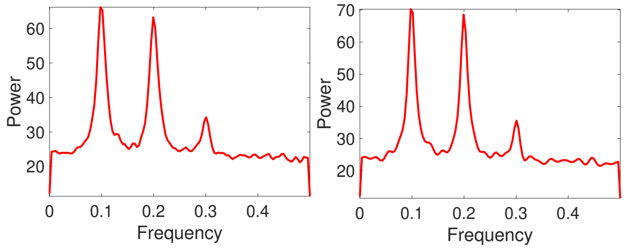
(a) Welch

(b) Yule-walker

Figure 2: Estimated PSD with full time domain data.

B. Methanol velocity data

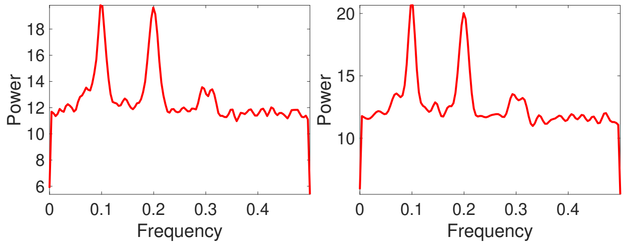
Next, we test our proposed methods on the velocity data of methanol collected from molecular dynamic simulations with LAMMPS [17]. This data contains 256 atoms, each with 1.9×10^4 uniform samples of the velocity. The maximal lag used to compute the ACF is set as 99. We set $\epsilon = 1.2969 \times 10^{-07}$, which is obtained by computing the variance of the estimated ACF. We present the estimated ACF and PSD of using only 4% and 1% of the velocity data in Figures 9 and 10, respectively. The red solid lines denote the estimated ACF and PSD obtained by using all of the 1.9×10^4 uniform velocity samples of each atom and we view this as the ground truth. It can be seen that both the dual and primal methods work well in spectral estimation. Moreover, they can roughly recover the PSD even when using only 1% of the velocity data.



(a) Welch

(b) Yule-walker

Figure 3: Estimated PSD using 20% of the time domain data.



(a) Welch

(b) Yule-walker

Figure 4: Estimated PSD using 9% of the time domain data.

IV. Conclusion

In this work, we consider the problem of estimating the PSD of a stationary stochastic process and develop a novel nuclear norm minimization-based method to accurately estimate the spectrum from subsampled data of the stochastic process, based on a parametric representation of the PSD as the sum of Lorentzians. Experiments conducted on both synthetic data and Methanol velocity data indicate that the proposed methods perform very well in estimating the PSD.

Acknowledgments

The authors thank Marc Thomson for providing the LAMMPS simulation data. MW and SL were supported by NSF grant CCF-1704204, and SB was supported by NSF grant 1819251.

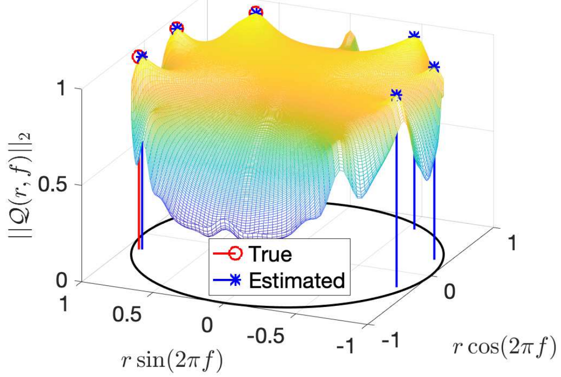
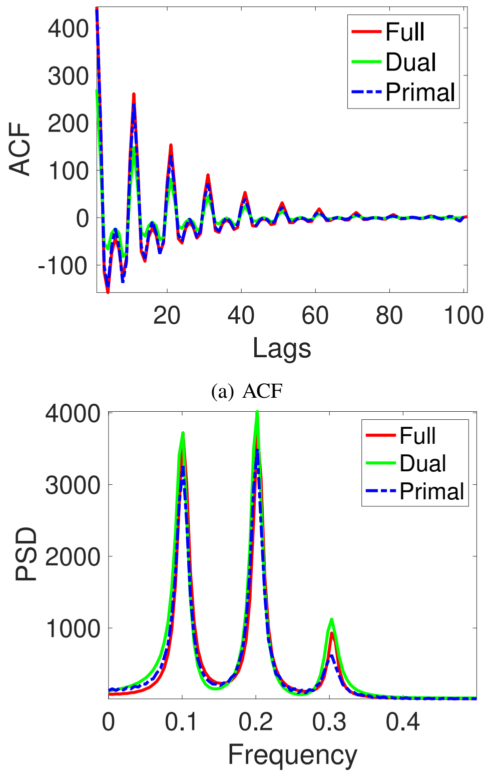


Figure 5: Dual polynomial obtained by solving (II.1) with 20% of the time domain data.



(a) ACF

(b) PSD

Figure 6: ACF and PSD estimated by solving (II.1) with 20% of the time domain data.

References

[1] K. Binder, *Monte Carlo and molecular dynamics simulations in polymer science*. Oxford University Press, 1995.

[2] D. C. Rapaport, *The art of molecular dynamics simulation*. Cambridge University Press, 2004.

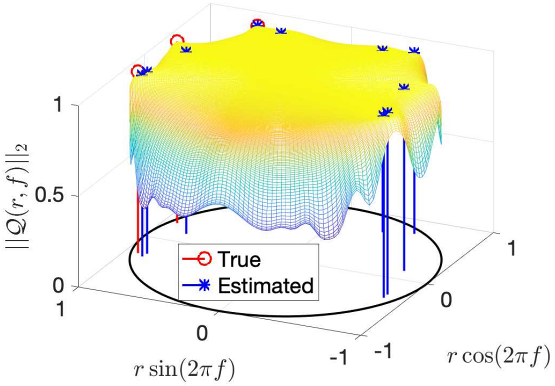
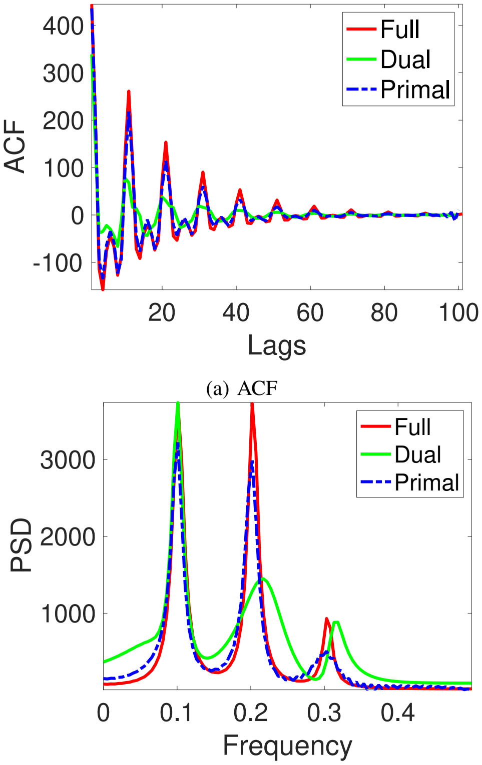


Figure 7: Dual polynomial obtained by solving (II.1) with 9% of the time domain data.



(a) ACF

(b) PSD

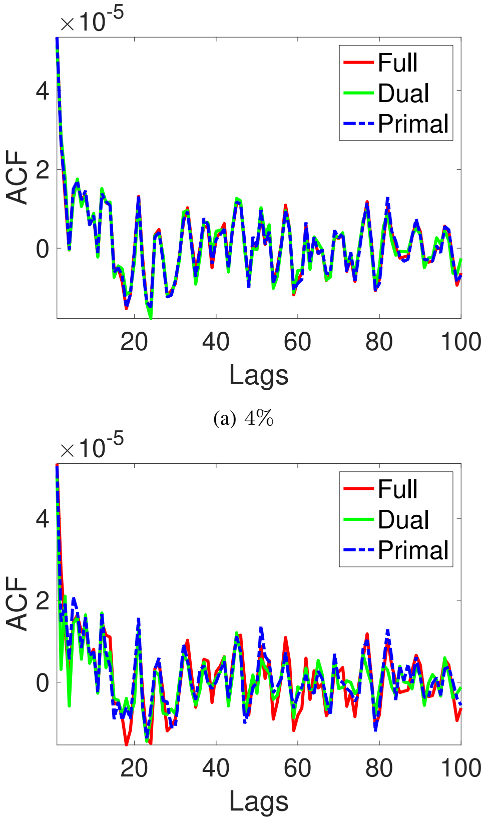
Figure 8: ACF and PSD estimated by solving (II.1) with 9% of the time domain data.

[3] Z. Huang and S. Becker, "Spectral estimation from simulations via sketching," *arXiv preprint arXiv:2007.11026*, 2020.

[4] R. J. Sadus, *Molecular simulation of fluids*. Elsevier, 2002.

[5] H. Alonso, A. A. Bliznyuk, and J. E. Gready, "Combining docking and molecular dynamic simulations in drug design," *Medicinal Research Reviews*, vol. 26, no. 5, pp. 531–568, 2006.

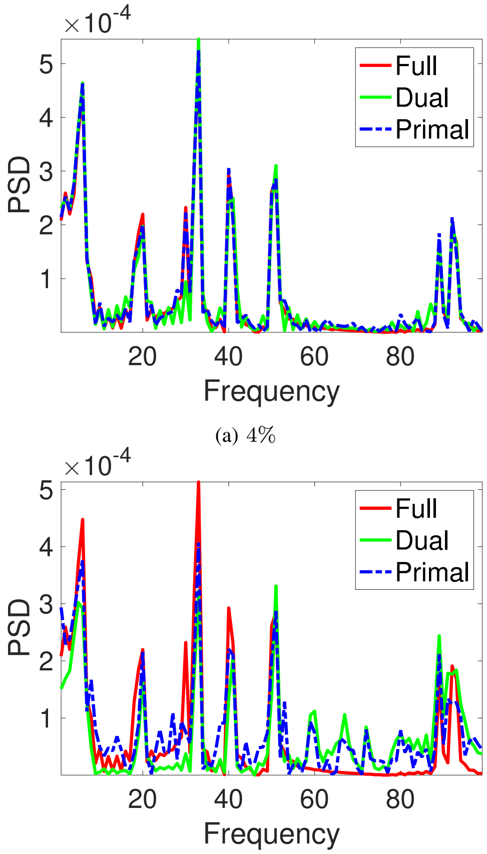
[6] A. P. Scott and L. Radom, "Harmonic vibrational frequencies:



(a) 4%

(b) 1%

Figure 9: Estimated ACF with full velocity data (red line), (a) 4% velocity data, and (b) 1% velocity data.



(a) 4%

(b) 1%

Figure 10: Estimated PSD with full velocity data (red line), (a) 4% velocity data, and (b) 1% velocity data.

an evaluation of hartree- fock, møller- plesset, quadratic configuration interaction, density functional theory, and semiempirical scale factors," *The Journal of Physical Chemistry*, vol. 100, no. 41, pp. 16502–16513, 1996.

[7] K. Yabana and G. Bertsch, "Time-dependent local-density approximation in real time," *Physical Review B*, vol. 54, no. 7, p. 4484, 1996.

[8] D.-W. Kim, N. Enomoto, Z.-e. Nakagawa, and K. Kawamura, "Molecular dynamic simulation in titanium dioxide polymorphs: Rutile, brookite, and anatase," *Journal of the American Ceramic Society*, vol. 79, no. 4, pp. 1095–1099, 1996.

[9] D. Varsano, L. A. Espinosa-Leal, X. Andrade, M. A. Marques, R. Di Felice, and A. Rubio, "Towards a gauge invariant method for molecular chiroptical properties in tddft," *Physical Chemistry Chemical Physics*, vol. 11, no. 22, pp. 4481–4489, 2009.

[10] E. Runge and E. K. Gross, "Density-functional theory for time-dependent systems," *Physical Review Letters*, vol. 52, no. 12, p. 997, 1984.

[11] M. Grant and S. Boyd, "Cvx: Matlab software for disciplined convex programming, version 2.1," 2014.

[12] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on pure and applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[13] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

[14] S. Li, M. B. Wakin, and G. Tang, "Atomic norm denoising for complex exponentials with unknown waveform modulations," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3893–3913, 2019.

[15] S. Li, D. Yang, G. Tang, and M. B. Wakin, "Atomic norm minimization for modal analysis from random and compressed samples," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1817–1831, 2018.

[16] S. Li, H. Mansour, and M. B. Wakin, "Recovery analysis of damped spectrally sparse signals and its relation to music," *Information and Inference: A Journal of the IMA*, 2020.

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.