Data efficiency, dimensionality reduction, and the generalized symmetric information bottleneck

K. Michael Martini[1, 3] **and Ilya Nemenman**[1, 2, 3]
[1] Department of Physics, Emory University, Atlanta, GA 30322.
[2] Department of Biology, Emory University, Atlanta, GA 30322.
[3] Initiative in Theory and Modeling of Living Systems, Emory University, Atlanta, GA 30322.

Keywords: Information Bottleneck, Symmetric Information Bottleneck, Dimensionality Reduction, Error Bounds, Data Efficiency

Abstract

The Symmetric Information Bottleneck (SIB), an extension of the more familiar Information Bottleneck, is a dimensionality reduction technique that simultaneously compresses two random variables to preserve information between their compressed versions. We introduce the Generalized Symmetric Information Bottleneck (GSIB), which explores different functional forms of the cost of such simultaneous reduction. We then explore the dataset size requirements of such simultaneous compression. We do this by deriving bounds and root-mean-squared estimates of statistical fluctuations of the involved loss functions. We show that, in typical situations, the simultaneous GSIB compression requires qualitatively less data to achieve the same errors compared to compressing variables one at a time. We suggest that this is an example of a more general principle that simultaneous compression is more data efficient than independent compression of each of the input variables.

1 Introduction

Recent years have seen an explosion of large-dimensional experimental data sets (de Vries et al., 2020; Siegle et al., 2021; Haghighi et al., 2022) and the parallel growth in the number of methods for *dimensionality reduction* (DR)—that is, for extracting low-dimensional structure from large-dimensional data (Carreira-Perpinán, 1997; Van Der Maaten et al., 2009; Nanga et al., 2021). Broadly speaking, we classify dimensionality reduction methods into two classes: unsupervised and supervised. Unsupervised DR methods seek a low-dimensional description, T_X, of a large-dimensional variable, X, that preserves its variance, entropy, or another measure of diversity of the data. Such methods include the familiar principal component analysis (PCA) (Hotelling, 1933), non-negative matrix factorization (Lee and Seung, 1999), multidimensional scaling (MDS) (Kruskal, 1964), t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008), Isomap (Tenenbaum et al., 2000), Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), autoencoders (Hinton and Salakhutdinov, 2006), and related techniques (Kingma and Welling, 2014). In contrast, supervised DR techniques aim to find a low-dimensional description, T_X, of a large dimensional X, while preserving T_X's ability to explain another variable Y, which provides an effective *relevance* or *supervision* signal. Common examples include variable selection in regression (Andersen and Bro, 2010; Kuo and Mallick, 1998), cross-encoders, Bayesian Ising Approximation (BIA) (Fisher and Mehta, 2015), and the Information Bottleneck (IB) (Tishby et al., 2000; Tishby and Slonim, 2000). A particularly interesting class of such supervised dimensionality reduction problems is when both the reduced variable X and the relevance variable Y are large-dimensional. In these situations, finding significant correlations within combinatorially many groups of components of X and Y is hard, suggesting parallel dimensionality reduction of both X and Y into T_X and T_Y, respectively.

We distinguish three classes of approaches to this problem. In the first, which we call the *Independent Unsupervised Dimensionality Reduction* (IUDR), one applies unsupervised DR methods to X and Y independently. One then searches for statistical dependencies between T_X and Y or T_Y and X or T_X and T_Y, but the dimensionality reduction itself is agnostic of this subsequent step. A familiar example of this is the Principal Components Regression, where the projections on the principal components of X are regressed against Y. We also distinguish *Independent Supervised Dimensionality Reduction* (ISDR), where T_X is produced by compressing X with Y as the supervision signal, while T_Y emerges from compressing Y with X as the supervision. The Information Bottleneck (IB) (Tishby et al., 2000), the Generalized and Deterministic Information Bottleneck (GIB) (Strouse and Schwab, 2017), and cross-encoders are examples of such approaches. Finally, *Simultaneous Supervised Dimensionality Reduction* (SSDR) is a class of methods where T_X and T_Y are produced simultaneously, typically being supervision signals of each other.[1] Examples of SSDR include the Canonical Correlation Analysis (CCA) (Hotelling, 1936; Yang et al., 2021) and its modern nonlinear neural network based generalizations (Andrew et al., 2013; Chapman and Wang, 2021), Partial Least Squares (PLS) (Wold, 1966; Wold et al., 2001), and the Symmetric version of the Information Bottleneck (SIB) (Slonim et al., 2006).[2] In this paper we introduce a Generalized version of the Symmetric Information Bottleneck (GSIB) by interpolating between the compression cost measured by entropy and information. This parallels for SSDR the introduction of the Generalized Information Bottleneck (GIB) for ISDR,

[1] SSDR methods are sometimes referred to as *dual* DR (Sponberg et al., 2015a). We believe that the terminology we propose here is better suited for classifying the breadth of DR approaches.

[2] Different sources refer to CCA or PLS as both supervised or unsupervised techniques or something inbetween (Holbrook et al., 2017; Scott and Crone, 2021; Zhuang et al., 2020). Within our classification scheme, these are supervised methods.

of which the Deterministic Information Bottleneck and the Information Bottleneck are limits (Strouse and Schwab, 2017).

We then argue that SSDR approaches can require a lot fewer data than their ISDR counterparts to achieve the same accuracy. We demonstrate this by comparing the bias and statistical fluctuations in the objective functions of independent GIB reductions of variables X and Y (ISDR approach) with the corresponding bias and fluctuations for the GSIB (SSDR approach). We show that the bias for the GSIB scales as the product of cardinalities of the compressed variables, while the bias for the GIB scales as the (typically much larger) product of cardinalities of the supervision signal and the compressed variable. We do the comparison for both typical fluctuations and for the upper bounds on the fluctuations. While our derivations are done for the IB approaches only, the intuitive explanation of the differences between the approaches suggests that SSDR methods are likely to require less data than their ISDR analogues more generally.

2 Background: Information Bottleneck and the Symmetric Information Bottleneck

2.1 Information Bottleneck and Its Generalizations

The goal of *Information Bottleneck* (IB) is to produce a compression, T_X of a random variable X, such that the compression retains as much information as possible about another random variable Y, which is called the *relevant* (or, in our language, the *supervising*) variable. The information is measured using Shannon's mutual information (Shannon, 1948), which quantifies the difference between the joint probability distribution $p(x, y)$ and the product of the marginal distributions $p(x)p(y)$:

$$I(X, Y) = \sum_{X,Y} p(x, y) \frac{\log(p(x, y))}{p(x)p(y)} = H(X) - H(X|Y), \tag{1}$$

where $H(X)$ is the entropy of the variable X and $H(X|Y)$ is the conditional entropy of X given Y, $H(X|Y) = \sum_Y p(y)H(X|Y = y) = \sum_Y p(y) \sum_X p(x|y) \log(p(x|y))$. Mutual information is symmetric, always non-negative, and is only zero when the random variables are independent (Cover, 1999).

To achieve its goal, IB produces a probabilistic mapping from X to T_X, $p(t_x|x)$, which minimizes a specific cost function. The cost function trades off preserving the information in the compression about the relevant variable, $I(T_X, Y)$, against losing the information about X (reducing the variable), $I(T_X, X)$:

$$L_{\text{IB}} = I(T_X, X) - \beta I(T_X, Y). \tag{2}$$

Here β is the trade-off parameter, which controls how important the compression $I(T_X, X)$ is compared to preserving the relevant information $I(T_X, Y)$. As $\beta \to \infty$, the cost function is minimized by having no compression, $X = T$. Recently a Generalized version of IB was proposed (GIB) (Strouse and Schwab, 2017), which changes the cost function to

$$L_{\text{GIB}} = H(T_X) - \alpha_x H(T_X|X) - \beta I(T_X, Y), \tag{3}$$

which has a formal solution

$$p(t_x|x) = \frac{1}{Z(\beta,\alpha)} \exp\left[\frac{1}{\alpha_x}\left(\log p(t_x) - \beta D_{\mathrm{KL}}(p(y|x)||p(y|t_x))\right)\right], \tag{4}$$

$$p(y|t_x) = \frac{1}{p(t_x)} \sum_X p(t_x|x) p(x,y), \tag{5}$$

where D_{KL} is the usual Kullback–Leibler divergence (Kullback and Leibler, 1951).

The original IB is recovered from GIB when $\alpha_x = 1$. In contrast, when $\alpha_x \to 0$, $I(T_X, X)$ is replaced with $H(T_X)$ in the cost function. This corresponds to replacing the cost of having a noisy channel encoding X into T_X with the cost of directly storing T_X. In this case, the formal solution results in a deterministic mapping between X and T_X, and the resulting problem is known as the *Deterministic Information Bottleneck* (DIB) (Strouse and Schwab, 2017).

If both X and Y are large-dimensional and require dimensionality reduction, one can apply IB to produce the mapping $X \to T_X$ with Y as the relevant variable, and then solve a separate IB problem to map $Y \to T_Y$ with X as the supervision. This approach would fall into the ISDR class in our nomenclature.

2.2 Symmetric Information Bottleneck and its Generalization

The Symmetric Information Bottleneck (SIB), introduced in Slonim et al. (2006), is an SSDR approach, where X and Y are compressed simultaneously, such that the compressed versions T_X, and T_Y contain the maximal amount of information about each other. This corresponds to optimizing the loss function:

$$L_{\mathrm{SIB}} = I(T_X;X) + I(T_Y;Y) - \beta I(T_X;T_Y), \tag{6}$$

where optimization is over all possible probabilistic compressions $p(t_x|x)$ and $p(t_y|y)$. As before, β determines the strength of the trade-off between the compression and preserving the relevant information.

For generality, here we propose a Generalized SIB (GSIB), which incorporates flexible compression terms, similar to how GIB was optained from IB. The new cost function is

$$\begin{aligned} L_{\mathrm{GSIB}} &= I_{\alpha_X}(T_X;X) + I_{\alpha_Y}(T_Y;Y) - \beta I(T_X;T_Y) & (7)\\ &= H(T_X) - \alpha_X H(T_X|X) + H(T_Y) - \alpha_Y H(T_Y|Y) - \beta I(T_X,T_Y). & (8)\end{aligned}$$

Here we defined shorthands $I_{\alpha_X}(T_X, X) = H(T_X) - \alpha_X H(T_X|X)$, and similarly for I_{α_Y}, and the cost function must be minimized with respect to $p(t_x|x)$ and $p(t_y|y)$. The parameters α_X and α_Y are what dictates how probabilistic the mapping between the uncompressed variables and their compressed versions is. In the limit $\alpha_x, \alpha_Y \to 0$, the mapping can be verified to be deterministic (see below), resulting in the Determistic SIB (DSIB). When $\alpha_X, \alpha_Y \to 1$, GSIB becomes the usual SIB.

Optimization of the cost function has a formal solution:

$$p(t_x|x) = \frac{\exp\left[\frac{1}{\alpha_X}\left(\ln p(t_x) - \beta D_{\mathrm{KL}}(p(t_y|x)||p(t_y|t_x))\right)\right]}{Z_x(x, \alpha_X, \beta)}, \tag{9}$$

$$p(t_y|y) = \frac{\exp\left[\frac{1}{\alpha_Y}\left(\ln p(t_y) - \beta D_{\mathrm{KL}}(p(t_x|y)||p(t_x|t_y))\right)\right]}{Z_y(y, \alpha_Y, \beta)}, \tag{10}$$

$$p(t_y|x) = \frac{\sum_Y p(t_y|y)p(x,y)}{p(x)}, \qquad p(t_y|t_x) = \frac{\sum_{X,Y} p(t_y|y)p(t_x|x)p(x,y)}{\sum_X p(t_x|x)p(x)}, \tag{11}$$

$$p(t_x|y) = \frac{\sum_X p(t_x|x)p(x,y)}{p(y)}, \qquad p(t_x|t_y) = \frac{\sum_{X,Y} p(t_y|y)p(t_x|x)p(x,y)}{\sum_Y p(t_y|y)p(y)}. \tag{12}$$

Similar to IB, this formal solution can be iterated starting from an initial guess for both $p(t_x|x)$ and $p(t_y|y)$.

Interestingly, parenthetically we note that, unlike for IB, there are now exponentially many, $\sim 2^{|T_X|+|T_Y|}$, trivial fixed points for this iteration scheme (here $|\cdot|$ denotes cardinality of the variable, so that the rest of our discussion focuses on random variables defined on discrete, finite sets of possible values). For example, a uniform distribution for both random mappings, $p(t_x|x) = 1/|T_X|$ and $p(t_y|y) = 1/|T_Y|$ is a fixed point of the iteration with the cost of zero, even though a uniform mapping, independent of the conditioning variable, is clearly not a useful compression. Furthermore, all distributions, where $p(t_x|x)$ is zero for several values of t_x and uniform otherwise, are also trivial fixed points. There are exponentially many distributions of this type. When $\alpha_x = \alpha_y = 1$, these distributions are part of a larger class of trivial fixed points, which includes all mappings independent of the data, i. e., $p(t_x|x) = A(t_x)$ and $p(t_y|y) = B(t_y)$. One can easily verify that the first derivative of L_{GSIB} vanishes for these solutions. The second derivative, which controls if these solutions are minima or maxima, is:

$$\frac{\partial^2 L_{\mathrm{GSIB}}}{\partial p(t_x|x)\partial p(t'_x|x')} = \frac{-p(x)}{A(t_x)}(p(x)-\alpha_X)\delta(x,x')\delta(t,t'_x) - \frac{p(x)p(x')}{A(t_x)}\delta(t_x,t'_x)(1-\delta(x,x')), \tag{13}$$

(with similar expression for the compression of Y). These trivial fixed points are maxima when $\alpha_x < p(x)$, and $\alpha_y < p(y)$. When $\alpha_x > p(x)$ and $\alpha_y > p(y)$, such as in the case of SIB, when $\alpha_X = \alpha_Y = 1$, the trivial fixed points are saddles. Thus solutions found by the iterative algorithm must be viewed with suspicion, and one should always verify if the algorithm got trapped by one of the trivial solutions. One may be worried that it would be difficult to find non-trivial solution of SIB among the sea of trivial fixed points. In fact, Ref. Abdelaleem et al. (2023a) shows that a variational version of SIB easily solves this problem.

In the limit of $\alpha_X, \alpha_Y \to 0$, the exponent in the formal solution blows up. As a result, one obtains a deterministic mapping from uncompressed variables to their compressions:

$$
\begin{aligned}
p(t_x|x) &= \delta(t_x, \tau_x(x)) && (14)\\
\tau_x(x) &= \text{argmax}_{t_x}\left[\ln p(t_x) - \beta D_{\mathrm{KL}}(p(t_y|x)||p(t_y|t_x))\right], && (15)\\
p(t_y|y) &= \delta(t_y, \tau_y(y)) && (16)\\
\tau_y(y) &= \text{argmax}_{t_y}\left[\ln p(t_y) - \beta D_{\mathrm{KL}}(p(t_x|y)||p(t_x|t_y))\right]. && (17)
\end{aligned}
$$

This is the Deterministic SIB (DSIB).

3 Results

To show that GSIB is more data efficient than two GIBs applied independently to X and to Y, we notice that, in practical applications, all of the information and entropy terms in the loss functions must be estimated from data. Estimation of information-theoretic quantities is a hard task, potentially as hard as estimating the underlying distributions themselves, largely due to the estimation bias (Antos and Kontoyiannis, 2001; Paninski, 2003). Crucially, for a DR algorithm to produce meaningful results, the empirically estimated loss function must accurately represent the true loss function, which is unknown to us. Thus the question of which algorithm is more data efficient is equivalent to a different question: for which of the considered IB algorithms does the estimate of the respective loss function converge faster to its true value as the sample size grows?

A lot of ink has been expended on the problem of mutual information estimation (Roulston, 1999; Kraskov et al., 2004; Goebel et al., 2005; Belghazi et al., 2018). Here we do not try to produce better estimation techniques. Instead we focus on discrete random variables with finite cardinalities, and we use the simplest estimator, known as plug-in, naive, or maximum likelihood estimator, for estimation of all of the terms in the loss functions (Roulston, 1999; Paninski, 2003). For this estimator, which we denote with $\hat{\cdot}$, the probability distribution $p(x)$ is estimated by its maximum likelihood (ML) value, namely the frequency of an outcome in the sample, $\hat{p}(x) = n(x)/N$, where $n(x)$ is the number of times x occurred, and N is the total number of samples. Then $\hat{H}$, $\hat{I}$, and $\hat{L}$ are all given by plugging in $\hat{p}$ instead of p in the expression for these quantities. Shamir et al. (2010) showed that, while the ML estimator of mutual information $\hat{I}(X,Y)$ is guaranteed to converge to the true value only when $N \gg |X||Y|$, the ML estimator of the loss function, $\hat{L}_{\mathrm{IB}}$, converges at much smaller N, making IB more practical than one would naively think.

Here we continue this line of analysis and examine the convergence properties of $\hat{L}_{\mathrm{GSIB}}$ and $\hat{L}_{\mathrm{GIB}}$ when both $|X|, |Y| \gg 1$ in two different ways. First, we extend the derivations of Shamir et al. (2010) and bound the error of estimating each information-theoretic term in each of the loss functions from data. This allows us to build bounds on how close L and $\hat{L}$ are, and we can compare these bounds for GSIB and GIBs. Second, inspired by Still and Bialek (2004), we calculate the standard deviation and bias of $L - \hat{L}$ for different versions of the IB. By both measures, for $|X|, |Y| \gg 1$, $\hat{L}_{\mathrm{GSIB}}$ will have a smaller bias then $\hat{L}_{\mathrm{GIB}}$. This is our main result, allowing us to claim that the symmetric version of IB is more data efficient.

3.1 Bounds on The Loss Functions

The loss functions L_{GSIB} and L_{GIB} consist of multiple mutual information and entropy terms. We calculate bounds on the fluctuations between each of these terms and their estimators, and then combine them into a single estimate of the fluctuations of each loss function. We do this below in detail for $I(T_X;X)$ and its estimator $\hat{I}(T_X;X)$. Analysis of the other terms is similar. Furthermore, for our analysis, only the distributions of x and y are unknown, and must be sampled from data. The distributions $p(t_x|x)$ and $p(t_y|y)$ are chosen by the algorithm and optimized over. That is, they are *known* in any particular iteration of the scheme. Thus they do not produce fluctuations in the loss function directly, but only through the induced $p(t_x,t_y)$, which fluctuate. This means that, as first noticed in Ref. Shamir et al. (2010), some terms do not contribute to the fluctuation bounds, simplifying the results. Crucially, our expressions below will hold for all mappings $p(t_x|x)$ and $p(t_y|y)$, and not just the mappings that minimize their respective loss functions.

To estimate $|I(T_X;X)-\hat{I}(T_X;X)|$, we compare both terms to the expected value of the empirical information $E(I(T_X;X))$:

$$\begin{aligned}|\hat{I}(T_X;X)-I(T_X;X)| &= |\hat{I}(T_X;X)-E(\hat{I}(T_X;X))+E(\hat{I}(T_X;X))-I(T_X;X)|\\ &\leq |\hat{I}(T_X;X)-E(\hat{I}(T_X;X))|+|I(T_X;X)-E(\hat{I}(T_X;X))|. \quad (18)\end{aligned}$$

This is analogous to the usual bias-variance decomposition for bounds on the magnitude of fluctuations, with the first term in Eq. (18) representing the absolute deviation of the estimator, and the second the bias. We now bound the absolute deviation and the the bias terms separately.

First we focus on the absolute deviation (first) term in Eq. (18). For this, we follow Shamir et al. (2010) and rely on the the McDiarmid's inequality. This concentration inequality bounds the probability of the difference between a function of an empirical sample and its expected value. The bound is constructed from bounds on the change in the function due to changes in individual data points:

$$P\left[\left|f(x_1,x_2,\ldots,x_N)-E\left(f(x_1,x_2,\ldots,x_N)\right)\right|\geq\epsilon\right]\leq 2\exp\left[-\frac{2\epsilon^2}{\sum c_i}\right]\equiv\delta_1, \quad (19)$$

$$\text{where}\quad |f(x_1,\ldots,x_i,\ldots,x_N)-f(x_1,\ldots,x_i',\ldots,x_N)|\leq c_i. \quad (20)$$

Thus, to use the inequality, we consider the maximum change in $\hat{I}$ if a single datum is changed. That is, suppose the data point (x,y) is replaced by another data point (x',y'). Then the maximum likelihood estimator at the point (x,y), $\hat{p}(x,y)$, decreases by $1/N$. In contrast, $\hat{p}(x',y')$ increases by $1/N$, and the estimate does not change at all other x, y values. Similarly, the marginals $\hat{p}(x)$, $\hat{p}(x')$, $\hat{p}(y)$, and $\hat{p}(y')$ change by at most $1/N$, while marginals at all other values remain the same. For a fixed compression mapping, we calculate $\hat{p}(t_x)=\sum_x p(t_x|x)\hat{p}(x)$. We see that, with a single datum moving, $\hat{p}(t_x)$ can change by at most $|p((t_x|x')-p(t_x|x))|/N\leq 1/N$ for each $t_x\in T_X$. Similarly $\hat{p}(t_y)$ can change by at most $1/N$ for each $t_y\in T_Y$.

We now express the relevant mutual information in terms of entropy, $\hat{I}_{\alpha_X}(T_X;X)=\hat{H}(T_X)-\alpha_X\hat{H}(T_X|X)$, where the entropy $\hat{H}(T_X)$ depends on the probability density

$\hat{p}(t_x)$:

$$\hat{H}(T_X) = -\sum_{t_x} \hat{p}(t_x) \log \hat{p}(t_x). \tag{21}$$

The change in entropy from moving a single datum can be bounded using the following inequality, again borrowed from Shamir et al. (2010):

$$|(a+\delta)\log(a+\delta) - a\log a| \leq \log(N)/N \tag{22}$$

for any positive integer N and for any $a \in [0, 1-1/N]$ and $\delta \leq 1/N$. We apply this identity for each term in the sum in Eq. (21) and find that the change in $\hat{H}(T_X)$ is bounded by $|T_X| \log N/N$.

We bound the change in $\hat{H}(T_X|X) = \sum_x \hat{p}(x) H(T_X|X=x)$. $H(T_X|X=x)$ only depends on $p(t_x|x)$, which we consider fixed. $\hat{p}(x)$ changes by at most $1/N$ for two values of x. Thus the largest change is $|H(T_X|x') - H(T_X|x)|/N \leq |\max(H(T_X|x'), H(T_X|x))|/N \leq \log|T_X|/N$. The last inequality comes from $H(T_X|X=x) \leq \log|T_X|$, with the bound achieved for the uniform distribution.

Finally, combining the results for both entropy terms, we see that $\hat{I}_{\alpha_X}(T_X;X)$ can change by at most $(|T_X|\log N + \alpha_X \log|T_X|)/N$. Now we apply the McDiarmid inequality, Eqs. (19, 20) to finally obtain that, with probability of at least $1-\delta_1$:

$$|\hat{I}_{\alpha_X}(T_X;X) - E(\hat{I}_{\alpha_X}(T_X;X))| \leq (|T_X|\log N + \alpha_X \log|T_X|)\frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}}. \tag{23}$$

This generalizes the result of Shamir et al. (2010) to $\alpha_X \neq 1$. Similarly, we get that, with probability of at least $1-\delta_1$,

$$|\hat{I}_{\alpha_Y}(T_Y;Y) - E(\hat{I}_{\alpha_Y}(T_Y;Y))| \leq (|T_Y|\log N + \alpha_Y \log|T_Y|)\frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}}. \tag{24}$$

This leaves us with the final bound on the difference between the ML estimators of various informations and their expectations, namely for $\hat{I}(T_X;T_Y)$; this quantity is not analysed in Shamir et al. (2010), but we proceed very similarly. First, we calculate how much this term changes from a single datum being moved by using the identity $\hat{I}(T_X;T_X) = \hat{H}(T_X) + \hat{H}(T_Y) - \hat{H}(T_X,T_Y)$. Luckily we already calculated that $\hat{H}(T_X)$ changes by, at most, $|T_X|\log N/N$, and $\hat{H}(T_Y)$ changes by, at most, $|T_Y|\log N/N$. We are left to calculate how much $\hat{H}(T_X,T_Y)$ can change. We write $\hat{H}(T_X,T_Y) = -\sum_{t_x,t_y} \hat{p}(t_x,t_y)\log\hat{p}(t_x,t_y)$, where $\hat{p}(t_x,t_y) = \sum_{x,y} p(t_x|x)p(t_y|y)\hat{p}(x,y)$. Therefore, $\hat{p}(t_x,t_y)$ can change by, at most, $1/N$ for all $(t_x,t_y) \in (T_X,T_Y)$. Thus, $\hat{H}(T_X,T_Y)$ can change by at most $|T_X||T_Y|\log N/N$. We again use the McDiarmid's inequality and we determine that, with probability of at least $1-\delta_1$, the difference between the ML estimate $\hat{I}(T_X;T_Y)$ and its expected value is bounded by

$$|\hat{I}(T_X;T_Y) - E(\hat{I}(T_X;T_Y))| \leq ((|T_X| + |T_Y| + |T_X||T_Y|)\log N)\frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}}. \tag{25}$$

Now we need to calculate bounds on the bias (second) terms in Eq. (18) and similar expressions for the other information quantities. For this, we use results from Paninski

(2003), namely:

$$|H(T_X) - E(\hat{H}(T_X))| \leq \log\left(1 + \frac{|T_X| - 1}{N}\right) \leq \frac{|T_X| - 1}{N}, \tag{26}$$

$$|H(T_Y) - E(\hat{H}(T_Y))| \leq \log\left(1 + \frac{|T_Y| - 1}{N}\right) \leq \frac{|T_Y| - 1}{N}, \tag{27}$$

$$|H(T_X, T_X) - E(\hat{H}(T_X, T_X))| \leq \log\left(1 + \frac{|T_X||T_Y| - 1}{N}\right) \leq \frac{|T_X||T_Y| - 1}{N}. \tag{28}$$

Since we consider mapping $p(t_x|x)$ as fixed and known for this analysis, there is no bias $H(T_X|X) - E(\hat{H}(T_X|X))$. This means that the bias $|I_{\alpha_X}(T_X; X) - \hat{I}_{\alpha_X}(T_X; X)|$ only comes from the $|H(T_X) - \hat{H}(T_X)|$ term and does not have an $|X|$ or α_x dependence.

Putting the bounds on deviations of the estimates from their expectations and of expectations from the true values together, we get bounds on fluctuations of various information quantities that contribute to the GSIB loss function

$$|I_{\alpha_X}(T_X; X) - \hat{I}_{\alpha_X}(T_X; X)| \leq (|T_X| \log N + \alpha_X \log |T_X|) \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}} + \frac{|T_X| - 1}{N}, \tag{29}$$

$$|I_{\alpha_Y}(T_Y; Y) - \hat{I}_{\alpha_Y}(T_Y; Y)| \leq (|T_Y| \log N + \alpha_Y \log |T_Y|) \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}} + \frac{|T_Y| - 1}{N}, \tag{30}$$

$$\begin{aligned} |I(T_X; T_Y) - \hat{I}(T_X; T_Y)| \leq & (|T_X| + |T_Y| + |T_X||T_Y|) \log N \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}} \\ & + \frac{|T_X| - 1}{N} + \frac{|T_Y| - 1}{N} + \frac{|T_X||T_Y| - 1}{N} \\ = & ((|T_X| + 1)(|T_Y| + 1) - 1) \log N \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}} \\ & + \frac{(|T_X| + 1)(|T_Y| + 1) - 4}{N}. \end{aligned} \tag{31}$$

For comparison, the term $|I_{\alpha_x}(T_X; X) - \hat{I}_{\alpha_x}(T_X; X)|$ in the error of the GIB loss function has the same bounds as the corresponding term in GSIB, Eq. (29). Further the term $|I(T_X; Y) - \hat{I}(T_X; Y)|$ in the error of the GIB loss function is the same as for the traditional IB. Shamir et al. (2010) calculated it to be:

$$|I(T_X; Y) - \hat{I}(T_X; Y)| \leq (3|T_X| + 2) \log N \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}} + \frac{(|T_X| + 1)(|Y| + 1) - 4}{N}. \tag{32}$$

All of these bounds have a similar structure. The term proportional to $1/\sqrt{N}$ comes from the absolute deviation of the estimators. Its contribution is controlled by δ_1, so that if a high certainty is required ($\delta_1 \to 0$), then these terms are large. The terms proportional to $1/N$ are the bias terms.

The most crucial observation is that, even though the data comes from the joint probability distribution $p(x, y)$, which has the cardinality of $|X||Y|$, the terms proportional to this joint cardinality do not appear in the bounds, similar to Shamir et al. (2010). In other words, one does not need to have the joint distribution well-sampled to apply any of the IB variants.

The second observation from the bounds is that the deterministic versions, $\alpha = \alpha_X = \alpha_Y = 0$, of both the SIB and the IB have slightly tighter bounds than their generalized counterparts, including the original IB versions with $\alpha = \alpha_X = \alpha_Y = 1$. The tightening does not affect the bias component of the bounds, but provides a small correction to the absolute deviation, eliminating the terms similar to $\alpha \log |T_X| \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}}$, which are subdominant in the size of the reduced representations compared to the terms like $|T_X| \log N \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}}$.

We now compare the data efficiency of GSIB with that of two GIBs applied to reduce X and Y independently. We do so by bounding the error of the estimates of the loss for the GSIB vs. for two GIBs run in parallel.

The GSIB loss function error is:

$$\begin{aligned}|L_{\text{GSIB}} - \hat{L}_{\text{GSIB}}| \leq \left((|T_X| + |T_Y|) \log N + \alpha_X \log |T_X| + \alpha_Y \log |T_Y| \right) \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}} \\ + \beta \left((|T_X| + 1)(|T_Y| + 1) - 1 \right) \log N \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}} \\ + \frac{|T_X| - 1}{N} + \frac{|T_Y| - 1}{N} + \beta \frac{(|T_X| + 1)(|T_Y| + 1) - 4}{N}. \end{aligned} \quad (33)$$

The combined loss of two GIBs reducing X and Y independently is:

$$\begin{aligned}|L_{\text{GIB}} - \hat{L}_{\text{GIB}}| \leq \left((|T_X| + |T_Y|) \log N + \alpha_X \log |T_X| + \alpha_Y \log |T_Y| \right) \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}} \\ + \beta \left(3|T_X| + 3|T_Y| + 4 \right) \log N \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}} \\ + \frac{|T_X| - 1}{N} + \frac{|T_Y| - 1}{N} + \beta \frac{(|T_X| + 1)(|Y| + 1) + (|T_Y| + 1)(|X| + 1) - 8}{N}. \end{aligned} \quad (34)$$

We see that the dominant contribution to the absolute deviation part of L_{GSIB} bound is $\beta |T_X||T_Y| \log N \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}}$. For two GIBs run in parallel, Eq. (34) says that the dominant contributions to the absolute deviation would be $3\beta(|T_X| + |T_Y|) \log N \frac{\sqrt{\log(2/\delta_1)}}{\sqrt{2N}}$. That is, the two GIBs have smaller absolute deviations than GSIB for all but the smallest cardinalities of the compressed variables. However, notice that the cardinality of the compressed variables is usually not large, almost by definition, so that this loosening of the bound may be too small to notice for realistic $N \gg 1$. The behavior of the bias contributions to the bounds is different. The leading term for GSIB is $|T_X||T_Y|/N$, while for two GIBs it is $(|T_X||Y| + |X||T_Y|)/N$. Thus, when $|X|, |Y| \sim N$, the GSIB can be *significantly* more efficient that GIBs. When $|X|, |Y| \gg N$, the bias bounds for GIBs become meaningless, but GSIB bounds do not depend on the cardinality of the data variables. This is the reason for our assertion that GSIB has better data efficiency than two GIBs run in parallel for realistic cardinalities of variables and sample sizes.

3.2 Mean error and Mean squared error

The error bounds for the mutual information estimators must hold for worst case underlying distributions. Thus there are many cases when the error is significantly smaller than the calculated bounds. To explore if typical errors are different from the worst case bounds, here we calculate the mean squared error of $L_{\text{GSIB}} - \hat{L}_{\text{GSIB}}$, and similarly for the GIB. As always, the mean squared error is the sum of the squared bias and the variance of the estimator

$$E(L_{\text{GSIB}} - \hat{L}_{\text{GSIB}})^2 = (L_{\text{GSIB}} - E(\hat{L}_{\text{GSIB}}))^2 + E((\hat{L}_{\text{GSIB}} - E(\hat{L}_{\text{GSIB}}))^2), \tag{35}$$

and similarly for the GIB. This expression is the bias-variance decomposition and is similar to the bias absolute deviation decomposition for the bounds, Eq. (18). However, instead of bounding terms, we now calculate them. For this, we decompose every mutual information term in the loss functions into the corresponding entropy components.

We use the notation $\delta h \equiv \hat{h} - h$ for any variable that is being estimated via the ML estimator. For the ML estimator of the probability distribution $p(x, y)$, multinomial counting statistics textbook results give

$$E(\delta p(x,y)) = 0, \tag{36}$$

$$E(\delta p(x,y)\delta p(x',y')) = \frac{p(x,y)\delta_{x,x'}\delta_{y,y'}}{N} - \frac{p(x,y)p(x',y')}{N}. \tag{37}$$

Expectations for fluctuations of marginal distributions can be obtained by marginalizing Eqs. (36, 37).

In what follows, we will focus on $N \gg 1$, so that fluctuations $\delta p(x, y)$ have a small relative variance. Then, to obtain expressions for the variance of entropies, we follow Still and Bialek (2004) and expand $\hat{H}$ around the true value H for small δp. For $H(X)$, we get (expressions for other entropy terms are similar):

$$\begin{aligned}\hat{H}(X) &= -\sum_X (p(x) + \delta p(x)) \log(p(x) + \delta p(x)) \\ &= -\sum_X \left[p(x)\log p(x) + (\log p(x) + 1)\delta p(x) + \sum_{n=2}^{\infty} \frac{(-1)^n (\delta p(x))^n}{n(n-1)p(x)^{n-1}} \right] \\ &= H(X) - \sum_X \left[(\log p(x) + 1)\delta p(x) + \sum_{n=2}^{\infty} \frac{(-1)^n (\delta p(x))^n}{n(n-1)p(x)^{n-1}} \right]. \end{aligned} \tag{38}$$

From this, it follows that $\delta H(X) = -\sum_X \Big[(\log p(x) + 1)\delta p(x) + \frac{(\delta p(x))^2}{2p(x)}$ $+O((\delta p(x))^3))]$. Noticing that terms first order in δp vanish under averaging with respect to δp, cf. Eq. (36), we immediately calculate $|E(\delta H(X))| = \frac{|X|-1}{2N}$ and $|E(\delta H(Y))| = \frac{|Y|-1}{2N}$. Similarly, because $p(t_x|x)$ is fixed, we get $|E(\delta H(X, T_X))| = \frac{|X|-1}{2N}$, $|E(\delta H(Y, T_Y))| = \frac{|Y|-1}{2N}$. Further, $|E(\delta H(T_X))| = \frac{\sum_{T_X, X} p(t_x|x)p(x|t_x) - 1}{2N} \leq \frac{|T_X|-1}{2N}$ and $|E(\delta H(T_Y))| =\leq \frac{|T_Y|-1}{2N}$ where the inequalities comes from $p(t_x|x), p(t_y|y) \leq 1$. Combining these and similar results, we get biases of estimators of mutual information terms, which enter the GSIB loss functions:

$$|E(\delta I_{\alpha_X}(X,T_X))| \leq \frac{|T_X|-1}{2N}, \tag{39}$$

$$|E(\delta I_{\alpha_Y}(Y,T_Y))| \leq \frac{|T_Y|-1}{2N}, \tag{40}$$

$$|E(\delta I(T_X,T_Y))| \leq \frac{(|T_X|+1)(|T_Y|+1)-4}{2N}. \tag{41}$$

For the terms in the GIB loss function, we similarly get

$$|E(\delta I(Y,T_X))| \leq \frac{(|Y|+1)(|T_X|+1)-4}{2N}, \tag{42}$$

$$|E(\delta I(Y,T_Y))| \leq \frac{(|X|+1)(|T_Y|+1)-4}{2N}. \tag{43}$$

Note that these biases, to the two leading orders in δp, are half of the bound on the biases obtained in the previous Section, Eqs. (29-32). Thus the same scaling analyses apply. Crucially, we again observe that the bias of the symmetric variant of GIB only depends on the cardinalities of the compressed variables and not the uncompressed ones. Hence it is much smaller than for two GIBs applied in parallel, where the bias depends on $|X||T_Y|$ and $|Y||T_X|$.

Similarly we now calculate the mean squared error (see Appendix for details):

$$E(\delta I(X,T_X)^2) = \\ = \frac{1}{N}\left[\sum_{X,T_X,T'_X} p(t_x|x)p(t'_x|x)p(x)\log\frac{p(x,t_x)}{p(x)p(t_x)}\log\frac{p(x,t'_x)}{p(x)p(t'_x)} - I(X,T_X)^2\right]. \tag{44}$$

This expression can be simplified in two important limits. First, we consider the trivial minimum of the loss function, discussed earlier. There the mapping is uniform, $p(t_x|x)=1/|T_X|$, so that also $p(t_x)=1/|T_X|$. We get:

$$E(I(X,T_X)-\hat{I}(X,T_X))^2 = \\ = \sum_{X,T_X,T'_X} \frac{p(x)}{N|T_X|^2}\log\frac{p(x)/|T_X|}{p(x)/|T_X|}\log\frac{p(x)/|T_X|}{p(x)/|T_X|} - \frac{0^2}{N} = 0. \tag{45}$$

That is, fluctuations vanish in this case. This is expected since there is no information between T_X and X, and measuring more data points does not result in a more accurate estimate of the mutual information.

The second interesting case is a "winner-take-all" mapping, $p(t_x|x)=\delta(t_x,\tau(x))$, which would correspond to a deterministic clustering of multiple values of x into one t_x. This results in

$$E(I(X,T_X)-\hat{I}(X,T_X))^2 = \frac{1}{N}\left[\sum_X p(x)\log\frac{1}{p(\tau(x))}\log\frac{1}{p(\tau(x))} - I(X,T_X)^2\right] \\ \leq \frac{1}{N}\left[\log(\min(|T_X|,|X|))^2 - I(X,T_X)^2\right]. \tag{46}$$

Thus, here the average squared error is bound by $\frac{\log|T_X|^2 - I(X,T_X)^2}{N} \leq \frac{\log|T_X|^2}{N}$, which means that the RMS error for $I(T_X, X)$ is $\leq \frac{\log|T_X|}{\sqrt{N}}$. Similarly, the RMS errors for $I(T_Y, Y)$ and $I(T_X, T_Y)$ are $\leq \frac{\log|T_Y|}{\sqrt{N}}$ and $\leq \frac{\log\min(|T_X|,|T_Y|)}{\sqrt{N}}$, respectively. For the traditional IB, the RMS error for $I(T, X)$ is $\leq \frac{\log|T|}{\sqrt{N}}$, and the RMS error for $I(T, Y)$ is $\leq \frac{\log|T|}{\sqrt{N}}$. Thus, the average fluctuations are small and are of the same order of magnitude for both the symmetric bottleneck and the traditional bottleneck. This means that the dominant term is the average bias. As we saw earlier, the latter can be much worse for the traditional IB than for the symmetric IB.

4 Conclusion

Here we defined the generalized symmetric version of the information bottleneck (GSIB). We calculated the error bounds for each term within the loss function of GSIB and of the loss functions of the traditional generalized information bottleneck (GIB). We showed that the bias in estimating the loss function, and hence the error in finding the solution to the optimization problem from a finite dataset, is smaller for the GSIB compared to applying traditional GIB to each of the input variables, in parallel. We also calculated the average error and RMS error for each of these terms, resulting in essentially the same conclusions. All of these results suggest that when the cardinality of the measured variables X and Y are both large, and both variables require compression, then simultaneous compression is more data efficient than independently compressing each of the input variables.

While making extrapolations from a simple discrete variable case to more complex scenarios is difficult, we hope that these results are only the first of many to demonstrate a more general point that *simultaneous* dimensionality reduction is typically more data efficient than *independent* dimensionality reduction. In fact, using numerical simulations, we recently demonstrated a very similar result for a class of linear dimensionality reduction techniques for continuous variables Abdelaleem et al. (2023b), as well as for variational autoencoders (an IDR method) and a variational version of SIB (an SDR method) for large-dimensional continuous variables Abdelaleem et al. (2023a). Collectively, these findings suggest a general paradigm for efficient dimensionality reduction in complex multivariate datasets. For example, since physical theories are often formulated in terms of collective, coarse-grained representations (e.g., magnetization or temperature, which are expectation values of microscopic spins or energies of molecules), existence of data efficient algorithms for finding such reduced representations bodes well for using data-driven approaches for building physical theories of complex systems. Similarly, in biology, many central questions can be formulated as finding relations between large dimensional datasets. For example, in neuroscience, one aims to relate neural activity to behavior (Steinmetz et al., 2021; Urai et al., 2022; Krakauer et al., 2017; Sponberg et al., 2015b), and in systems biology, one looks to relate the gene expression state of a cell to its phenotypic profile (Clark et al., 2013; Zheng et al., 2017; Svensson et al., 2018; Huntley et al., 2015; Lorenzi et al., 2018). Our analysis suggests that methods based on the simultaneous dimensionality reduction can have a substantial impact on these fields as well.

Acknowledgments

The authors are grateful to Eslam Abdelaleem and Ahmed Roman for useful discussions and Sean Ridout for providing feedback on the manuscript. This work was supported, in part, by the Simons Investigator award and NSF Grant No. 2010524.

5 Appendix

5.1 Appendix: Derivation of the Generalized Symmetric Bottleneck

In what follows, we will derive the formal solution for the generalized symmetric bottleneck for $p(t_x|x)$. The formal solution is found by minimizing the cost function, Eq. (8) with respect to $p(t_x|x)$, subject to the normalization constraint. For this, we calculate the following useful derivatives:

$$\frac{\partial p(t_x)}{\partial p(t'_x|x')} = \frac{\partial}{\partial p(t'_x|x')} \sum_X p(t_x|x)p(x) = \delta(t_x, t'_x)p(x'), \tag{47}$$

$$\frac{\partial p(t_y)}{\partial p(t'_x|x')} = 0, \tag{48}$$

$$\frac{\partial p(t_x, t_y)}{\partial p(t'_x|x')} = \frac{\partial}{\partial p(t'_x|x')} \sum_X p(t_x|x)p(x, t_y) = \delta(t_x, t'_x)p(x', t_y). \tag{49}$$

To enforce the normalization of $p(t_x|x)$, we add a Lagrange multiplier λ times the normalization constraint to the cost function. With the helpful identities above, we now find the first derivative:

$$\begin{aligned}
&\frac{\partial(L_{\text{GSIB}} + \lambda(\sum_{X,T_X} p(t_x|x)p(x) - 1))}{\partial p(t'_x|x')} = \\
&= \frac{\partial}{\partial p(t'_x|x')}\left[-\sum_{T_X} p(t_x)\ln p(t_x) + \alpha_x \sum_{X,T_X} p(x)p(t_x|x)\ln p(t_x|x)\right. \\
&\quad - \sum_{T_Y} p(t_y)\ln p(t_y) + \alpha_y \sum_{Y,T_Y} p(y)p(t_y|y)\ln p(t_y|y) \\
&\quad \left. -\beta \sum_{T_X,T_Y} p(t_x,t_y)\ln\frac{p(t_x,t_y)}{p(t_x)p(t_y)} + \lambda\left(\sum_{X,T_X} p(t_x|x)p(x) - 1\right)\right] \\
&= -p(x')\ln p(t'_x) - p(x') + \alpha_x[p(x')\ln p(t'_x|x') + p(x')] \\
&\quad - \beta\sum_{T_Y} p(x',t_y)\ln\frac{p(t'_x,t_y)}{p(t'_x)p(t_y)} + \lambda p(x') \\
&= -p(x')\left[\ln p(t'_x) + 1 - \lambda - \alpha_x\left(\ln p(t'_x|x') + 1\right)\right. \\
&\quad \left. +\beta\sum_{T_Y} p(t_y|x')\ln\frac{p(t_y|t'_x)}{p(t_y)}\frac{p(t_y|x')}{p(t_y|x')}\right] \\
&= -p(x')\left[\ln p(t'_x)) + 1 - \lambda - \alpha_x\left(\ln p(t'_x|x') + 1\right)\right. \\
&\quad \left. +\beta\sum_{T_Y} p(t_y|x')\ln\frac{p(t_y|t'_x)p(t_y|x')}{p(t_y)p(t_y|x')}\right] \\
&= -p(x')\left[\ln p(t'_x) + 1 - \lambda - \alpha_x\left(\ln p(t'_x|x') + 1\right)\right. \\
&\quad \left. +\beta\sum_{T_Y} p(t_y|x')\left(\ln\frac{p(t_y|x')}{p(t_y)} - \ln\frac{p(t_y|x')}{p(t_y|t'_x)}\right)\right] \\
&= -p(x')\left[\ln p(t'_x) + 1 - \lambda - \alpha_x\left(\ln p(t'_x|x') + 1\right)\right. \\
&\quad \left. +\beta D_{\text{KL}}(p(t_y|x')||p(t_y)) - \beta D_{\text{KL}}(p(t_y|x')||p(t_y|t'_x))\right]. \qquad (50)
\end{aligned}$$

We now find the minimum of the cost function subject to the constraint that $p(t_x|x)$ is normalized by setting this derivative to zero and solving for $p(t'_x|x')$. Doing this, we find a formal solution:

$$p(t'_x|x') = \frac{\exp\left[\frac{1}{\alpha_x}\left(\ln p(t'_x) - \beta D_{\text{KL}}(p(t_y|x')||p(t_y|t'_x))\right]}{Z_x(x',\alpha_x,\beta)}, \qquad (51)$$

where $Z_x(x',\alpha_x,\beta) = \exp\left[-1 + \lambda + \alpha_x - \beta D_{\text{KL}}(p(t_y|x')||p(t_y))\right]$, and λ is chosen such that $p(t'_x|x')$ is normalized. Notice that the normalization constant Z_x is independent of t_y and t'_x. It only depends on x', α_x, and β. The same procedure can be followed to find the solution of the generalized symmetric information bottleneck for $p(t_y|y)$.

$$p(t'_y|y') = \frac{\exp\left[\frac{1}{\alpha_y}\left(\ln p(t'_y) - \beta D_{\text{KL}}(p(t_x|y')||p(t_x|t'_y)\right)\right]}{Z_y(y',\alpha_y,\beta)}, \qquad (52)$$

5.2 Appendix: Mean Error

Here we make explicit the calculations started in Section 3.2. Using Eq. (38) from the main text we, find the expected bias for X to depend on the cardinality $|X|$ and to be:

$$\begin{aligned}|E(\delta H(X))| &= \sum_X \frac{E(\delta p(x)^2)}{2p(x)} = \sum_X \frac{E(\sum_Y \delta p(x,y))^2}{2\sum_Y p(x,y)} \\ &= \sum_X \frac{\sum_{Y,Y'} E(\delta p(x,y)\delta p(x,y'))}{2\sum_Y p(x,y)} \\ &= \sum_X \frac{\sum_Y p(x,y) - \sum_{Y,Y'} p(x,y)p(x,y')}{2N\sum_Y p(x,y)} \\ &= \sum_X \frac{p(x)-p^2(x)}{2Np(x)} = \frac{|X|-1}{2N}. \end{aligned} \tag{53}$$

Similarly, $|E(\delta H(Y))| = \frac{|Y|-1}{2N}$, and $|E(\delta H(X,T_X))| = \frac{|X|-1}{2N}$.

Now we write:

$$\begin{aligned}|E(\delta H(T_X))| &= \sum_{T_X} \frac{E(\delta p(t_x)^2)}{2p(t_x)} = \sum_{T_X} \frac{E(\sum_{X,Y} \delta p(t_x|x)p(x,y))^2}{2\sum_{X,Y} p(x,y)} \\ &= \sum_{T_X} \frac{\sum_{X,X',Y,Y'} E(p(t_x|x)p(t_x|x')\delta p(x,y)\delta p(x',y'))}{2\sum_{X,Y} p(t_x|x)p(x,y)} \\ &= \sum_{T_X} \frac{\sum_{X,Y} p(t_x|x)^2 p(x,y) - \sum_{X,X',Y,Y'} p(t_x|x)p(t_x|x')p(x,y)p(x',y')}{2N\sum_{X,Y} p(t_x|x)p(x,y)} \\ &= \sum_{T_X} \frac{\sum_X p(t_x|x)^2 p(x) - p(t_x)^2}{2Np(t_x)} = \sum_{T_X} \frac{\sum_X p(t_x|x)p(x|t_x) - p(t_x)}{2N} \\ &= \frac{\sum_{T_X,X}[p(t_x|x)p(x|t_x)] - 1}{2N} \le \frac{|T_X|-1}{2N}, \end{aligned} \tag{54}$$

where the inequality comes from $p(t|x) \le 1$, so that $p(t|x)^2 p(x) \le p(t|x)p(x)$.

We can combine these results to find the overall bias for $\hat{I}(X,T_X)$:

$$\begin{aligned}|E(\delta I(X,T_X))| &= |E(\delta H(X)) + E(\delta H(T_X)) - E(\delta H(X,T_X))| \\ &= \frac{|X|-1}{2N} + \frac{\sum_{T_X,X} p(t_x|x)p(x|t_x) - 1}{2N} - \frac{|X|-1}{2N} \\ &= \frac{\sum_{T_X,X} p(t_x|x)p(x|t_x) - 1}{2N} \le \frac{|T_X|-1}{2N}. \end{aligned} \tag{55}$$

Similarly,

$$\begin{aligned}|E(\delta I(Y,T_Y))| &= |E(\delta H(Y)) + E(\delta H(T_Y)) - E(\delta H(Y,T_Y))| \\ &= \frac{|Y|-1}{2N} + \frac{\sum_{T_Y,Y} p(t_y|y)p(y|t_y) - 1}{2N} - \frac{|Y|-1}{2N} \\ &= \frac{\sum_{T_Y,Y} p(t_y|y)p(y|t_y) - 1}{2N} \le \frac{|T_Y|-1}{2N}. \end{aligned} \tag{56}$$

Finally, we calculate the bias for $\hat{I}(T_X, T_Y)$:

$$\begin{aligned}|E(\delta I(T_X,T_Y))| =&|E(\delta H(T_X)) + E(\delta H(T_Y)) - E(\delta H(T_X,T_Y))| \\ =&\frac{\sum_{T_X,X} p(t_x|x)p(x|t_x) - 1}{2N} + \frac{\sum_{T_Y,Y} p(t_y|y)p(y|t_y) - 1}{2N} \qquad (57)\\ &- \frac{\sum_{T_X,T_Y,X,Y} p(t_x,t_y|x,y)p(x,y|t_x,t_y) - 1}{2N}, \qquad (58)\end{aligned}$$

and

$$\begin{aligned}|E(\delta I(T_X,T_Y))| &\leq |E(\delta H(T_X))| + |E(\delta H(T_Y))| + |E(\delta H(T_X,T_Y))| \\ &\leq \frac{|T_X| - 1}{2N} + \frac{|T_Y| - 1}{2N} + \frac{|T_X||T_Y| - 1}{2N}. \qquad (59)\end{aligned}$$

We can perform similar calculations for the original bottleneck to obtain:

$$\begin{aligned}|E(\delta I(Y,T))| &\leq |E(\delta H(Y))| + |E(\delta H(T))| + |E(\delta H(Y,T))| \\ &= \frac{|Y| - 1}{2N} + \frac{\sum_{T,X} p(t|x)p(x|t) - 1}{2N} + \sum_{Y,T} \frac{\sum_X p(t|x)p(x|t,y)}{2N} - \frac{1}{2N} \\ &\leq \frac{|Y| - 1}{2N} + \frac{|T| - 1}{2N} + \frac{|Y||T| - 1}{2N}. \qquad (60)\end{aligned}$$

5.3 Appendix: Mean Squared Error

Using a method inspired by Still and Bialek (2004), we start by calculating the expected squared error for the mutual information between two arbitrary variables A and B, where the estimated probabilities are different from the true ones by a small error δ, $\hat{p}(a,b) = p(a,b) + \delta p(a,b)$, $\hat{p}(a) = p(a) + \delta p(a)$ and $\hat{p}(b) = p(b) + \delta p(b)$. First, let's calculate the mutual information to the first order in δp:

$$
\begin{aligned}
\hat{I}(A,B) &= \sum_{A,B}(p(a,b)+\delta p(a,b))\log\frac{p(a,b)+\delta p(a,b)}{(p(a)+\delta p(a))(p(b)+\delta p(b))} \\
&= \sum_{A,B}(p(a,b)+\delta p(a,b))\log\left[\frac{p(a,b)}{p(a)p(b)}\frac{1+\delta p(a,b)/p(a,b)}{(1+\delta p(a)/p(a))(1+\delta p(b)/p(b))}\right] \\
&= \sum_{A,B}(p(a,b)+\delta p(a,b))\left[\log\frac{p(a,b)}{p(a)p(b)}+\log\left(1+\frac{\delta p(a,b)}{p(a,b)}\right)\right. \\
&\qquad \left.-\log\left(1+\frac{\delta p(a)}{p(a)}\right)-\log\left(1+\frac{\delta p(b)}{p(b)}\right)\right] \\
&\approx \sum_{A,B}(p(a,b)+\delta p(a,b))\left[\log\frac{p(a,b)}{p(a)p(b)}+\frac{\delta p(a,b)}{p(a,b)}-\frac{\delta p(a)}{p(a)}-\frac{\delta p(b)}{p(b)}+\dots\right] \\
&\approx \sum_{A,B}\left[\delta p(a,b)\log\frac{p(a,b)}{p(a)p(b)}+p(a,b)\left(\frac{\delta p(a,b)}{p(a,b)}-\frac{\delta p(a)}{p(a)}-\frac{\delta p(b)}{p(b)}+\dots\right)\right] \\
&= \sum_{A,B}\left[\delta p(a,b)\log\frac{p(a,b)}{p(a)p(b)}+(\delta p(a,b)-p(b|a)\delta p(a)-p(a|b)\delta p(b))+\dots)\right] \\
&\qquad +I(A,B) \\
&= \sum_{A,B}\delta p(a,b)\log\frac{p(a,b)}{p(a)p(b)}+\sum_{A,B}\delta p(a,b)-\sum_{A}\delta p(a)-\sum_{B}\delta p(b)+\dots \\
&\qquad +I(A,B) \\
&= I(A,B)+\sum_{A,B}\delta p(a,b)\left(\log\frac{p(a,b)}{p(a)p(b)}-1\right)+\dots \qquad (61)
\end{aligned}
$$

Where in the last two lines, we used $\sum_B p(b|a)=1$, $\sum_A p(a|b)=1$, and $\delta p(a)=\sum_B \delta p(a,b)$, $\delta p(b)=\sum_A \delta p(a,b)$, respectively.

Thus, we see that $\delta I(A,B)=\sum_{A,B}\delta p(a,b)(\log\frac{p(a,b)}{p(a)p(b)}-1)$ to first order in $\delta p(a,b)$. We can now calculate the average squared error:

$$
\begin{aligned}
E[\delta I(A,B)^2] &= E\left[\sum_{A,B}\delta p(a,b)\left(\log\frac{p(a,b)}{p(a)p(b)}-1\right)\right. \\
&\qquad \left.\times\sum_{A',B'}\delta p(a',b')\left(\log\frac{p(a',b')}{p(a')p(b')}-1\right)\right] \\
&= \sum_{A,B,A',B'}E\left[\delta p(a,b)\delta p(a',b')\right] \\
&\qquad \times\left(\log\frac{p(a,b)}{p(a)p(b)}-1\right)\left(\log\frac{p(a',b')}{p(a')p(b')}-1\right). \qquad (62)
\end{aligned}
$$

We can use this generic expression to find the squared error for the estimator of information between the variables X and T_X, where $\delta p(x,t_x)=p(t_x|x)\delta p(x)$, and

$E(\delta p(x)\delta p(x')) = 1/N[\delta(x,x')p(x) - p(x)p(x')]$. We calculate $E[\delta I(X,T_X)^2]$ as follows:

$$
\begin{aligned}
&E[\delta I(X,T_X)^2] \\
&= \sum_{X,T,X',T'} E\left[\delta p(x,t_x)\delta p(x',t'_x)\right]\left(\log\frac{p(x,t_x)}{p(x)p(t_x)} - 1\right)\left(\log\frac{p(x',t'_x)}{p(x')p(t'_x)} - 1\right) \\
&= \sum_{X,T_X,X',T'_X} p(t_x|x)p(t'_x|x')E\left[\delta p(x)\delta p(x')\right]\left(\log\frac{p(x,t_x)}{p(x)p(t_x)} - 1\right) \\
&\qquad \times \left(\log\frac{p(x',t'_x)}{p(x')p(t'_x)} - 1\right) \\
&= \sum_{X,T_X,X',T'_X} p(t_x|x)p(t'_x|x')\frac{p(x)\delta(x,x') - p(x)p(x')}{N} \\
&\qquad \times \left(\log\frac{p(x,t_x)}{p(x)p(t_x)} - 1\right)\left(\log\frac{p(x',t'_x)}{p(x')p(t'_x)} - 1\right) \\
&= \frac{1}{N}\left[\sum_{X,T_X,T'_X} p(t_x|x)p(t'_x|x)p(x)\right. \\
&\qquad \left.\times\left(\log\frac{p(x,t_x)}{p(x)p(t_x)}\log\frac{p(x,t'_x)}{p(x)p(t'_x)} - \log\frac{p(x,t_x)}{p(x)p(t_x)} - \log\frac{p(x,t'_x)}{p(x)p(t'_x)} + 1\right)\right] \\
&- \frac{1}{N}\left[\sum_{X,T_X} p(t_x|x)p(x)\left(\log\frac{p(x,t_x)}{p(x)p(t_x)} - 1\right)\right. \\
&\qquad \left.\times\sum_{X',T'_X} p(t'_x|x')p(x')\left(\log\frac{p(x',t'_x)}{p(x')p(t_x)'} - 1\right)\right] \\
&= \frac{1}{N}\left[\sum_{X,T_X,T'_X} p(t_x|x)p(t'_x|x)p(x)\log\frac{p(x,t_x)}{p(x)p(t_x)}\log\frac{p(x,t'_x)}{p(x)p(t'_x)}\right. \\
&\qquad \left. -2I(X,T_X) + 1 - (I(X,T_X)-1)^2\right] \\
&= \frac{1}{N}\left[\sum_{X,T_X,T'_X} p(t_x|x)p(t'_x|x)p(x)\log\frac{p(x,t_x)}{p(x)p(t_x)}\log\frac{p(x,t'_x)}{p(x)p(t'_x)} - I(X,T_X)^2\right]. \qquad (63)
\end{aligned}
$$

Now let's look at two limits when we can simplify the above expression. In the first limit, we assume that the mapping is uniform, $p(t_x|x) = 1/|T_X|$, which means that $p(t_x) = 1/|T_X|$ as well. Then

$$
E[(I(X,T_X) - \hat{I}(X,T_X))^2] = \sum_{X,T_X,T'_X} \frac{p(x)}{|T_X|^2}\log\frac{p(x)/|T_X|}{p(x)/|T_X|}\log\frac{p(x)/|T_X|}{p(x)/|T_X|}\frac{1}{N} - \frac{0^2}{N} = 0. \qquad (64)
$$

In the other limit, we assume a "winner-take-all" mapping, where $p(t_x|x) = \delta(t_x, \tau(x))$.

We can reduce the expression to:

$$
\begin{aligned}
&E[\delta I(X,T_X)^2] = \\
&\quad = \frac{1}{N}\Bigg[\sum_{X,T_X,T'_X} \delta(t_x,\tau(x))\delta(t'_x,\tau(x))p(x)\log\frac{\delta(t_x,\tau(x))}{p(t_x)}\log\frac{\delta(t'_x,\tau(x))}{p(t'_x)} \\
&\qquad\quad -I(X,T_X)^2\Bigg] \\
&\quad = \frac{1}{N}\left[\sum_X p(x)\log\frac{1}{p(\tau(x))}\log\frac{1}{p(\tau(x))} - I(X,T_X)^2\right] \\
&\quad \leq \frac{1}{N}\left[\log(\min(|T_X|,|X|))^2 - I(X,T_X)^2\right] \leq \frac{1}{N}\left[\log(\min(|T_X|,|X|))^2\right]. \qquad (65)
\end{aligned}
$$

The result for $E[\delta I(Y,T_Y)^2]$ is similar to that for $E[\delta I(X,T_X)^2]$, Eq. (63:

$$
\begin{aligned}
&E[\delta I(Y,T_Y)^2] = \\
&= \frac{1}{N}\left[\sum_{Y,T_Y,T'_Y} p(t_y|y)p(t'_y|y)p(y)\log\frac{p(y,t_y)}{p(y)p(t_y)}\log\frac{p(y,t'_y)}{p(y)p(t'_y)} - I(Y,T_Y)^2\right]. \qquad (66)
\end{aligned}
$$

Finally we can calculate the covariance of fluctuations in the compressed variables, T_X and T_Y. Here $\delta p(t_x,t_y) = \sum_{X,Y} p(t_x|x)p(t_y|y)\delta p(x,y)$, and

$$
\begin{aligned}
E[\delta p(t_x,t_y)\delta p(t'_x,t'_y)] &= E\left[\sum_{X,Y} p(t_x|x)p(t_y|y)\delta p(x,y)\sum_{X',Y'} p(t'_x|x')p(t'_y|y')\delta p(x',y')\right] \\
&= \sum_{X,Y,X',Y'} p(t_x|x)p(t_y|y)p(t'_x|x')p(t'_y|y')E[\delta p(x,y)\delta p(x,y)] \\
&= \sum_{X,Y,X',Y'} p(t_x|x)p(t_y|y)p(t'_x|x')p(t'_y|y') \\
&\qquad \times \frac{p(x,y)\delta(x,x')\delta(y,y') - p(x,y)p(x',y')}{N} \\
&= \left[\frac{\sum_{X,Y} p(t_x|x)p(t_y|y)p(t'_x|x)p(t'_y|y)p(x,y)}{N}\right] \\
&\qquad - \left[\frac{p(t_x,t_y)p(t'_x,t'_y)}{N}\right]. \qquad (67)
\end{aligned}
$$

Using the previous result and Eq. (62), we find:

$$
\begin{aligned}
E[\delta I(T_X,T_Y)^2] = &\sum_{T_X,T_Y,T'_X,T'_Y} E[\delta p(t_x,t_y)\delta p(t'_x,t'_y)] \left(\log\frac{p(t_x,t_y)}{p(t_x)p(t_y)} - 1\right) \\
&\qquad \times \left(\log\frac{p(t'_x,t'_y)}{p(t'_x)p(t'_y)} - 1\right) \\
= &\sum_{T_X,T_Y,T'_X,T'_Y} \left[\frac{\sum_{X,Y} p(t_x|x)p(t_y|y)p(t'_x|x)p(t'_y|y)p(x,y)}{N}\right. \\
&\qquad \left. \times \log\frac{p(t_x,t_y)}{p(t_x)p(t_y)}\log\frac{p(t'_x,t'_y)}{p(t'_x)p(t'_y)}\right] - I(T_X,T_Y)^2/N. \qquad (68)
\end{aligned}
$$

In the "winner-take-all" limit, where $p(t_x|x) = \delta(t_x,\tau_x(x))$, and $p(t_y|y) = \delta(t_y,\tau_y(y))$, we find:

$$
\begin{aligned}
&E[\delta I(T_X,T_Y)^2] = \\
&= \sum_{T_X,T_Y,T'_X,T'_Y} \left[\frac{\sum_{X,Y} \delta(t_x,\tau_x(x))\delta(t_y,\tau_y(y))\delta(t'_x,\tau_x(x))\delta(t'_y,\tau_y(y))p(x,y)}{N}\right. \\
&\qquad \left. \times \log\frac{p(t_x,t_y)}{p(t_x)p(t_y)}\log\frac{p(t'_x,t'_y)}{p(t'_x)p(t'_y)}\right] - I(T_X,T_Y)^2/N \\
&= \sum_{X,Y}\left[\frac{p(x,y)}{N}\log\left(\frac{p(\tau_x(x),\tau_y(y))}{p(\tau_x(x))p(\tau_y(y))}\right)^2\right] - I(T_X,T_Y)^2/N \\
&\le \frac{1}{N}\log\left(\min(|T_X|,|T_Y|)\right)^2 - I(T_X,T_Y)^2/N. \qquad (69)
\end{aligned}
$$

Here we have calculate the average bias and variance for each term in the GSIB and the GIB. We found, in general, that the variance decays as $1/N$ and depends only on the cardinality of the compressed variables $|T_X|$ and $|T_Y|$. The expected bias for the GSIB depends on the cardinality of the compressed variables, while the bias for the GIB can depend on both the cardinality of the compressed variables and the cardinality of the uncompressed supervisor variables $|X|$ and $|Y|$.

References

Abdelaleem, E., Nemenman, I., and Martini, K. M. (2023a). Deep variational multivariate information bottleneck–a framework for variational losses. *arXiv preprint arXiv:2310.03311*.

Abdelaleem, E., Roman, A., Martini, K. M., and Nemenman, I. (2023b). Simultaneous dimensionality reduction: A data efficient approach for multimodal representations learning. *arXiv preprint arXiv:2310.04458*.

Andersen, C. M. and Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of chemometrics*, 24(11-12):728–737.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1247–1255, Atlanta, Georgia, USA. PMLR.

Antos, A. and Kontoyiannis, I. (2001). Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. (2018). Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR.

Carreira-Perpinán, M. A. (1997). A review of dimension reduction techniques. *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, 9:1–69.

Chapman, J. and Wang, H.-T. (2021). Cca-zoo: A collection of regularized, deep learning based, kernel, and probabilistic cca methods in a scikit-learn style framework. *Journal of Open Source Software*, 6(68):3823.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. (2013). The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

de Vries, S. E., Lecoq, J. A., Buice, M. A., Groblewski, P. A., Ocker, G. K., Oliver, M., Feng, D., Cain, N., Ledochowitsch, P., Millman, D., et al. (2020). A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature neuroscience*, 23(1):138–151.

Fisher, C. K. and Mehta, P. (2015). Bayesian feature selection for high-dimensional linear regression via the ising approximation with applications to genomics. *Bioinformatics*, 31(11):1754–1761.

Goebel, B., Dawy, Z., Hagenauer, J., and Mueller, J. (2005). An approximation to the distribution of finite sample size mutual information estimates. In *IEEE Int Conf Commun (ICC)*, volume 2, pages 1102–1106. IEEE.

Haghighi, M., Caicedo, J. C., Cimini, B. A., Carpenter, A. E., and Singh, S. (2022). High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nature methods*, 19(12):1550–1557.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Holbrook, A., Vandenberg-Rodes, A., Fortin, N., and Shahbaba, B. (2017). A bayesian supervised dual-dimensionality reduction model for simultaneous decoding of lfp and spike train signals. *Stat*, 6(1):53–67.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*.

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., and O'Donovan, C. (2015). The goa database: gene ontology annotation updates for 2015. *Nucleic acids research*, 43(D1):D1057–D1063.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3):480–490.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys Rev E*, 69(6):066138.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Lorenzi, M., Altmann, A., Gutman, B., Wray, S., Arber, C., Hibar, D. P., Jahanshad, N., Schott, J. M., Alexander, D. C., Thompson, P. M., et al. (2018). Susceptibility of brain atrophy to trib3 in alzheimer's disease, evidence from functional prioritization in imaging genetics. *Proceedings of the National Academy of Sciences*, 115(12):3162–3167.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Nanga, S., Bawah, A. T., Acquaye, B. A., Billa, M.-I., Baeta, F. D., Odai, N. A., Obeng, S. K., and Nsiah, A. D. (2021). Review of dimension reduction methods. *Journal of Data Analysis and Information Processing*, 9(3):189–231.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253.

Roulston, M. (1999). Estimating the errors on measured entropy and mutual information. *Physica D*, 125(3-4):285–294.

Scott, E. R. and Crone, E. E. (2021). Using the right tool for the job: the difference between unsupervised and supervised analyses of multivariate ecological data. *Oecologia*, 196(1):13–25.

Shamir, O., Sabato, S., and Tishby, N. (2010). Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T. K., Choi, H., Luviano, J. A., et al. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92.

Slonim, N., Friedman, N., and Tishby, N. (2006). Multivariate information bottleneck. *Neural computation*, 18(8):1739–1789.

Sponberg, S., Daniel, T. L., and Fairhall, A. L. (2015a). Dual dimensionality reduction reveals independent encoding of motor features in a muscle synergy for insect flight control. *PLoS Computational Biology*, 11(4):e1004168.

Sponberg, S., Daniel, T. L., and Fairhall, A. L. (2015b). Dual dimensionality reduction reveals independent encoding of motor features in a muscle synergy for insect flight control. *PLOS Computational Biology*, 11(4):1–23.

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588.

Still, S. and Bialek, W. (2004). How many clusters? an information-theoretic perspective. *Neural computation*, 16(12):2483–2506.

Strouse, D. and Schwab, D. (2017). The deterministic information bottleneck. *Neural computation*, 29(6):1611–1630.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Tishby, N., Pereira, F., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Tishby, N. and Slonim, N. (2000). Data clustering by markovian relaxation and the information bottleneck method. *Adv Neural Inf Proc Syst*, 13.

Urai, A. E., Doiron, B., Leifer, A. M., and Churchland, A. K. (2022). Large-scale neural recordings call for new insights to link brain and behavior. *Nature neuroscience*, 25(1):11–19.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Van Der Maaten, L., Postma, E. O., van den Herik, H. J., et al. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(66-71):13.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, pages 391–420.

Wold, S., Sjöström, M., and Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130. PLS Methods.

Yang, X., Liu, W., Liu, W., and Tao, D. (2021). A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2349–2368.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049.

Zhuang, X., Yang, Z., and Cordes, D. (2020). A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, 41(13):3807–3833.