Dimension Reduction for High-dimensional Small Counts with KL Divergence

Abstract

Dimension reduction for high-dimensional count data with a large proportion of zeros is an important task in various applications. As a large number of dimension reduction methods rely on the proximity measure, we develop a dissimilarity measure that is well-suited for small counts based on the Kullback-Leibler divergence. We compare the proposed measure with other widely used dissimilarity measures and show that the proposed one has superior discrimination ability when applied to high-dimensional count data having an excess of zeros. Extensive empirical results, on both simulated and publicly-available real-world datasets that contain many zeros, demonstrate that the proposed dissimilarity measure can improve a wide range of dimension reduction methods.

1 INTRODUCTION

High-dimensional count data, especially those with a large proportion of zeros, are omnipresent in various fields, such as ecology and genomics [Warton, 2018, Townes et al., 2019]. Dimension reduction (DR) techniques are used to extract useful information from high-dimensional count data, by eliminating noisy/uninformative dimensions of the data. Owing to the mean-variance dependency that is often observed in count data, it is inappropriate to apply standard DR methods that are optimal under the normality assumption, such as principal component analysis (PCA) [Pearson, 1901, Hotelling, 1933, Tipping and Bishop, 1999] and Gaussian process latent variable model (GPLVM) [Lawrence, 2005].

Hence, to perform DR on count data, a number of specific strategies/methods have been proposed. A common strategy is to first apply a variance-stabilizing transformation (VST) to the data [Rober, 1938, Cameron and Trivedi, 2013, Bartlett, 1947, Anscombe, 1948], aiming to make the data more Gaussian-like, and then feed the transformed data into standard DR approaches. The transformation function is specifically chosen to remove the mean-variance dependency. Popular transformation functions include the square root, logarithm, and inverse hyperbolic sine functions. Despite the widespread use of the VSTs, they can only be guaranteed to work well with large counts [Cameron and Trivedi, 2013, Bartlett, 1947, Anscombe, 1948] and cannot reasonably be expected to stabilise the variance of small counts containing a large faction of zeros [Yu, 2009, Warton, 2018]. Rather than focusing on making count data more normally distributed, several approaches have been developed to model the original data directly. With the assumption that count data follow the exponential family distributions, PCA variants maximise the likelihood of the observed data to get the low-dimensional representation [Collins et al., 2002, Mohamed et al., 2008, Li and Tao, 2013, Smallman et al., 2020]. By adopting the same distributional assumption, a robust estimator of the covariance matrix is derived and the data of reduced dimension are obtained by the eigendecomposition of this estimator [Liu et al., 2018]. Nonnegative matrix factorisation (NMF) acquires the low-dimensional representation by factorising count data matrix into two nonnegative matrices of low rank [Dhillon and Sra, 2005, Févotte and Idier, 2011, Cichocki et al., 2011]. Despite the popularity of NMF and PCA variants, it is still unclear whether they perform well on count data having an excess of zeros.

Unlike the aforementioned works, we focus on developing measures that can reliably quantify the pairwise dissimilarity for small count data, motivated by the importance of proximity matrix in common DR frameworks. Specifically, many DR approaches seek to preserve properties of a proximity matrix of high-dimensional data when reducing the dimension of the data. Examples of such approaches include PCA with the Euclidean distance matrix and the Gram matrix [Torgerson, 1952, Van Der Maaten et al., 2009, Lawrence, 2005], GPLVM with the Gram matrix [Lawrence, 2005], multidimensional scaling (MDS) with an input dissimilarity matrix [Torgerson, 1952], and

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

t-distributed stochastic neighbor embedding (tSNE) with the matrix of the Gaussian kernels [Maaten and Hinton, 2008]. Therefore, a proximity measure that properly quantifies the dissimilarity between small-count data points could benefit a wide range of DR methods.

The two core contributions of this paper can be summarised as follows. First, we develop two dissimilarity measures for small count data based on the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] and the assumption that the data follow either Poisson or negative binomial (NB) distributions. We take both Poisson and NB distributions into account, as it is common to model count data with these two types of distribution [Chan and Vasconcelos, 2009, Lindén and Mäntyniemi, 2011, Zeileis et al., 2008, Townes et al., 2019, Kim et al., 2020]. Furthermore, to reliably calculate the KL divergence, we propose to use empirical Bayes estimators to estimate the distributional parameters. Secondly, we propose an index to evaluate the discrimination abilities of different dissimilarity measures and show that the measure developed with the NB assumption has superior discrimination ability compared with other widely used dissimilarity measures for high-dimensional small counts, in terms of their statistical behaviours. Moreover, consistent with our statistical investigation, the experimental results, on both real and simulated count data, also demonstrate that the measure obtained with the NB assumption is superior to other measures when handling small counts. Additionally, representative DR methods with the proposed superior measure outperform the variants of PCA and NMF.

The rest of this paper can be summarised as follows. First, we present standard transformations for count data in Section 2.1. We then derive two new dissimilarity measures with the KL divergence and the empirical Bayes estimators in Section 2.2. Secondly, we propose an index to evaluate the discriminate ability of a dissimilarity measure (Section 3.1) and compare different dissimilarity measures according to the proposed index. It is shown that, when applied to small counts, the Euclidean distance of the transformed data exhibits better discrimination ability than the original Euclidean distance, although the corresponding VST is unable to stabilise the variances (Section 3.2). More importantly, the measure obtained with the NB assumption is expected to perform the best when used for separating different distributions of small count data including many zeros (Section 3.3). Lastly, we present the experimental results of representative DR methods with different measures, NMF, and variants of PCA, on both real and simulated datasets (Section 4).

2 DISSIMILARITY MEASURES FOR COUNT DATA

In this section, we first present widely used VSTs for count data and then derive two new dissimilarity measures with the KL divergence and the empirical Bayes estimators for small counts.

2.1 VARIANCE-STABILIZING TRANSFORMATIONS (VSTS)

A VST is a data transformation that applies to data such that the variance of the transformed data is independent of their mean. Most VSTs for count data are developed by assuming data follow an either Poisson or NB distribution [Anscombe, 1948, Cameron and Trivedi, 2013, Bartlett, 1947]. Let y be the raw counts. The square root transformation

$$g_r(y) = \sqrt{(y + \frac{3}{4})} \tag{1}$$

is a popular technique for stabilising the variance of a Poisson random variable. For an NB random variable with $\mathrm{Var}(y) = \mathbb{E}(y) + \frac{1}{r}\mathbb{E}(y)^2$, a prevalent transformation is

$$g_{\mathrm{asin}}(y) = \mathrm{arcsinh}\sqrt{\frac{y + \frac{3}{8}}{r - \frac{3}{4}}}, \tag{2}$$

where $\mathrm{arcsinh}$ is the inverse hyperbolic sine function. Since $g_{\mathrm{asin}}(y)$ requires an approximate knowledge of r and in some cases it cannot be estimated well enough, a simpler logarithm transformation with a pseudocount 1 is preferred in practice, which is given by

$$g_{\log}(y) = \log(y + 1). \tag{3}$$

As mentioned before, these transformations fail to stabilise the variance of small counts. Thus, there is no guarantee that the Euclidean distances of the data transformed from raw counts by these VSTs perform well on small counts.

2.2 TWO NEW DISSIMILARITY MEASURES DEVELOPED WITH KL DIVERGENCE

The KL divergence is a statistical measure of how one probability distribution is different from a second, reference probability distribution [Kullback and Leibler, 1951]. For discrete probability distributions P and Q defined on the same probability space $\mathbf{Z}$, the KL divergence is defined as $D_{\mathrm{KL}}(P \mid Q) = \sum_{z \in \mathbf{Z}} P(z) \log \frac{P(z)}{Q(z)}$. The KL divergence for continuous random variables can be defined similarly by replacing the sum with the integral. For a pair of univariate normal distributions, $P : \mathcal{N}(\mu_x, \sigma^2)$ and $Q : \mathcal{N}(\mu_y, \sigma^2)$,we have $D_{\mathrm{KL}}\left[\mathcal{N}(\mu_x, \sigma^2) \mid \mathcal{N}(\mu_y, \sigma^2)\right] = \frac{(\mu_x - \mu_y)^2}{2\sigma^2}$. The squared Euclidean distance D_E^2 between two vectors $\mathbf{x} = [x_1, \ldots, x_p]^T, \mathbf{y} = [y_1, \ldots, y_p]^T \in \mathbb{R}^p$ is equivalent to the sum of the KL divergence between two univariate normal distributions across dimensions up to a constant $\frac{1}{2\sigma^2}$ and the mean values of the distributions are estimated by MLEs. This equivalence is shown by the fol-

lowing equation:

$$\sum_{i=1}^{p} \hat{D}_{\mathrm{KL}}\left[\mathcal{N}(\mu_{ix}, \sigma^2) \mid \mathcal{N}(\mu_{iy}, \sigma^2)\right] = \sum_{i=1}^{p} \frac{(\hat{\mu}_{ix} - \hat{\mu}_{iy})^2}{2\sigma^2} = \sum_{i=1}^{p} \frac{(x_i - y_i)^2}{2\sigma^2},$$

where x_i and y_i are the MLEs of mean parameters of the normal distributions on the i-th dimension of $\mathbf{x}$ and $\mathbf{y}$, respectively, when there is only one realisation observed for each distribution.

Stimulated by the equivalence between D_E^2 and the KL divergence, we propose to quantify the pairwise dissimilarity for count data by the KL divergence. To calculate the KL divergence, the distribution type and the corresponding parameter values are required to be specified. For the distribution type, we assume the observed data follow either Poisson or NB distributions, which are commonly used for modelling count data. Regarding the parameter estimation, a straightforward estimator is the MLE. However, the MLE incurs a numerical problem in practice. To clarify this problem, we derive two dissimilarity measures with the MLEs for Poisson and NB distributions, respectively. Suppose x_i and y_i follow $\mathrm{Pois}(\lambda_{ix})$ and $\mathrm{Pois}(\lambda_{ix})$, respectively. The respective MLEs of λ_{ix} and λ_{iy} are x_i and y_i. The KL divergence between $\mathbf{x}$ and $\mathbf{y}$ with these MLEs is thus

$$\sum_{i=1}^{p} \hat{D}_{KL}\left[\mathrm{Pois}(\lambda_{ix}) \mid \mathrm{Pois}(\lambda_{iy})\right] = \sum_{i=1}^{p} \left[y_i - x_i + x_i \log \frac{x_i}{y_i}\right]. \tag{4}$$

Analogously, we suppose x_i and y_i follow $\mathrm{NB}(r, p_{ix})$ and $\mathrm{NB}(r, p_{iy})$, respectively, with known r. Note that there are multiple definitions of the NB distribution and we use the following ones: the PMF of $\mathrm{NB}(r, p_{ix})$ for x_i is $\binom{x_i+r-1}{x_i}(1-p_{ix})^r p_{ix}^{x_i}$ and similarly for $\mathrm{NB}(r, p_{iy})$, where p_{ix} and p_{iy} are the probabilities of success, r represents the number of failures, and x_i and y_i count the numbers of successes. The respective MLEs of p_{ix} and p_{iy} are $\frac{x_i}{x_i+r}$ and $\frac{y_i}{y_i+r}$. The KL divergence between $\mathbf{x}$ and $\mathbf{y}$ with the MLEs is given by

$$\sum_{i=1}^{p} \hat{D}_{KL}\left[\mathrm{NB}(r, p_{ix}) \mid \mathrm{NB}(r, p_{iy})\right] = \sum_{i=1}^{p} r\log\frac{y_i + r}{x_i + r} + x_i\log\frac{x_i(y_i + r)}{y_i(x_i + r)}. \tag{5}$$

The dissimilarity measures presented in Equation (4) and Equation (5) both involve the logarithm terms, and thus zeros in count data would result in the numerical problem. Further, since the MLEs are close to the true values of parameters only if the number of observations is sufficiently large, the MLEs calculated from one observation respectively are unreliable, so are $\hat{D}_{KL}\left[\mathrm{Pois}(\lambda_{ix}) \mid \mathrm{Pois}(\lambda_{iy})\right]$ and $\hat{D}_{KL}\left[\mathrm{NB}(r, p_{ix}) \mid \mathrm{NB}(r, p_{iy})\right]$.

To address these issues, we propose to use the empirical Bayes estimators rather than the MLEs. The conjugate priors of Poisson and NB distributions are employed for estimating the parameters ($\lambda_{ix},\ \lambda_{iy},\ p_{ix},\ p_{iy}$). In addition, the hyperparameters of these priors are learned from data themselves, sidestepping the difficulty of specifying proper priors to some degree. Concretely, we specify a Gamma prior distribution $G(m_i, 1)$, where the shape parameter m_i is the mean value of the i-th dimension across all data points and the other parameter is the scale parameter, for the Poisson means ($\lambda_{ix},\ \lambda_{iy}$). For the probability parameters of NB distributions ($p_{ix},\ p_{iy}$), we specify a Beta prior distribution $B(m_i, r)$ for them. Note that the mean value can be thought of an additional observation. With the priors, we obtain the posterior distribution of λ_{ix} is $G(m_i + x_i, \frac{1}{2})$ and that of λ_{iy} is $G(m_i + y_i, \frac{1}{2})$. The posterior means, which are $\frac{m_i+x_i}{2}$ and $\frac{m_i+y_i}{2}$, respectively, are used as the estimated distributional parameters. Analogously, we obtain the posterior mean $\frac{m_i+x_i}{m_i+x_i+2r}$ from the posterior distribution $B(m_i + x_i, 2r)$ and $\frac{m_i+y_i}{m_i+x_i+2r}$ from $B(m_i + y_i, 2r)$, as the estimated probability parameters. Now we obtain the KL divergence between $\mathbf{x}$ and $\mathbf{y}$ with the Bayes estimators (posterior means):

$$\begin{aligned}
&\hat{D}_{KL}^{Bayes}\left[\mathrm{Pois}(\lambda_{ix}) \mid \mathrm{Pois}(\lambda_{iy})\right] \\
&= \frac{1}{2}\sum_{i=1}^{p} y_i - x_i + (x_i + m_i)\log\frac{x_i + m_i}{y_i + m_i}, \\
&\hat{D}_{KL}^{Bayes}\left[\mathrm{NB}(r, p_{ix}) \mid \mathrm{NB}(r, p_{iy})\right] \\
&= \sum_{i=1}^{p}\Bigg[r\log\frac{y_i + m_i + 2r}{x_i + m_i + 2r} \\
&+ \frac{x_i + m_i}{2}\log\frac{(x_i + m_i)(y_i + m_i + 2r)}{(y_i + m_i)(x_i + m_i + 2r)}\Bigg].
\end{aligned} \tag{6}$$

The logarithm terms in Equation (6) are well defined for $m_i > 0$, which is easily satisfied in practice as only meaningless features in the form of all zeros have $m_i = 0$. Owing to the asymmetry of the KL divergence, we propose to use

$$\begin{aligned}
D_P^2 &= \hat{D}_{KL}^{Bayes}\left[\mathrm{Pois}(\lambda_{ix}) \mid \mathrm{Pois}(\lambda_{iy})\right] \\
&+ \hat{D}_{KL}^{Bayes}\left[\mathrm{Pois}(\lambda_{iy}) \mid \mathrm{Pois}(\lambda_{ix})\right] \\
D_{NB}^2 &= \hat{D}_{KL}^{Bayes}\left[\mathrm{NB}(r, p_{ix}) \mid \mathrm{NB}(r, p_{iy})\right] \\
&+ \hat{D}_{KL}^{Bayes}\left[\mathrm{NB}(r, p_{iy}) \mid \mathrm{NB}(r, p_{ix})\right]
\end{aligned} \tag{7}$$

to measure the pairwise dissimilarity for count data. Table 1 lists the dissimilarity measures that we take into account in this paper. Note that for D_P and D_{NB} we ignore their multiplicative constant $\frac{1}{2}$ for conciseness.

Table 1: Dissimilarity measures and their equations.

Measure	Equation
$D_E(\mathbf{x},\mathbf{y})$	$\left[\sum_{i=1}^{p}(x_i-y_i)^2\right]^{\frac{1}{2}}$
$D_r(\mathbf{x},\mathbf{y})$	$\left[\sum_{i=1}^{p}(g_r(x_i)-g_r(y_i))^2\right]^{\frac{1}{2}}$
$D_{\text{asin}}(\mathbf{x},\mathbf{y})$	$\left[\sum_{i=1}^{p}(g_{\text{asin}}(x_i)-g_{\text{asin}}(y_i))^2\right]^{\frac{1}{2}}$
$D_{\log}(\mathbf{x},\mathbf{y})$	$\left[\sum_{i=1}^{p}(g_{\log}(x_i)-g_{\log}(y_i))^2\right]^{\frac{1}{2}}$
$D_P(\mathbf{x},\mathbf{y})$	$\left[\sum_{i=1}^{p}(\log(x_i+m_i)-\log(y_i+m_i))(x_i-y_i)\right]^{\frac{1}{2}}$
$D_{NB}(\mathbf{x},\mathbf{y})$	$\left[\sum_{i=1}^{p}\left(\log\frac{x_i+m_i}{x_i+m_i+2r}-\log\frac{y_i+m_i}{y_i+m_i+2r}\right)(x_i-y_i)\right]^{\frac{1}{2}}$

3 COMPARISON OF DISSIMILARITY MEASURES FOR HIGH-DIMENSIONAL SMALL COUNTS

In this section, we compare different measures listed in Table 1 according to their abilities to distinguish distributions that tend to produce small counts. First, we propose an index to quantify the discrimination abilities of different dissimilarity measures. Then, we investigate and compare the statistical behaviours of different measures based on the proposed index when the dimension is high and the count data consist of many zeros.

3.1 EVALUATION INDEX

The main goal of DR is to eliminate noisy or uninformative dimensions of high-dimensional data and to assist classification/clustering algorithms to uncover meaningful classes in the data. Different classes of count data can be characterised by different distributions, and thus a dissimilarity measure that benefits DR techniques should distinguish these distributions well. In this subsection, we propose an index to evaluate how well a dissimilarity measure separates data points generated from different distributions and groups those from the same distribution. The definition of the proposed index is given in Definition 1.

Definition 1 *Suppose there are two count data distributions, denoted by $F_{\mathbf{x}}$ and $F_{\mathbf{y}}$, respectively. Let $S_X = \{\mathbf{x}_1,\ldots,\mathbf{x}_{n_x}\}$ be the set of samples generated from $F_{\mathbf{x}}$ and $S_Y = \{\mathbf{y}_1,\ldots,\mathbf{y}_{n_y}\}$ the set of samples from $F_{\mathbf{y}}$. For a given dissimilarity measure $D(\cdot,\cdot)$, the proposed index $R(F_{\mathbf{x}},F_{\mathbf{y}})$ is defined as follows:*

$$\frac{\sum_{\mathbf{x}\in S_X,\mathbf{y}\in S_Y} D^2(\mathbf{x},\mathbf{y})/(n_x n_y)}{\sum_{\mathbf{x}_i,\mathbf{x}_j\in S_x,\mathbf{x}_i\neq\mathbf{x}_j}\frac{2D^2(\mathbf{x}_j,\mathbf{x}_i)}{(n_x-1)n_x}+\sum_{\mathbf{y}_i,\mathbf{y}_j\in S_y,\mathbf{y}_i\neq\mathbf{y}_j}\frac{2D^2(\mathbf{y}_j,\mathbf{y}_i)}{(n_y-1)n_y}}. \tag{8}$$

The proposed index $R(F_{\mathbf{x}},F_{\mathbf{y}})$ assesses the variation between data points from the same distribution and that from different distributions simultaneously, which are measured by $D(\cdot,\cdot)$. In particular, $R(F_{\mathbf{x}},F_{\mathbf{y}})>1$ implies that using the corresponding dissimilarity measure makes the separation between $F_{\mathbf{x}}$ and $F_{\mathbf{x}}$ greater than the within-distribution variation. Note that here we consider only two distributions for simplicity, and to facilitate the following analysis we use the squared dissimilarity function. In the following, the subscript $*$ of $R_*(F_{\mathbf{x}},F_{\mathbf{y}})$ will be that of the corresponding dissimilarity measure. It is straightforward to see that the higher the value of $R(F_{\mathbf{x}},F_{\mathbf{y}})$, the more powerful the discrimination ability of the dissimilarity measure. Particularly, a higher value of $R(F_{\mathbf{x}},F_{\mathbf{y}})$ indicates that the corresponding measure is more robust in the presence of noisy dimensions, which possibly reduce the inter-distribution separation degree and increase the within-distribution variation.

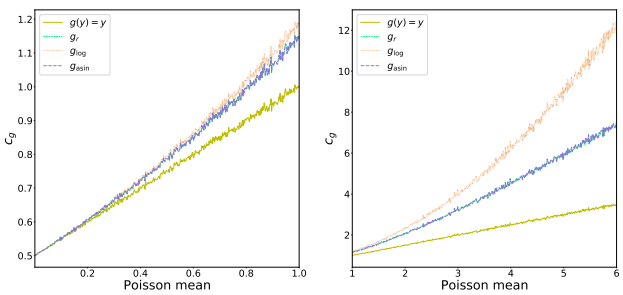
Before we dive into the comparison of measures using $R(F_{\mathbf{x}},F_{\mathbf{y}})$, the statistical behaviour of $R(F_{\mathbf{x}},F_{\mathbf{y}})$ in the high-dimensional space should be clarified. Proposition 1 presents the behaviour of $R(F_{\mathbf{x}},F_{\mathbf{y}})$ for the dissimilarity functions in a generic form: $D^2(x,y) = [f(x)-f(y)][g(x)-g(y)]$, which covers all the measures presented in Table 1. Proposition 1 shows that $R(F_{\mathbf{x}},F_{\mathbf{y}})$ moves towards a constant as dimension d grows, irrespective of the number of samples from distributions. The covariance of two increasing functions (f,g) of a random variable is positive [Schmidt, 2003], and thus the constant which $R(F_{\mathbf{x}},F_{\mathbf{y}})$ converges to would be greater than $\frac{1}{2}$ iff $[\mathrm{E}f(x)-\mathrm{E}f(y)][\mathrm{E}g(x)-\mathrm{E}g(y)] > 0$, which is readily satisfied in practice. The convergence will still hold under some mild conditions, such as with dependent dimensions and non-identical distributions.

Proposition 1 *Suppose points in $S_X\cup S_Y$ are independent, and each coordinate of points in S_X and S_Y are independently drawn from 1-dimensional non-degenerate data distributions F_x and F_y, respectively. For $D^2(x,y) = [f(x)-f(y)][g(x)-g(y)]$ with $x\sim F_x$, $y\sim F_y$, where $f(\cdot)$ and $g(\cdot)$ are predetermined functions, if $\mathbb{E}[D^2(x,y)]$, $\mathbb{E}[D^2(x_1,x_2)]$ and $\mathbb{E}[D^2(y_1,y_2)]$ exist for independent samples $x_1,x_2,x\sim F_x$, $y_1,y_2,y\sim F_y$, we have*

$$R(F_x,F_y)\xrightarrow{p}\frac{1}{2}+\frac{1}{2}\frac{[\mathrm{E}f(x)-\mathrm{E}f(y)][\mathrm{E}g(x)-\mathrm{E}g(y)]}{\mathrm{Cov}(f(x),g(x))+\mathrm{Cov}(f(y),g(y))}.$$

3.2 COMPARE THE EUCLIDEAN DISTANCES W/O VSTS

In the following, we compare D_E of original data with the Euclidean distances of the transformed data, according to their behaviours when dealing with small counts in the space of high dimension; that is, we compare them in terms of the respective constants that their $R(F_{\mathbf{x}},F_{\mathbf{y}})$'s converge to as dimension diverges to infinity. We first examine the discrimination ability of D_E when count data are small. Corollary 1 provides the sufficient and necessary condition for $R_E(F_x,F_y)\xrightarrow{p}c_E>1$. This condition suggests that for any pairs of Poisson distributions that generate small counts



(a) (b)

Figure 1: c_g for different Poisson distributions and different transforms.

with mean values less than 1, we obtain $c_E < 1$; that is, D_E cannot distinguish the two distributions well. Therefore, D_E is expected to perform poorly when handling small counts. An example showing D_E is unable to distinguish two Poisson distributions with different patterns of small counts is provided in Section B of Appendix.

Corollary 1 *With the same assumptions and notation as those in Proposition 1, for $D(\cdot,\cdot) = D_E(\cdot,\cdot)$, we have*

1. $R_E(F_x, F_y) \xrightarrow{p} c_E \geq \frac{1}{2}$ *for some constant* $c_E > 0$. *The equality holds iff* $\mathbb{E}(x) = \mathbb{E}(y)$ *for* $x \sim F_x$, $y \sim F_y$.
2. $c_E > 1$ *iff* $[\mathbb{E}(x) - \mathbb{E}(y)]^2 > \mathrm{Var}(x) + \mathrm{Var}(y)$.

We then investigate the behaviours of $R(F_{\mathbf{x}}, F_{\mathbf{y}})$'s of the Euclidean measures based on the VSTs when either F_x or F_y generates small counts. Without loss of generality, we assume F_x outputs small counts and F_y is an arbitrary distribution. Suppose there is a VST characterised by an increasing transformation function $g(\cdot)$, such that $g(y) \geq 0$ for $y \geq 0$. Note that $g(\cdot)$ covers $g_r(y)$, $g_{\text{asin}}(y)$, and $g_{\log}(y)$. Let D_E of the data transformed from raw counts by $g(\cdot)$ be D_g and the corresponding index $R_g(F_{\mathbf{x}}, F_{\mathbf{y}})$. Corollary 2 provides the difference between c_g and c_E when the proportion of zeros of each data point in S_X moves towards 1. It shows that D_g is better suited for distinguishing data points as μ_x approaches 0 in comparison with D_E iff $\frac{[g(0)-\mathrm{E}(y)]^2}{\mathrm{Var}[g(y)]} - \frac{\mathrm{E}^2(y)}{\mathrm{Var}(y)} > 0$.

To illustrate the advantages of applying VSTs to small counts, we obtain values of $c_g = \frac{[g(0)-\mathrm{E}(y)]^2}{\mathrm{Var}[g(y)]}$ for Poisson distributions with different mean values and transformation functions by numerical computation. Figure 1 supplies the numerical results and shows that g_r, $g_{\log}$, and g_{asin} always result in a c_g that is no less than c_E. In particular, it is observed from Figure 1(a) that c_g's exceed 1 when the Poisson mean is higher than 0.8, indicating that the corresponding measures distinguish better between data points with large proportions of zeros compared with D_E. Note that we assign a large value to r in $c_{g_{\text{asin}}}$ ($r = 1000$) since an arbitrary $\mathrm{NB}(r, p)$ approximates a Poisson distribution when r approaches infinity. The above analysis suggests that although the VSTs are unable to stabilise variances when count data are small, they improve the discrimination ability over D_E. Proofs of Proposition 1 and Corollary 2 are provided in Section A of Appendix.

Corollary 2 *Suppose that x and y are non-negative random variables. Let the expectation of F_x be μ_x. With the same assumptions and notation as those in Proposition 1, we have*

$$\lim_{\mu_x \to 0} (c_g - c_E) = \frac{1}{2}\left[\frac{[\mathrm{E}g(0) - \mathrm{E}g(y)]^2}{\mathrm{Var}[g(y)]} - \frac{\mathrm{E}^2(y)}{\mathrm{Var}(y)}\right],$$

where c_g is the constant that $R_g(F_{\mathbf{x}}, F_{\mathbf{y}})$ approaches.

3.3 COMPARE THE TWO PROPOSED MEASURES WITH OTHER DISSIMILARITY MEASURES

In the following, we will compare the proposed measures (D_P,D_{NB}) with the other dissimilarity measures in terms of the proposed index $R(F_{\mathbf{x}}, F_{\mathbf{y}})$ computed in the high-dimensional space. Note that the estimate $\hat{R}(F_x, F_y)$ would be close enough to the constant that $R(F_x, F_y)$ approaches as long as the dimension is high enough. Given a pair of distributions (F_x, F_y) and that of measures (D_{NB}, D_E), we believe D_{NB} is superior to D_E for distinguishing between F_x and F_y if $\hat{R}_{NB}(F_x, F_y) > \hat{R}_E(F_x, F_y)$. Further, to thoroughly evaluate their discrimination abilities for a specific distribution type, we compare their performances in terms of the fraction that $\hat{R}_{NB}(F_x, F_y) > \hat{R}_E(F_x, F_y)$ for different configurations of parameters. The fraction greater than 0.5 suggests D_{NB} is better suited for distinguishing this distribution type than D_E and vice versa. The distributions types (F_x, F_y) taken into account are the broadly used Poisson and NB distributions. F_x and F_y are of the same distribution type but with different parameter configurations. More details on the simulations and the calculation of different measures are presented in Section C of Appendix.

After simulations, we get $\hat{R}(F_x, F_y)$'s for a wide scope of parameter configurations and different distribution types. As mentioned before, the measures are compared in terms of $\hat{R}(F_x, F_y)$'s. Table 2 and Table 3 supply the comparison of the measures when data follow Poisson and NB distributions, respectively. For small count data following Poisson distributions, $D_{\log}$ achieves the best performance according to its highest average fraction value, followed by D_{NB}. Although $D_{\log}$ is superior to D_{NB} for Poisson-distributed count data, D_{NB} achieves a much higher value than $D_{\log}$ when count data are negative-binomially distributed. In addition, for small count data following NB distributions, D_P exhibits the second best discrimination ability. D_E, as anticipated, performs much worse than the other measures. The

simulation results show that D_{NB} distinguishes both the equidispersed and overdispersed distributions well, and thus we expect that D_{NB} outperforms the others when integrated into standard DR methods.

Table 2: Fraction that $\hat{R}(F_x, F_y)$ of a measure is greater than that of another measure for Poisson distributions. The value in entry (i, j) represents the fraction that $\hat{R}$ of the measure on the i-th row is greater than that of the measure on the j-th column. Top two measures are shown in bold.

Measures	D_E	D_r	D_{asin}	$D_{\log}$	D_P	D_{NB}	Ave
D_E	-	0.060	0.060	0.060	0.060	0.060	0.060
D_r	0.940	-	0.060	0.100	0.240	0.240	0.316
D_{asin}	0.940	0.940	-	0.100	0.240	0.240	0.492
$D_{\log}$	0.940	0.900	0.900	-	0.740	0.740	**0.844**
D_P	0.940	0.760	0.760	0.260	-	0.080	0.560
D_{NB}	0.940	0.760	0.760	0.260	0.920	-	**0.728**

Table 3: Fraction that $\hat{R}(F_x, F_y)$ of a measure is greater than that of another measure for NB distributions.

Measures	D_E	D_r	D_{asin}	$D_{\log}$	D_P	D_{NB}	Ave
D_E	-	0.252	0.300	0.270	0.146	0.148	0.223
D_r	0.748	-	0.542	0.500	0.092	0.092	0.396
D_{asin}	0.700	0.458	-	0.364	0.118	0.046	0.337
$D_{\log}$	0.730	0.500	0.636	-	0.150	0.078	0.419
D_P	0.854	0.908	0.882	0.850	-	0.178	**0.734**
D_{NB}	0.852	0.904	0.954	0.922	0.822	-	**0.891**

4 EXPERIMENTAL RESULTS

In this section, we present experimental results of representative DR methods with different dissimilarity measures, PCA variants, and NMF on both real and simulated high-dimensional count datasets with large fractions of zeros.

4.1 DATASETS

The high-dimensional count data considered in this paper is the single cell RNA sequencing (scRNA-seq) data with unique molecular identifiers (UMI). scRNA-seq data offer a unique opportunity to investigate the stochastic heterogeneity of complex issues at a near-genome-wide scale [Saliba et al., 2014, Shapiro et al., 2013, Kolodziejczyk et al., 2015]. scRNA-seq data with UMI are often modelled by NB or Poisson distributions [Townes et al., 2019, Kim et al., 2020, Svensson, 2020] and exhibit large proportions of zero counts. We run experiments on both real and simulated scRNA-seq datasets. These datasets contain large proportions of zeros, ranging from 0.6 to 0.97. The characteristics of the real scRNA-seq datasets used in this paper are summarised in Table 4. We simulate three additional scRNA-seq datasets by using the R's Splatter package [Zappia et al., 2017] with most of the parameters learned from real datasets except for differential expression factors, which determine the difference between groups of cells and the number of clusters. The information of the simulated datasets and the corresponding real datasets used for simulations is summarised in Table 5. Note that different clusters presented by the datasets represent different cell types.

Table 4: Real scRNA-seq datasets used in this paper.

Dataset	#clusters	#cells	#genes	prop of zeros
sc-CEL-seq2[Tian et al., 2019]	3	274	22060	0.678
sc-CEL-seq2-5cl-p1[Tian et al., 2019]	5	297	15564	0.608
sc-CEL-seq2-5cl-p2[Tian et al., 2019]	5	307	14078	0.598
sc-CEL-seq2-5cl-p3[Tian et al., 2019]	5	305	13426	0.643
Zheng8eq [Zheng et al., 2017]	8	3994	13301	0.957

Table 5: Simulated scRNA-seq datasets used in this paper.

Dataset	#clusters	#cells	#genes	prop of zeros	corresponding real dataset
sim-Zheng8eq	8	3994	13770	0.969	Zheng8eq[Zheng et al., 2017]
sim-manno-vm	5	1977	19416	0.899	manno-ESCs[La Manno et al., 2016]
sim-manno-ESCs	5	1715	19459	0.834	manno-ventral-midbrain[La Manno et al., 2016]

4.2 EVALUATION

We compare the proposed measures with two different types of methods that deal with count data: the VSTs and the DR methods designed specifically for original non-negative data. The proposed measures and D_E's of the data transformed by the VSTs are compared based on their performances when integrated into representative DR approaches, including PCA, tSNE, and GPLVM. In addition, the generalised PCAs (GPCAs) [Collins et al., 2002] and NMF [Dhillon and Sra, 2005], which handle original non-negative data, are included in comparison. GPCAs assume observed data follow either Poisson or NB distributions, denoted by GPCA-P and GPCA-NB, respectively. Note that r in D_{NB} and D_{asin} is set to the common NB dispersion parameter estimated by the R's edgeR package [Robinson et al., 2009].

We evaluate the performance of DR from two perspectives: visualisation and clustering. We project the high-dimensional dataset onto a two-dimensional (2D) space with a DR method and assess the 2D visualisation by visual inspection. A good visualisation should exhibit well-separated groups of data. The k-means algorithm [Lloyd, 1982, MacQueen, 1967] is used for inferring the cluster labels of data in the space of reduced dimension, and the number of clusters is set to be the ground truth. The clustering performance is assessed in terms of the adjusted rand index (ARI) between the cluster labels from the original publication/simulation and the inferred ones. The higher the ARI, the better the performance. Normally, tSNE and GPLVM map high-dimensional data to a 2D space, and thus we consider only 2D projections when evaluating their clustering performance. Since the results of the tSNE algorithm, GPCA-P, and GPCA-NB could be variable, we replicate the procedure, of first performing tSNE/GPCA-P/GPCA-NB and then applying k-means, for 10 times on the datasets for a more reliable comparison. We evaluate the clustering

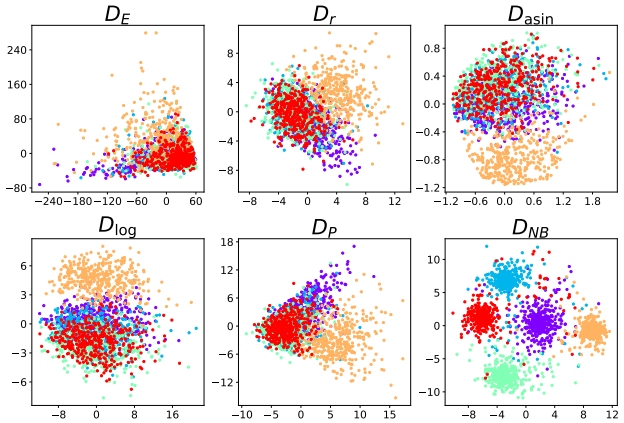


Figure 2: Visualisation of the sim-manno-vm dataset obtained by GPLVM with different measures. Different clusters are shown in different colours.

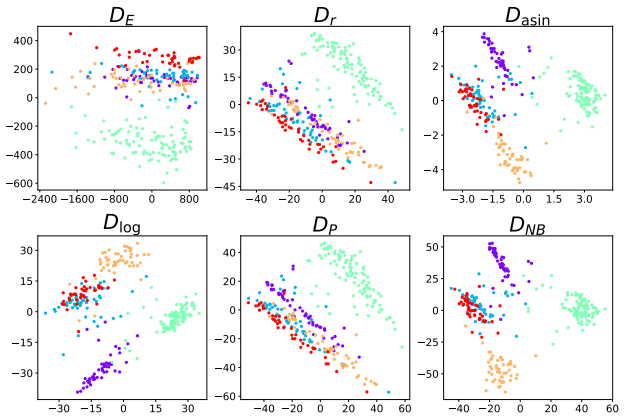


Figure 3: Visualisation of the sc-CEL-seq2-5cl-p1 dataset obtained by PCA with different measures.

performances of GPCAs and NMF in the spaces of different dimensions for a thorough comparison, ranging from 2 to 7. More details are provided in Section D of Appendix.

4.3 VISUALISATION

In this subsection, we examine whether the use of D_{NB} can produce better visualisation.

First, we visualise the dimension-reduced data obtained by different dissimilarity measures. Figure 2 and Figures 6-12 of Appendix provide the visualisation results obtained by GPLVM with different measures. It is observed that GPLVMs with D_{NB}, D_r, D_{asin}, and $D_{\log}$ perform equally well on most datasets except for the sim-manno-ESCs dataset and the sim-manno-vm dataset. For the sim-manno-ESCs dataset (Figure 12 of Appendix), GPLVMs with D_P and D_{NB} display well-grouped data in the 2D space while those with the other measures fail to do so. Furthermore, only the GPLVM with D_{NB} can distinguish different groups of data points on the sim-manno-vm dataset (Figure 2).

Figures 13-19 of Appendix supply the visualisation obtained by tSNE with different measures. tSNEs with D_{NB}, D_r, D_{asin}, and $D_{\log}$ produce well-distinguished groups of data on most datasets except the sim-manno-ESCs dataset (Figure 18 of Appendix) and the sim-manno-vm dataset (Figure 19 of Appendix), where all the measures fail to recognise the groups of data. By comparing PCA results with different measures in the 2D space (Figure 3 and Figures 20-26 of Appendix), we found that PCAs with D_{NB}, D_{asin}, $D_{\log}$ produce more distinguished groups of data on the sc-CEL-seq2-5cl-p1 dataset (Figure 3) and the sc-CEL-seq2-5cl-p2 dataset (Figure 20 of Appendix). For the sc-CEL-seq2-5cl-p3 dataset (Figure 21 of Appendix), the PCA with $D_{\log}$ presents 3 distinguished groups while those with the other measures display only 2 groups. Both PCAs with D_{NB} and D_{asin} separate the groups in the sc-CEL-seq2 dataset (Figure 22 of Appendix). PCAs with the proposed measures and VSTs perform equally well on the Zheng8eq dataset (Figure 23 of Appendix) and the sim-Zheng8eq dataset (Figure 24 of Appendix). For the sim-manno-ESCs dataset (Figure 25 of Appendix), the uses of D_r, D_P, and D_{NB} lead to more distinguished groups in the data. Furthermore, it is observed in Figure 26 of Appendix that only the PCA with D_{NB} can separate groups of data to some degree on the sim-manno-vm dataset while those with the other measures fail to do so.

Secondly, we assess the visualisation produced by GPCAs and NMF (Figure 27 and Figures 28-34 of Appendix) and compare them with those obtained by the PCA with D_{NB} (Figure 3). It is observed that the PCA with D_{NB} display more well-grouped data than GPCAs and NMF on most datasets except the Zheng8eq dataset (Figure 31 of Appendix) and sim-Zheng8eq dataset (Figure 32 of Appendix). For these two datasets, GPCAs distinguish the groups existing in the data more clearly compared with the PCA with D_{NB}.

To sum up, the representative DR approaches with D_{NB} outperform those with the other measures, and the use of D_{NB} often results in visualisation better than those produced by GPCAs and NMF.

4.4 CLUSTERING RESULTS

We first compare the proposed dissimilarity measures with those based on the VSTs in terms of the clustering performances in the reduced spaces. The k-means clustering results with GPLVM and different dissimilarity measures are shown in Figure 35 of Appendix. It is observed that the GPLVM with D_{NB} performs consistently well on most datasets except the sc-CEL-seq2-5cl-p2, where its ARI value is slightly lower than those obtained by the VSTs. Further, the GPLVM with D_{NB} obtains much higher values than

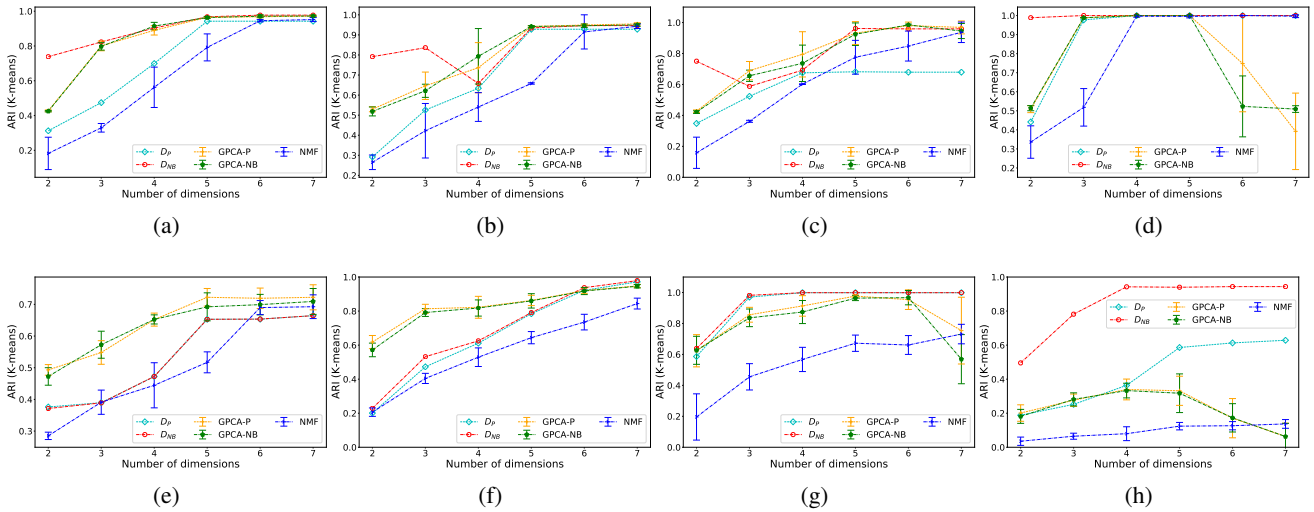


Figure 4: ARI of k-means with GPCAs in comparison with that with PCA and the proposed measures on the datasets: (a) sc-CEL-seq2-5cl-p1, (b) sc-CEL-seq2-5cl-p2, (c) sc-CEL-seq2-5cl-p3, (d) sc-CEL-seq2, (e) Zheng8eq, (f) sim-Zheng8eq, (g) sim-manno-ESCs, and (h) sim-manno-vm.

the other measures on the sim-manno-ESCs dataset and the sim-manno-vm dataset. To sum up, D_{NB} outperforms the other measures when integrated into GPLVM.

Figure 36 of Appendix presents the k-means clustering results with tSNE and different measures. The clustering performances obtained by tSNEs on the datasets differ greatly from those by GPLVMs. For the sc-CEL-seq2-5cl-p1 dataset, the sc-CEL-seq2-5cl-p2 dataset, and the sc-CEL-seq2 dataset, tSNEs with the VSTs and the proposed measures perform equally well according to their average ARI values. For the sc-CEL-seq2-5cl-p3 dataset, D_{NB} and $D_{\log}$ achieve more stable and higher ARI values than the others. Unlike GPLVM, the tSNE with D_E achieves the highest ARI value on both the Zheng8eq dataset and the sim-manno-ESCs dataset, as well as a value approximately the same as those with D_P and D_{NB} on the sim-Zheng8eq dataset. All measures perform poorly on the sim-manno-vm dataset. There is no measure that maintains the superiority over the other measures across the datasets when input into tSNE.

It is observed in Figure 37 of Appendix, D_{NB} is superior to the other measures when combined with PCA, irrespective of the number of dimensions, on most datasets except for the sc-CEL-seq2-5cl-p3 dataset and the Zheng8eq dataset. Furthermore, D_{NB} outperforms most measures when the dimension is greater than 5 on the the sc-CEL-seq2-5cl-p3 dataset and the Zheng8eq dataset.

Secondly, we compare the proposed measures, which are input into PCA, with GPCAs and NMF according to the clustering results provided in Figure 4. On the sc-CEL-seq2 dataset and the sim-manno-vm dataset, the PCA with D_{NB} not only outperforms GPCAs and NMF by a large margin when dimension is 2 but also achieves the highest ARI value with the lowest dimension. The value of ARI obtained by D_{NB} is much higher than those by GPCAs and NMF when dimension equals 2 on the sc-CEL-seq2-5cl-p1 dataset. For the sim-manno-ESCs dataset, the ARI values of GPCAs dramatically decrease as the dimension increases from 6 to 7, while those obtained by PCAs with D_{NB} and D_P are higher and more stable. Although the ARI values of D_{NB} do not rise as the dimension grows from 2 to 4 on the sc-CEL-seq2-5cl-p1 dataset and the sc-CEL-seq2-5cl-p3 dataset, PCA with D_{NB} outperforms GPCAs and NMF by a wide margin in the 2D space. For the Zheng8eq dataset and the sim-Zheng8eq dataset, GPCA-P accomplishes the highest value with the lowest dimension. It is found that the results of GPCAs could be highly variable.

Overall, the clustering performances obtained by the representative DR methods with D_{NB} are superior to those with the other measures. Moreover, the PCA with D_{NB} outperforms GPCAs and NMF.

5 CONCLUSION

This paper investigates how to perform DR for high-dimensional small count data. We propose a dissimilarity measure D_{NB} that is well-suited for count data with many zeros. The statistical behaviours of different dissimilarity measures when the dimension is high enough are investigated in terms of a proposed index. It is found that the proposed measure D_{NB} is superior to the other measures in the sense that it distinguishes data points from different distributions better. Consistent with the statistical comparison, the experimental results demonstrate that D_{NB} enhances a variety of standard DR approaches.

References

F. J. Anscombe. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948. ISSN 00063444.

M. S. Bartlett. The use of transformations. *Biometrics*, 3(1): 39–52, 1947. ISSN 0006341X, 15410420.

A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge University Press, 2013.

A. B. Chan and N. Vasconcelos. Bayesian Poisson regression for crowd counting. In *2009 IEEE 12th International Conference on Computer Vision*, pages 545–551, 2009.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13 (1):134–170, 2011.

Michael Collins, S. Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

Inderjit S Dhillon and Suvrit Sra. Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in Neural Information Processing systems*, volume 18, 2005.

Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. Demystifying "drop-outs" in single-cell UMI data. *Genome Biology*, 21(1):196, Aug 2020. ISSN 1474-760X.

Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. The technology and biology of single-cell RNA sequencing. *Molecular Cell*, 58(4):610 – 620, 2015. ISSN 1097-2765.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86, 1951.

Gioele La Manno, Daniel Gyllborg, Simone Codeluppi, Kaneyasu Nishimura, Carmen Salto, Amit Zeisel, Lars E Borm, Simon RW Stott, Enrique M Toledo, J Carlos Villaescusa, et al. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell*, 167(2): 566–580, 2016.

Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(Nov): 1783–1816, 2005.

Jun Li and Dacheng Tao. Simple exponential family pca. *IEEE Transactions on Neural Networks and Learning Systems*, 24(3):485–497, 2013.

Andreas Lindén and Samu Mäntyniemi. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92(7):1414–1421, 2011.

Lydia T Liu, Edgar Dobriban, and Amit Singer. ePCA: high dimensional exponential family PCA. *The Annals of Applied Statistics*, 12(4):2121–2150, 2018.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family pca. In *Advances in Neural Information Processing Systems*, volume 21, pages 1089–1096, 2008.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Dorfman Rober. A note on the δ-method for finding variance formulae. *Biometric Bulletin*, 1938.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 11 2009. ISSN 1367-4803.

Antoine-Emmanuel Saliba, Alexander J. Westermann, Stanislaw A. Gorski, and Jörg Vogel. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, 42(14):8845–8860, 2014.

Klaus D Schmidt. *On the covariance of monotone functions of a random variable*. 2003.

Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9): 618, 2013.

Luke Smallman, William Underwood, and Andreas Artemiou. Simple Poisson PCA: an algorithm for (sparse) feature extraction with simultaneous dimension determination. *Computational Statistics*, 35(2):559–577, 2020.

Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, Feb 2020. ISSN 1546-1696.

Luyi Tian, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S. Weber, Azadeh Seidi, Jafar S. Jabbari, Shalin H. Naik, and Matthew E. Ritchie. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods*, 16 (6):479–487, Jun 2019. ISSN 1548-7105.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61 (3):611–622, 1999. ISSN 13697412, 14679868.

Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):295, 2019. ISSN 1474-760X.

L. Van Der Maaten et al. Dimensionality reduction: a comparative review. *Technical report, Tilburg University, TiCC-TR 2009-005*, 2009.

David I. Warton. Why you cannot transform your way out of trouble for small counts. *Biometrics*, 74(1):362–368, 2018.

Guan Yu. Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. *Statistics & Probability Letters*, 79(14):1621–1629, 2009. ISSN 0167-7152.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, 2017. ISSN 1474-760X.

Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in R. *Journal of Statistical Software*, 27(8):1–25, 2008.

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1): 14049, Jan 2017. ISSN 2041-1723.

A PROOFS

In this section, we provide the proofs of Proposition 1 and Corollary 2 in the main paper.

A.1 PROOF OF PROPOSITION 1

Proof 1 *First note that* $R(F_x, F_y)$ *converges to* $\frac{\mathbb{E}[D^2(x,y)]}{\mathbb{E}[D^2(x_1,x_2)]+\mathbb{E}[D^2(y_1,y_2)]}$ *in probability, as* $d \to \infty$, *based on the weak law of large numbers and the Slutsky's theorem. Now we have*

$$\begin{aligned}\mathbb{E}[D^2(x,y)] &= \mathrm{Cov}\left(f(x)-f(y), g(x)-g(y)\right) \\ &+ \left[\mathrm{E}f(x)-\mathrm{E}f(y)\right]\left[\mathrm{E}g(x)-\mathrm{E}g(y)\right] \\ &= \mathrm{Cov}\left(f(x), g(x)\right) + \mathrm{Cov}\left(f(y), g(y)\right) \\ &+ \left[\mathrm{E}f(x)-\mathrm{E}f(y)\right]\left[\mathrm{E}g(x)-\mathrm{E}g(y)\right],\end{aligned} \tag{9}$$

and

$$\begin{aligned}&\mathbb{E}[D^2(x_1,x_2)] + \mathbb{E}[D^2(y_1,y_2)] = 2\mathrm{E}\left[f(x)g(x)\right] \\ &- 2\mathrm{E}f(x)\mathrm{E}g(x) + 2\mathrm{E}\left[f(y)g(y)\right] - 2\mathrm{E}f(y)\mathrm{E}g(y) \\ &= 2\mathrm{Cov}\left(f(x), g(x)\right) + 2\mathrm{Cov}\left(f(y), g(y)\right).\end{aligned} \tag{10}$$

Combining results from Equation (9) and Equation (10), we obtain

$$\begin{aligned}\frac{\mathbb{E}[D^2(x,y)]}{\mathbb{E}[D^2(x_1,x_2)]+\mathbb{E}[D^2(y_1,y_2)]} &= \frac{1}{2} \\ &+ \frac{1}{2}\frac{\left[\mathrm{E}f(x)-\mathrm{E}f(y)\right]\left[\mathrm{E}g(x)-\mathrm{E}g(y)\right]}{\mathrm{Cov}\left(f(x), g(x)\right)+\mathrm{Cov}\left(f(y), g(y)\right)},\end{aligned}$$

which completes the proof.

A.2 PROOF OF COROLLARY 2

Proof 2 *First note that* $\lim_{\mu_x \to 0} p(x = 0) = 1$ *for non-negative random variable* x *and* $\lim_{\mu_x \to 0} \mathrm{Var}(x) = 0$. *Further, we obtain* $\lim_{\mu_x \to 0} \mathrm{E}\left[g(x)\right] = g(0)$ *and* $\lim_{\mu_x \to 0} \mathrm{Var}\left[g(x)\right] = 0$. *Now consider the limiting difference between* c_g *and* c_E*:*

$$\begin{aligned}\lim_{\mu_x \to 0} c_g - c_E &= \lim_{\mu_x \to 0}\left[\frac{1}{2}\frac{\left[\mathrm{E}g(x)-\mathrm{E}g(y)\right]^2}{\mathrm{Var}\left[g(x)\right]+\mathrm{Var}\left[g(y)\right]}\right. \\ &\left.- \frac{1}{2}\frac{\left[\mathrm{E}(x)-\mathrm{E}(y)\right]^2}{\mathrm{Var}(x)+\mathrm{Var}(y)}\right] \\ &= \frac{1}{2}\left[\frac{\left[g(0)-\mathrm{E}g(y)\right]^2}{\mathrm{Var}\left[g(y)\right]} - \frac{\mathrm{E}^2(y)}{\mathrm{Var}(y)}\right].\end{aligned}$$

B AN EXAMPLE SHOWING THE EUCLIDEAN DISTANCE PERFORMS BADLY ON SMALL COUNTS

Suppose we have two Poisson distributions with mean values 0.01 and 0.96, respectively. Although both distributions outputs small counts, the PMFs of these two distributions, provided in Figure 5, display disparate patterns: the mass of one distribution concentrates in 0, while the frequent values of the other one lie in $\{0, 1, 2\}$. It is highly likely that $R_E(F_x, F_y) < 1$ in the space of high dimension according to Corollary 1 due to the mean-variance dependency, indicating that D_E cannot distinguish these two distributions well.

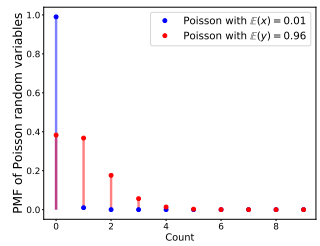


Figure 5: PMFs of two Poisson distributions.

C SIMULATION DETAILS

To obtain $\hat{R}(F_x, F_y)$, we first generate high-dimensional samples in S_X and S_Y with each coordinate i.i.d., and then calculate $\hat{R}(F_x, F_y)$ in Equation (8) with different measures. As we compared D_E with the Euclidean distances of the transformed data, we let F_x be a count distribution that frequently generates zeros. Specifically, F_x is either $\mathrm{Pois}(0.05)$ or $\mathrm{NB}(r, 0.05)$ that has the same r as that of F_y. Note that the probability of getting a zero from $\mathrm{Pois}(0.05)$ is around 0.951 and that from $\mathrm{NB}(r, 0.05)$ is 0.95^r. The parameter spaces for F_y are provided in Table 6. Note that the parameters for simulating NB and Poisson distributions are selected in an evenly spaced way from the intervals provided in Table 6.

With the mean value increasing, the proportion of zeros produced from a Poisson distribution decreases to 0. We hence bound the mean value from above by 1 to generate small counts and the expected fractions of zeros in the Poisson-distributed count data lie in the range $[0.368, 0.905]$. When r approaches infinity, the NB distribution approximates an equi-dispersed Poisson distribution. Thus, the upper bound of r for simulating NB distributions should be small in order to generate overdispersed distributions. When simulating data from NB distributions, the values of r for both F_x and F_y are the same, and we use the true value of r for calculating D_{NB} and D_{asin}. Since $r < \frac{3}{4}$ results in the numerical problem when calculating D_{asin}, we only take values that is larger than $\frac{3}{4}$ for r . Again, we restrict the range of p in NB distributions with the aim of simulating small counts. The proportion of zeros in the simulated NB data ranges from 0.363 to 0.963. For Poisson distributions, we set r to 1000 in both D_{NB} and D_{asin} due to the approximation of NB distributions to Poisson distributions when r is large.

Table 6: Parameter spaces for F_y.

Distributions	Parameter 1	Parameter 2
Poisson	means: $[0.05, 1]$ #samples: 50	variances: the same as the space of means
NB	r (# failures): $(\frac{3}{4}, 5]$; #samples: 50	p (probability of success): $[0.05, 0.2]$; #samples: 10

The dimension for each data point is set to be 5000 which is high enough to obtain a reliable estimation of the constant that $R(F_x, F_y)$ converges to. Although the convergence of $R(F_x, F_y)$ holds regardless of the number of samples in each distribution, small values of n_x and n_y possibly result in the meaningless coordinates, in the form of vectors of zeros. Therefore, we set n_x and n_y to 200 to avoid such cases.

D EVALUATION DETAILS

When applying tSNE with the proposed measures to the datasets, we replace the Euclidean distance with the proposed ones to characterise the dissimilarities between data points for the construction of conditional probabilities. When performing PCA with the proposed measures, we first get a pseudo Gram matrix by double-centring the corresponding pairwise dissimilarity matrix and then get the low-dimensional components by eigen-decomposing the pseudo Gram matrix. We empirically found that making the pseudo Gram matrix positive semi-definite would produce better results when combined with GPLVM. Thus, we first eigen-decompose the pseudo Gram matrix and keep all the eigenvectors with non-negative eigenvalues. We then feed the modified pseudo Gram matrix $Q\Lambda Q^T$ into the GPLVM, where Q is the matrix whose columns are the kept eigenvectors and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues.

E MORE VISUALISATION RESULTS

E.1 VISUALISATION OF GPLVM RESULTS

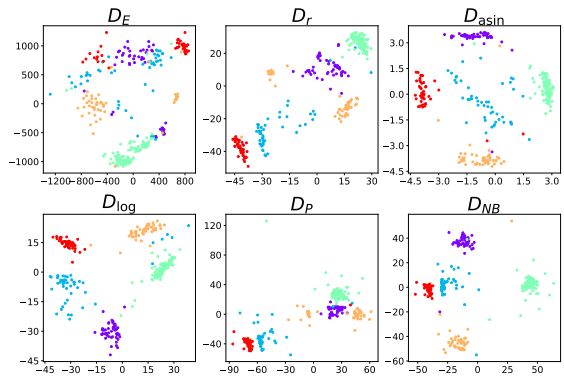


Figure 6: Visualisation of the sc-CEL-seq2-5cl-p1 dataset obtained by GPLVM with different measures.

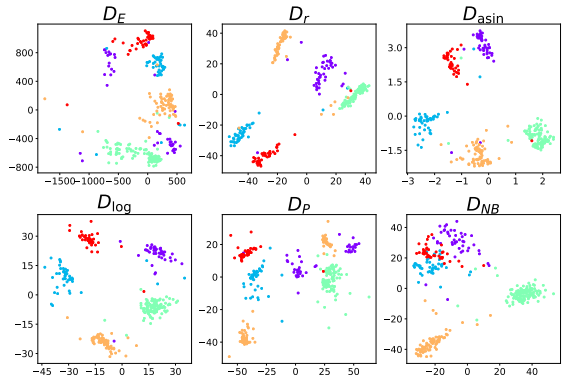


Figure 7: Visualisation of the sc-CEL-seq2-5cl-p2 dataset obtained by GPLVM with different measures.

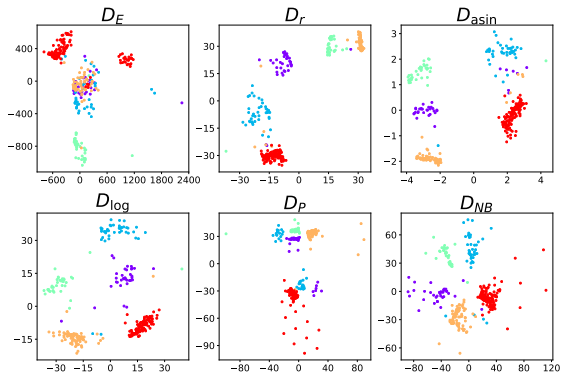


Figure 8: Visualisation of the sc-CEL-seq2-5cl-p3 dataset obtained by GPLVM with different measures.

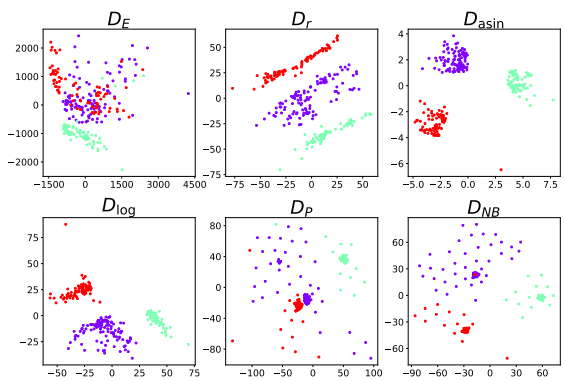


Figure 9: Visualisation of the sc-CEL-seq2 dataset obtained by GPLVM with different measures.

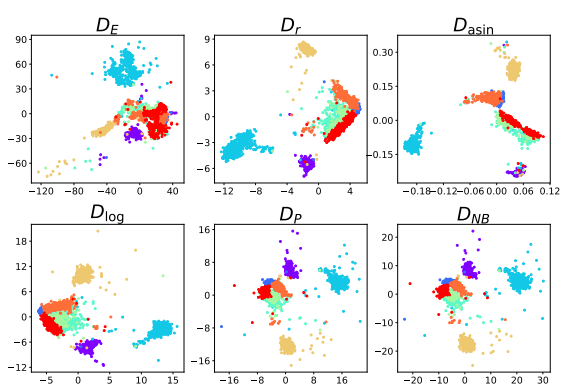


Figure 10: Visualisation of the Zheng8eq dataset obtained by GPLVM with different measures.

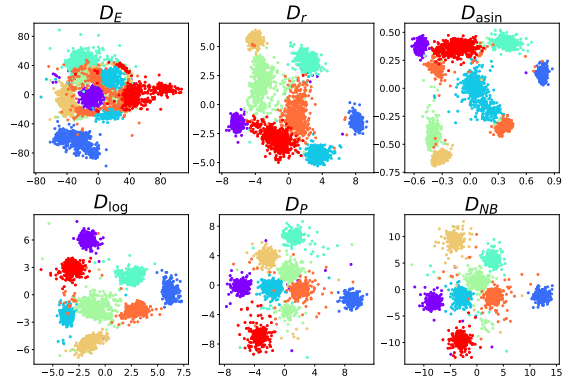


Figure 11: Visualisation of the sim-Zheng8eq dataset obtained by GPLVM with different measures.

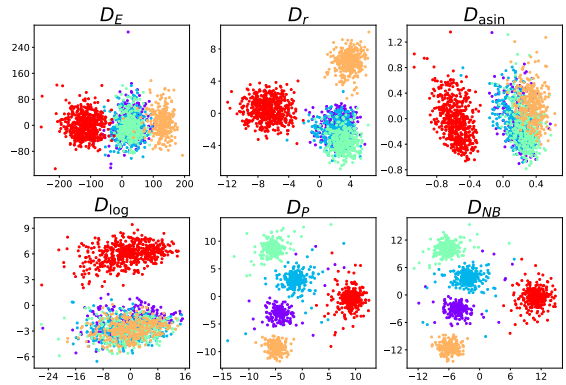


Figure 12: Visualisation of the sim-manno-ESCs dataset obtained by GPLVM with different measures.

E.2 VISUALISATION OF TSNE RESULTS

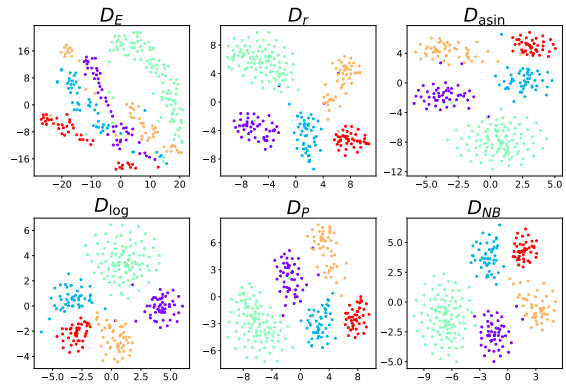


Figure 13: Visualisation of the sc-CEL-seq2-5cl-p1 dataset obtained by tSNE with different measures.

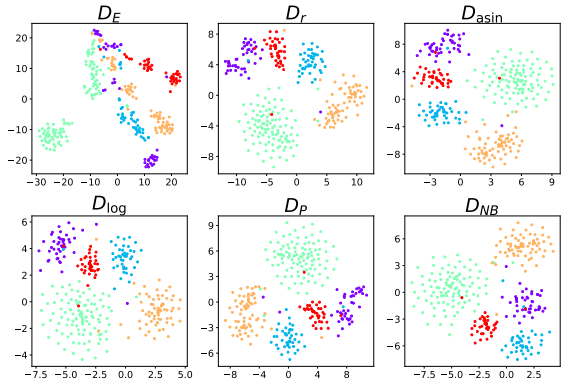


Figure 14: Visualisation of the sc-CEL-seq2-5cl-p2 dataset obtained by tSNE with different measures.

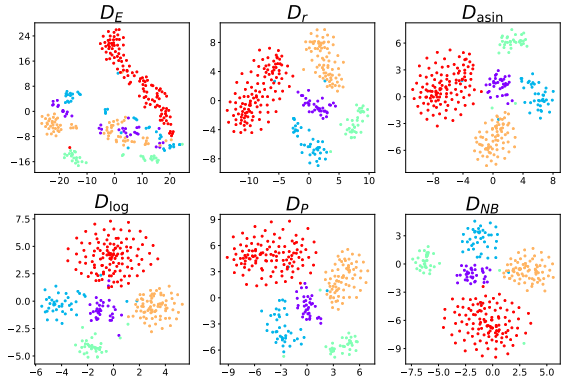


Figure 15: Visualisation of the sc-CEL-seq2-5cl-p3 dataset obtained by tSNE with different measures.

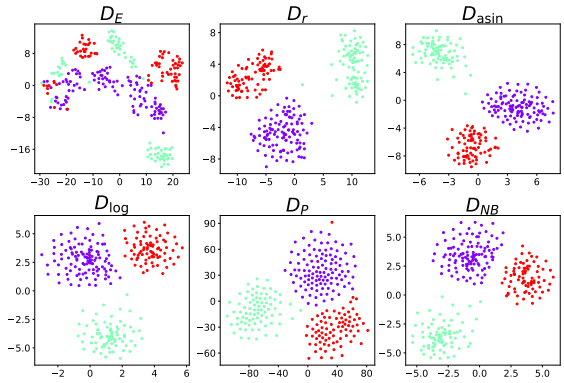


Figure 16: Visualisation of the sc-CEL-seq2 dataset obtained by tSNE with different measures.

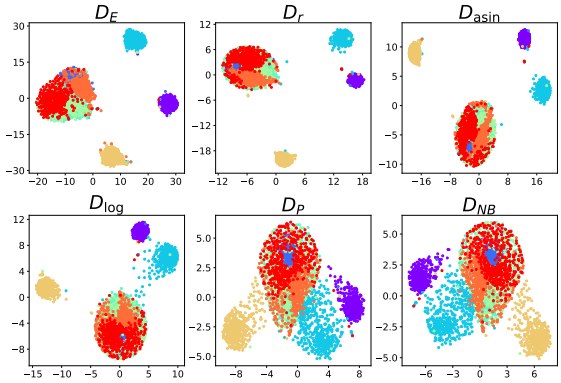


Figure 17: Visualisation of the Zheng8eq dataset obtained by tSNE with different measures.

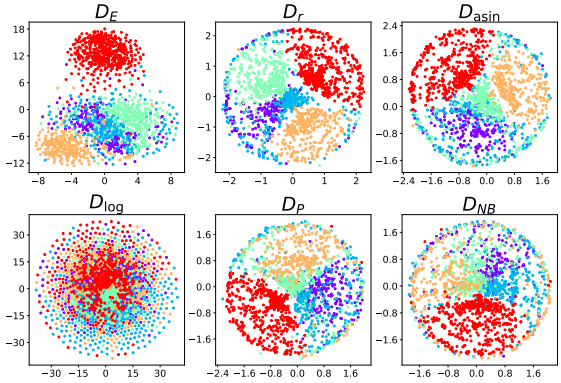


Figure 18: Visualisation of the sim-manno-ESCs dataset obtained by tSNE with different measures.

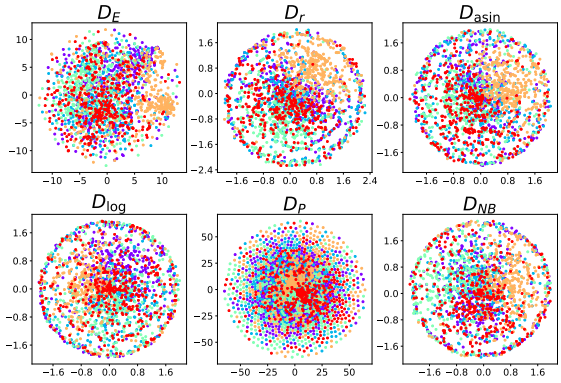


Figure 19: Visualisation of the sim-manno-vm dataset obtained by tSNE with different measures.

E.3 VISUALISATION OF PCA RESULTS

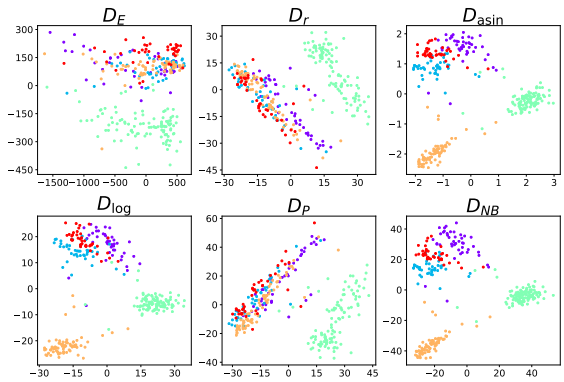


Figure 20: Visualisation of the sc-CEL-seq2-5cl-p2 dataset obtained by PCA with different measures.

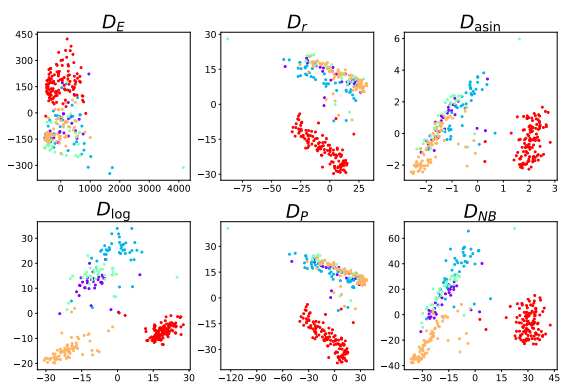


Figure 21: Visualisation of the sc-CEL-seq2-5cl-p3 dataset obtained by PCA with different measures.

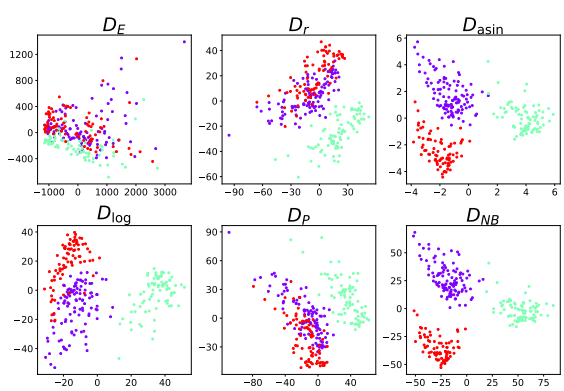


Figure 22: Visualisation of the sc-CEL-seq2 dataset obtained by PCA with different measures.

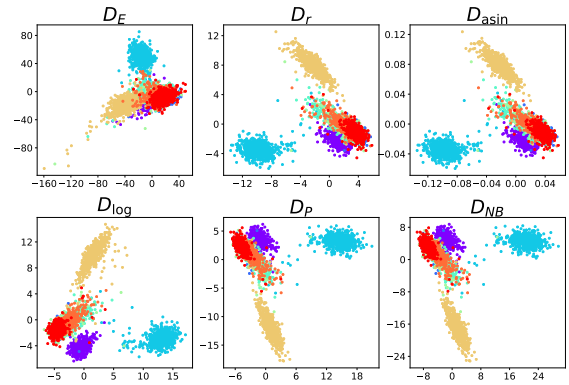


Figure 23: Visualisation of the Zheng8eq dataset obtained by PCA with different measures.

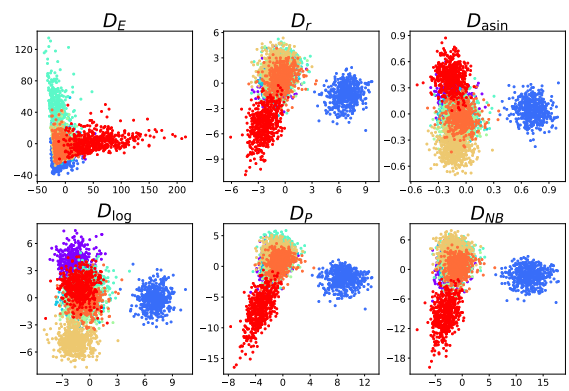


Figure 24: Visualisation of the sim-Zheng8eq dataset obtained by PCA with different measures.

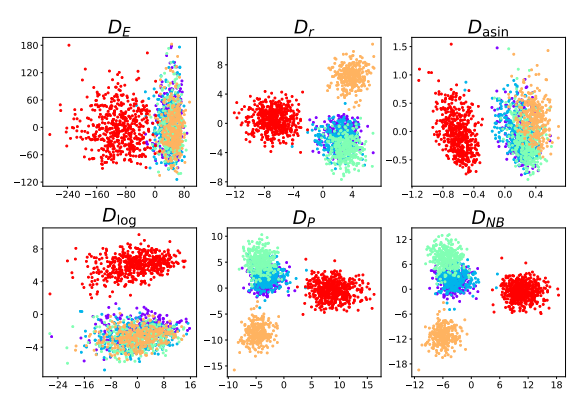


Figure 25: Visualisation of the sim-manno-ESCs dataset obtained by PCA with different measures.

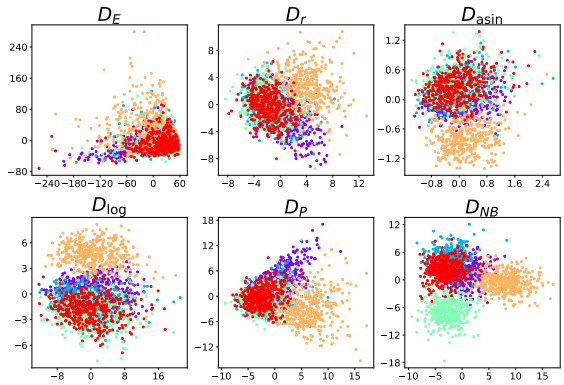


Figure 26: Visualisation of the sim-manno-vm dataset obtained by PCA with different measures.

E.4 VISUALISATION OF GPCAS AND NMF RESULTS

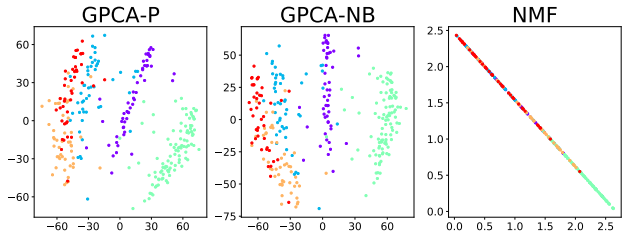


Figure 27: Visualisation of the sc-CEL-seq2-5cl-p1 dataset obtained by GPCAs and NMF.

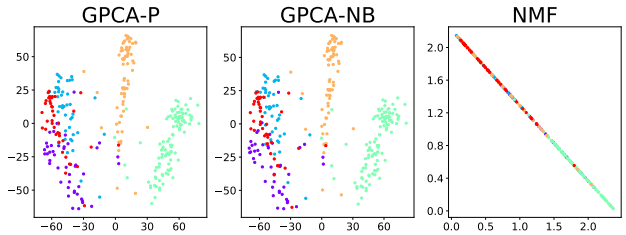


Figure 28: Visualisation of the sc-CEL-seq2-5cl-p2 dataset obtained by GPCAs and NMF.

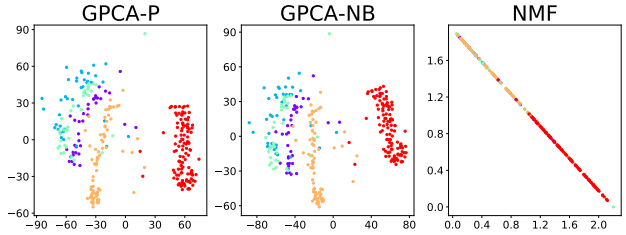


Figure 29: Visualisation of the sc-CEL-seq2-5cl-p3 dataset obtained by GPCAs and NMF.

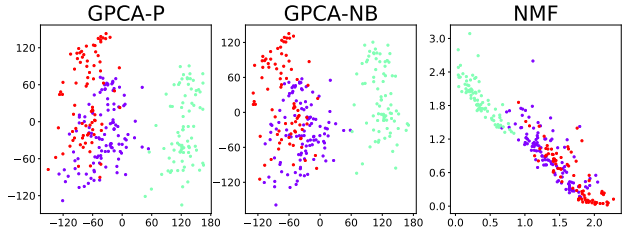


Figure 30: Visualisation of the sc-CEL-seq2 dataset obtained by GPCAs and NMF.

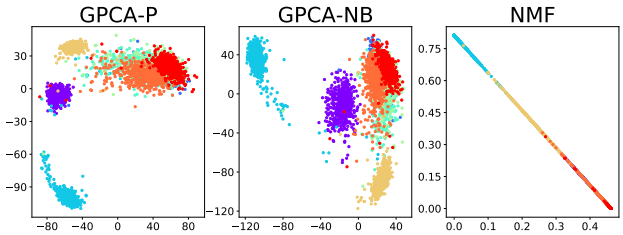


Figure 31: Visualisation of the Zheng8eq dataset obtained by GPCAs and NMF.

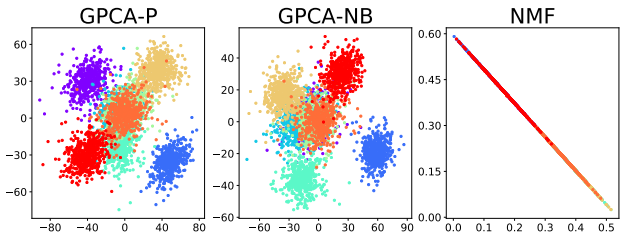


Figure 32: Visualisation of the sim-Zheng8eq dataset obtained by GPCAs and NMF.

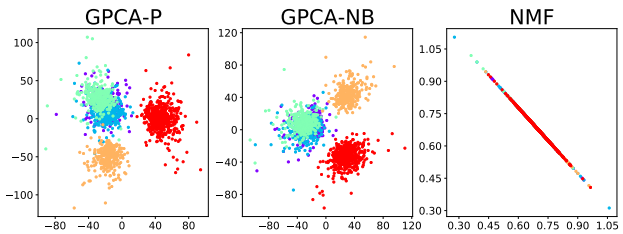
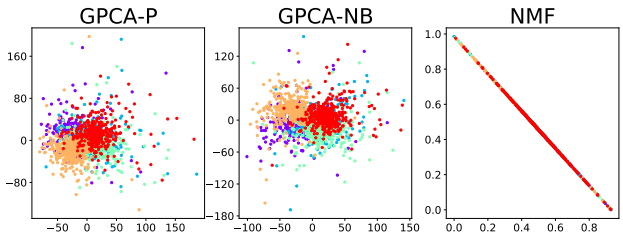


Figure 33: Visualisation of the sim-manno-ESCs dataset obtained by GPCAs and NMF.



(a)

Figure 34: Visualisation of the sim-manno-vm dataset obtained by GPCAs and NMF.

F MORE K-MEANS RESULTS

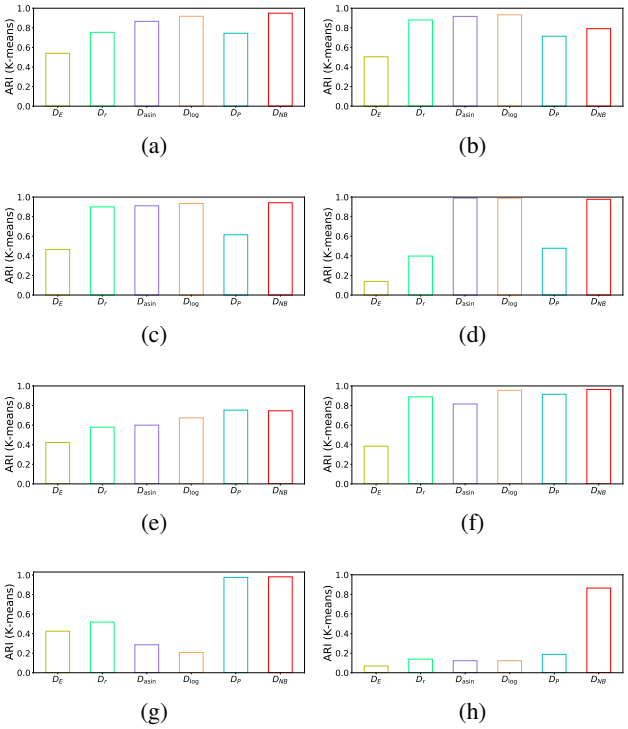


Figure 35: ARI of k-means with GPLVM and different dissimilarity measures on the datasets: (a) sc-CEL-seq2-5cl-p1, (b) sc-CEL-seq2-5cl-p2, (c) sc-CEL-seq2-5cl-p3, (d) sc-CEL-seq2, (e) Zheng8eq, (f) sim-Zheng8eq, (g) sim-manno-ESCs, and (h) sim-manno-vm.

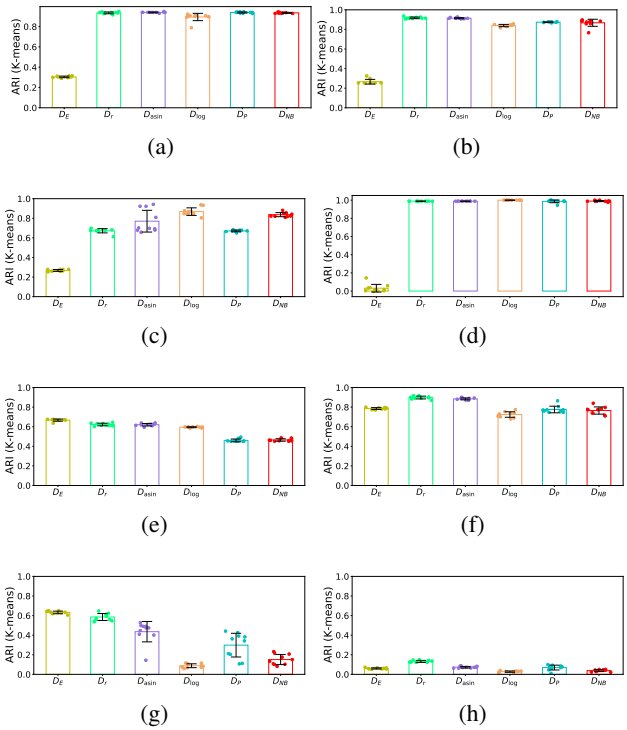


Figure 36: ARI of k-means with tSNE and different dissimilarity measures on the datasets: (a) sc-CEL-seq2-5cl-p1, (b) sc-CEL-seq2-5cl-p2, (c) sc-CEL-seq2-5cl-p3, (d) sc-CEL-seq2, (e) Zheng8eq, (f) sim-Zheng8eq, (g) sim-manno-ESCs, and (h) sim-manno-vm.

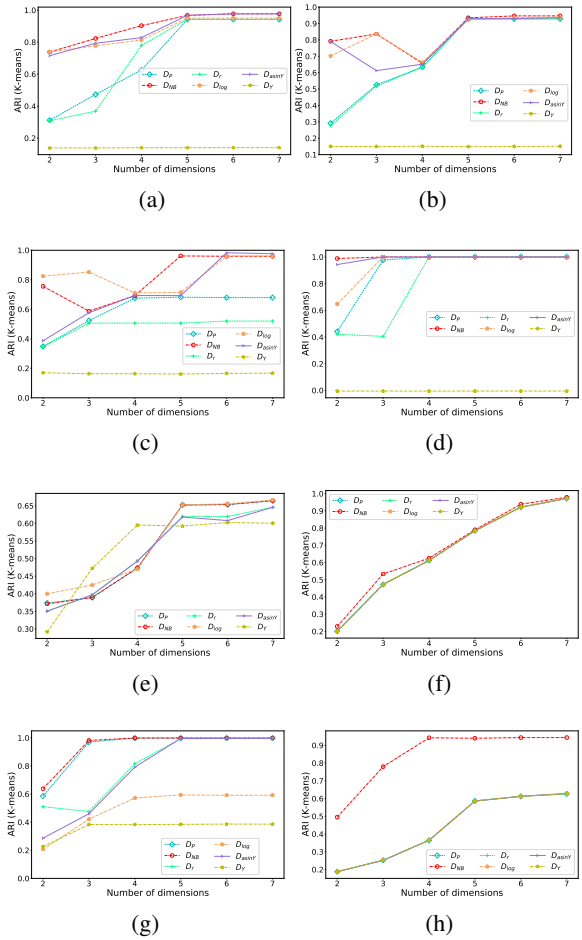


Figure 37: ARI of k-means with PCA and different dissimilarity measures on the datasets: (a) sc-CEL-seq2-5cl-p1, (b) sc-CEL-seq2-5cl-p2, (c) sc-CEL-seq2-5cl-p3, (d) sc-CEL-seq2, (e) Zheng8eq, (f) sim-Zheng8eq, (g) sim-manno-ESCs, and (h) sim-manno-vm.