Testing and Understanding the Jackknife Approach to Mutual Information Estimation

Grishma Mainali

This manuscript was compiled on June 14, 2025

Estimating Mutual Information (MI) from limited data remains a recurrent challenge in information theory and practical data science. Zeng, Xia, and Tong's 2018 PNAS publication offered a bias-corrected estimator employing the Jackknife resampling method, which served as the foundation for this study. We developed this estimator in Python and tested its performance on synthetic datasets with varied dependency strengths, sample sizes, and noise levels. The Jackknife MI estimates were compared to Scikit-learn's built-in MI estimator and a manually created histogram-based MI approach. We also investigated multivariate dependence estimation using Jackknife MI and compared its effectiveness to bootstrap-based resampling. In addition, we looked into its potential as an alternative to Wasserstein distance in noise-sensitive and sample-sized scenarios. Our findings show that Jackknife MI consistently produces more stable and less biased estimates in low-sample settings and across a variety of dependency architectures. These findings lend support to the usage of Jackknife MI in feature selection, dependency detection, and learning tasks that require accurate information estimation.

Mutual Information | Jackknife Resampling | Wasserstein Distance |

Significance Statement

Mutual Information (MI) is an important measure in statistics and machine learning for discovering interdependence between variables. However, estimating MI from limited samples remains a significant difficulty due to inherent bias. This study extends and evaluates the Jackknife-based technique developed by Zeng, Xia, and Tong across a variety of simulated dependence structures, sample sizes, and noise levels. In addition to comparing Jackknife MI to standard estimators and Scikit-learn's implementation, we evaluate its performance in multivariate scenarios and against the Wasserstein distance, a popular alternative that is known to decline in reliability with small sample sizes. Our data reveal that Jackknife MI produces strong and consistent estimates, even in low-sample regimes, indicating its value as a reliable tool for feature selection, dependence detection, and sample-sensitive tasks.

Mutual Information (MI) is an important concept in information theory that quantifies the dependence of random variables. MI, unlike correlation coefficients, can identify both linear and nonlinear correlations, making it a useful tool in a variety of fields including neurology, feature selection, statistical learning, and complex systems modeling(1).

Despite its theoretical appeal, effective estimation of MI from limited datasets remains a difficult task due to its vulnerability to sampling variability and bias. Common estimate methods, such as histogram-based approaches and kernel density estimators, frequently exhibit substantial bias and instability when applied to small or noisy datasets. While k-nearest neighbor (KNN) and neural network-based approaches have improved, they are still prone to sample limits, have significant processing costs, and require complicated tuning(2).

To address these constraints, this study focuses on a resampling-based correction method called Jackknife resampling for estimating MI. Building on Zeng, Xia, and Tong's (3) technique, we developed a Jackknife-corrected MI estimator in Python and assessed its performance on a variety of synthetic datasets. These datasets differed in dependency strength, noise levels, and sample sizes, offering a demanding test ground for estimator reliability.

We compared the Jackknife MI estimator against two baselines: Scikit-learn's built-in MI function, which calculates entropy using nearest neighbors, and a customized histogram-based MI estimator. Further research looked into the Jackknife method's suitability for multivariate dependence estimation and its effectiveness against bootstrap resampling techniques. We also assessed the estimator's performance as an alternative to the Wasserstein distance, which has gained favor as a robust measure of statistical dissimilarity but suffers in low-sample settings.

The experimental results show that the Jackknife MI estimator generates more stable and less biased estimates than previous approaches, especially in low-sample and noise-prone

circumstances. This robustness shows that Jackknife-based MI estimation can be used as a dependable tool for data-driven activities requiring precise dependency quantification.

Literature Review

Mutual Information (MI) is a fundamental notion in information theory that describes the amount of information shared by two or more random variables. MI was first presented by Claude Shannon in his fundamental 1948 work on entropy and communication theory, and it has since been central to fields such as as statistical learning, neurology, computational biology, and machine learning. Unlike correlation-based measures, which only capture linear associations, MI takes into account both linear and nonlinear dependencies, giving it a more comprehensive and robust measure of statistical association(4).

Conventional MI Estimation Techniques. Estimating MI from actual data is a long-standing problem, especially in finite-sample scenarios. Classical methods include histogram-based estimators, which divide continuous variables into bins and compute joint and marginal entropies. Although straightforward and simple to use, these estimators are sensitive to bin size and can introduce significant bias, particularly when data is sparse. To address these constraints, Kernel Density Estimation (KDE) approaches were developed to estimate probability density curves more smoothly(5). However, KDE approaches are computationally intensive and strongly reliant on bandwidth adjustment, making them unsuitable for high-dimensional or noisy datasets(6).

Kraskov (2) presented the k-nearest neighbors (KNN) technique, which significantly improved MI estimation by responding to local data density rather than relying on explicit density estimates. This method is still commonly used because of its capacity to handle continuous variables in intermediate dimensions. However, it still suffers in very high-dimensional environments and in the presence of noise, making neighborhood estimates less stable(6).

Recent attempts have also investigated MI estimation using neural networks, with the goal of maximizing variational lower bounds of MI. While these deep learning-based algorithms show promise in complicated circumstances, they are prone to overfitting, require significant processing resources, and are extremely sensitive to hyperparameter tuning(7).

Bias Correction and Resampling Approaches. Because MI estimators are biased in small samples, resampling methods have been proposed as correction measures. The bootstrap method, which uses random sampling with replacement, is often used to calculate estimator variance and confidence intervals. Although powerful, the bootstrap's performance varies greatly depending on the data structure and may introduce its own sampling noise(8).

The Jackknife resampling method, originally proposed by Quenouille and then expanded by Tukey in the mid-twentieth century, provides a computationally easier option(9)(10). It systematically removes one observation at a time and recalculates the estimator for all potential subsamples. The data is then aggregated to decrease bias and evaluate variability. The Jackknife test is non-parametric, simple to use, and especially useful in small-sample circumstances.

Zeng, Xia, and Tong introduced a Jackknife-based MI estimator that directly overcomes the bias inherent in entropy-based MI estimations(3). Their methodology uses leave-one-out entropy estimates to correct bias in both marginal and joint entropy components, resulting in a theoretically valid and experimentally effective solution. Despite its potential, the approach has not been fully tested over a range of sample sizes, noise levels, and dependency structures, a gap that this study seeks to fill.

Wasserstein Distance and Alternative Measures. In recent years, optimal transport-based metrics, such as the Wasserstein distance, have grown in prominence as measures of statistical dissimilarity(11). Unlike MI, Wasserstein distance gives a geometric description of the distance between probability distributions and is especially useful when comparing empirical distributions(11). However, its reliability can suffer in low-sample situations, and it may fail to identify certain types of dependency where MI is still sensitive. We have also tried to directly compared MI-based measures to Wasserstein distance in terms of robustness and stability under practical restrictions such as sample noise and large variation.

Methodology

This research project used simulation-based experimental methods to assess the effectiveness of Jackknife resampling Mutual Information (MI) estimate. The approach was broken down into four major phases: estimator implementation, dataset production, comparative evaluation, and expansion to multivariate and alternative measures.

Estimator Implementation. The main focus of this study was to develop a Jackknife-corrected MI estimator in Python, based on Zeng, Xia, and Tong (3) technique. Mutual information between two random variables, X and Y, is defined as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \qquad [1]$$

where $H(\cdot)$ represents the Shannon entropy. To estimate the entropies, we used histogram-based probability estimates and applied Jackknife resampling to reduce bias. Specifically, leave-one-out entropy values were computed for each sample and aggregated to obtain bias-corrected estimates of the marginal and joint entropies. These corrected entropies were then used to derive the final MI score. All estimators were implemented using `NumPy` and `SciPy` to ensure numerical efficiency and reproducibility.

Synthetic Dataset Generation. To systematically evaluate estimator performance under known conditions, we generated synthetic bivariate datasets with well-defined dependency structures. Specifically, three categories were modeled: (i) *independent* variables, drawn from uniform distributions with no mutual interaction; (ii) *weak dependencies*, created through linear or nonlinear functional mappings perturbed by noise; and (iii) *strong dependencies*, based on deterministic or near-deterministic transformations (e.g., $Y = X$, $Y = X^2 + \epsilon$).

Each experimental condition was tested at varying sample sizes (ranging from $n = 50$ to $n = 1000$), and additive Gaussian noise was introduced to simulate real-world measurement variability. Multiple trials were conducted under each scenario to account for the stochastic nature of data sampling and ensure statistical robustness in the evaluation.

Comparative Evaluation. To ensure a fair comparison across methods, all MI estimators were evaluated using identical synthetic datasets under varying dependency structures. Key performance metrics included: (i) *bias*, defined as the deviation from known theoretical MI values; (ii) *variance*, capturing sensitivity across repeated trials; and (iii) *noise stability*, reflecting robustness to additive Gaussian noise. These metrics allowed for a comprehensive assessment of estimator reliability in practical, data-limited conditions.

Multivariate Extension. We extended the Jackknife MI estimator to handle multivariate settings by computing joint entropy terms across multiple input variables and their mutual interaction with an output. This involved applying the same leave-one-out strategy across subsets of variables to obtain corrected joint MI estimates. The consistency and accuracy of the estimator were tested on controlled synthetic datasets with predefined interaction structures, enabling ground-truth validation.

Bootstrap and Wasserstein Comparisons. To contextualize the effectiveness of the Jackknife MI estimator, we compared its performance to two alternatives: the Bootstrap MI estimator, which employs random resampling with replacement to reduce estimator variability, and the Wasserstein distance, an optimal transport-based metric for comparing empirical distributions. All methods were applied to identical datasets across a range of noise levels and sample sizes. The comparison focused on each method's ability to detect statistical dependence, particularly in challenging scenarios involving small sample sizes and nonlinear or noisy dependencies.

Results and Analysis

This section summarizes the findings from a series of experiments that assessed the performance of Jackknife-based Mutual Information (MI) estimation under a variety of dependence structures, across multiple sample sizes, and in comparison to alternative estimation methods including naïve entropy difference, histogram-based, Scikit-learn's estimator, Bootstrap-based MI, and the Wasserstein distance.

Performance Across Dependency Structures. To evaluate estimator accuracy under varying dependency, synthetic datasets were created for five types of variable relationships: independent, perfect, linear with noise, nonlinear, and partially dependent situations. Each setup employed 200 samples across ten trials. All estimators accurately returned low MI (< 0.03) for the independent scenario. In the perfect dependency situation, naïve and histogram estimators produced values near 1.99. However, Jackknife lowered this to approximately 0.997, illustrating its bias correction. In noisy linear relationships, naive estimates ranged from 1.5 to 1.56, while Jackknife consistently adjusted this to approximately 0.778. Similarly, for nonlinear transformations (e.g., $Y = X^2 + \epsilon$), naive estimates peaked around 2.52, while Jackknife yielded a more conservative at approximately 1.26. These findings support Jackknife's capacity to reduce inflation in biased estimators, especially in noisy or small-sample circumstances.

Bias Quantification Against Theoretical MI. To assess estimator bias relative to known ground truth values, two controlled scenarios were analyzed: independent variables (true MI = 0 bits) and perfect dependency (true MI = $\log_2(4) = 2.0$ bits). Over 30 trials, estimators were evaluated based on mean error and standard deviation. While all estimators somewhat overstated MI under independence, Jackknife exhibited the smallest bias. In perfect dependency, most estimators were accurate, but Jackknife somewhat undercorrected (bias ≈ -1.0), demonstrating its conservative character in highly dependent contexts. These findings indicate that Jackknife is particularly good at detecting weak or moderate dependency, but may overcorrect when real MI is significant.

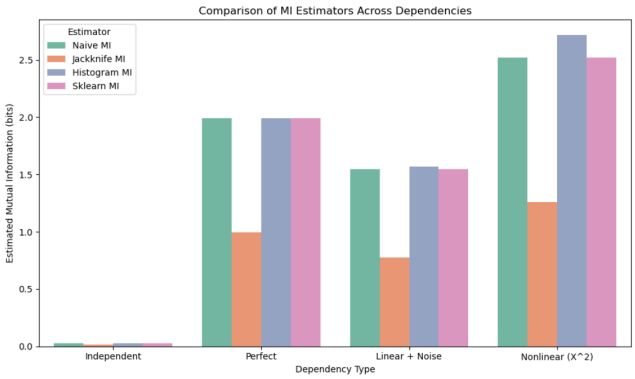

Fig. 1. Sample mutual information comparison across estimators.

Sensitivity to Sample Size. MI estimators were tested on datasets ranging from 50 to 1000 samples, assuming completely dependent data with a theoretical MI of 2.0 bits. At small sample sizes, all estimators underestimated MI, with histogram-based estimates having the lowest bias. Jackknife MI showed higher variability at smaller sizes but improved significantly with more data. At moderate-to-large sample sizes ($n = 500$–1000), all estimators converged, with Jackknife exhibiting minimal bias and variance. These findings confirm that estimator stability and accuracy improve with sample size, and Jackknife is more dependable in bigger datasets.

Comparison with Bootstrap Resampling. Jackknife was compared to bootstrap MI estimation under four different dependent circumstances. For independent data, Jackknife frequently outperforms bootstrap in terms of bias and mean squared error (MSE), particularly for small sample sizes. In perfect dependency scenarios, bootstrap outperformed Jackknife, which had somewhat greater underestimate. Jackknife produced lower bias and MSE in noisy linear and nonlinear dependencies in the majority of cases. Although both techniques improved with higher sample sizes, Jackknife converged faster and was more stable, especially for weak to moderate MI values.

Multivariate Dependency Estimation. The Jackknife MI estimator was expanded to multivariate situations by determining the relationship between a multivariate input vector $X = [X_1, X_2]$ and an output variable Y. Across six multivariate configurations, from clear dependence to complete independence, the estimator displayed trustworthy performance: In dependent circumstances (e.g., $Y = X_1 + X_2$), MI was consistently high (2.3 bits) with minimal fluctuation. MI remained robust (1.73 bits) with realistic and noisy inputs, suggesting tolerance to partial randomness. In entirely independent

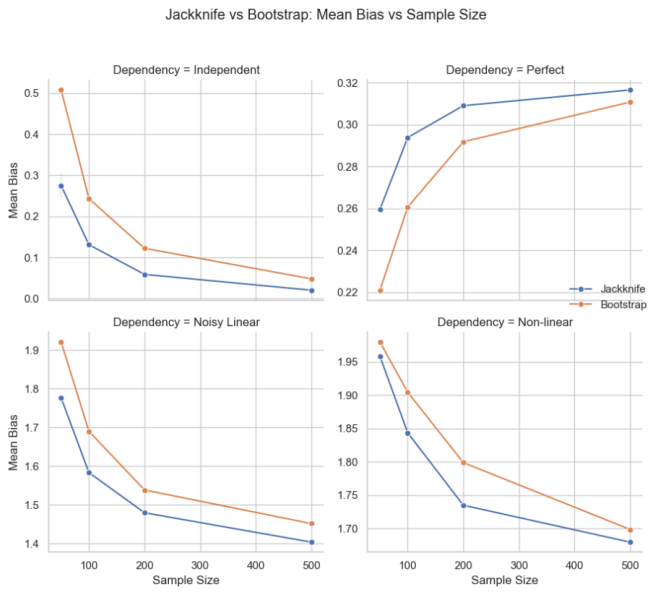

Fig. 2. Jackknife vs Bootstrap performance comparison

data, MI dropped with sample size, from 0.76 at 100 samples to 0.03 at 2000 samples. Jackknife provided moderate MI (1.1 bits) for weak dependencies, accurately capturing subtle patterns. MI attained approximately 4.57 bits in non-linear settings using trigonometric/log transformations, indicating the method's ability to record complex interactions. Noisy linear dependencies resulted in MI 1.98 bits, with Jackknife estimates indicating little volatility but general stability.

These findings verify the Jackknife estimator's suitability for high-dimensional and real-world data, where dependencies are frequently complex, noisy, or nonlinear.

Comparative Analysis: Jackknife MI vs. Wasserstein Distance. To further understand the effectiveness of the Jackknife Mutual Information (MI) estimator, we extended our research by comparing it to the Wasserstein Distance, a well-known metric from optimal transport theory that is frequently used to quantify differences between probability distributions. While the Wasserstein Distance has grown in popularity due to its intuitive geometric interpretation and stability in a variety of statistical applications, it is known to be sample size sensitive, particularly when data is limited or noisy. In contrast, Jackknife MI has demonstrated theoretical advantages in small-sample settings due to bias correction via resampling. Motivated by this discrepancy, we carried out a series of controlled tests to compare both techniques at different noise levels, sample sizes, and dependency structures. The goal was to see how well each method captures significant correlations between variables and to evaluate their strengths and weaknesses in real-world data sets.

Performance Across Noise Levels In a controlled bivariate situation, increasing Gaussian noise was introduced into a dependent relationship between variables X and Y. As noise grew, Jackknife MI values decreased consistently from 3.15 to around 0.39 bits, showing weaker reliance, whereas Wasserstein Distance increased monotonically, indicating divergence in the marginal distributions. Notably, Jackknife MI accurately caught the decline of dependency with high sensitivity and low variance, but WD failed to distinguish between structural dependency and simple distributional mismatch.

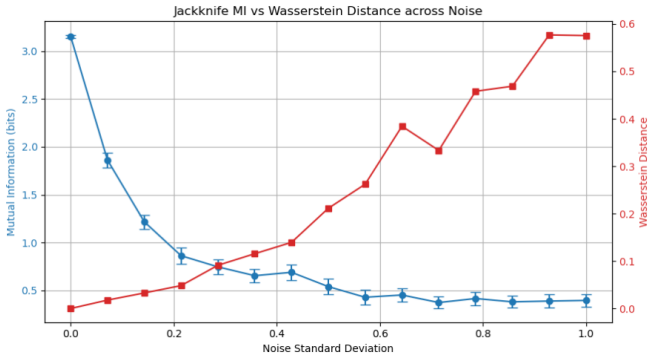

Fig. 3. Jackknife vs Wasserstein across Noise levels

Sensitivity to Sample Size We then investigated estimator robustness across a wide range of sample sizes (20–1000). The Jackknife MI increased gradually from 4.2 to 9.96 bits, appropriately indicating greater dependency signals with more data. In contrast, Wasserstein Distance fluctuated and remained unstable at reduced sample numbers, even under established dependency, indicating its unreliability in data-poor conditions. Jackknife MI exhibited low variation, demonstrating its robustness to small-sample estimation.

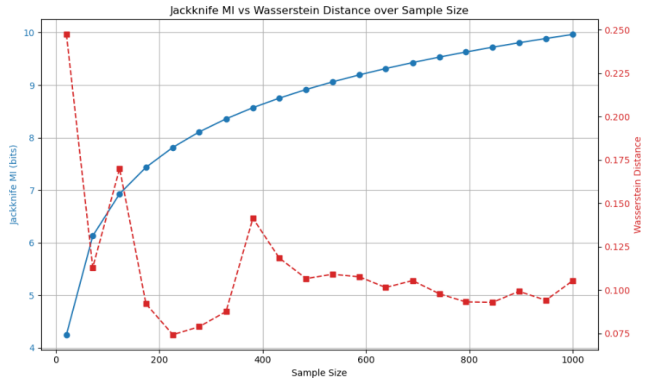

Fig. 4. Jackknife vs Wasserstein across Sample sizes

When WD Detects Shape, but MI Detects Dependency Three experiments were designed in which the variables were independent but had significantly different distribution shapes:
Mean-shifted distributions (WD = 2.03, MI ≈ 0.11)
Different Variance: WD = 3.03, MI = 0.12.
Gaussian vs. Uniform: WD = 0.66, MI = 0.11.

In all situations, Jackknife MI properly identified the absence of dependency, but WD reported a substantial distance due to form differences. These findings demonstrate that WD is appropriate for distribution comparison but not for dependence discovery.

When MI Captures Dependency, but WD Does Not

In contrast, two investigations indicated that the Wasserstein Distance fails to capture meaningful relationships:
Nonlinear dependency: $Y = sin(X) + \epsilon \rightarrow$ MI = 1.44, WD = 3.14

Threshold Function: $Y = 1if X > 0.7$; otherwise, 0: MI = 0.88, WD = 0.29.

In all situations, Jackknife MI accurately assessed the reliance despite distribution mismatch, whereas Wasserstein Distance was unresponsive to the underlying link.

Perfect Dependency Case Finally, in a binary situation with Y=X, Jackknife MI approached the theoretical maximum of one bit, while Wasserstein Distance decreased to zero, confirming both estimators under perfect conditions. However, only MI offers meaningful interpretation in dependency modeling scenarios.

Feature Selection via Jackknife MI To test the effectiveness of Jackknife MI for feature selection, we used synthetic datasets in which the target variable was solely dependent on a subset of input features. Jackknife MI correctly recognized useful features, while assigning near-zero MI to irrelevant noise features. Notably, MI standard error was higher for more informative characteristics, indicating greater sensitivity to their fluctuation. Jackknife MI discovered common information between redundant features, indicating the need for extra filtering after MI selection.

Discussion

The research findings in this paper provide crucial insights on the performance and reliability of Jackknife-based Mutual Information estimate. Across a wide range of dependency scenarios, from perfect linear relationships to noisy, nonlinear, and threshold-based dependencies, Jackknife MI showed strong consistency in detecting underlying statistical relationships, even in cases with small sample sizes or complex distributions.

Our findings highlight Jackknife MI's robustness in the presence of noise and limited datasets. While traditional estimators suffer from excessive variance or bias in such situations, the Jackknife technique efficiently reduces overestimation by leave-one-out resampling. This makes it particularly reliable in situations where complicated dependencies could normally be obscured by sample noise.

In contrast, the Wasserstein Distance, while well-known for characterizing distributional differences, had difficulties when applied to dependency estimating tasks. Our comparison research revealed that Wasserstein Distance can produce high values even in the absence of statistical dependence—particularly when distributions differ in shape, location, or variance but have no relevant informational overlap. These data emphasize a key difference between the two metrics: Wasserstein Distance can detect differences in distributional shape or support, making it useful in tasks like generative model evaluation and domain shift detection. However, it lacks the sensitivity required to detect information-theoretic links, which Jackknife MI is specifically designed to quantify.

Despite its strong performance, one major limitation that we witnessed with Jackknife MI was in circumstances of perfect dependency with small discrete domains, where Jackknife may overcorrect and underestimate the genuine MI value, as demonstrated by our bias analysis. However, in general, our study supports the use of Jackknife MI as a robust, bias-corrected alternative to traditional estimators, especially in small-sample or noisy settings. Further research is needed to investigate hybrid techniques, enhance computational efficiency, and test findings using real-world, high-dimensional datasets.

Conclusion

This research assessed the efficiency of Jackknife-based Mutual Information (MI) estimation in quantifying variable dependency under a variety of scenarios and compared it to the Wasserstein Distance, a prominent measure for comparing probability distributions. Through systematic trials with synthetic data, different noise levels, sample sizes, and dependency structures, we demonstrated that Jackknife MI consistently gives a more stable and interpretable measure of reliance, particularly in noisy or small-sample scenarios.

Jackknife MI proved especially useful in situations involving nonlinear, or threshold-based interactions, precisely characterizing informational reliance that Wasserstein Distance failed to discover. While the Wasserstein Distance correctly recognized variations in distributional shape or spread, it frequently provided misleadingly high results when the variables were statistically independent. This contrast emphasizes an important conclusion: Wasserstein Distance and Mutual Information address fundamentally different questions: distributional difference versus informational dependence, and hence, should be used accordingly.

Overall, this research highlights the usefulness of resampling-based information estimators and provides a good foundation for using Jackknife MI in applications that need precise, noise-resistant dependency analysis.

Data Archival. This report used synthetic datasets generated programmatically to simulate various dependency structures. The Python code and data generation scripts developed for this research are openly available on GitHub at: https://github.com/WSU-Data-Science/project-IMGrishma17.git.

ACKNOWLEDGMENTS. This project was carried out as part of the author's postgraduate coursework project. The author gratefully acknowledges the supervision and insightful feedback provided by Prof. Paul Hurley throughout the research process.

1. W Li, Mutual information functions versus correlation functions. *J. Stat. Phys.* **60**, 823–837 (1990).
2. A Kraskov, H Stögbauer, P Grassberger, Estimating mutual information. *Phys. review. E, Stat. nonlinear, soft matter physics* **69**, 066138 (2004).
3. G Zeng, Y Xia, X Tong, A jackknife approach to mutual information estimation. *Entropy* **20**, 828 (2018).
4. G Zeng, A unified definition of mutual information with applications in machine learning. *Math. Probl. Eng.* **2015** (2015).
5. YI Moon, B Rajagopalan, U Lall, Estimation of mutual information using kernel density estimators. *Phys. review. E, Stat. physics, plasmas, fluids, related interdisciplinary topics* **52**, 2318–2321 (1995).
6. L Paninski, Estimation of entropy and mutual information. *Neural Comput.* **15**, 1191–1253 (2003).
7. B Poole, S Ozair, A Oord, A Alemi, G Tucker, On variational bounds of mutual information (2019).
8. J Beirlant, E Dudewicz, L Gyor, E Meulen, Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.* **6** (1997).
9. M Quenouille, *Notes on Bias in Estimation.* pp. 309–316 (2001).
10. J Tukey, Bias and confidence in not quite large samples. *Annals Math. Stat.* **29**, 614 (1958).
11. G Peyré, M Cuturi, *Computational Optimal Transport.* (2019).