BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. DO NOT EXCEED FIVE PAGES.

NAME: Silverman, Justin D

eRA COMMONS USER NAME (credential, e.g., agency login): JSILVE24

POSITION TITLE: Assistant Professor

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE (if applicable)	START DATE MM/YYYY	COMPLETION DATE MM/YYYY	FIELD OF STUDY
Johns Hopkins University, Baltimore, Maryland	BS	09/2007	05/2011	Physics and Biophysics
Duke University, Durham, North Carolina	PHD	07/2014	05/2019	Computational Biology and Bioinformatics
Duke University, Durham, North Carolina	MD	08/2012	05/2020	N/A

A. Personal Statement

I am a statistician (PhD) and physician (MD). My research focuses on developing robust and efficient statistical methods for the analysis of complex biomedical data. From a methodological perspective I focus on the analysis imperfect data that lead to partially identified statistical models. This includes but is not limited to challenges stemming from unmeasured confounding, informative missingness, and measurement bias. From an applied perspective I have particular expertise in the analysis of non-standard high-dimensional data such as compositional data, functional data, and shape data. I have applied my work widely in fields ranging from genomics to neuroscience.

I have successfully mentored numerous graduate students in a variety of programs ranging from statistics to animal sciences. Within the past year, three of my mentees have been awarded graduate fellowships and/or funding through T32 funded training grants. I have a long-standing collaboration with Dr. Cheng who is co-sponsoring this proposal. Some of our recent work modeling the shape of blood cells imaged by micro-CT was recently published in eLife.

Overall I have the interest, technical expertise, and mentoring experience needed to facilitate the successful completion of the proposed work and assist Andrew in achieving his long-term career goals.

1. Yakovlev,Maksim,A, Liang,Ke,, Zaino,Carolyn,R, Vanselow,Daniel,J, Sugarman,Andrew,L, Lin,Alex,Y, La Riviere,Patrick ,J, Xheng,Yuxi,, Silverman,Justin,D, Leichty,John,C, Huang,Sharon,X, Cheng,Keith,C. Quantitative Geometric Modeling of Blood Cells from X-ray Histotomograms of Whole Zebrafish Larvae. eLife [Preprint]. 2023 May 23. Available from: https://elifesciences.org/reviewed-preprints/89432v1/reviews#tab-content DOI: 10.7554/eLife.89432.1
2. Silverman JD, Roche K, Holmes ZC, David LA, Mukherjee S. Bayesian Multinomial Logistic Normal Models through Marginally Latent Matrix-T Processes. Journal of Machine Learning Research. 2022 February 01; 23(7):1-42. Available from: http://jmlr.org/papers/v23/19-882.html
3. Silverman JD, Hupert N, Washburne AD. Using influenza surveillance networks to estimate state-specific prevalence of SARS-CoV-2 in the United States. Sci Transl Med. 2020 Jul 29;12(554) PubMed Central PMCID: PMC7319260.

4. Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. Elife. 2017 Feb 15;6 PubMed Central PMCID: PMC5328592.

B. Positions and Honors

Positions and Scientific Appointments

2020 -	Assistant Professor, Penn State University, State College, PA

Honors

2022	Dean's Circle of Teaching Excellence, The Pennsylvania State University
2017	Mitchell Meritorious Research Travel Award, Duke University
2017	Best Young Presentation, Compositional Data Analysis Workshop
2011	H. Keffer Hartline Award for Outstanding Scholarship in Biophysics, Johns Hopkins University
2011	Donald E. Kerr Memorial Award for Excellence in Physics, Johns Hopkins University
2011	Phi Beta Kappa, Johns Hopkins University
2010	Barry M. Goldwater Scholarship, Barry Goldwater Foundation
2010	Provost Undergraduate Research Award, Johns Hopkins

C. Contribution to Science

1. **Bayesian count-compositional modeling of sequence count data.** While popular, direct log-ratio transformation of sequence count data are plagued by largely heuristic approaches to handling zero values and an inability to account for variation due to counting also present in these data. To address these limitations I have developed many Bayesian count-compositional models demonstrating how core concepts from the field of compositional data analysis could be applied as different choices of link functions in Bayesian hierarchical multinomial logistic-normal models. Beyond linear models, I have developed longitudinal extensions of these models based upon generalized dynamic linear models, extensions for non-linear regression based upon generalized Gaussian process regression models, and dimensionality reduction methods based upon joint modeling of multi-omic studies. I have also developed popular software for inferring these models including fido and Songbird. Underpinning fido, I developed the theory of Marginally Latent Matrix-t Processes as a means of developing scalable and accurate Posterior estimation for these models that is often 10,000-100,000 times more efficient than MCMC with provable error bounds. These methods have been adopted by numerous research groups and are being used in projects across the NIH and NSF. Beyond methodological development, I have also demonstrated how these models, combined with novel experimental designs can be used to mitigate PCR bias and have established best practices for modeling zeros in sequence count data.

    a. Silverman JD, Roche K, Holmes ZC, David LA, Mukherjee S. Bayesian Multinomial Logistic Normal Models through Marginally Latent Matrix-T Processes. Journal of Machine Learning Research. 2022 February 01; 23(7):1-42. Available from: http://jmlr.org/papers/v23/19-882.html

    b. Silverman JD, Bloom RJ, Jiang S, Durand HK, Dallow E, Mukherjee S, David LA. Measuring and mitigating PCR bias in microbiota datasets. PLoS Comput Biol. 2021 Jul;17(7):e1009113. PubMed Central PMCID: PMC8284789.

    c. Silverman JD, Roche K, Mukherjee S, David LA. Naught all zeros in sequence count data are the same. Comput Struct Biotechnol J. 2020;18:2789-2798. PubMed Central PMCID: PMC7568192.

    d. Silverman JD, Durand HK, Bloom RJ, Mukherjee S, David LA. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. Microbiome. 2018 Nov 12;6(1):202. PubMed Central PMCID: PMC6233358.

2. **Developments in Scale Reliant Inference** I am responsible for the conceptualization and the

theoretical foundations underlying the emerging field of Scale Reliant Inference (SRI). By posing SRI as an estimation problem involving partially identified models I was able to prove the first non-axiomatic limits on the analysis of multivariate survey data where the variation in the sum of each measurement is arbitrary and does not reflect the scale (i.e., size) of the underlying systems being surveyed. With application to the analysis of sequence count data, these results have resolved long-standing debates within the bioinformatics community regarding the extend to which modeling assumptions (e.g., normalizations) can overcome the limitations of the observed data. Through theoretical and empirical studies, I have shown that avoiding a troubling statistical phenomena called unacknowledged bias requires considering uncertainty in those modeling assumptions themselves. To facilitate such analyses I developed a specialized type of Bayesian hierarchical models called Scale Simulation Random Variables (SSRVs). I have shown that by addressing unacknowledged bias SSRVs can often reduce false positive rates while retaining statistical power in a wide range of applications from differential abundance analysis to inter-gene correlation analysis. Numerous groups have started to apply tools and theory from SRI to their studies of sequence count data and SSRVs are now integrated into popular modeling tools such as ALDEx2.

a. Nixon MP, Gloor GB, Silverman JD. Beyond Normalization: Incorporating Scale Uncertainty in Microbiome and Gene Expression Analysis. bioRxiv. 2024 Apr 2; PubMed Central PMCID: PMC11014594.
b. McGovern KC, Nixon MP, Silverman JD. Addressing erroneous scale assumptions in microbe and gene set enrichment analysis. PLoS Comput Biol. 2023 Nov;19(11):e1011659. PubMed Central PMCID: PMC10695402.
c. Nixon,Michelle,P, Letourneau,Jeffrey,, David,Lawrence,A, Lazar,Nicole,A, Mukherjee,Sayan,, Silverman,Justin,D. Scale Reliant Inference. arXiv [Preprint]. 2022 January 10 [revised 2023 February 10]. Available from: https://arxiv.org/abs/2201.03616 DOI: 10.48550/arXiv.2201.03616

3. **Merging phylogenetics and compositional data analysis to enhance the analysis of microbiome data.** In addition to informing the relative abundance of taxa, sequence similarity in microbiome data can be used to estimate the evolutionary relationships between taxa. Just as dogs and wolves are more likely to be found in similar environments utilizing similar resources than dogs and dolphins, evolutionary structure in microbial communities can provide critical insights on the forces that structure these communities. Early in my career I developed the phylogenetic isometric log-ratio transform (PhILR). PhILR uses the phylogenetic relationships to overcome limitations with existing compositional data analysis methods. Under this transform, microbial compositions are projected from the compositional simplex into Real space with coordinates that take on a phylogenetic interpretation: the balance in the abundance between neighboring phylogenetic clades. While this provided new avenues for interpreting these data, this transform itself had a number mathematical advantages over the more standard additive or centered log-ratio transforms at the heart of compositional data analysis. Building on this work I participated in the development of the generalized phylofactorization model which is a compositional factor model with factors constrained by the phylogenetic relationships between microbial taxa. Together these methods have been widely used to identify forces structuring both host-associated and environmental microbial communities. The PhILR transform has also been used by a number of groups as a pre-processing step to improve the performance of black-box machine learning algorithms.

a. Washburne AD, Silverman JD, Morton JT, Becker DJ, Crowley D, Mukherjee S, David LA, Plowright RK. Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data. Ecological monographs. 2019 February 19; 98(2):e01353. DOI: 10.1002/ecm.1353
b. Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. Elife. 2017 Feb 15;6 PubMed Central PMCID: PMC5328592.

c. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ. 2017;5:e2969. PubMed Central PMCID: PMC5345826.

4. **Developed novel surveillance systems for quantifying and tracking COVID-19 burden.** Stating in January 2020, my expertise in multivariate time-series analysis and clinical medicine were called upon to aid in the response to the COVID-19 pandemic. Mounting an effective, early responses to an emerging infectious disease requires an accurate picture of the prevalence of the disease and the geographic spread. Yet this type of tracking is often hindered by resource limitations that hinder the widespread use of specific molecular tests. Early in 2020, we found that the rate at which patients present to their physician complaining of influenza-like symptoms could be used to quantify and track the burden of COVID-19 at the state-level in the United States. Using a large surveillance network designed by the CDC to track influenza (ILINet) we provided some of the earliest evidence that the vast majority of cases were going undetected. Between March 8 and March 28th, we found that approximately one in every 80 cases were identified and included in nationally reported cases counts. This work made national news and led to a number of collaborations with local, state, and national governmental organizations helping to direct the response to the COVID-19 pandemic. Later, as key variants of concern were emerging out of England, South Africa, and Brazil, I developed a number of statistical tools that were used to quantify the relative fitness, transmissible, and burden of these variants. Through this work we provided some of the earliest evidence that the B.1.1.7 variant emerging in England was more transmissible and conferred higher mortality risk than previously dominant the wild type.

   a. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, Pearson CAB, Russell TW, Tully DC, Washburne AD, Wenseleers T, Gimma A, Waites W, Wong KLM, van Zandvoort K, Silverman JD, Diaz-Ordaz K, Keogh R, Eggo RM, Funk S, Jit M, Atkins KE, Edmunds WJ. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science. 2021 Apr 9;372(6538) PubMed Central PMCID: PMC8128288.
   b. Silverman JD, Hupert N, Washburne AD. Using influenza surveillance networks to estimate state-specific prevalence of SARS-CoV-2 in the United States. Sci Transl Med. 2020 Jul 29;12(554) PubMed Central PMCID: PMC7319260.

5. **Open Source Scientific Software.** I have authored and currently maintain 13 open source software packages including multiple software packages for statistical analysis of sequence count data (e.g., PhILR, *RcppHungarian*, *RcppCoDA*, *NVC*, *driver*, and *fido*). I am committed to long-term support and maintenance of my software. I authored *PhILR* at the start of 2016 and continue to routinely collaborate with the community to enhance this software: improving documentation, reviewing pull requests, authoring new features, and fixing bugs. *PhILR* is published on Bioconductor and is downloaded approximately 3000 times per year. To date *PhILR* has been used in numerous research projects including studying the role of gut microbiota in body composition (Ang et al. eLife, 2021) to uncovering the interaction between host genetics and microbiota in autism spectrum disorder (Buffington et al. Cell, 2021), and even inferring changing ecology of human oral microbiota over the past 100,000 years (Yates et al. PNAS, 2021). I am also the author of *RcppHungarian*, a fast C++ header library with bindings to the R programming language for solving minimum cost bipartite matching problems. *RcppHungarian* is published on CRAN and has supported a number of studies including novel methods for large-scale analysis of spectral imaging data (Paradis, International Journal of Applied Earth OBservation and Geoinformation, 2022). One of my post popular software packages, *fido*, is a fast and flexible implementation of a wide range of Bayesian Multinomial Logistic Normal models. While still in a beta release, *fido* has already been applied in a number of studies ranging from identifying the role of short-chain fatty acid production in pediatric obesity (Holmes et al. mBio, 2020) to biomarker discovery in Parkinson's disease (Pereira et al. bioRxiv, 2021). Recently, we even used *fido* quantify the influenza-like illness surge at the start of the COVID-19 epidemic in the United States (Silverman et al. Science Translational Medicine, 2020). Of note, *fido* is more commonly

known as *stray* yet was recently renamed due to a name collision with another package on CRAN.

D. Scholastic Performance