Estimating mutual information under measurement error

Cong Ma[1] and Carl Kingsford*[1]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University,
5000 Forbes Ave., Pittsburgh, PA

November 22, 2019

Abstract

Mutual information is widely used to characterize dependence between biological signals, such as co-expression between genes or co-evolution between amino acids. However, measurement error of the biological signals is rarely considered in estimating mutual information. Measurement error is widespread and non-negligible in some cases. As a result, the distribution of the signals is blurred, and the mutual information may be biased when estimated using the blurred measurements. We derive a corrected estimator for mutual information that accounts for the distribution of measurement error. Our corrected estimator is based on the correction of the probability mass function (PMF) or probability density function (PDF, based on kernel density estimation). We prove that the corrected estimator is asymptotically unbiased in the (semi-) discrete case when the distribution of measurement error is known. We show that it reduces the estimation bias in the continuous case under certain assumptions. On simulated data, our corrected estimator leads to a more accurate estimation for mutual information when the sample size is not the limiting factor for estimating PMF or PDF accurately. We compare the uncorrected and corrected estimator on the gene expression data of TCGA breast cancer samples and show a difference in both the value and the ranking of estimated mutual information between the two estimators.

Keywords: measurement error, mutual information, kernel density estimation, unbiased estimator.

*To whom correspondence should be addressed: carlk@cs.cmu.edu

1 Introduction

Mutual information is an important measure to evaluate the degree of dependence between biological signals. It compares the joint probability of two signals with their marginal probability and is able to capture both linear and non-linear dependencies. However, when the biological signals are measured with error, the error blurs the probability distribution of the signals and leads to inaccurate mutual information estimates. Current studies on mutual information, both from a theoretical perspective and from an application perspective, mostly assume the observed signals are accurate. Ignoring the potential measurement error leads to biases in the estimated mutual information between the signals.

Applications of mutual information in computational biology include analyzing the co-evolution relationship between amino acids or nucleotides [1], inferring co-occurrence patterns of protein domains [2], constructing gene regulatory networks [3], studying neural connectivity circuits [4], and so on. An accurate estimation of mutual information is critical to these studies.

Measurement error for some biological signals is non-negligible, especially when high resolution of the measurement is needed. For example, inferring transcript-level abundances tends to be more error-prone than inferring gene-level abundances [5]. However, transcript-level abundances are able to reveal more detailed changes between samples such as differential isoform usage [6]. Recent single-cell measurements [7–12] suffer from more measurement error, of which the causes include doublets [13] and dropouts [14]. The increasing degree of measurement error poses a challenge of accurately estimating the mutual information of the true biological signals. In general, the high noise-to-signal ratio is challenging for revealing patterns in many fields of analyses.

Measurement error has been modeled and used in analyses in some areas, but it is not known how to incorporate these errors into the mutual information estimation. For example, a correction term for measurement error has been developed for Pearson correlation by Spearman [15]. In expression quantification, the measurement error has been modeled and integrated into the analysis of detecting differentially expressed genes/transcripts. Some quantification methods [16, 17] use bootstrapping or Gibbs sampling strategies to estimate a series of possible abundances to represent the scale of the measurement error. As shown by Pimentel et al. [18] and Zhu et al. [19], incorporating a model of measurement error into the differential expression detection methods improves the accuracy. This inspires the incorporation of the measurement error in other analyses.

Many theoretical studies about mutual information focus on correcting biases due to small sample sizes or deriving better estimation for probability density function (PDF). However, error-free measurements are usually implicitly assumed in these studies. Basharin [20] focused on discrete distributions and derived a correction term for the estimation bias due to small sample sizes in estimating probability mass function (PMF), which is further used in mutual information calculation. Moon et al. [21] proposed a mutual information estimator when the random variables follow continuous distributions. Their estimator uses kernel density estimation (KDE) with the bandwidth suggested in Silverman [22] to estimate PDF. Kraskov et al. [23] later proposed a k-nearest neighbor (KNN) approach for estimating PDF in mutual information calculation. Khan et al. [24] compared KNN-based mutual information estimator and the KDE-based one and characterized the cases where one is better than the other. Their comparison includes the case where signals are measured with error, but adaptation or correction for the error is not proposed. Holmes and Nemenman [25] based their work on the KNN mutual information estimator and presented a strategy to use bootstrapping of samples to estimate the error bars of estimated mutual information. Zeng et al. [26] developed a mutual information estimator based on copula density estimation [27] with the Jackknife approach.

In this work, we derive a corrected mutual information estimator that reduces the inaccuracy caused by measurement error. Our corrected estimator is based on a correction for the estimated PMF in the (semi-) discrete case or for the KDE in the continuous case using the distribution of measurement error. We prove that in the (semi-) discrete case our corrected estimator is asymptotically unbiased when the measurement error distribution is known. We discuss the assumptions under which the corrected estimator in the continuous case leads to a reduced bias in mutual information estimation. Using simulated data, we show that our corrected estimator is more accurate than the baseline estimator that plugs in the average bootstraps of observations in the (semi-) discrete or KDE-based mutual information estimators. The result shows that the derivation is correct and our corrected estimator effectively reduces the bias of mutual information estimation. Using the corrected and uncorrected estimator to calculate the pairwise mutual information between gene expression estimates on 1168 breast cancer samples in The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov), we observe a high Spearman correlation between the results of both estimators in general. However, the specific values of estimated mutual information vary as much as 40%, and the sets of genes with high estimated mutual information differ. Therefore, when the values of mutual information are of interest or the ranking of the subset of genes with the highest mutual information is needed, the effect of measurement error cannot be ignored.

2 Methods

2.1 Problem setup and baseline solution

Mutual information $I(X, Y)$ between random variables X and Y is defined by the Kullback–Leibler divergence between the joint distribution and the multiplication of its two marginal distributions:

$$I(X, Y) = \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \, \mathrm{d}x \, \mathrm{d}y \tag{1}$$

where $P(x, y)$ is the joint probability density or probability mass function, $P(x)$ and $P(y)$ are the marginal probability density or mass functions of random variables X and Y. $P(x, y)$ is usually not known and is estimated from samples. Mutual information captures the probability dependence between two random variables, specifically, whether the value of one random variable changes the belief of what value the other random variable takes.

In the presence of measurement error, we assume the true values and observed values of two biological signals (ξ_x and ξ_y) are generated as follows. The joint distribution $P_{\mu_x \mu_y}$ describes the likelihood of true values of signals ξ_x and ξ_y across all samples. Let (μ_{xs}, μ_{ys}) be the true intensities of the two signals in sample s. (μ_{xs}, μ_{ys}) is drawn from the distribution $P_{\mu_x \mu_y}$. The measurement error follows the distribution $P_{\epsilon_x \epsilon_y}$ and assumed to affect the true signals through addition. Measurement errors of signal ξ_x and of ξ_y may affect each other, and thus we do not assume the independence between ϵ_x and ϵ_y. Suppose the true signal (μ_{xs}, μ_{ys}) of sample s is measured multiple times and let $(\epsilon_{xsj}, \epsilon_{ysj})$ be the error of the j^{th} measurement, the j^{th} observation is

$$(X_{sj}, Y_{sj}) = (\mu_{xs} + \epsilon_{xsj}, \mu_{ys} + \epsilon_{ysj})$$

Given the observation (X_{sj}, Y_{sj}) of S samples each measured B times, we would like to estimate the dependence between the true signals $I(\mu_x, \mu_y)$.

Treating the observed signal values (X_{sj}, Y_{sj}) as the true values (μ_{xs}, μ_{ys}) leads to an inaccurate estimation of mutual information $I(\mu_x, \mu_y)$. Because the probability density or mass function of the observation is different from that of the true signal values. Let P_{xy} be the probability density or mass function of the observation. It has the following relationship with the distribution of the true signal values:

$$p_{xy}(X_{sj} = x, Y_{sj} = y) = \int p_{\mu_x \mu_y}(\mu_{xs} = a, \mu_{ys} = b) p_{\epsilon_x \epsilon_y}(\epsilon_{xsj} = x - a, \epsilon_{ysj} = y - b) \, \mathrm{d}a \, \mathrm{d}b. \quad (2)$$

Similarly, the marginal distributions of X_{sj} and Y_{sj} are different from that of μ_{xs} and μ_{ys}. Since P_{xy}, P_x, and P_y differ from $P_{\mu_x \mu_y}$, P_{μ_x}, and P_{μ_y}, the mutual information calculated using the distribution of the observation also differs from $I(\mu_x, \mu_y)$.

Our goal is to derive an accurate estimation of the mutual information between μ_x and μ_y using the observations X_{sj} and Y_{sj}. We measure the accuracy by whether the bias of estimation is close to zero, where the bias is the difference between the expectation of the estimated and the true mutual information.

We further assume that the measurement error has zero expectation and that follows the same distribution across all samples. Based on these assumptions, we derive a corrected mutual information estimator for both the semi-discrete case and the continuous case separately. Finally, we relax the assumption that the error distribution is the same and extend our corrected estimators.

2.2 Semi-discrete case: correcting the probability mass function using a transition matrix

In the semi-discrete case, we assume that both the true signals (μ_x, μ_y) and the observed signals (X, Y) are real-valued. Observed signals (X, Y) are a perturbation of (μ_x, μ_y) by adding measurement errors. The real-valued space is partitioned into several categories. Let $g_x : \mathbb{R} \to \{C_{x1}, \cdots, C_{xr_x}\}$ be the category mapping function that maps the real-valued signal of ξ_x to its corresponding categories, and $g_y : \mathbb{R} \to \{C_{y1}, \cdots, C_{yr_y}\}$ be the category mapping function for signal ξ_y. The two category mapping functions can be combined by $g : \mathbb{R} \times \mathbb{R} \to \{C_1, \ldots, C_n\}$, which is defined as $g(a, b) = (g_x(a), g_y(b))$ for any signal intensity a of ξ_x and intensity b of ξ_y, and the pairs of categories of (C_{xj}, C_{yk}) are reparameterized using C_i. g informs the categories of both ξ_x and ξ_y simultaneously, and the probability mass function (PMF) of $g(X, Y)$ (or $g(\mu_x, \mu_y)$) informs both the PMF of $g_x(X)$ (or $g_x(\mu_x)$) and the PMF of $g_y(Y)$ (or $g_y(\mu_y)$). The mutual information between the categories, $g_x(\mu_x)$ and $g_y(\mu_y)$, is of interest. For example, μ_x and μ_y are the probability of a logistic regression that has two categories, high and low (Figure 1). X and Y are estimated probabilities, which may fall into a different category from the true signals. We assume that g is known, which is a reasonable assumption when the partition of the real-valued space is biologically meaningful. Our goal is to derive an estimator for the mutual information between $g_x(\mu_x)$ and $g_y(\mu_y)$ using the observation $g_x(X)$ and $g_y(Y)$.

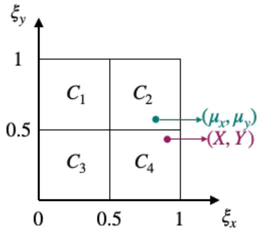

Figure 1: Example of real-valued (μ_x, μ_y) and the partition of the space into four categories. Observation (X, Y) is different from (μ_x, μ_y) due to measurement error. Here $g(\mu_x, \mu_y) = C_2$ and $g(X, Y) = C_4$.

Given sample s and bootstrap measure b, the probability of the observation $g(X_{sb}, Y_{sb})$ has the following relationship with $g(\mu_{xs}, \mu_{ys})$:

$$\mathbb{P}(g(X_{sb}, Y_{sb}) = C_i) = \sum_j \mathbb{P}(g(X_{sb}, Y_{sb}) = C_i \mid g(\mu_{xs}, \mu_{ys}) = C_j)\mathbb{P}(g(\mu_{xs}, \mu_{ys}) = C_j). \tag{3}$$

Define the following matrix $F \in \mathbb{R}^{n \times n}$, and PMF $P \in \mathbb{R}^n$ and $Q \in \mathbb{R}^n$:

$$\begin{aligned} F_{ji} &= \mathbb{P}(g(X_{sb}, Y_{sb}) = C_i \mid g(\mu_{xs}, \mu_{ys}) = C_j) \\ P &= (\mathbb{P}(g(\mu_{xs}, \mu_{ys}) = C_1),\ \ldots,\ \mathbb{P}(g(\mu_{xs}, \mu_{ys}) = C_n))^T \\ Q &= (\mathbb{P}(g(X_{sb}, Y_{sb}) = C_1),\ \ldots,\ \mathbb{P}(g(X_{sb}, Y_{sb}) = C_n))^T . \end{aligned} \tag{4}$$

Using the above matrix and vector notations, equation (3) is equivalent to $Q^T = P^T F$, where matrix F can be viewed as a transition matrix. We use this equality to derive the following corrected estimator.

Theorem 1. *Given transition matrix F as defined in equation (4), measurements of signals $\{X_{sb}\}$ and $\{Y_{sb}\}$ for $1 \leq s \leq S$ and $1 \leq b \leq B$, let $\hat{Q} = \frac{1}{S}\sum_{s=1}^{S}\left(\mathbb{1}(g(X_{s1}, Y_{s1}) = C_1), \ldots, \mathbb{1}(g(X_{s1}, Y_{s1}) = C_n)\right)^T$, where $\mathbb{1}$ is the indicator function. If $\hat{Q}^T F^{-1} \geq 0$ element-wise, and the mutual information is estimated using joint PMF of $\hat{Q}^T F^{-1}$ and the corresponding marginal PMFs, then the estimated mutual information is an unbiased estimator for the target mutual information of P^T asymptotically as $S \to \infty$.*

Sketch of proof. We use the property that mutual information is a linear combination of three entropy terms

$$I(X, Y) = H(X, Y) - H(X) - H(Y),$$

where the entropy of a (joint) discrete random variable with PMF $Q = (q_1, \ldots, q_n)^T$ is defined as

$$H = -\sum_i q_i \log(q_i).$$

If the entropy estimator $\hat{H} = -\sum_i (\hat{Q}^T F^{-1})_i \log(\hat{Q}^T F^{-1})_i$ is an unbiased estimator for $H = -\sum_i p_i \log(p_i)$, then the mutual information calculated using $\hat{Q}^T F^{-1}$ is an unbiased estimator for the mutual information calculated using P, where $(\hat{Q}^T F^{-1})_i$ is the i^{th} element of vector $\hat{Q}^T F^{-1}$.

Basharin [20] proved that entropy calculated using $\hat{Q}$ is an unbiased estimator for the entropy calculated using Q as the number of samples goes to infinity. We use their proof as a template and derive that the entropy of $\hat{Q}^T F^{-1}$ is an unbiased estimator of $Q^T F^{-1} = P$ as the number of samples goes to infinity. See Lemma S2 in Appendix Section S1.1 for the details of proof that entropy of $\hat{Q}^T F^{-1}$ is an unbiased estimator of the entropy of $Q^T F^{-1}$. □

With a little abuse of notation, the unbiasedness also holds for the following estimator.

Corollary 1.1. *Given measurements of signals $\{X_{sb}\}$ and $\{Y_{sb}\}$ for $1 \leq s \leq S$ and $1 \leq b \leq B$, let $\bar{X}_{s\bullet} = \frac{1}{B}\sum_{b=1}^{B} X_{sb}$ and $\bar{Y}_{s\bullet} = \frac{1}{B}\sum_{b=1}^{B} Y_{sb}$. Let transition matrix $F \in \mathbb{R}^{n \times n}$ have the following entries*

$$F_{ji} = \mathbb{P}(g(\bar{X}_{s\bullet}, \bar{Y}_{s\bullet}) = C_i \mid g(\mu_{xs}, \mu_{ys}) = C_j).$$

Let the estimated PMF be $\hat{Q} = \frac{1}{S}\sum_{s=1}^{S}\left(\mathbb{1}(g(\bar{X}_{s\bullet}, \bar{Y}_{s\bullet}) = C_1), \ldots, \mathbb{1}(g(\bar{X}_{s\bullet}, \bar{Y}_{s\bullet}) = C_n)\right)^T$, where $\mathbb{1}$ is the indicator function. If $\hat{Q}^T F^{-1} \geq 0$ element-wise, and the mutual information is estimated using joint PMF of $\hat{Q}^T F^{-1}$ and the corresponding marginal PMFs, then estimated mutual information is an unbiased estimator for the target mutual information of P^T asymptotically as $S \to \infty$.

The above theorem and corollary hold for both the semi-discrete case and the discrete case. However, matrix F is generally not known beforehand, and an estimation of F is required. We provide an estimation for F based on the real-valued assumption in the semi-discrete case.

We assume the distribution of (μ_x, μ_y) within each category is a uniform distribution and approximate the measurement error by a Gaussian distribution. A Gaussian distribution is appropriate to approximate the measurement error because the Central Limit Theorem states that $(\bar{X}_{s\bullet}, \bar{Y}_{s\bullet})$ converges in distribution to a Gaussian distribution as $B \to \infty$. Let $N(0, \hat{\Sigma}_\epsilon)$ be the limiting distribution for $(\bar{X}_{s\bullet} - \mu_{xs}, \bar{Y}_{s\bullet} - \mu_{ys})$. Let $U_i = [l_{xi}, u_{xi}] \times [l_{yi}, u_{yi}]$ be the real-valued range corresponding to $g(\mu_x, \mu_y) = C_i$ and $U_j = [l_{xj}, u_{xj}] \times [l_{yj}, u_{yj}]$ be the ranges corresponding to $g(\mu_x, \mu_y) = C_j$. The approximated transition probability is:

$$\hat{F}_{ji} = \frac{1}{(u_{xj} - l_{xj})(u_{yj} - l_{yj})} \int_{(x_i, y_i) \in U_i} \int_{(x_j, y_j) \in U_j} \frac{1}{2\pi\sqrt{|\hat{\Sigma}_\epsilon|}} \exp\left\{ - \begin{pmatrix} x_i - x_j \\ y_i - y_j \end{pmatrix}^T \hat{\Sigma}_\epsilon^{-1} \begin{pmatrix} x_i - x_j \\ y_i - y_j \end{pmatrix} \right\} \mathrm{d}x_j \, \mathrm{d}y_j \, \mathrm{d}x_i \, \mathrm{d}y_i.$$

In the discrete case, correcting the distribution of (μ_x, μ_y) using transition matrix F has the same form as image denoising or deblurring [28]. Nevertheless, the goal of image deblurring is to correct the individual observation, but our estimator is based on correcting the distribution. The estimated PMF is proportional to the counts in the discrete case. The counts can be viewed as an observation from an multinomial distribution. From this perspective, correcting the counts is equivalent to correcting an multinomial observation, which explains the similarity between image deblurring and PMF correction.

2.3 Continuous case: correcting estimated PDF using kernel density estimation (KDE)

In this section, we aim to reduce the bias in the KDE-based mutual information estimation in the continuous case. We first introduce the formula of the KDE-based mutual information estimator and derive its estimation bias. We then derive a correction for the estimated density and prove the unbiasedness of the corrected density estimation. We finally discuss the scenarios when the corrected density reduces the error in mutual information estimation.

Moon et al. [21] develops a mutual information estimator that uses the kernel density estimation to estimate the PDF. Given true signal values (μ_{xs}, μ_{ys}) and using a diagonal bandwidth matrix, the kernel density estimation given by Silverman [22] is

$$p_{\mu_x \mu_y}(\mu_x, \mu_y) = \frac{1}{S h_x h_y} \sum_{s=1}^{S} \frac{1}{2\pi} \exp\left\{ -\frac{1}{2} \begin{pmatrix} \mu_x - \mu_{xs} \\ \mu_y - \mu_{ys} \end{pmatrix}^T \begin{pmatrix} \frac{1}{h_x^2} & 0 \\ 0 & \frac{1}{h_y^2} \end{pmatrix} \begin{pmatrix} \mu_x - \mu_{xs} \\ \mu_y - \mu_{ys} \end{pmatrix} \right\}, \tag{5}$$

where h_x and h_y are the bandwidth for the two axes separately. The estimation of mutual information is the sample average of the differences among the log of probability densities:

$$\hat{I}(\mu_x, \mu_y) = \frac{1}{S} \sum_{s=1}^{S} \left(\log p_{\mu_x \mu_y}(\mu_{xs}, \mu_{ys}) - \log p_{\mu_x}(\mu_{xs}) - \log p_{\mu_y}(\mu_{ys}) \right). \tag{6}$$

With given true values (μ_{xs}, μ_{ys}), these estimators are unbiased as $S \to \infty$ except for the boundaries or tails of the distribution [29].

In presence of measurement error, (μ_{xs}, μ_{ys}) is not observed. The observation (X_{sj}, Y_{sj}) is a summation of the true signal and error, where the error distribution is estimable from the bootstraps. We assume the distribution of measurement error is the same for all samples. The density of the observation p_{xy} is different from the density of the true signals $p_{\mu_x \mu_y}$ as shown by equation (2). Our goal is to derive a corrected estimator for mutual information of $I(\mu_x, \mu_y)$ with a reduced bias.

The bias of each $\log$ term in mutual information estimator in equation (6) has the following upper bound.

Lemma 1. *Given the true probability density* p *and fixed point* (x, y)*, for any estimator of the density at the point* $\hat{p}(x, y)$*, if the true and estimated density are lower bounded by* $\delta > 0$*, that is,* $p(x, y) \geq \delta$ *and* $\hat{p}(x, y) \geq \delta$*, then the bias of* $\log(\hat{p})$ *at the point is upper bounded by:*

$$|\mathbb{E}(\log \hat{p}(x, y)) - \log p(x, y)| \leq \frac{1}{p(x, y)} |\mathbb{E}(\hat{p}(x, y)) - p(x, y)| + \frac{\mathrm{Var}(\hat{p}(x, y)) + (\mathbb{E}(\hat{p}(x, y)) - p(x, y))^2}{\delta^2}. \tag{7}$$

Sketch of proof. The three main steps of deriving the inequality are: applying Taylor expansion on the log function with mean value form of the remainder, then taking the expectation on both sides, and finally replacing the quadratic term with bias-variance decomposition. See Appendix Section S1.2 for the details. □

According to this lemma, if the density estimator $\hat{p}$ has both small bias and small variance at a non-zero density position (x, y), the $\log$ terms in equation (6) will have small bias. The points evaluated in equation (6) are true signal values (μ_{xs}, μ_{ys}), and thus the true densities and the estimated densities by KDE using the observed signals (X_{sj}, Y_{sj}) at these points are non-zero. We focus on reducing the bias of density estimation, $\mathbb{E}(\hat{p}(x, y)) - p(x, y)$, and derive the following (asymptotically) unbiased estimators for each summation term in the KDE density formula (5).

Theorem 2. *Let* $\bar{X}_{s\bullet} = \frac{1}{B} \sum_{j=1}^{B} X_{sj}$ *and* $\bar{Y}_{s\bullet} = \frac{1}{B} \sum_{i=1}^{B} Y_{sj}$*. Let* W *be an diagonal bandwidth matrix used in KDE,* $W = \mathrm{diag}(\frac{1}{h_x^2}, \frac{1}{h_y^2})$*. Assuming the measurement error of* $\left(\bar{X}_{s\bullet} - \bar{X}_{t\bullet}, \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet}\right)^T$ *follows a Gaussian distribution* $N((\mu_{xs} - \mu_{xt}, \mu_{ys} - \mu_{yt})^T, \frac{2}{B}\Sigma_\epsilon)$*, let* P *and* $\{\zeta_1, \zeta_2\}$ *be the eigenvectors and eigenvalues of* $\frac{1}{B}\Sigma_\epsilon^{\frac{1}{2}} W \Sigma_\epsilon^{\frac{1}{2}}$*. Let* $\lambda_i = \frac{\zeta_i}{1-2\zeta_i}$*,* $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$*, and* $A = \frac{B}{2}\Sigma_\epsilon^{-\frac{1}{2}} P^T \Lambda P \Sigma_\epsilon^{-\frac{1}{2}}$*. When* ζ_i *satisfies* $\zeta_i < \frac{1}{2}$ *for all* i*, the following estimator,*

$$\left(\prod_{i=1}^{2} \sqrt{1 + 2\lambda_i}\right) \exp\left\{-\begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}^T A \begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}\right\}, \tag{8}$$

is an unbiased estimator for the KDE term in equation (5)*:*

$$\exp\left\{-\frac{1}{2}\begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}^T W \begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}\right\}.$$

The proof of the theorem is in Appendix Section S1.3.

Given sample s, the covariance of error Σ_ϵ is usually not known or the error may not be a Gaussian distribution. Nevertheless, a Gaussian limiting distribution is a good approximate to model the error as shown in Central Limit Theorem:

$$\begin{pmatrix} \bar{X}_{s\bullet} \\ \bar{Y}_{s\bullet} \end{pmatrix} \xrightarrow{d} N(\begin{pmatrix} \mu_{xs} \\ \mu_{ys} \end{pmatrix}, \frac{1}{B}\hat{\Sigma}_\epsilon),$$

where d means convergence in distribution and the estimated $\hat{\Sigma}_\epsilon$ is calculated by

$$\hat{\Sigma}_\epsilon = \frac{1}{SB-1}\sum_{s=1}^{S}\sum_{i=1}^{B}\begin{pmatrix} X_{sj}-\bar{X}_{s\bullet} \\ Y_{sj}-\bar{Y}_{s\bullet}\end{pmatrix}\begin{pmatrix} X_{sj}-\bar{X}_{s\bullet} \\ Y_{sj}-\bar{Y}_{s\bullet}\end{pmatrix}^T. \tag{9}$$

We prove that using the estimated error distribution in equation (9) leads to an asymptotically unbiased estimator for the corresponding KDE term.

Theorem 3. *Let P and $\{\zeta_1, \zeta_2\}$ be the eigenvectors and eigenvalues of $\frac{1}{B}\hat{\Sigma}_\epsilon^{\frac{1}{2}} W \hat{\Sigma}_\epsilon^{\frac{1}{2}}$. Let $\lambda_i = \frac{\zeta_i}{1-2\zeta_i}$, $\Lambda = \begin{pmatrix}\lambda_1 & 0 \\ 0 & \lambda_2\end{pmatrix}$, and $\hat{A} = \frac{B}{2}\hat{\Sigma}_\epsilon^{-\frac{1}{2}} P^T \Lambda P \hat{\Sigma}_\epsilon^{-\frac{1}{2}}$. When ζ_i satisfies $\zeta_i < \frac{1}{2}$ for all i, the following estimator is an asymptotically unbiased estimator for the KDE term in equation* (5)*:*

$$\left(\prod_{i=1}^{2}\sqrt{1+2\lambda_i}\right)\exp\left\{-\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}^T \hat{A}\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}\right\}. \tag{10}$$

See Appendix Section S1.4 for the proof.

Corollary 3.1. *Let $\hat{\sigma}^2_{\epsilon x}$ be the variance of measurement error on signal ξ_x, which is the top left element in covariance matrix $\hat{\Sigma}_\epsilon$. When $\frac{\hat{\sigma}^2_{\epsilon x}}{B} \leq \frac{h_x^2}{2}$, the following estimator is an unbiased estimator for $\exp\left\{-\frac{1}{2}\frac{(\mu_{xs}-\mu_{xt})^2}{h_x^2}\right\}$:*

$$\frac{h_x}{\sqrt{h_x^2 - \frac{2}{B}\hat{\sigma}^2_{\epsilon x}}}\exp\left\{-\frac{1}{2}\frac{(\bar{X}_{s\bullet}-\bar{Y}_{s\bullet})^2}{h_x^2 - \frac{2\hat{\sigma}^2_{\epsilon x}}{B}}\right\}. \tag{11}$$

A corrected mutual information estimator can be derived by plugging in the corrected density estimator (10) and (11) into equation (6). We claim that when λ_i is small, the expectation of $\hat{p}$ dominates the second order moment around 0, which further dominates $\mathrm{Var}(\hat{p}(x, y))$.

Claim 1. *There exists a small positive number δ such that when $0 \leq \lambda_i \leq \delta$,*

$$\mathbb{E}\left(\left(\prod_{i=1}^{2}\sqrt{1+2\lambda_i}\right)\exp\left\{-\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}^T A\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}\right\}\right) \geq$$
$$\mathbb{E}\left[\left(\left(\prod_{i=1}^{2}\sqrt{1+2\lambda_i}\right)\exp\left\{-\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}^T A\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}\right\}\right)^2\right].$$

See Appendix Section S1.5 for the proof. In the case where the expectation of an estimator dominates over the variance, reducing the bias tends to be more effective than reducing the variance and keeping a large bias. Using the corrected density estimator, the bias of density term in equation (7) is zero, $\mathbb{E}(\hat{p}(x,y)) - p(x,y) = 0$. With a zero bias and a small variance, the error in mutual information estimation is also small.

The condition of $\zeta_i < \frac{1}{2}$ in Theorem 2 and small λ_i in Claim 1 may not hold with a predefined bandwidth. We take a heuristic strategy of shrinking the error covariance by a scalar and using the shrunk covariance to construct the estimator. Using this strategy, the corrected density estimator is no longer asymptotically unbiased. Nevertheless, the corrected estimator with the shrinking covariance strategy still performs better than the baseline that directly uses $(\bar{X}_{s\bullet}, \bar{Y}_{s\bullet})$ in the KDE (5) and mutual information estimation (6), as we show in Results Section.

2.4 Relaxing the assumption of the same error distribution across samples

Biological datasets usually do not have the same distribution of measurement errors across all samples or datasets. Even within datasets, the error distribution varies because of the domain of the measurements or the constraints of the computational methods. For example, gene expression is constrained to be non-negative. Thus the standard deviation of lowly expressed genes is lower compared to that of highly expressed genes. Therefore, we adapt our corrected estimators for the case where different samples have different measurement errors.

In the semi-discrete case, the error distribution assumption can be relaxed so that each joint category has its unique error distribution. Under the relaxed error distribution assumption, the entry in transition matrix F_{ji} can be calculated using the error distribution for category C_j. The corrected PMF $\hat{Q}^T F^{-1}$ is still an asymptotically unbiased estimator.

In the continuous case using KDE, the error distribution assumption can be relaxed so that each sample has a unique error distribution. Each KDE term $\exp\left\{-\frac{1}{2}\begin{pmatrix}\mu_{xs}-\mu_{xt}\\ \mu_{ys}-\mu_{yt}\end{pmatrix}^T W \begin{pmatrix}\mu_{xs}-\mu_{xt}\\ \mu_{ys}-\mu_{yt}\end{pmatrix}\right\}$ can be corrected using the sum of error covariance in sample s and t. However, the estimated error covariance may be less accurate because there are only B bootstraps to estimate the sample-specific error distribution, compared to estimating the covariance for all samples from SB bootstraps.

3 Results

3.1 Corrected PMF leads to more accurate mutual information estimation in simulated semi-discrete signals

We simulate the semi-discrete signals by the following procedure. Each of ξ_x and ξ_y have 5 categories with real-valued measurements corresponding to $[i-1, i)$ for $1 \leq i \leq 5$. A ground truth PMF is simulated from the 5×5 categories with a Dirichlet prior. The true signals within each category follows a uniform distribution in the space of $[i-1, i) \times [j-1, j)$. Each joint category has its unique measurement error distribution, which is a mixture of two Gaussian distributions. Using these probability distributions, we simulate observations using various numbers of samples (100, 500, 1000, 10 000 and 100 000) and various numbers of bootstrap measurements (10, 20, 50 and 100). We compare our corrected estimator with the baseline estimator that treats $\bar{X}_{s\bullet}$ and $\bar{Y}_{s\bullet}$ as true signals for estimating PMF and mutual information.

With a fixed bootstrap size and a large sample size, our correction in more accurate in estimating PMF as well as mutual information (Figure 2A, C). Specifically, when the sample size is larger than or equal to 500, the corrected estimator shows its improvement compared to the baseline in both PMF and mutual information estimation. As the sample size grows, the improvement becomes more noticeable. With a small sample size, the sample size is the bottleneck of accurately estimating the PMF, and therefore the corrected estimator does not show an improvement.

Fixing the bootstrap size, the estimation error for mutual information increases as the number of samples increases for both corrected and baseline estimator (Figure 2C). The baseline PMF estimation converges to a different distribution from $P_{\mu_x\mu_y}$ and thus is more biased, which explains its increasing error. The estimation error of the corrected estimator increases more slowly than the baseline. However, it does not converge to the true mutual information either, which is possibly because the error is non-Gaussian and suffers from

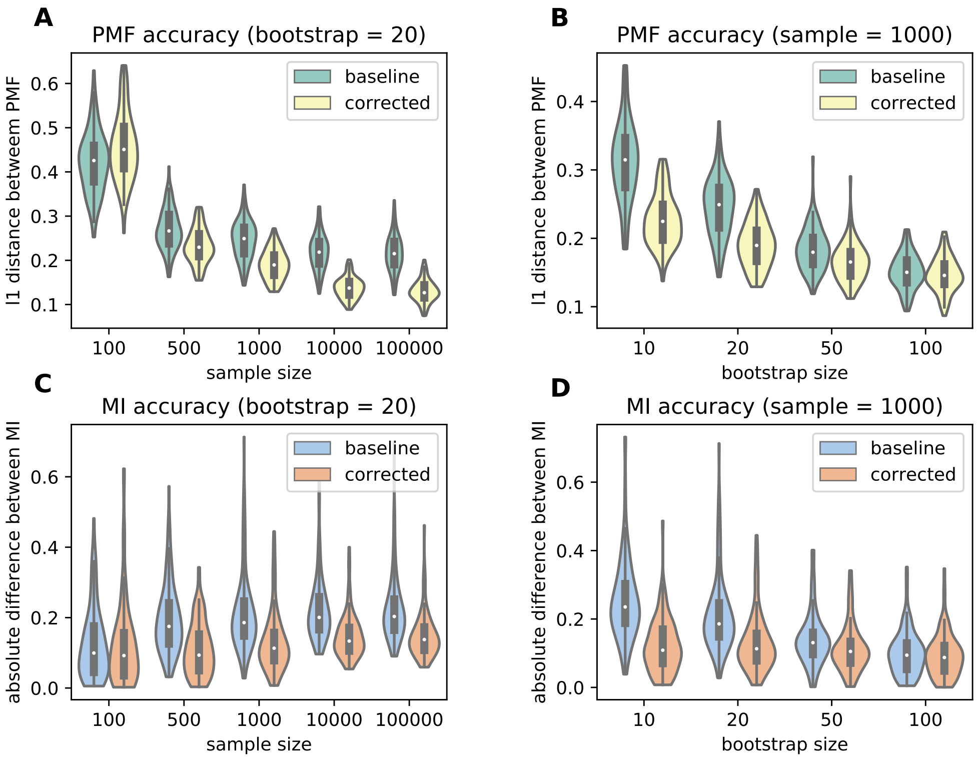

Figure 2: (A–B) ℓ_1 distance between the true PMF and estimated PMF when using baseline (average of bootstraps) estimator and the corrected estimator. (A) The number of bootstraps is fixed to be 20 and the number of samples are indicated by x axis. (B) The number of samples is fixed to be 1000 and the number of bootstraps are indicated by x axis. (C–D) Absolute difference between true mutual information and estimated mutual information when using baseline PMF and corrected PMF. (C) The number of bootstraps is fixed to be 20 and the number of samples is indicated by x axis. (D) The number of samples is fixed to be 1000 and the number of bootstraps is indicated by x axis.

estimation error with a fixed bootstrap size.

With a fixed sample size, increasing the number of bootstraps reduces the estimation error for both baseline and corrected estimator (Figure 2B, D). This can be explained by that both the baseline and the corrected estimator are asymptotically unbiased as the number of bootstraps goes to infinity. However, before the number of bootstraps is sufficiently large, the corrected estimator achieves a smaller estimation error.

3.2 Corrected KDE leads to more accurate mutual information estimation for simulated Gaussian mixtures

The true signal (μ_x, μ_y) is simulated from a mixture of bi-variated Gaussian distributions. In each of the mixture, the measurement error is a mixture of two Gaussian distributions that have smaller covariances than the Gaussian covariance matrix of (μ_x, μ_y). We simulate various numbers of mixtures (2, 5, 10 and 20), various numbers of samples (500, 1000, 5000 and 10 000), and various numbers of bootstraps (20, 50 and 100). Since there is no closed-form expression of mutual information for a mixture of Gaussian distributions,

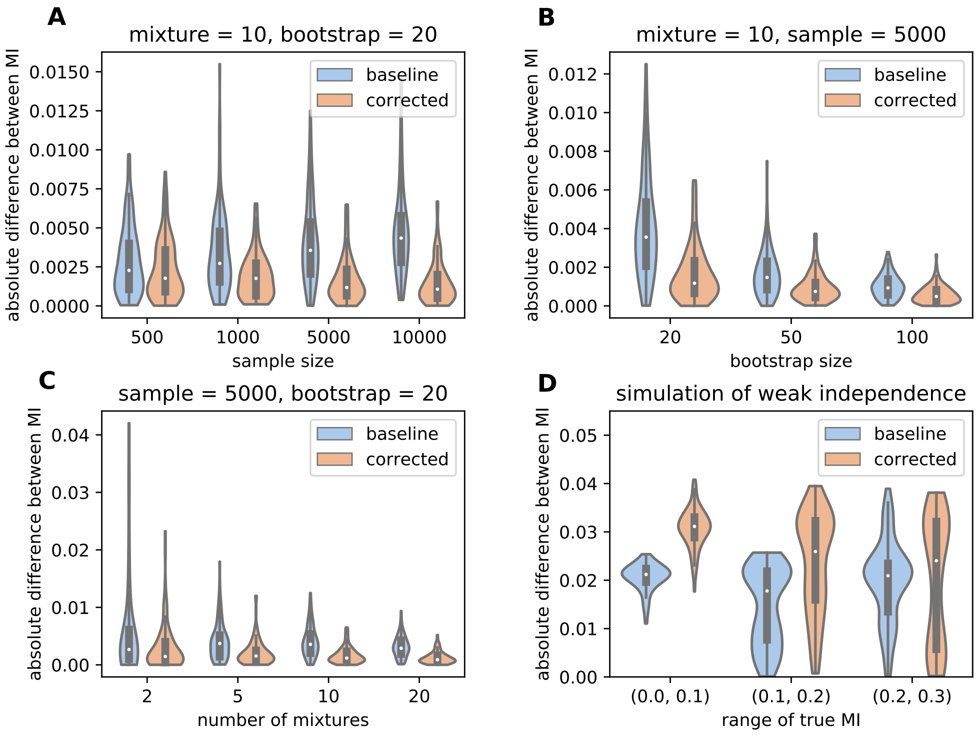

Figure 3: Absolute difference between the true mutual information and estimated mutual information using the baseline estimator and the corrected estimator. (A) The mixture size is 10, the bootstrap size is 20, and the sample size is indicated by x axis. (B) The mixture size is 10, the sample size is 5000, and the bootstrap size is indicated by x axis. (C) The sample size is 5000, bootstrap size is 20, and the mixture size is indicated by x axis. (D) Simulation of weakly dependent random variables. The true signals follows a bi-variate Gaussian distribution. The sample size is 5000 and the bootstrap size is 20. X axis is the true mutual information.

we apply the KDE-based mutual information estimation (6) on the simulated true signals (μ_{xs}, μ_{ys}) and use it as the true mutual information. When calculating the corrected estimation in equation (10), we split the samples into clusters by applying K-means clustering on the sample-specific error covariance and estimate an error covariance using all samples in each cluster. The number of K-means clusters is set to the same as the number of Gaussian mixtures. The error covariance matrix is shrunk so that $\zeta_i < 0.25$ to satisfy the conditions in Theorem 3 and Claim 1. We compare to a baseline estimator that uses the average of bootstraps in the KDE-based mutual information estimator in (5) and (6). The bandwidths are set to be $h_x = (\frac{4\hat{\sigma_x}^5}{3S})^{\frac{1}{5}}$ and $h_y = (\frac{4\hat{\sigma_y}^5}{3S})^{\frac{1}{5}}$ as suggested by Silverman [22] for both estimators.

With a fixed number of bootstraps, we observe the same pattern as in the semi-discrete case: the corrected mutual information estimator achieves smaller biases than the baseline (Figure 3A). When the sample size becomes larger, the improvement of the corrected estimator becomes more apparent because the estimated density converges to its expectation.

With a fixed number of samples, increasing the number of bootstraps reduces the mutual information estimation error for both estimators (Figure 3B). The improvement of the corrected estimator is visible even with 100 bootstraps, especially in terms of the smaller variance of estimator error.

We observe that the correction is effective for various numbers of mixtures (Figure 3C), but the effectiveness is more significant under a larger number of mixtures. Using Mann-Whitney one-sided U test to compare the accuracy between the baseline and corrected mutual information, the p-value is 0.002 for two mixtures, 1.41×10^{-7} for five mixtures, and 4.46^{-16} for ten mixtures under 5000 samples and 20 bootstraps. Nevertheless, a wider range of mixture sizes needs to be tested in order to study the effect of the number of mixtures on mutual information estimation in more detail.

The corrected estimator is able to reveal strong dependencies when they are shadowed by the measurement error. However, when the dependence is weak or even the two signals are independent, the correction may lift the estimated mutual information and falsely show a small dependence (Figure 3D). Therefore, the corrected mutual information estimator should be only applied when the dependence is large after correction.

3.3 Genes with largest mutual information with known cancer genes are slightly changed by corrected estimator

We applied the corrected mutual information estimator on 1168 breast cancer samples from TCGA to investigate what genes have high mutual information with known cancer genes. TCGA RNA-seq samples are quantified by Salmon [17] with 100 bootstraps for evaluating the measurement error. With 10 arbitrarily chosen cancer genes from COSMIC [30] (cancer.sanger.ac.uk), we estimate the mutual information using both the original KDE-based estimator (uncorrected) with the default single-point Salmon quantification and our corrected mutual information.

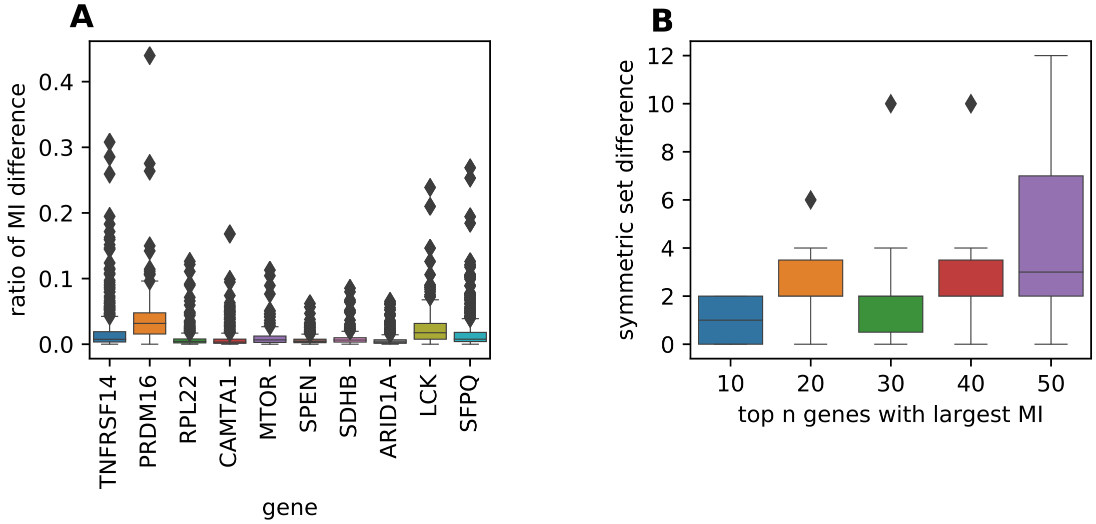

Figure 4: (A) The ratio of change from the uncorrected mutual information estimate to the corrected estimate for the top 500 genes with the largest uncorrected estimates. X axis is the 10 known cancer genes and the box is plotted for the 500 genes with the largest uncorrected mutual information. (B) The symmetric difference between the set of top several genes with largest uncorrected mutual information and the set of top genes with the largest corrected mutual information. X axis is the size of each set and the box is plotted for 10 known cancer genes.

In general, the uncorrected and corrected estimator agree with each other. For each of the selected cancer gene, we compute the corrected and uncorrected mutual information between this gene and all other genes. The Spearman correlation between uncorrected and corrected mutual information is greater than

0.98. However, a detailed analysis of the top genes with high mutual information reveals the difference between the two estimators.

Among the 500 genes with highest uncorrected mutual information values for each selected cancer gene, the corrected estimations between the selected genes and those 500 genes vary as much as 10%–40% from the uncorrected ones (Figure 4A, the range of the largest y axis values across selected genes shown on x axis). The sets of genes with the highest estimated mutual information between the two estimators are different when fixing the set size (Figure 4B). Therefore, when the specific values of mutual information are used or when the subset of genes with the highest mutual information is of interest, the measurement error has an effect on the result and the corrected mutual information estimator should be considered.

We show two example genes where the corrected estimator decreases its ranking of mutual information. The rank of estimated mutual information between gene *CASP9* and known cancer gene *SDHB* decreases from 91 to 67 (the estimate changes from 0.198 to 0.211). *SDHB* affects mitochondrial function and *CASP9* helps cell apoptosis [31]. Mitochondria plays a role in cell apoptosis [32], providing support that the expressions of the two genes possibly are dependent. The rank of the estimate between genes *GRAP2* and *LCK* decreases from 86 to 71 (the estimate changed from 0.483 to 0.509). Both *GRAP2* and *LCK* are involved in T-cell-receptor signaling pathway [33].

In this analysis, there is no ground truth about the true mutual information or the ranking of the pairwise mutual information. In addition, the number of biological samples may not be large enough to reveal the true dependence between genes. A larger number of samples and a better validation are needed for a more comprehensive evaluation of the performances of the corrected and uncorrected estimator.

4 Discussion

We derive a corrected mutual information estimator to account for the measurement error for both semi-discrete and continuous measurements. Our corrected estimator is based on estimating the probability mass function (PMF) or probability density function (PDF) in an asymptotically unbiased way. We prove that in the semi-discrete case the corrected mutual information estimator derived from the unbiased PMF estimation is asymptotically unbiased. We give conditions under which our corrected estimator in the continuous case reduces the bias of mutual information estimation. On simulated data, the corrected estimators for both semi-discrete and continuous cases are more accurate compared to the baseline of using the bootstrap average in mutual information estimation.

We compare our corrected estimator to the uncorrected estimator on detecting the genes with high dependence with known cancer-related genes using TCGA breast cancer gene expression data. The estimated mutual information is generally consistent between the corrected and uncorrected estimator. However, the values of the estimated mutual information may change up to 40% for the top 500 dependent genes. The sets of a fixed number of top dependent genes differ by a few number of genes between the two estimators. The observation suggests that the measurement error has an effect on the mutual information estimation and should be accounted for when carrying out analyses based on the value or ranking of estimated mutual information.

Our corrected estimator for continuous random variables reduces the bias of mutual information estimation only under certain assumptions. Whether there is an unbiased estimator and the form of the estimator remains to be determined. In addition, a correction for the KNN-based mutual information estimator to correct for measurement error is also a useful future direction.

Our corrected estimators assume that the measurement error has zero mean and does not have systematic biases. However, this assumption is unrealistic for certain data types. Modeling and correcting the systematic biases is also critical for accurately estimating the mutual information, and this is an important direction for future work.

We focus on reducing the bias of mutual information estimation. However, low variance or other properties may also be desired. Designing new corrections with the consideration of measurement error and studying their statistical properties will be interesting theoretical future directions.

Acknowledgements

This research is funded in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4554 to C.K., by the US National Science Foundation (CCF-1256087, CCF-1319998) and by the US National Institutes of Health (R01GM122935). This work was partially funded by The Shurl and Kay Curci Foundation. This project is funded, in part, under a grant (#4100070287) with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions. The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC) [34]. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

We thank Guillaume Marçais, Hongyu Zheng, and Amir Alavi for useful suggestions.

Disclosure Statement

C.K. is a co-founder of Ocean Genomics, Inc.

References

[1] Cristina Marino Buslje, Javier Santos, Jose Maria Delfino, and Morten Nielsen. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, 25(9):1125–1131, 2009.

[2] Jung Eun Shim and Insuk Lee. Weighted mutual information analysis substantially improves domain-based functional network models. *Bioinformatics*, 32(18):2824–2830, 2016.

[3] Xiujun Zhang, Xing-Ming Zhao, Kun He, Le Lu, Yongwei Cao, Jingdong Liu, Jin-Kao Hao, Zhi-Ping Liu, and Luonan Chen. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, 28(1):98–104, 2011.

[4] Abhinav Singh and Nicholas A Lesica. Incremental mutual information: a new method for characterizing the strength and dynamics of connections in neuronal circuits. *PLoS Computational Biology*, 6 (12):e1001035, 2010.

[5] Charlotte Soneson, Michael I Love, and Mark D Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 2015.

[6] Marek Cmero, Nadia M Davidson, and Alicia Oshlack. Using equivalence class counts for fast and accurate testing of differential transcript usage. *F1000Research*, 8, 2019.

[7] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093, 2013.

[8] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A Sperling, Alon Goren, David A Weitz, and Bradley E Bernstein. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165, 2015.

[9] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486, 2015.

[10] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife OShaughnessy-Kirwan, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):59, 2017.

[11] Junyue Cao, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361 (6409):1380–1385, 2018.

[12] Avi Srivastava, Laraib Malik, Tom Smith, Ian Sudbery, and Rob Patro. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biology*, 20(1):65, 2019.

[13] Christopher S McGinnis, Lyndsay M Murrow, and Zev J Gartner. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Systems*, 8(4):329–337, 2019.

[14] Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, 2015.

[15] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[16] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

[17] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417, 2017.

[18] Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, 14(7):687, 2017.

[19] Anqi Zhu, Avi Srivastava, Joseph G Ibrahim, Rob Patro, and Michael I Love. Nonparametric expression analysis using inferential replicate counts. *Nucleic Acids Research*, 47(18):e105–e105, 08 2019.

[20] Georgij P Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications*, 4(3):333–336, 1959.

[21] Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

[22] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[23] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

[24] Shiraj Khan, Sharba Bandyopadhyay, Auroop R Ganguly, Sunil Saigal, David J Erickson III, Vladimir Protopopescu, and George Ostrouchov. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2):026209, 2007.

[25] Caroline M Holmes and Ilya Nemenman. Estimation of mutual information for real-valued data with error bars and controlled bias. *arXiv preprint arXiv:1903.09280*, 2019.

[26] Xianli Zeng, Yingcun Xia, and Howell Tong. Jackknife approach to the estimation of mutual information. *Proceedings of the National Academy of Sciences*, 115(40):9956–9961, 2018.

[27] Gery Geenens, Arthur Charpentier, Davy Paindaveine, et al. Probit transformation for nonparametric kernel estimation of the copula density. *Bernoulli*, 23(3):1848–1873, 2017.

[28] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):259–279, 1986.

[29] Wolfgang Härdle. *Applied nonparametric regression*. Number 19. Cambridge university press, 1990.

[30] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2018.

[31] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2015.

[32] Chunxin Wang and Richard J Youle. The role of mitochondria in apoptosis. *Annual Review of Genetics*, 43:95–118, 2009.

[33] Morgan Huse. The T-cell-receptor signaling network. *J Cell Sci*, 122(9):1269–1273, 2009.

[34] Nicholas A Nystrom, Michael J Levine, Ralph Z Roskies, and J Scott. Bridges: a uniquely flexible HPC resource for new communities and data analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, page 30. ACM, 2015.

[35] Patrick Chareka. The converse to Curtiss' theorem for one-sided moment generating functions. *arXiv preprint arXiv:0807.3392*, 2008.

S1 Proofs

S1.1 The corrected estimator in semi-discrete case is unbiased in estimating entropy

Let $H(\hat{Q}^T F^{-1})$ be the entropy calculated using PMF $\hat{Q}^T F^{-1}$ and similarly $H(Q^T F^{-1})$ be the entropy calculated using $Q^T F^{-1}$. Let $G = F^{-1}$ and g_{ij} be the entry of G on the i^{th} row j^{th} column.

Lemma S2. *Let Q be the underlying PMF of $g(X_{s1}, Y_{s1})$. Given S samples of (X_{s1}, Y_{s1}). Let $\hat{Q}$ be the estimator of Q given by*

$$\hat{Q} = \frac{1}{S}\sum_{s=1}^{S}\left(\mathbb{1}(g(X_{s1}, Y_{s1}) = C_1), \dots, \mathbb{1}(g(X_{s1}, Y_{s1}) = C_n)\right).$$

Given matrix $G \in \mathbb{R}^{n\times n}$, the entropy calculated using $\hat{Q}^T G$ is an unbiased estimator of the entropy calculated using $Q^T G$ as $S \to \infty$

$$H(\hat{Q}^T G) \xrightarrow{S\to\infty} H(Q^T G)$$

Proof. We use the proof from Basharin [20] as a template. Basharin [20] derived the Taylor expansion of the entropy:

$$\begin{aligned} H(\hat{Q}) = H(Q) - \sum_{i=1}^{n}(\hat{q}_i - q_i)(1 + \ln q_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{(\hat{q}_i - q_i)^2}{q_i} \\ + \frac{1}{6}\sum_{i=1}^{n}\frac{(\hat{q}_i - q_i)^3}{q_i^2} - \frac{1}{12}\sum_{i=1}^{n}\frac{(\hat{q}_i - q_i)^4}{(q_i + \theta(\hat{q}_i - q_i))^3}. \end{aligned}$$

They derive the following asymptotic values of the expectation of each term in the above equation:

$$\begin{aligned} &\mathbb{E}(\hat{q}_i) = q_i \\ &\mathbb{E}(\hat{q}_i - q_i)^2 = \frac{q_i(1-q_i)}{S} \\ &\mathbb{E}(\hat{q}_i - q_i)(\hat{q}_j - q_j) = \frac{q_i q_j}{S} \\ &\mathbb{E}(\hat{q}_i - q_i)(\hat{q}_j - q_j)(\hat{q}_u - q_u) = O(\frac{1}{S^2}) \\ &\mathbb{E}(\hat{q}_i - q_i)(\hat{q}_j - q_j)(\hat{q}_u - q_u)(\hat{q}_v - q_v) = O(\frac{1}{S^2}) \\ &\mathbb{E}(\frac{(\hat{q}_i - q_i)^4}{(q_i + \theta(\hat{q}_i - q_i))^3}) \le \mathbb{E}(\frac{(\hat{q}_i - q_i)^4}{q_i^3(1-\theta)^3}) = O(\frac{1}{S^2}). \end{aligned}$$

Using their derivation, the Taylor expansion of $H(\hat{Q}^T G)$ is

$$\begin{aligned} H(\hat{Q}^T G) = H(Q^T G) - \sum_{i=1}^{n}((\hat{Q}^T G)_i - (Q^T G)_i)(1 + \ln(Q^T G)_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{((\hat{Q}^T G)_i - (Q^T G)_i)^2}{(Q^T G)_i} \\ + \frac{1}{6}\sum_{i=1}^{n}\frac{((\hat{Q}^T G)_i - (Q^T G)_i)^3}{(Q^T G)_i^2} - \frac{1}{12}\sum_{i=1}^{n}\frac{((\hat{Q}^T G)_i - (Q^T G)_i)^4}{((Q^T G)_i + \theta((\hat{Q}^T G)_i - (Q^T G)_i))^3}. \end{aligned} \tag{12}$$

The term $(\hat{Q}^T G)_i$ and $(Q^T G)_i$ can be explicitly expressed as $(\hat{Q}^T G)_i = \sum_{j=1}^n \hat{q}_j g_{ji}$ and $(Q^T G)_i = \sum_{j=1}^n q_j g_{ji}$. Using the explicit expression, the expectation of the terms in the Taylor expression is

$$\begin{aligned}
&\mathbb{E}((\hat{Q}^T G)_i) = \mathbb{E}(\sum_{j=1}^n \hat{q}_j g_{ji}) = \sum_{j=1}^n \mathbb{E}(\hat{q}_j) g_{ji} = \sum_{j=1}^n q_j g_{ji} = (Q^T G)_i \\
&\mathbb{E}((\hat{Q}^T G)_i - (Q^T G)_i)^2 = \mathbb{E}(\sum_{j=1}^n (\hat{q}_j - q_j) g_{ji})^2 = \\
&\qquad \sum_{j,u} \mathbb{E}(\hat{q}_j - q_j)(\hat{q}_u - q_u) g_{ji} g_{ui} = \sum_{j,u} O(\frac{1}{S}) g_{ji} g_{ui} = O(n^2) O(\frac{1}{S}) \\
&\mathbb{E}((\hat{Q}^T G)_i - (Q^T G)_i)((\hat{Q}^T G)_j - (Q^T G)_j) = \mathbb{E}(\sum_{u=1}^n (\hat{q}_u - q_u) g_{ui})(\sum_{u=1}^n (\hat{q}_u - q_u) g_{uj}) = O(n^2) O(\frac{1}{S}) \\
&\mathbb{E}(\hat{Q}^T G)_i - (Q^T G)_i)(\hat{Q}^T G)_j - (Q^T G)_j)(\hat{Q}^T G)_u - (Q^T G)_u) = O(n^3) O(\frac{1}{S^2}) \\
&\mathbb{E}(\hat{Q}^T G)_i - (Q^T G)_i)(\hat{Q}^T G)_j - (Q^T G)_j)(\hat{Q}^T G)_u - (Q^T G)_u)(\hat{Q}^T G)_v - (Q^T G)_v) = O(n^4) O(\frac{1}{S^2}) \\
&\mathbb{E}(\frac{(\hat{Q}^T G)_i - (Q^T G)_i)^4}{((Q^T G)_i + \theta(\hat{Q}^T G)_i - (Q^T G)_i))^3}) \leq \mathbb{E}(\frac{(\hat{Q}^T G)_i - (Q^T G)_i)^4}{((Q^T G)_i(1-\theta))^3}) = O(n^4) O(\frac{1}{S^2})
\end{aligned} \tag{13}$$

Plugging the equations and inequalities in (13) in the expectation of equation (12), the expectation of the entropy estimator is

$$\mathbb{E}(H(\hat{Q}^T G)) = H(Q^T G) + O(n^4) O(\frac{1}{S^2}).$$

Since the number of categories n is fixed and the number of samples S goes to infinity,

$$\mathbb{E}(H(\hat{Q}^T G)) \xrightarrow{S \to \infty} H(Q^T G).$$

Hence, we proved that the entropy $H(\hat{Q}^T G)$ is an unbiased estimator for entropy $H(Q^T G)$ as the number of samples S goes to infinity. □

S1.2 Estimation bias of log term using estimated density $\hat{p}$

Lemma 1. *Given the true probability density p and fixed point (x, y), for any estimator of the density at the point $\hat{p}(x, y)$, if the true and estimated density are lower bounded by $\delta > 0$, that is, $p(x, y) \geq \delta$ and $\hat{p}(x, y) \geq \delta$, then the bias of $\log(\hat{p})$ at the point is upper bounded by:*

$$|\mathbb{E}(\log \hat{p}(x, y)) - \log p(x, y)| \leq \frac{1}{p(x, y)} |\mathbb{E}(\hat{p}(x, y)) - p(x, y)| + \frac{\mathrm{Var}(\hat{p}(x, y)) + (\mathbb{E}(\hat{p}(x, y)) - p(x, y))^2}{\delta^2}$$

Proof. The Taylor expansion on the log function with the mean value form states that there exists $\theta \in (0, 1)$ such that

$$\log \hat{p}(x, y) = \log p(x, y) + \frac{1}{p(x, y)} (\hat{p}(x, y) - p(x, y)) - \frac{1}{(\theta \hat{p}(x, y) + (1-\theta) p(x, y))^2} (\hat{p}(x, y) - p(x, y))^2.$$

Using the lower bound δ of $p(x,y)$ and $\hat{p}(x,y)$, the second-order term can be bounded by

$$-\frac{(\hat{p}(x,y)-p(x,y))^2}{\delta^2} \leq \frac{(\hat{p}(x,y)-p(x,y))^2}{(\theta\hat{p}(x,y)+(1-\theta)p(x,y))^2} \leq \frac{(\hat{p}(x,y)-p(x,y))^2}{\delta^2},$$

and thus the following inequalities hold:

$$-\frac{(\hat{p}(x,y)-p(x,y))^2}{\delta^2} \leq \log\hat{p}(x,y) - \log p(x,y) - \frac{1}{p(x,y)}(\hat{p}(x,y)-p(x,y)) \leq \frac{(\hat{p}(x,y)-p(x,y))^2}{\delta^2}.$$

Taking the expectation on both sides, the equation can be arranged in the following way:

$$\begin{aligned}-\frac{\mathbb{E}[(\hat{p}(x,y)-p(x,y))^2]}{\delta^2} \leq \mathbb{E}(\log\hat{p}(x,y))) - \log p(x,y) - \frac{1}{p(x,y)}(\mathbb{E}(\hat{p}(x,y)) - p(x,y)) \\ \leq \frac{\mathbb{E}[(\hat{p}(x,y)-p(x,y))^2]}{\delta^2}.\end{aligned}$$

The two inequalities can be combined using the absolute value form:

$$|\mathbb{E}(\log\hat{p}(x,y))) - \log p(x,y) - \frac{1}{p(x,y)}(\mathbb{E}(\hat{p}(x,y)) - p(x,y))| \leq \frac{\mathbb{E}[(\hat{p}(x,y)-p(x,y))^2]}{\delta^2}.$$

Applying the triangle inequality of the absolute value on the left hand side, we have:

$$|\mathbb{E}(\log\hat{p}(x,y))) - \log p(x,y)| - |\frac{1}{p(x,y)}(\mathbb{E}(\hat{p}(x,y)) - p(x,y))| \leq \frac{\mathbb{E}[(\hat{p}(x,y)-p(x,y))^2]}{\delta^2}.$$

Replacing the term $\mathbb{E}(\hat{p}(x,y)-p(x,y))^2$ with the bias-variance decomposition formula, therefore, we derive the upper bound of the estimation bias of the $\log$ term:

$$|\mathbb{E}(\log\hat{p}(x,y))) - \log p(x,y)| \leq \frac{1}{p(x,y)}|\mathbb{E}(\hat{p}(x,y)) - p(x,y)| + \frac{\mathrm{Var}(\hat{p}(x,y)) + (\mathbb{E}(\hat{p}(x,y)) - p(x,y))^2}{\delta^2}$$

□

S1.3 The corrected density estimator in the continuous case is unbiased in estimating the exponential term in KDE

We use the following lemma to prove the unbiasedness of density estimator.

Lemma S3. *Given n-dimensional Gaussian random variable $X \sim N(\mu, \Sigma)$ and matrix $A \in \mathbb{R}^{n\times n}$, the moment generating function for $Q = X^TAX$ is*

$$\mathbb{E}(\exp(tQ)) = \left(\prod_{j=1}^{n}\frac{1}{\sqrt{1-2t\lambda_j}}\right)\exp\left\{\mu^T\Sigma^{-\frac{1}{2}}P^T\begin{pmatrix}\frac{t\lambda_1}{1-2t\lambda_1} & & \\ & \ddots & \\ & & \frac{t\lambda_n}{1-2t\lambda_n}\end{pmatrix}P\Sigma^{-\frac{1}{2}}\mu\right\} \tag{14}$$

where P and λ_i are the eigenvectors and eigenvalues of $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$, that is, $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}} = P^T\Lambda P$, and t is constrained by $1-2t\lambda_i > 0$ for all i.

Proof. The proof was originally given by user "kjetil b halvorsen" on StackExchange (https://stats.stackexchange.com/questions/262604/what-is-the-moment-generating-function-of-the-generalized-multivariate-chi-squ/318908#318908). We reproduce it here for completeness to make our argument self contained.

P and λ_i are the eigenvectors and eigenvalues of $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$, they can be equivalently expressed by the equation of $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}} = P^T\Lambda P$, where P is orthonormal and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

Let $U = P(\Sigma^{-\frac{1}{2}}X - \Sigma^{-\frac{1}{2}}\mu)$ and $b = P\Sigma^{-\frac{1}{2}}\mu$. $U + b$ is an affine function of Gaussian random variable X, thus the distribution of $U + b$ is a also Gaussian distribution. The expectation is

$$\mathbb{E}(U+b) = \mathbb{E}(P(\Sigma^{-\frac{1}{2}}X - \Sigma^{-\frac{1}{2}}\mu) + P\Sigma^{-\frac{1}{2}}\mu) = P\Sigma^{-\frac{1}{2}}\mathbb{E}X = P\Sigma^{-\frac{1}{2}}\mu.$$

The covariance matrix is

$$\mathrm{Cov}(U+b) = \mathrm{Cov}(P(\Sigma^{-\frac{1}{2}}X - \Sigma^{-\frac{1}{2}}\mu) + P\Sigma^{-\frac{1}{2}}\mu) = \mathrm{Cov}(P\Sigma^{-\frac{1}{2}}X) = P\Sigma^{-\frac{1}{2}}\Sigma(\Sigma^{-\frac{1}{2}})^T P^T = I.$$

Therefore, $U + b$ follows a Gaussian distribution where each entry is independent of the others:

$$U + b \sim N(P\Sigma^{-\frac{1}{2}}\mu, I).$$

Connecting $U + b$ back to Q, one can verify that $Q = X^TAX = (U+b)^T\Lambda(U+b) = \sum_i \lambda_i(U+b)_i^2$, which is the weighted sum of n independent non-central chi-squared random variables. Therefore,

$$\mathbb{E}(\exp(tQ)) = \mathbb{E}(\prod_i \exp(t\lambda_i(U+b)_i^2)) = \prod_i \mathbb{E}(\exp(t\lambda_i(U+b)_i^2)).$$

The product can be moved out of the expectation due to the independence between the chi-squared random variables $(U + b)_i$. Each of the summation term is the moment generating function of the non-central chi-squared random variable evaluated at point $t\lambda_i$. Let $\nu_i = (\mathbb{E}(U + b))_i = (P\Sigma^{-\frac{1}{2}}\mu)_i$. When $t\lambda_i < \frac{1}{2}$, the moment generating function evaluated at $t\lambda_i$ is

$$\mathbb{E}(\exp(t\lambda_i(U+b)_i^2)) = \frac{1}{\sqrt{1-2t\lambda_i}}\exp(\frac{2t\lambda_i\nu_i^2}{1-2t\lambda_i}).$$

When the inequality $t\lambda_i < \frac{1}{2}$ holds for all i, all the moment generating functions is finite at the corresponding point, and all terms in the production is well-defined. Plugging the expression of each term in the production,

$$\begin{aligned}
\mathbb{E}(\exp(t\lambda_i(U+b)_i^2)) &= \left(\prod_i \frac{1}{\sqrt{1-2t\lambda_i}}\right)\exp\left\{\sum_i \frac{2t\lambda_i\nu_i^2}{1-2t\lambda_i}\right\} \\
&= \left(\prod_i \frac{1}{\sqrt{1-2t\lambda_i}}\right)\exp\left\{\nu^T\begin{pmatrix}\frac{2t\lambda_1}{1-2t\lambda_1} & & \\ & \ddots & \\ & & \frac{2t\lambda_n}{1-2t\lambda_n}\end{pmatrix}\nu\right\} \\
&= \left(\prod_i \frac{1}{\sqrt{1-2t\lambda_i}}\right)\exp\left\{\mu^T\Sigma^{-\frac{1}{2}}P^T\begin{pmatrix}\frac{2t\lambda_1}{1-2t\lambda_1} & & \\ & \ddots & \\ & & \frac{2t\lambda_n}{1-2t\lambda_n}\end{pmatrix}P\Sigma^{-\frac{1}{2}}\mu\right\}
\end{aligned}$$

□

Theorem 2. *Let $\bar{X}_{s\bullet} = \frac{1}{B}\sum_{j=1}^{B} X_{sj}$ and $\bar{Y}_{s\bullet} = \frac{1}{B}\sum_{i=1}^{B} Y_{sj}$. Let W be an diagonal bandwidth matrix used in KDE. Assuming the measurement error of $\left(\bar{X}_{s\bullet} - \bar{X}_{t\bullet}, \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet}\right)^T$ follows a Gaussian distribution $N((\mu_{xs} - \mu_{xt}, \mu_{ys} - \mu_{yt})^T, \frac{2}{B}\Sigma_\epsilon)$, let P and $\{\zeta_1, \zeta_2\}$ be the eigenvectors and eigenvalues of $\frac{1}{B}\Sigma_\epsilon^{\frac{1}{2}} W \Sigma_\epsilon^{\frac{1}{2}}$. Let $\lambda_i = \frac{\zeta_i}{1-2\zeta_i}$, $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, and $A = \frac{B}{2}\Sigma_\epsilon^{-\frac{1}{2}} P^T \Lambda P \Sigma_\epsilon^{-\frac{1}{2}}$. When ζ_i satisfies $\zeta_i < \frac{1}{2}$ for all i, the following estimator,*

$$\left(\prod_{i=1}^{2} \sqrt{1+2\lambda_i}\right) \exp\left\{-\begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}^T A \begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}\right\},$$

is an unbiased estimator for the KDE term in equation (5)

$$\exp\left\{-\frac{1}{2}\begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}^T W \begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}\right\}.$$

Proof. Plugging in $t = -1$ and $X = \begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}$ in Lemma S3. After calculating the eigenvalues and eigenvectors of $(\frac{2}{B}\Sigma_\epsilon)^{\frac{1}{2}} A (\frac{2}{B}\Sigma_\epsilon)^{\frac{1}{2}}$, we have the eigenvalues are $\{\lambda_1, \lambda_2\}$ and the eigenvectors are P. When expressing ζ_i using λ_i, the constraint $\zeta_i < \frac{1}{2}$ under $t = -1$ indicates that $1 - 2t\lambda_i > 0$, that is, the condition of the above lemma holds. Thus, the expectation of the estimator is:

$$\begin{aligned}
&\mathbb{E}(\left(\prod_{j=1}^{2} \sqrt{1+2\lambda_i}\right) \exp\{-\begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}^T A \begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}\}) \\
&= \left(\prod_{j=1}^{2} \sqrt{1+2\lambda_i}\right)\left(\prod_{j=1}^{2} \frac{1}{\sqrt{1+2\lambda_j}}\right) \\
&\qquad \exp\left\{\begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}^T (\frac{2}{B}\Sigma_\epsilon)^{-\frac{1}{2}} P^T \begin{pmatrix} \frac{-\lambda_1}{1+2\lambda_1} & \\ & \frac{-\lambda_2}{1+2\lambda_2} \end{pmatrix} P (\frac{2}{B}\Sigma_\epsilon)^{-\frac{1}{2}} \begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}\right\} \\
&= \exp\left\{\frac{B}{2}\begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}^T \Sigma_\epsilon^{-\frac{1}{2}} P^T \begin{pmatrix} -\zeta_1 & \\ & -\zeta_2 \end{pmatrix} P \Sigma_\epsilon^{-\frac{1}{2}} \begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}\right\} \\
&= \exp\left\{\frac{B}{2}\begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}^T \Sigma_\epsilon^{-\frac{1}{2}} (-\frac{1}{B}\Sigma_\epsilon^{\frac{1}{2}} W \Sigma_\epsilon^{\frac{1}{2}}) \Sigma_\epsilon^{-\frac{1}{2}} \begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}\right\} \\
&= \exp\left\{-\frac{1}{2}\begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}^T W \begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}\right\}
\end{aligned}$$

□

S1.4 Corrected density estimator is asymptotically unbiased when using estimated error covariance

Lemma S4. *Given a series of n-dimensional Gaussian random variables $X_n \xrightarrow{d} N(\mu, \Sigma)$ where each of X_n has a finite moment generating function for $t \in (a, b)$. Given matrix $A \in \mathbb{R}^{n \times n}$, the moment generating* *function for $Q_n = X_n^T A X_n$ is*

$$\mathbb{E}(\exp(tQ_n)) \longrightarrow \left(\prod_{j=1}^{n} \frac{1}{\sqrt{1-2t\lambda_j}}\right) \exp\left\{\mu^T \Sigma^{-\frac{1}{2}} P^T \begin{pmatrix} \frac{t\lambda_1}{1-2t\lambda_1} & & \\ & \ddots & \\ & & \frac{t\lambda_n}{1-2t\lambda_n} \end{pmatrix} P\Sigma^{-\frac{1}{2}}\mu\right\} \tag{15}$$

pointwise for each t values within the region $(\cap_i(0, \frac{1}{2\lambda_i})) \cap (a, b)$, where P and λ_i are the eigenvectors and eigenvalues of $\Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}$.

Proof. By the same construction of $U_n = P(\Sigma^{-\frac{1}{2}} X_n - \Sigma^{-\frac{1}{2}}\mu)$ and $b = P\Sigma^{-\frac{1}{2}}\mu$, the limiting distribution is also a Gaussian distribution by an affine transformation of the random variable:

$$U_n + b \xrightarrow{d} N(P\Sigma^{-\frac{1}{2}}\mu, I).$$

The limiting distribution for $(U_n + b)_i^2$ is a chi-squared distribution because of the following. Let $\nu = P\Sigma^{-\frac{1}{2}}\mu$. The cumulative distribution function of $(U_n + b)_i^2$ is:

$$F(t) = \mathbb{P}((U_n + b)_i^2 \le t) = \mathbb{P}(-\sqrt{t} \le (U_n + b)_i \le \sqrt{t}) \to \Phi(t - \nu_i) - \Phi(-t - \nu_i),$$

where the right hand side is the same as the cumulative distribution function of a non-central chi-squared distribution with degrees of freedom equal to 1.

Chareka [35] has proved that when the moment generating function of each $(U_n + b)_i$ is finite at $t \in (a, b)$, the convergence of distribution implies the convergence of moment generating function. Using the moment generating function calculated in Lemma S3, $\mathbb{E}(\exp(tQ_n))$ converges to the moment generating function of the summation of chi-squared random variables in the limiting distribution. □

Theorem 3. *Let P and $\{\zeta_1, \zeta_2\}$ be the eigenvectors and eigenvalues of $\frac{1}{B}\hat{\Sigma}_\epsilon^{\frac{1}{2}} W \hat{\Sigma}_\epsilon^{\frac{1}{2}}$. Let $\lambda_i = \frac{\zeta_i}{1-2\zeta_i}$, $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, and $\hat{A} = \frac{B}{2}\hat{\Sigma}_\epsilon^{-\frac{1}{2}} P^T \Lambda P \hat{\Sigma}_\epsilon^{-\frac{1}{2}}$. When ζ_i satisfies $\zeta_i < \frac{1}{2}$ for all i, the following estimator is an asymptotically unbiased estimator for the KDE term in equation* (5)*:*

$$\left(\prod_{i=1}^{2} \sqrt{1+2\lambda_i}\right) \exp\left\{-\begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}^T \hat{A} \begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}\right\}.$$

Proof. The proof is almost the same as the proof for Theorem 2 except that Lemma S4 is used for calculating the moment generating function of the limiting distribution. Set $t = -1$ and $X = \begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}$. The expectation of the estimator approaches to

$$\mathbb{E}(\left(\prod_{j=1}^{2} \sqrt{1+2\lambda_i}\right) \exp\{-\begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}^T \hat{A} \begin{pmatrix} \bar{X}_{s\bullet} - \bar{X}_{t\bullet} \\ \bar{Y}_{s\bullet} - \bar{Y}_{t\bullet} \end{pmatrix}\}) \to$$

$$\exp\left\{\begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}^T (\frac{2}{B}\Sigma_\epsilon)^{-\frac{1}{2}} P^T \begin{pmatrix} \frac{-\lambda_1}{1+2\lambda_1} & \\ & \frac{-\lambda_2}{1+2\lambda_2} \end{pmatrix} P(\frac{2}{B}\Sigma_\epsilon)^{-\frac{1}{2}} \begin{pmatrix} \mu_{xs} - \mu_{xt} \\ \mu_{ys} - \mu_{yt} \end{pmatrix}\right\}$$

Substituting the value of λ_i and the equality relationship between P, Σ_ϵ, and ζ_i, the above expectation converges to the KDE term of the true signal values:

$$\mathbb{E}(\left(\prod_{j=1}^{2}\sqrt{1+2\lambda_i}\right)\exp\{-\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}^T \hat{A}\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s1}-\bar{Y}_{t\bullet}\end{pmatrix}\}) \to \exp\left\{-\frac{1}{2}\begin{pmatrix}\mu_{xs}-\mu_{xt}\\ \mu_{ys}-\mu_{yt}\end{pmatrix}^T W\begin{pmatrix}\mu_{xs}-\mu_{xt}\\ \mu_{ys}-\mu_{yt}\end{pmatrix}\right\}$$

□

S1.5 The expectation of corrected density estimator is larger than its second order moment when λ_i is small

Claim 1. *There exists a small positive number δ such that when $0 \le \lambda_i \le \delta$,*

$$\mathbb{E}\left(\left(\prod_{i=1}^{2}\sqrt{1+2\lambda_i}\right)\exp\left\{-\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}^T A\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}\right\}\right) \ge$$

$$\mathbb{E}\left[\left(\left(\prod_{i=1}^{2}\sqrt{1+2\lambda_i}\right)\exp\left\{-\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}^T A\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}\right\}\right)^2\right].$$

Proof. The expectation, as the first step of proof of Theorem 2, is

$$\mathbb{E}(\left(\prod_{j=1}^{2}\sqrt{1+2\lambda_i}\right)\exp\{-\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}^T A\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}\}) =$$

$$\exp\left\{\begin{pmatrix}\mu_{xs}-\mu_{xt}\\ \mu_{ys}-\mu_{yt}\end{pmatrix}^T (\frac{2}{B}\Sigma_\epsilon)^{-\frac{1}{2}}P^T\begin{pmatrix}\frac{-\lambda_1}{1+2\lambda_1} & \\ & \frac{-\lambda_n}{1+2\lambda_n}\end{pmatrix}P(\frac{2}{B}\Sigma_\epsilon)^{-\frac{1}{2}}\begin{pmatrix}\mu_{xs}-\mu_{xt}\\ \mu_{ys}-\mu_{yt}\end{pmatrix}\right\}$$

The second order moment around 0 has the expression of

$$\left(\prod_{j=1}^{2}(1+2\lambda_i)\right)\mathbb{E}(\exp\{-2\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}^T A\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}\}).$$

This can be calculated by the same steps as in the proof of Theorem 2 except setting $t=-2$:

$$\left(\prod_{j=1}^{2}(1+2\lambda_i)\right)\mathbb{E}(\exp\{-2\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}^T A\begin{pmatrix}\bar{X}_{s\bullet}-\bar{X}_{t\bullet}\\ \bar{Y}_{s\bullet}-\bar{Y}_{t\bullet}\end{pmatrix}\})$$

$$=\left(\prod_{j=1}^{2}(1+2\lambda_i)\right)\left(\prod_{j=1}^{2}\frac{1}{\sqrt{1+4\lambda_j}}\right)$$

$$\exp\left\{\begin{pmatrix}\mu_{xs}-\mu_{xt}\\ \mu_{ys}-\mu_{yt}\end{pmatrix}^T (\frac{2}{B}\Sigma_\epsilon)^{-\frac{1}{2}}P^T\begin{pmatrix}\frac{-2\lambda_1}{1+4\lambda_1} & \\ & \frac{-2\lambda_2}{1+4\lambda_2}\end{pmatrix}P(\frac{2}{B}\Sigma_\epsilon)^{-\frac{1}{2}}\begin{pmatrix}\mu_{xs}-\mu_{xt}\\ \mu_{ys}-\mu_{yt}\end{pmatrix}\right\}$$

Let $\begin{pmatrix}t_1\\ t_2\end{pmatrix} = (P(\frac{2}{B}\Sigma_\epsilon)^{-\frac{1}{2}}\begin{pmatrix}\mu_{xs}-\mu_{xt}\\ \mu_{ys}-\mu_{yt}\end{pmatrix})_1$. The ratio between the expectation and the second order moment is

$$\prod_j\sqrt{\frac{1+4\lambda_j}{1+4\lambda_j+4\lambda_j^2}}\exp\{t_j^2(\frac{-\lambda_j}{1+2\lambda_j}-\frac{-2\lambda_j}{1+4\lambda_j})\}.$$

We need to prove that within a small positive region $(0, \delta]$, each product term in the above equation is greater than 1. Or equivalently, we need to prove the rearranged form:

$$\frac{-\lambda_j}{1+2\lambda_j} - \frac{-2\lambda_j}{1+4\lambda_j} \geq \frac{1}{2t_j^2} \log \frac{1+4\lambda_j+4\lambda_j^2}{1+4\lambda_j}$$

Let function $f(x) = \frac{-x}{1+2x} - \frac{-2x}{1+4x} - C \log \frac{1+4x+4x^2}{1+4x}$, where $C = \frac{1}{2t_j^2}$. The function is 0 when x takes value of 0, that is, $f(0) = 0$. When $x > 0$, the derivative is

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{1-8Dx}{(1+2x)^2(1+4x)^2},$$

where $D = 1 + C(1+4x)(1+2x)$. Thus, there exists a positive region $(0, \delta]$ such that when $x \in (0, \delta]$, $\frac{\mathrm{d}f}{\mathrm{d}x} \geq 0$. And in this region,

$$f(x) \geq f(0) = 0.$$

This is equivalent to:

$$\frac{-x}{1+2x} - \frac{-2x}{1+4x} \geq C \log \frac{1+4x+4x^2}{1+4x}$$

Therefore, when $\lambda_1, \lambda_2 \in (0, \delta]$, the ratio between the expectation and the second order moment is larger than 1, and the second order moment is upper bounded by the expectation.

□