Semiparametric Estimation of Solar Generation

Vivak Patel
Department of Statistics
University of Chicago
Chicago, IL 60637
Email: vp314@uchicago.edu

Daniel A. Maldonado, *Member, IEEE*
MCS Division
Argonne National Laboratory
Lemont, IL 60439
Email: maldonadod@anl.gov

Mihai Anitescu, *Member, IEEE*
MCS Division
Argonne National Laboratory
Lemont, IL 60439
Email: anitescu@mcs.anl.gov

***Abstract*—Because of the high variability of solar generation, integrating solar energy into the grid requires estimating the total energy demand and total solar generation behind a node using such available data as load measurements at the node and irradiance data. In this work, we propose a flexible, robust semiparametric solar generation estimation methodology with uncertainty bounds, which nonparametrically estimates the total energy demand and parametrically estimates the solar generation based on a photovoltaic performance model. We demonstrate the effectiveness of our methodology on experiments that combine publicly available hourly load and per minute irradiance data with simulated data. We also demonstrate the high accuracy achieved by our methodology if per minute load measurements became available.**

November 7, 2017

I. Introduction

As the manufacturing costs of photovoltaic (PV) panels decrease, integrating solar energy into the grid remains a critical challenge owing to solar generation's high variability caused by soiling, weather variability, and cloud coverage [1]. Therefore, in order for operators to better plan and cost-effectively dispatch existing energy resources, solar energy generation must be accurately predicted from available information such as load measurements and irradiance (i.e., solar penetration) measurements.

Previously, [2] used micro-synchrophasor measurements at the substation and a solar generator to estimate the amount of solar energy generation in real time. In particular, [2] created an accurate linear regression-based methodology for estimating solar energy generation that was robust to the high uncertainty of the plane-of-array irradiance. However, the main difficulties with the approach in [2] were that it used a restrictive parametric model for the load and had poor accuracy above a five-minute sampling rate.

In this work, we offer a robust and flexible semiparametric solar generation estimation methodology with uncertainty bounds using publicly available load and irradiance measurements. Specifically, we allow for the total energy demand to be modeled using a highly flexible nonparametric model [3], which requires fewer assumptions than the approach in [2]. Based on the fact that the total energy demand dynamics are *slower* than the dynamics of solar energy generators, we can estimate the total energy demand, solar energy generation, and proportion of energy demand met by solar using only hourly load and irradiance measurements. Moreover, we show that with even more frequent observations, such as per minute observations, the accuracy of this estimation greatly increases.

II. Modeling and Estimation

Here, we formalize our model, our estimation procedure, and our assumptions. The notation used throughout the remainder of this work is summarized in Table I.

A. Modeling Assumptions

We define the measured or net load at the substation as the total energy demand minus the energy generated by solar:

$$E_g(t) = E_r(t) - E_i(t), \tag{1}$$

where $E_r(t) \geq E_g(t),$ which implies $E_i(t) \geq 0$.

We will assume that $E_r(t)$ is a continuous function of time, and we will estimate it using a nonparametric Haar wavelet estimator with a resolution dependent on the rate at which load measurements and solar irradiance data are collected [3]. For example, when $E_g(t)$ and $I_R(t)$ are known at hourly intervals, we can use a resolution of 45 minutes, which means that changes occurring within a 45-minute span cannot be resolved by the estimator.

For $E_i(t)$, an appropriate model would describe the contribution of a collection of rooftop PV panels connected to the node with different performance models and incident irradiance that varies over space. However, such a model would introduce too many parameters, resulting in nonidentifiability given hourly or even per minute load measurements. Therefore, a simpler model must be used. To elaborate the simpler model, we will use the performance model described in [4], [5], denoted $f(I_R)$, for all rooftop PV generation, where I_R is the solar irradiance incident on the panel. The performance model, f, depends on the number of panels in series, the number of panels in parallel, and several other

TABLE I
Notation

$E_g(t)$	Net load over $[t-1, t]$ (MW-hr)
$E_i(t)$	Generated solar energy $[t-1, t]$ (MW-hr)
$E_r(t)$	Total energy demand over $[t-1, t]$ (MW-hr)
$I_R(t)$	Average net solar irradiance over $[t-1, t]$ (W/m^2)
$f(I_R)$	Performance model for solar generation
$\epsilon(t)$	Error term over $[t-1, t]$

parameters. However, from the functional form of f in [4], [5], for most operating conditions and net solar penetrations, $f(I_R) \approx d' I_R^{1.21}$, where d' accounts for nearly all of the other parameters needed to describe f. Thus, if all the generators experience the same solar penetration, $I_R(t)$, then a reasonable aggregate model for $E_i(t)$ would be $E_i(t) = d(f \circ I_R)(t)$, where d is an unknown parameter that reflects the penetration level of solar energy and $(f \circ I_R)(t) = f(I_R(t))$. While not all solar generators experience the same irradiance (and we use varying irradiances in our experiments), we will see that this aggregate model is adequate for capturing the correct behavior.

B. Well-Posedness Assumptions

At a resolution of 45 minutes with hourly estimates, the nonparametric estimator introduces ill-posedness. Specifically, over the span of one day, our nonparametric estimator would have 64 unknown parameters, and the solar generation component would have one unknown parameter; and these 65 parameters would need to be determined from the 24 measurements available over the span of a day. Such a problem is clearly ill-posed.

We may think that this ill-posedness can be resolved by including more days in our models. For each included day, however, the model would add over 60 more parameters and the number of observations would increase by 24. Hence, we consider other solutions.

First, we can increase the number of observations taken within a day. This would certainly fix several difficulties including some that we will discuss later in this work. At the moment, however, load data is available only at hourly intervals from the grid. Another option is to regularize the estimation by, for example, using a Tikhonov term with prior estimates of the unknown parameters [6]. Here, we assume that the total energy demand at any moment is realized as a mean-zero additive random perturbation of some mean underlying total energy demand, and we regularize this underlying demand by requiring it to be periodic at daily intervals over a three-day period, resulting in a model with 65 unknown parameters and 72 measurements and thus alleviating the ill-posedness.

Under this regularization, the mean total demand is restricted to the space of continuous, periodic functions, while the actual realized total energy demand is *not* required to be periodic. Therefore, the model allows for flexibility in representing the total demand and requires the selection of only one parameter, the resolution level (e.g., 45 minutes). Moreover, because of the periodicity, large deviations in the actual total demand from the mean total demand at the same hour of the day are then a consequence of the error process or, if the irradiance data allows it, can be used to identify the solar generation coefficient. Therefore, this modeling framework offers a flexible representation of the total demand and requires choosing only a temporal scale. This nonparametric bent is likely to increase the robustness of the model.

C. Estimation Problem Formulation

Letting $\epsilon(t)$ represent this additive random perturbation and, with an abuse of notation, letting $E_r(t)$ now represent the underlying total energy demand, we have as our model

$$E_g(t) = E_r(t) - d(f \circ I_R)(t) + \epsilon(t), \tag{2}$$

subject to $E_r(t) + \epsilon(t) \geq E_g(t)$ and $d \geq 0$.

Then, we can formulate estimates for $E_r(t)$ and d by solving the following corresponding constrained least-squares problem,

$$\begin{aligned} \min_{E \in \mathcal{H}_5, d} \quad & \sum_{k=1}^{72} \frac{1}{2} \left[E_g(k) - E(k) + d(f \circ I_R)(k) \right]^2 \\ \text{subject to} \quad & E(k) = E(k+24), \ k = 1, \ldots, 48 \\ & E(k) + \xi(k) \geq E_g(k), \ k = 1, \ldots, 72 \\ & \xi(k) \geq 0, \ k = 1, \ldots, 72 \\ & \sqrt{\sum_{k=1}^{72} E_g(k)^2} \geq \sum_{k=1}^{72} \xi(k)^2 \\ & d \geq 0, \end{aligned} \tag{3}$$

where t is measured in hours and $\mathcal{H}_5$ represents the class of all nonparametric Haar wavelet estimators of resolution corresponding to 45 minutes. We note that we added an additional constraint on the size of the slack, $\xi(k)$, whose purpose is to prevent the estimated values of $E_r(k)$ from deviating appreciably from the constraints imposed by $E_g(k)$.

III. Experimental Setup

We now test our model and estimation procedure using the National Oceanic and Atmospheric Administration's per minute solar penetration data for Bondville, Illinois [7], and Ameren Corporation's Illinois hourly load data [8] from March 31, 2016, to March 30, 2017. We start with an overview of how this data is used, and we then describe the details.

A. Overview

We use the per minute solar penetration data to simulate $E_i(t)$ and to calculate the average hourly solar irradiance, $I_{hour}(t)$. We use the hourly load data as our realized value of total energy demand. Using this total energy demand and the simulated $E_i(t)$, we compute $E_g(t)$. Then, using this $E_g(t)$ and the average hourly solar penetration $I_{hour}(t)$, we solve (3) to estimate $\hat{E}_r(t)$ and $\hat{d}$. Using these estimates, we estimate the energy demand met by solar and compare it with the true value used to generate E_g.

B. Solar Generation Simulation

We start by randomly placing $N = 100$ solar generators (uniformly distributed) in a unit square. For each generator, we randomly assign the number of panels between one and ten. Then, using the per minute solar irradiance data, for each minute we simulate the irradiance at each generator by drawing from a positive-orthant truncated, multivariate normal distribution centered at the solar penetration data at

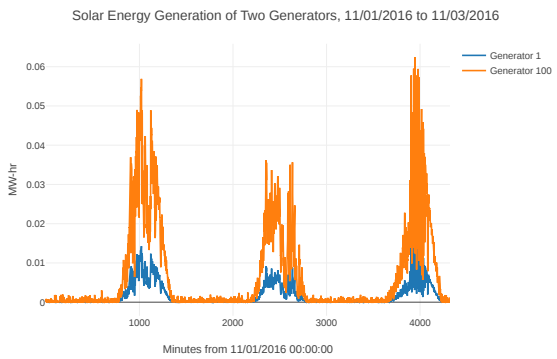


Fig. 1. Energy generated by two simulated solar generators using per minute solar irradiance data from November 1, 2016, to November 3, 2016.

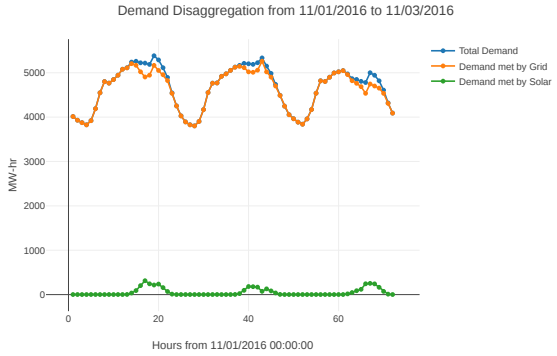


Fig. 2. Disaggregation of the total energy demand into the energy demand met by the grid and the energy demand met by solar generation.

the specified minute and a covariance matrix, C, whose entries are given by $C(i,j) = \sigma^2 \exp\left(\log(\rho)\left\|x(i) - x(j)\right\|_2\right)$, where $x(i)$ is the position of the ith solar generator and $\rho \in (0,1)$. To enforce a sufficient amount of variability and dependence between the generators, we choose $\sigma^2 = 200$ and $\rho = 0.9$. Figure 1 shows the energy generated by two simulated generators using per minute solar penetration data from November 1, 2016 to November 3, 2016.

C. Grid Demand Simulation

Using the load data as the total energy demand and using the total of the energy generated by the N simulated solar generators, we can simulate the energy demand met by the grid by taking the difference of these two quantities. See Figure 2 for an example with 100 generators operating between November 1, 2016, and November 3, 2016.

D. Estimation

For the estimation, we use solar and load data in four specific (weekday) time periods, where each is in the middle of the four seasons: May 5, 2016, to May 7, 2016 (spring); August 1, 2016, to August 3, 2016 (summer); November 1, 2016, to November 3, 2016 (autumn); and January 31, 2017, to February 2, 2017 (winter). Using this data, we simulate solar generation and grid demand as described above.

Using the simulated hourly values of $E_g(t)$ and the average hourly values of $I_{hour}(t)$, we computed the estimators of total energy demand and the solar generation scaling constant, $\hat{E}_r$ and $\hat{d}$, by solving an equivalent formulation of (3) that was amenable to the Julia programming language's "IPOPT" package [9]. Using these values, we estimated the proportion of energy demand met by solar generation, $\theta(t) = E_i(t)/E_r(t)$, and we computed a bias correction and 15% and 85% quantiles using the Jackknife procedure [10].

IV. Results for Hour-Interval Data

Figure 3 reports the estimated proportions compared with the true proportion of energy demand met by solar. We make several comments. First, the 15% to 85% quantile estimates are able capture the real scenario. However, there are winter quantile estimates that are negative. This is a consequence of the Jackknife procedure, which computes the parameter estimates by leaving out one observation at a time. Consequently, when we do the calculation for the proportion, we include the removed observation with the Jackknife parameter estimates, which allows for possibly negative proportions to occur. Although these values are not physically realistic and we could simply truncate at zero, we included the actual behavior for transparency.

Second, the median estimates are less accurate in the summer and winter than in the spring and autumn. We believe that this is the result of the high correlation (both positive and negative) between the total energy demand and net solar irradiance on the days used in the estimation procedure. When such a correlation exists, the two components of the estimation become colinear, in a sense, and such colinearity allows for the effects to be captured by either estimators. On the other hand, for the days used in the spring and autumn estimation, little correlation exists between the total energy demand and net solar irradiance, which is why we perform so well.

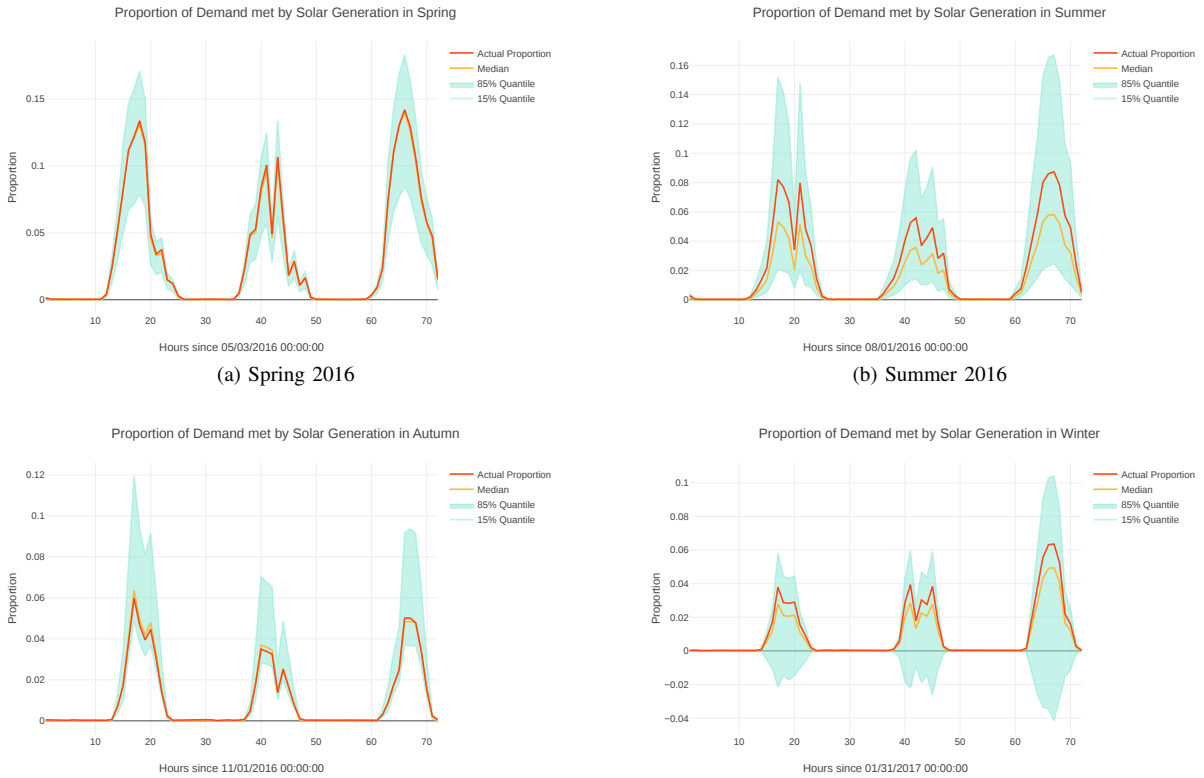
Table II reports correlations coefficients between the total energy demand and solar irradiance over the entirety of the seasons (not just the days considered in the experiment). Thus, in general, for the current rate at which load observations are made public, this procedure will be most accurate for autumn. However, if we are able to collect high-frequency observations, say at the minute scale, we can do much better, as we show in the next section.

V. Using Higher-Frequency Load Observations

Here, we explore our ability to estimate the proportion of energy demand met by solar generation when given per minute measurements of the net load on the grid. We follow the simulation procedure from the data sources above with one modification: in order to compute per minute load data from

TABLE II
Seasonal Load and Solar Penetration Correlation

Winter	Spring	Summer	Autumn
-0.21	0.25	0.5	0.06



(a) Spring 2016 (b) Summer 2016

(c) Autumn 2016 (d) Winter 2017

Fig. 3. Comparison of the quantile estimates for the proportion of demand met by solar generation.

the Ameren Illinois hourly load data, we interpolate between hourly points using a Brownian bridge. Figures 4a and 4b give examples of the data-based simulated disaggregation for two cases of standard deviations used in the Brownian bridge interpolation: (1) a standard deviation on the order of the variation of the solar generation and (2) a standard deviation that is one-third this value. Again, we note that it is unrealistic to consider the total demand and the net load on the grid to have so much noise; doing so, however, allows us to probe the impact of this uncertainty on the estimates and quantiles.

We now fit the same model described in III with two exceptions. First, we increase the nonparametric resolution to a 5.625-minute interval. Second, because the problem is well-posed owing to the per minute observations, we can do away with the periodicity assumption and limit ourselves to a one-day span of observations. Now, the minimization problem is

$$\min_{E\in\mathcal{H}_8,d}\sum_{k=1}^{1440}\frac{1}{2}\left[E_g(k)-E(k)-d(f\circ I_r)(k)\right]^2, \qquad (4)$$

where k is measured in minutes and $\mathcal{H}_8$ represents the class of all nonparametric Haar wavelet estimators of resolution corresponding to the 5.625-minute interval. We use the estimates $\hat{E}_r$ and $\hat{d}$ to compute $\hat{\theta}$. The estimates of θ and the 15% to 85% bootstrap-based quantile intervals (see [10]) are reported in Figure 4.

For the mildly noisy interpolation our estimates of the proportion of demand met by solar, even on a sunny day in the middle of the summer, are excellent—a significant increase in the accuracy of the hour-based estimates and uncertainty bounds. For the moderate noise case, however, our ability to accurately estimate the proportion of energy generated by solar decreases. This result indicates that it is the variability of the solar generation compared with the total energy demand that allows us such a high accuracy in our estimation in the low-noise cases.

VI. Conclusion

In this work we introduced a solar generation estimation methodology that uses a highly flexible, nonparametric model for the total energy demand. Even using only hourly load and irradiance data, our uncertainty bounds are able to capture the true underlying signal. Moreover, at the same measurement rate, we show that we can accurately estimate the underlying signal when the total energy demand and irradiance are uncorrelated, a situation that occurs most frequently in autumn. We also show that we can overcome this challenge with higher frequency load estimates, in which case we can accurately estimate the solar generation at any time of the year. Although such data is not currently available publicly, we expect that it will become public under a high renewable penetration scenario, since it could be market related (note that we do not use voltage or other quantities that would also

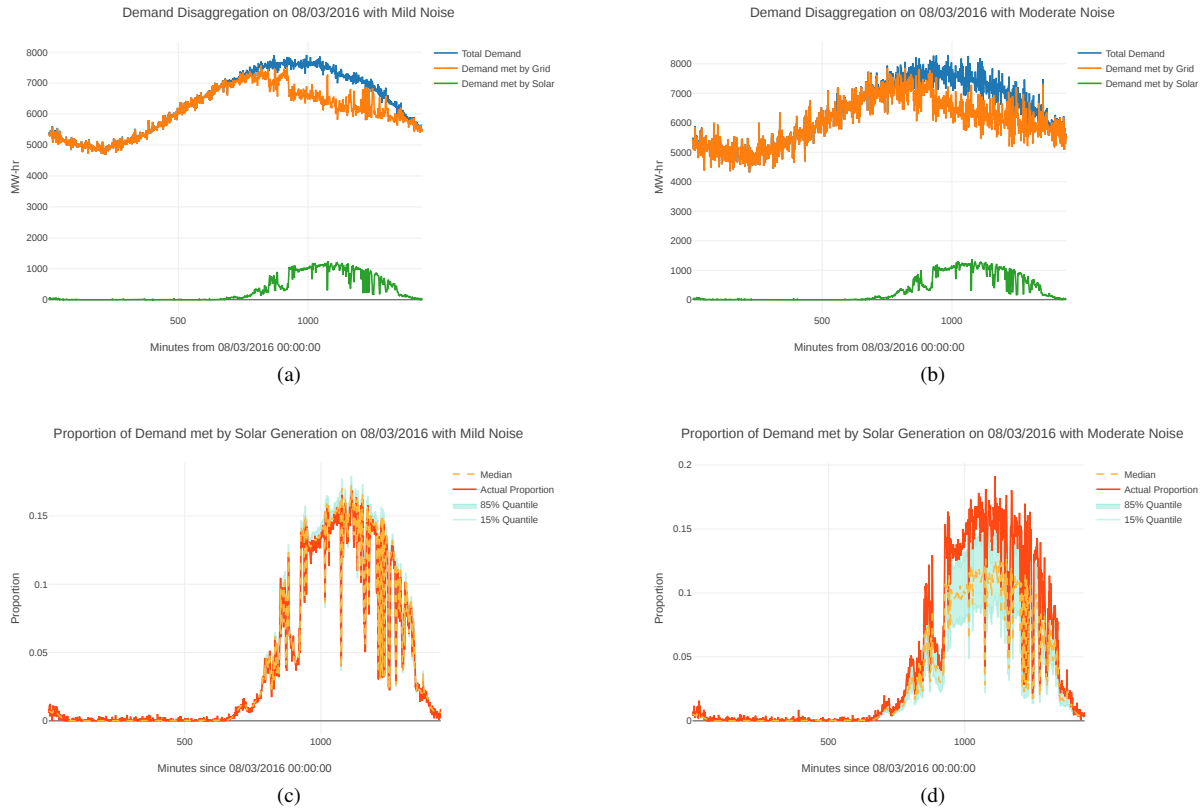


Fig. 4. The top row shows the data-based simulation of energy disaggregation. The bottom row shows the recovery of the proportion of the demand met by solar for the different noise values used in the interpolation.

touch on security issues). Our ability to accurately estimate the quantities of interest relies on the slower dynamics of the total energy demand compared with solar generation.

In future efforts we hope to understand the spatial resolution power of our approach, to improve the inverter model of the solar generation, and to improve our simulation of plane-of-array irradiance to further evaluate the methodology's ability to accurately recover the total energy demand, solar energy generation, and proportion of these two quantities.

ACKNOWLEDGMENTS

This material was based upon work supported by the U.S. Department of Energy, Office of Science, under Contract DE-AC02-06CH11347. V.P. and M. A. acknowledge partial NSF funding through awards FP061151-01-PR and CNS-1545046.

REFERENCES

[1] R. Haaren, M. Morjaria, and V. Fthenakis, "Empirical assessment of short-term variability from utility-scale solar pv plants," *Progress in Photovoltaics: Research and Applications*, vol. 22, no. 5, pp. 548–559, 2014.

[2] E. C. Kara, C. M. Roberts, M. Tabone, L. Alvarez, D. S. Callaway, and E. M. Stewart, "Towards real-time estimation of solar generation from micro-synchrophasor measurements," *arXiv preprint arXiv:1607.02919*, 2016.

[3] L. Wassermann, *All of nonparametric statistics*. Springer Science+ Business Media, New York, 2006.

[4] M. Athari and M. Ardehali, "Operational performance of energy storage as function of electricity prices for on-grid hybrid renewable energy system by optimized fuzzy logic controller," *Renewable Energy*, vol. 85, pp. 890–902, 2016.

[5] W. Zhou, H. Yang, and Z. Fang, "A novel model for photovoltaic array performance prediction," *Applied energy*, vol. 84, no. 12, pp. 1187–1198, 2007.

[6] A. N. Tikhonov, V. I. Arsenin, and F. John, *Solutions of ill-posed problems*. Winston Washington, DC, 1977, vol. 14.

[7] NOAA. (2017) Surfrad daily data for bondville, il. [Online]. Available: ftp://aftp.cmdl.noaa.gov/data/radiation/surfrad

[8] Ameren-Corp. (2017) Actual system load. [Online]. Available: https://www2.ameren.com/RetailEnergy/IcwpGeneral

[9] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017.

[10] B. Efron, *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.