Research Statement

Paxton Turner, Harvard University

Today's ever-growing datasets present formidable challenges to the statistician that arise from high dimensionality, heterogeneity, and the computational costs of processing. My research focuses on addressing these challenges through the development and analysis of statistical procedures for rich models. My goal is to identify mathematical structures underlying machine learning problems and to leverage them to design highly accurate and computationally efficient estimators. Along with high-dimensional statistics, discrete mathematics plays an especially important role in my research; for example in the analysis of count and network datasets as well as through the application of combinatorial tools to inference problems.

Below I discuss my research on data compression [1, 2], hypothesis testing of discrete datasets [3, 4], and covariate balancing in high dimension [5, 6, 7]. In my Ph.D. I examined fundamental trade-offs between data compression and statistical accuracy, focusing on a popular framework known as *coresets* that involves summarizing a larger dataset with a small, representative subsample [1]. During my postdoc, I investigated the detection of latent structure in heterogeneous discrete data, including (i) *testing for diversity* in heteroskedastic count datasets, such as word-document matrices [3], and (ii) *community detection* in degree-hetergeneous networks [4]. A third branch of my research focuses on *balancing covariates* in high-dimensional datasets [5, 6, 7], a combinatorial optimization problem with applications to the design of randomized control trials.

1. Coresets and estimation

A coreset is a small, representative subset of a dataset. Coresets improve the efficiency of data processing by serving as a more tractable proxy for a larger dataset. Inspired by this approach to improving computational efficiency, many recent works investigate running machine learning algorithms on coresets [8, 9, 10, 11, 12]. However, the performance of estimators run on coresets for basic statistical tasks, such as learning an unknown density from observations, is largely unexplored. In [1] we address this gap in understanding by developing a statistical perspective on coreset density estimation. Our results provide (i) a quantitative understanding of how many datapoints are needed in the coreset to attain a desired statistical accuracy, and (ii) a novel coreset construction method that is computationally efficient, resulting in a natural weighted kernel density estimator whose statistical accuracy is near-optimal.

Consider the problem of estimating an unknown probability density function f given observations $\mathcal{D} = \{X_1, \ldots, X_n\}$ sampled from the probability distribution associated to f. A coreset $\mathcal{C}$ is a data-dependent subset of $\mathcal{D}$, and we define a *coreset-based estimator* $\hat{f}_{\mathcal{C}}$ to be an estimator that only depends on the data points in $\mathcal{C}$. Our first contribution characterizes the optimal rate of estimation of smooth densities via coreset-based estimators. For compactly supported d-dimensional smooth densities with β bounded derivatives, we prove that the minimax rate of estimation is $|\mathcal{C}|^{-\beta/d}$, up to logarithmic factors. Moreover, we show that a weighted coreset kernel density estimator of the form

$$\hat{f}_{\mathcal{C}}(y) = \sum_{X_i \in \mathcal{C}} \lambda_i \, K\left(X_i - y\right), \tag{1}$$

is near-optimal, where K is a smoothing kernel with appropriate bandwidth, and the weights λ_i are nonnegative and sum to 1. The weights and coreset in (1) are obtained using Carathéodory's theorem, a classical theorem in discrete geometry, applied to a Fourier embedding of the kernel functions $\{K(X_i - \cdot)\}$ into a finite-dimensional space.

In follow-up work [2], we continue the investigation of data compression for statistical tasks and develop a fast evaluation method for generic nonparametric estimators. Our scheme interpolates a black-box estimator over a combinatorially structured mesh and extends prior works focusing on fast evaluation of kernel density estimators [13, 14, 15, 16].

2. Testing for diversity in count data

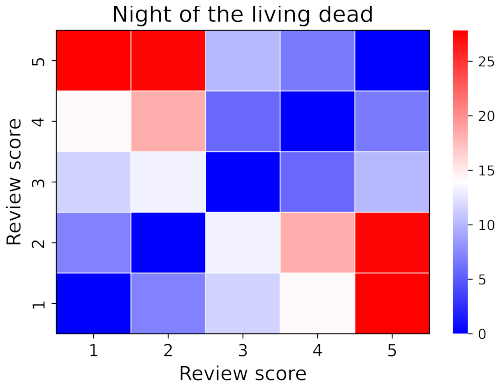
An important task in statistics is to develop methods that perform well in more realistic models extending beyond the simplified i.i.d. setting. During my postdoc I examined two high-dimensional testing problems in heteroskedastic settings involving multinomial [3] and network [4] datasets. For each problem, we identify a simple and practical test statistic that has a tractable parameter-free limiting distribution and prove that it is powerful against a broad class of alternatives. In this section I focus on multinomial testing [3].

Our work on multinomial testing [3] is inspired by questions about quantifying diversity in real-world discrete datasets, such as word counts of text data. For example, do consumers have widely differing reviews about a certain film? We aim to provide a data-driven answer using the word counts of that movie's reviews on Amazon. In [3], we present a general framework for testing the diversity of such types of count datasets, as summarized below.

We model the dataset as independent multinomial observations $X_1, \dots, X_n \in \mathbb{R}^p$ divided into K known *groups*. Here X_i is a p-dimensional vector of counts resulting from N_i random words drawn from a distribution on a dictionary of size p that has probability mass function $\Omega_i \in \mathbb{R}^p$. The *group mean* μ_k is the average probability mass function for the k^{th} group and provides a summary of the distribution on words in that group. Our goal is to test if the K group means are all equal to each other (i.e., $\mu_1 = \cdots = \mu_K$) or not. In the movie review example above, the null hypothesis indicates that the reviews are all similar to each other, while the alternative signifies diversity of the reviews.

This problem involves significant heterogeneity (the number of words N_i and population word frequency vectors Ω_i may vary, even within a group) and high dimensionality (the number of groups K, the number of observations n, and the dictionary size p may grow at different rates). To address these challenges, we propose a moment-based test statistic called the *debiased and length-adjusted variability estimator (DELVE)*.

Under mild assumptions, we prove that DELVE is asymptotically normal under the null hypothesis and that it achieves the optimal detection boundary. The flexibility of our setting allows for a wide array of experiments on real data. We apply our method to a dataset of abstracts in statistics journals [17] and another of Amazon movie reviews [18]. In the figure, we partition the reviews of a classic horror film by review score (which ranges from 1 to 5) and evaluate DELVE with $K = 2$ on the two corpora corresponding to each pair of scores. We observe significant polarization between the 1–2 score reviews and the 3–5 score reviews.



3. Covariate balancing

My works [5, 6] and [7] investigate a combinatorial optimization problem known as *covariate balancing* (commonly known as *discrepancy minimization* in the computer science literature) that involves dividing a collection of vectors $X_1, \dots, X_n \in \mathbb{R}^d$ into two groups

S_1 and S_2 that are well-balanced in the sense that the metric $\|\sum_{i\in S_1} X_i - \sum_{i\in S_2} X_i\|_\infty$ is small. Algorithmic questions in covariate balancing have been the subject of intensive study in theoretical computer science over the past decade (see *e.g.* [19, 20, 21]), and more recently applications of covariate balancing to the design of randomized control trials have been recognized [22, 23, 5, 24]. It is known that if the treatment and control groups balance the covariates of participants, then the two groups are statistically similar to each other, and treatment effects can be more accurately estimated.

Equipped with these motivations, we investigate in [5] a statistical variant of discrepancy minimization where the input vectors $X_1, \ldots, X_n$ are i.i.d. standard Gaussian vectors. We prove that the optimal balance has value $\sqrt{\frac{\pi}{2}} \cdot \sqrt{n}\, 2^{-n/d}$ asymptotically when $n \gg d$. We also develop an algorithm that achieves balance $n^{-(\log n)/d}$, which decays faster than any polynomial and establishes the best known guarantee when $2 \le d = O(1)$. In certain randomized control trial setups, designs based on our results lead to significantly improved inference of treatment effects [25].

In recent work [7], we study the *ellipsoid fitting problem*, which is the task of interpolating i.i.d. standard Gaussian points $v_1, \ldots, v_n \sim N(0, I_d)$ with an ellipsoid. This basic geometric question is related to the semidefinite programming relaxation (SDP) of covariate balancing as well as problems in machine learning, such as independent component analysis and low-rank matrix decompositions [26]. A well-known conjecture of [27] states that ellipsoid fitting is possible when $n \ll d^2/4$ and impossible when $n \gg d^2/4$.[1] In [7] we resolve this conjecture up to logarithmic factors, proving that ellipsoid fitting is possible if $n \le d^2/\mathsf{polylog}(d)$. As a corollary, our ellipsoid fitting result implies that when $n \asymp d$, SDP-based methods are unable to design randomized control trials with highly accurate treatment effect estimators.

References

[1] Paxton Turner, Jingbo Liu, and Philippe Rigollet. A statistical perspective on coreset density estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 2512–2520. PMLR, 2021.

[2] Paxton Turner, Jingbo Liu, and Philippe Rigollet. Efficient interpolation of density estimators. In *International Conference on Artificial Intelligence and Statistics*, pages 2503–2511. PMLR, 2021.

[3] T. Tony Cai, Tracy Ke, and Paxton Turner. Testing high-dimensional multinomials with applications to text analysis. *In submission, available on arXiv*, 2023.

[4] Jiashun Jin, Tracy Ke, Paxton Turner, and Anru Zhang. Phase transition for detecting a small community in a large network. *To appear at International Conference on Learning Representations*, 2023.

[5] Paxton Turner, Raghu Meka, and Philippe Rigollet. Balancing gaussian vectors in high dimension. In *Conference on Learning Theory*, pages 3455–3486. PMLR, 2020.

[6] Sinho Chewi, Patrik Gerber, Philippe Rigollet, and Paxton Turner. Gaussian discrepancy: A probabilistic relaxation of vector balancing. *Discrete Applied Mathematics*, 322:123–141, 2022.

[7] Aaron Potechin, Paxton Turner, Prayaag Venkat, and Alex Wein. Near-optimal fitting of ellipsoids to random points. *In submission, available on arXiv*, 2022.

[8] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the*

[1] A straightforward dimension-counting argument establishes impossibility when $n \gg d^2/2$, and this is the best known lower bound [27].

twenty-fourth annual ACM-SIAM symposium on Discrete algorithms, pages 1434–1453. SIAM, 2013.

[9] Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Advances in Neural Information Processing Systems*, pages 4080–4088, 2016.

[10] Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1975–1993, Phoenix, USA, 25–28 Jun 2019. PMLR.

[11] Jeff M. Phillips and Wai Ming Tai. Near-optimal coresets of kernel density estimates. In *34th International Symposium on Computational Geometry, SoCG 2018, June 11-14, 2018, Budapest, Hungary*, pages 66:1–66:13, 2018.

[12] Raaz Dwivedi and Lester Mackey. Kernel thinning. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1753–1753. PMLR, 15–19 Aug 2021.

[13] Leslie Greengard and John Strain. The fast gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94, 1991.

[14] Moses Charikar and Paris Siminelakis. Hashing-based-estimators for kernel density in high dimensions. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1032–1043. IEEE, 2017.

[15] David W Scott and Simon J Sheather. Kernel density estimation with binned data. *Communications in Statistics-Theory and Methods*, 14(6):1353–1359, 1985.

[16] M Chris Jones. Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, 84(407):733–741, 1989.

[17] Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

[18] Dharmendra Maurya. Web data: Amazon movie reviews. https://www.kaggle.com/datasets/dm4006/amazon-movie-reviews, 2018.

[19] Nikhil Bansal. Constructive algorithms for discrepancy minimization. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 3–10, Washington, DC, USA, 2010. IEEE Computer Society.

[20] Shachar Lovett and Raghu Meka. Constructive discrepancy minimization by walking on the edges. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, FOCS '12, pages 61–67, Washington, DC, USA, 2012. IEEE Computer Society.

[21] Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The gram-schmidt walk: a cure for the banaszczyk blues. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 587–597, 2018.

[22] Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112, 2018.

[23] A M Krieger, D Azriel, and A Kapelner. Nearly random designs with greatly improved balance. *Biometrika*, 106(3):695–701, 05 2019.

[24] Christopher Harshaw, Fredrik Sävje, Daniel Spielman, and Peng Zhang. Balancing covariates in randomized experiments using the gram–schmidt walk. *CoRR*, abs/1911.03071, 2019.

[25] Paxton Mark Turner. *Combinatorial Methods in Statistics*. PhD thesis, Massachusetts Institute of Technology, 2021.

[26] James Saunderson, Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012.

[27] James Saunderson, Pablo A Parrilo, and Alan S Willsky. Diagonal and low-rank decompositions and fitting ellipsoids to random points. In *52nd IEEE Conference on Decision and Control*, pages 6031–6036. IEEE, 2013.