

---

# PROJECTION PURSUIT WITH APPLICATIONS TO SCRNA SEQUENCING DATA

---

**Elvis Cui**

Department of Biostatistics  
University of California, Los Angeles  
[elviscuihan@g.ucla.edu](mailto:elviscuihan@g.ucla.edu)

**Heather Zhou**

Department of Statistics  
University of California, Los Angeles  
[heatherjzhou@ucla.edu](mailto:heatherjzhou@ucla.edu)

December 18, 2019

## ABSTRACT

In this paper, we explore the limitations of PCA as a dimension reduction technique and study its extension, projection pursuit (PP), which is a broad class of linear dimension reduction methods. We first discuss the relevant concepts and theorems and then apply PCA and PP (with negative standardized Shannon's entropy as the projection index) on single cell RNA sequencing data.

## 1 PCA and its issues

PCA is a popular dimension reduction technique commonly applied to scRNA sequencing data. There are several ways of deriving PCA, such as Karhunen-Loeve transform, Hotelling transform, and minimizing square error (chapter 7 and 16 of [2]). Despite of huge success in practice, we will illustrate three drawbacks of PCA. Due to these drawbacks, it is reasonable for us to seek alternatives of PCA.

### 1.1 Asymptotic distributions of eigenvalues

It is well known that the eigenvalues of sample covariance matrix is not consistent in high dimensional cases. Suppose we observe  $\mathbf{X} = (x_1, \dots, x_n)^T, x_i \in \mathbb{R}^d$ . If both  $n, d \rightarrow \infty$ , then we have the famous **quarter-circle law** (a.k.a. Marcenko-Pastur) in statistical physics:

**Theorem 1.1** (Marcenko-Pastur Law). Suppose  $\mathbf{X}^T \mathbf{X} \sim \mathcal{W}_d(n, \mathbf{I}_d)$ . Define the **empirical spectral distribution**

$$G_d(k) = \frac{1}{d} \# \{\hat{\lambda}_j \leq k\}$$

Where  $\hat{\lambda}_j$ 's are eigen-values of  $\frac{\mathbf{X}^T \mathbf{X}}{n}$ . If  $\frac{d}{n} \rightarrow \gamma \in (0, \infty)$ , then,  $G_d(k) \rightarrow G(k)$  almost surely, where the limiting distribution  $G(k)$  has density  $g(k) = G'(k)$ :

$$g(k) = \frac{\sqrt{(b_+ - k)(k - b_-)}}{2\pi\gamma k}, b_{\pm} = (1 \pm \sqrt{\gamma})^2$$

Therefore, in high dimensions, eigenvalues and eigen-vectors of sample covariance matrix are not consistent and PCA fails naturally.

### 1.2 Components other than uncorrelatedness

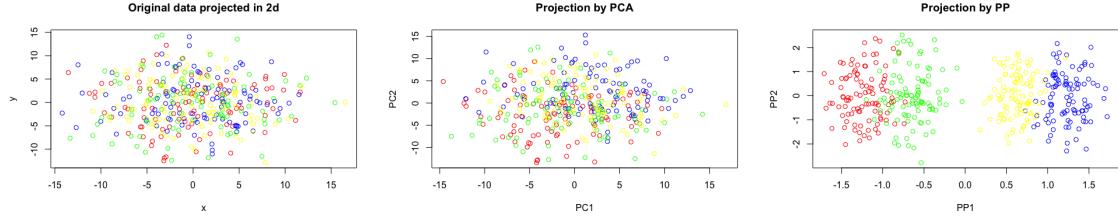
Every principal component is uncorrelated with each other but not independent. What if we want independence ? Thus, we should search for other criterions other than "maximizing variance". It turns out that "mutual information" in information theory will be a perfect alternative as we will illustrate in next section.

Besides, if we want sparsity on the support of eigen-vectors, then we may consider the so-called "sparse PCA" ([5]). If we want our estimation to be more robust, then we may consider robust PCA or other robust estimations ([11]).

Moreover, what if we are facing a supervised learning problem instead of just dimension reduction ? All in a word, there is a beautiful framework that unifies all these aspects known as **projection pursuit**. We will study it briefly in next section.

### 1.3 PCA is not suitable for clustering

It is worth noting that PCA is **NOT** designed for clustering, given its objective function. Therefore, it is not surprise that PCA behaves poorly in some situations for clustering. The following figure shows such situation (left: original data, middle: PCA, right: PP).



## 2 Projection pursuit: basic concepts and properties

There are many non-linear alternatives for PCA, such as SNE, t-SNE ([3]), kernel PCA, principal curves and surfaces ([2] and [6]), etc. However, linear dimension reduction is still the most important since **Fourier inversion theorem** ([1] and [4]) tells us that the distribution of any random vector is uniquely determined by its 1-d projection.

More generally speaking, suppose we are interested in a "lower dimensional projection" (not just 1-d) of our high dimensional data. Here "interesting" refers to different kinds of criterions. For instance, if we are interested in finding a direction such that the variance of projection is maximized, then we get ordinary PCA. Also, such lower dimensional projection should bypass "the curse of dimensionality" since there are few points in a high dimensional space. Besides, robustness and computational efficiency should be taken into account. All these ideas can be found in Huber ([11]) and lead us to the definition of **projection pursuit**.

### 2.1 Definition

First, we need a "loss function" or "index" that measures how well our projection is, this is called **projection index** in literature.

**Definition 2.1** (Projection index [10]). Let  $X$  be a random vector and  $A$  is a matrix. A **projection index** is a functional  $Q : F_A \rightarrow \mathbb{R}$  where  $F_A$  is the distribution of  $Z = AX$ . We will denote projection index as  $Q(Z)$  or  $Q(F_A)$ .

Second, given a specific projection index  $Q$  and data  $\mathbf{X}$ , we need to find a direction that maximizes such index. This is known as **projection pursuit** in literature.

**Definition 2.2** (Projection pursuit (PP) [11]). **Projection pursuit** (PP) searches for a projection  $A$  maximizing (or minimizing) a projection index  $Q(Z)$ . Usually, the projection will be less or equal to 3 dimensions for visualization.

### 2.2 Diaconis-Freedman theorem

**Theorem 2.1** (Diaconis and Freedman, 1984; Bickel et al, 2018). Suppose  $x_j$ 's are i.i.d. and the projection  $z$  was assigned the uniform distribution in  $\mathbb{S}^{d-1}$ , then as  $d, n \rightarrow \infty, \forall \epsilon > 0$

$$\mathbb{P}[\rho(\widehat{G}_z, \Phi) < \epsilon] \rightarrow 1$$

where  $\rho$  is the Levy-Prohorov metric:

$$\rho(\mu, \nu) := \inf \{\epsilon > 0 | \mu(A) \leq \nu(A^\epsilon) + \epsilon \text{ and } \nu(A) \leq \mu(A^\epsilon) + \epsilon \ \forall A \in \mathcal{B}(\Omega)\}$$

Thus, non-Gaussian projections would indeed be rare and interesting.

### 2.3 Bickel-Kur-Nadler theorem

A natural question on the properties of PP is: given an arbitrary cumulative distribution function  $G$  (i.e. mixtures of multivariate Gaussians), how well can our projected data approximates such distribution? Such question is answered elegantly by Bickel, Kur and Nadler.

**Theorem 2.2** (Bickel-Kur-Nadler [8]). Suppose  $\frac{d}{n} \rightarrow \infty$ . Let  $G(t)$  be an **arbitrary cumulative distribution function** with mean 0. There there  $\exists$  a sequence of projections  $z = z_n \in \mathbb{S}^{d-1}$  s.t. the following holds:

$$\lim_{n \rightarrow \infty} \|\widehat{G}_Z - G\|_\infty = 0$$

That is,  $\widehat{G}_Z$  converges uniformly (thus, weakly) to  $G$ .

Basically, this theorem says: if different types of cells have different gene expression, they there **EXISTs** a projection pursuit program s.t. we could visualize high dimensional clustered data in one or two or three dimensions.

## 3 Examples of projection indexes

In this section, we present some examples of various projection indexes, among which some are original. Those indexes can be classified as 2 types: non-entropy based and entropy based indexes. The former includes a blanket of classical statistical methods such as Fisher's LDA, PCA and canonical correlation analysis. The latter has strong connection with another technique called **independent component analysis** (ICA).

### 3.1 Non-entropy based indexes

**Example 3.1** (Sample mean [11]). Take  $Q(a^T X) = \mathbb{E}(a^T X) = a^T \mu$  s.t.  $\|a\|_2 = 1$ . This is a 1-d projection and a natural estimation is  $\widehat{Q}(a^T X) = \frac{1}{n} \sum_{i=1}^n a^T x_i$ . Using Lagragian multiplier, this index is maximized by  $a_0 = \frac{\mu}{\|\mu\|}$ . Thus, sample mean can be derived via PP as  $a_0 Q(a_0^T X)$ .

**Example 3.2** (Principal component analysis). Let  $Q(a^T X) = \text{Var}(a^T X)$  s.t.  $\|a\|_2 = 1$ . Then we get first principal component. Next, take  $Q(b^T X) = \text{Var}(b^T X)$  s.t.  $\|b\|_2 = 1, \langle a, b \rangle = 0$ . We get second principal component.

**Example 3.3** (Canonical correlation analysis (CCA)). Suppose  $X \in \mathbb{R}^{d_1}$  and  $Y \in \mathbb{R}^{d_2}$ , then

$$Q(a^T X, b^T Y) = \text{Corr}(a^T X, b^T Y)$$

corresponds to CCA. Note that CCA is affine invariant, that is , if we take  $Z = (a^X, b^T Y)^T$ , then  $Q(Z) = Q(t_1 Z + t_2)$ . This is known as **class III** index in [11].

**Example 3.4** (Fisher's linear discriminant analysis (LDA)). Suppose both  $X, Y \in \mathbb{R}^d$  and they share the same covariance matrix  $\text{Var}(X) = \text{Var}(Y) = \Sigma$ , then  $Q(a^T X, a^T Y) = \frac{(a^T \mathbb{E}(X) - a^T \mathbb{E}(Y))^2}{a^T \Sigma a}$  corresponds to the famous Fisher's linear discriminant function. To generalize this to multi-dimensional cases, we could take

$$Q(AX, AY) = \frac{\|A \mathbb{E}(X) - A \mathbb{E}(Y)\|_2^2}{\text{Tr}(A \Sigma A)} \text{ s.t. } A^T A = I$$

Both trace and  $L_2$  norm can be replaced by other suitable measurements (i.e. Frobenious norm) and det corresponds to multiple discriminant analysis (MDA). Note that the constraint  $A^T A = I$  can be solved alternatively. For instance, solving 1-d LDA gives us

$$a = \frac{\Sigma^{-1}(\mu_1 - \mu_2)}{\|\Sigma^{-1}(\mu_1 - \mu_2)\|_2}$$

**Example 3.5** (Johnson-Lindenstrauss embedding [5]). Supoose we want to preserve inner product or Euclidean distance among points after projection, we could take

$$Q(AX) = Q(Z) = \sum_{i,j} |\|z_i - z_j\|_2^2 - \|x_i - x_j\|_2^2|$$

And note that in an inner product space, we have  $4\langle x, y \rangle = \|x + y\|^2 - \|x - y\|^2$  where  $\|\cdot\|$  is any norm (for details, see [1]). Therefore, if distance induced by norm is preserved, then inner product is preserved automatically. However, maximizing this projection index requires solving a non-linear system which is computationally expansive (when both  $n$  and  $d$  is large, this is impossible). Therefore, we resort to a weaker constraint: "the distance/inner product among points are **almost** preserved".

More precisely, given a tolerance level  $\delta$  and a confidence level  $1 - \epsilon$ , we want to find a linear projection s.t.

$$(1 - \delta)\|x_i - x_j\|_2^2 \leq \|z_i - z_j\|_2^2 \leq (1 + \delta)\|x_i - x_j\|_2^2 \text{ holds w.p. } 1 - \epsilon$$

Johnson-Lindenstrauss embedding provides a perfect solution to this set-up: Let  $A \in \mathbb{R}^{r \times d}$  be filled with sub-Gaussian elements (i.e. normal, Rademacher, etc.), then  $A$  is the linear projection we want. For more details of this embedding, see chapter 2 of [5].

**Example 3.6** (Linear SNE). SNE originates from [15] which is a non-linear projection and non-convex problem. However, if we force the lower dimensional representation to be a linear projection of original data, then we can get the projection index

$$\begin{aligned} Q(AX) &= - \sum_{i=1}^n \sum_{j=1}^n p_{ij} \log(q_{ij}) \\ p_{ij} &= \frac{\phi(\frac{\|x_i - x_j\|_2}{\sigma_i})}{\sum_{k \neq i} \phi(\frac{\|x_i - x_k\|_2}{\sigma_i})}, q_{ij} = \frac{\phi(\|Ax_i - Ax_j\|_2)}{\sum_{k \neq i} \phi(\|Ax_i - Ax_k\|_2)} \end{aligned}$$

Where  $\phi(\cdot)$  denotes the pdf of a standard normal random variable. Here  $p_{ij}$ 's are constants and this is a weighted negative sum of  $\log q_{ij}$ . Note that the denominator of  $p_{ij}$  is not involved in optimization, Moreover, if we replace  $\sum_{j=1}^n$  by integration, then we do not need to normalize  $q_{ij}$ . The resulting projection index corresponds to "finding a projection that is mostly similar to high dimensional Gaussian distribution", which is the opposite of entropy-based indexes discussed in next subsection.

**Example 3.7** (Liquid association [3]). Taking  $Q(a^T X, b^T Y, c^T Z) = \mathbb{E}(a^T X b^T Y c^T Z)$  gives us a combination of PP and liquid association. For more details, see [3].

**Example 3.8** (Artifial neural networks [2]). PP for regression has a beatiful connection with Komogorov's universal approximation theorem: any continuous function with a compact support in  $\mathbb{R}^d$  can be approximated by the following conditional expectation  $\mathbb{E}(Y|X)$ . Let  $Y$ , a single output variable and  $X$ , a random vector in  $\mathbb{R}^d$  be modeled as

$$\begin{aligned} Y &= a_0 + \sum_{j=1}^t f_j(\beta_{0j} + X^T \beta_j) + \epsilon \\ \mathbb{E}(\epsilon) &= 0, \text{Var}(\epsilon) = \sigma^2 \end{aligned}$$

For instance, if  $\mathbb{E}(Y|X) = X_1 X_2$ , then we can rewrite  $X_1 X_2 = \frac{1}{4}(X_1 + X_2)^2 + \frac{1}{4}(X_1 - X_2)^2$ , a linear combination of projections of  $(X_1, X_2)$ . In the language of statistics,  $f_j(\cdot)$  is called **ridge function** while **activation function** is used in the field of machine learning. In fact,  $\mathbb{E}(Y|X)$  has the same form as a two-layer artifial neural network. Then, define our projection index to be

$$Q(\beta^T X + \beta_0) = \sum_{i=1}^n \left\{ y_i - a_0 + \sum_{j=1}^t f_j(\beta_{0j} + x_i^T \beta_j) \right\}^2$$

Where  $\beta = (\beta_1, \dots, \beta_d)$  and  $\beta_0 = (\beta_{01}, \dots, \beta_{0d})^T$ . Such index can be maximized by the so-called **back-propagation** algorithms. For more details, see [14], [12] and [2].

**Example 3.9** (Density approximation using Hellinger distance [11]). Suppose we are interested in approximating the density function  $f$  in  $\mathbb{R}^d$ , then we could use **multiplicative decompositions** ([11])

$$f_k(x) = f_0(x) \prod_{i=1}^k h_i(a_i^T X)$$

to approximate  $f$  where  $f_0$  is a standard density in  $\mathbb{R}^d$  (e.g. multivariate Gaussian). To measure the distance of  $f_k$  and  $f$ , we need a metric in the space of probability densities, such metric can be taken as **Hellinger distance**:

$$Q(AX) = \text{Hellinger}(f, f_k) = \int (\sqrt{f} - \sqrt{f_k})^2 dx$$

### 3.2 Entropy based indexes

Although a lot of linear models can be viewed as PP, most literatures concerned about **entropy based indexes** and their approximations. The reasons are:

- Due to Diaconis-Freedman theorem, non-Gaussian projections would indeed be rare and interesting.
- There is a strong connection between PP and independent component analysis (ICA) where the latter uses information theory as its theoretical foundations.
- In many unsupervised applications, PCA is suffice to give good visualizations (though in theory it may not perform good).

### 3.2.1 Entropy and its properties

To get into more technical details, we need some basic concepts from information theory. For details, see **appendix** and [19].

Applications of information theory to PP (based two lemmas on entropy in appendix) are Fisher's information and mutual information.

**Example 3.10** (Fisher's information). Let  $X$  have continuous differentiable density  $f$  w.r.t.  $\lambda_d$ , then the Fisher's information projection index is

$$Q(X) = \sigma^2 \int \left( \frac{f'}{f} \right)^2 f d\lambda_d - 1$$

Or equivalently,

$$Q(X) = \sigma^2 \int \left( \frac{f'}{f} - \frac{\phi'}{\phi} \right)^2 f d\lambda_d$$

where  $\sigma^2 = \text{Var}(X)$  and  $\phi(\cdot)$ , as usual, the pdf of  $\mathcal{N}(0, 1)$ . Then this index is affine invariant due to lemma 5.1.

**Example 3.11** (Mutual information (MI) [20],[19]). The general definition of mutual information is given in [20]. The intuition of MI is "the amount of information contained in  $X$  about  $Y$ ". Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space and  $X, Y$  are a pair of  $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable random variables. Then the **mutual information** of  $X$  and  $Y$  is

$$I(X, Y) = \int_{\mathbb{R}^2} \mathbb{P}_{X,Y}(dxdy) \log \frac{\mathbb{P}_{X,Y}(dxdy)}{\mathbb{P}_X(dx)\mathbb{P}_Y(dy)}$$

where  $\mathbb{P}_{X,Y}$  is the probability measure induced by  $(X, Y)$ . Note that,  $I(X, Y) \geq 0$  and equality is attained iff  $X$  and  $Y$  are independent. Thus, mutual information can be used as a projection index for measuring independence. Now suppose  $X$  is random vector and  $A \in \mathbb{R}^{d \times d}$  and define

$$Q(AX) = - \int_{\mathbb{R}^d} \mathbb{P}_{AX}(dx_1 \cdots dx_d) \log \frac{\mathbb{P}_{AX}(dx_1 \cdots dx_d)}{\mathbb{P}_{a_1^T X}(dx_1) \cdots \mathbb{P}_{a_d^T X}(dx_d)}$$

Then  $Q(AX) \leq 0$  with equality attained iff all projected directions are independent. Note that this is different from PCA since the latter only gives uncorrelated projections.

Equipped with new weapons, we will introduce a frequently used entroy based index: stardardized negative Shannon's entropy . Then we will discuss how to approximate it in practice. Methods based on cumulant and non-polynomial functions will be introduced.

**Example 3.12** (Stardardized negative Shannon's entropy). Suppose we have some prior knowledges about our data (i.e.  $\mathbb{E}(X) = \mu, \mathbb{E}(\sin(\beta^T X)) = c$ ) and we want our data is away from "chaos", which is measured by **entropy** in information theory, then due to the following theorem, we would expect our projected data to be as far away from a family of particular distributions as possible. Therefore, we choose our projection index to be

$$Q(a^T X) = - \int \log\left(\frac{\phi}{f}\right) f dx$$

where  $\phi(\cdot)$ , the pdf of  $\mathcal{N}(0, 1)$ ,  $f$ , the pdf of projected data and  $\sigma = \sqrt{\text{Var}(a^T X)}$ . One reason we want to set  $f$  to be far away from Gaussian is Diaconis-Freedman and another is the following.

**Theorem 3.1** (Maximum entropy principle [7],[18],[19]). Suppose  $X$  is a random vector with density  $f$  w.r.t. counting or Lebesgue measure and we have the following constraints:

$$\int_{\mathbb{R}^d} f(x) \mu(dx) = 1, \int_{\mathbb{R}^d} f(x) r_j(x) \mu(dx) = \alpha_j, 1 \leq j \leq k$$

Then  $f$  has the following form (w.r.t.  $\mu$ )

$$f(x) = A \exp\left\{ \sum_{j=1}^k \eta_j r_j(x) \right\}$$

where  $A$  and  $\eta_j$  are constants s.t.  $f$  is a probability density. If we take  $k = 2$ ,  $r_1(x) = x$  and  $r_2(x) = x^2$ , then

$$f(x) \propto \exp\{\eta_1 x + \eta_2 x^2\}$$

which is the density of a univariate Gaussian distribution.

However, we cannot evaluate  $\int f \log f dx$  directly since this is an integration. Also, estimation of density  $f$  is difficult (kernel estimators will be very poor prone [12]). Thus, we have to approximate the entropy directly by some numerical methods. This leads us to the following 2 subsubsections.

### 3.2.2 Cumulant-based approximations

The idea is to use high-order cumulants based on the Hermite polynomials, a complete orthogonal polynomial basis in  $L_2(\mathbb{P})$  where  $\mathbb{P}$  denotes the standard Gaussian distribution. They are given by  $H_0(x) = 1$  and

$$H_n(x) = (-1)^n \frac{\phi^{(n)}}{\phi}(x) \text{ for } n \geq 1$$

$$\int \phi(x) H_n(x) H_m(x) dx = n! \delta_{nm}$$

where  $\delta_{nm}$  is Kronecker delta. For technical details, see chapter 8 of [1]. If we cut off this expansion at the first 2 non-constants, then we get an approximation of the density function  $f(x)$ . Then the standardized negative Shannon's entropy of  $a^T X$  can be approximated by (e.g. [12])

$$\widehat{Q}(a^T X) = \widehat{Q}(Z) = \frac{(\kappa_3(Z))^2}{12} + \frac{(\kappa_4(Z))^2}{48}$$

where  $\kappa_3$  and  $\kappa_4$  are the usual cumulants. For cumulants in high dimensions, see [13].

### 3.2.3 Approximation based on non-polynomial functions

Cumulants are not robust to outliers so we need some other approximations. One mostly used approximation is based on non-polynomial functions ([17],[14],[12],[2],[6]). Satisfying a group of constraints (e.g. section 5.6 of [12], section 15.3 of [2]), we can approximate  $f(x)$  by

$$\widehat{f}(x) = \phi(x) \left( 1 + \sum_{i=1}^n c_i G_i(x) \right)$$

Where  $c_i = \mathbb{E}(G_i(x))$  and  $G_i$ 's are non-polynomial functions. Then an approximation of projection index is

$$\widehat{Q}(a^T X) = \widehat{Q}(Z) = \frac{1}{2} \sum_{i=1}^n (\mathbb{E}(G_i(Z)))^2$$

And we replace  $\mathbb{E}$  by sample mean or other common estimators. In section 4, we will let  $n=1$  and choose  $G(x) = \frac{1}{\alpha} \log \cosh(\alpha x)$  to approximate standardized negative Shannon's netropy.

## 4 Application to scRNA-seq data

### 4.1 Data pre-processing

We start with the raw UMI count data from PBMC experiments [21] (data downloaded from Broad Institute Single Cell Portal). First, we separate the data from the first PBMC experiment into 7 matrices (cell by gene) based on the sequencing method. Second, we filter out genes with zero expression in more than 80% cells. Lastly, we quantile normalize each count matrix so that in each count matrix, the genes have the same empirical distribution in every cell. The dimensions of the resulting count matrices are as follows:

1. pbmc1\_CEL\_Seq2:  $253 \times 3,576$
2. pbmc1\_10xChromiumv2A:  $3,222 \times 623$
3. pbmc1\_10xChromiumv2B:  $3,222 \times 736$
4. pbmc1\_10xChromiumv3:  $3,222 \times 1,696$

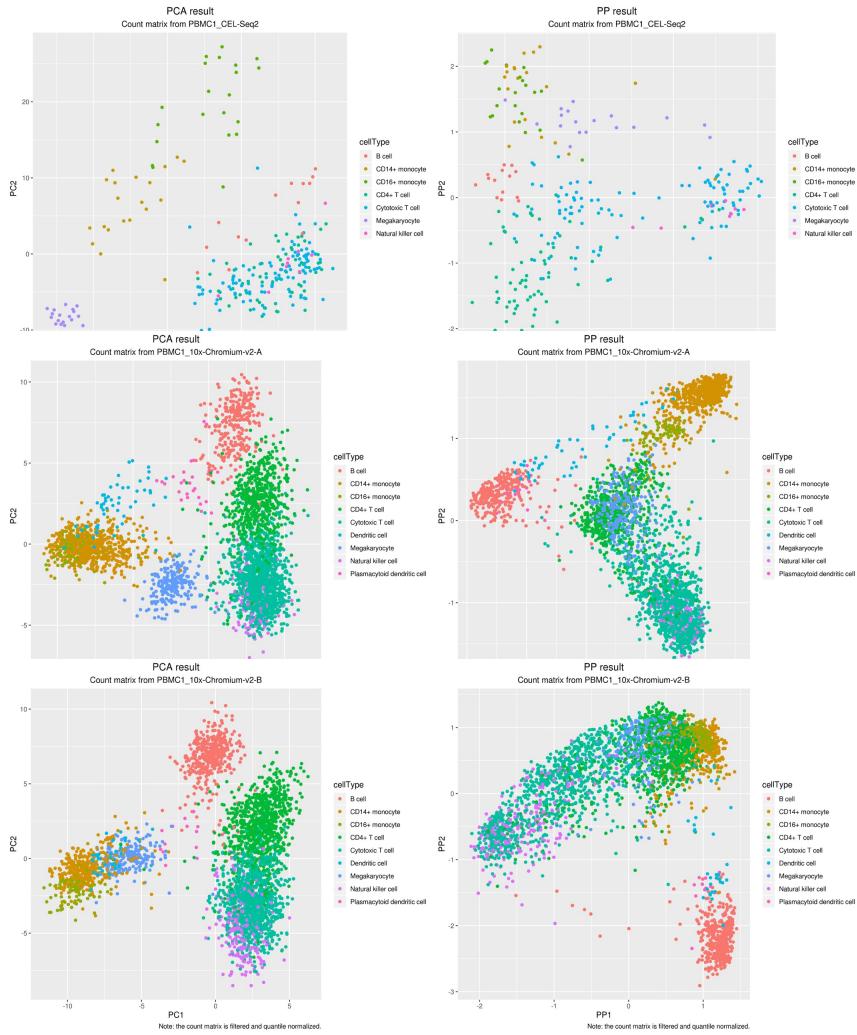
5. pbmc1\_Drop\_seq:  $3,222 \times 516$
6. pbmc1\_Seq\_Well:  $3,222 \times 337$
7. pbmc1\_inDrops:  $3,222 \times 212$

In addition, we extract the true cell type labels of the cells from meta.txt on the Single Cell Portal, which are derived using marker genes by [21]. We made sure that the cells in the count matrices matched with the cells in meta.txt.

## 4.2 Results

For each of the 7 pre-processed count matrices, we apply PCA and PP (with negative standardized Shannon's entropy as the projection index) and plot the results in Figure 1. Each row corresponds to one count matrix from one particular sequencing method. The left column is results from PCA, and the right column is results from PP. Each point in a plot is a cell, and the points are colored based on the true cell type.

We observe from Figure 1 that PCA actually produces better results than PP in terms of keeping cells of the same cell type close together. For example, for the count matrix from 10x Chromium V2-B (third row of Figure 1), the data points of the same color tend to stay close together in the PCA plot (left) more than they do in the PP plot (right). For our future work, we can explore PP methods with other projection indices that are potentially more suitable for clustering. In particular, we will need to study how to execute the optimization of the projection indices and implement the methods, since many of them are not already implemented.



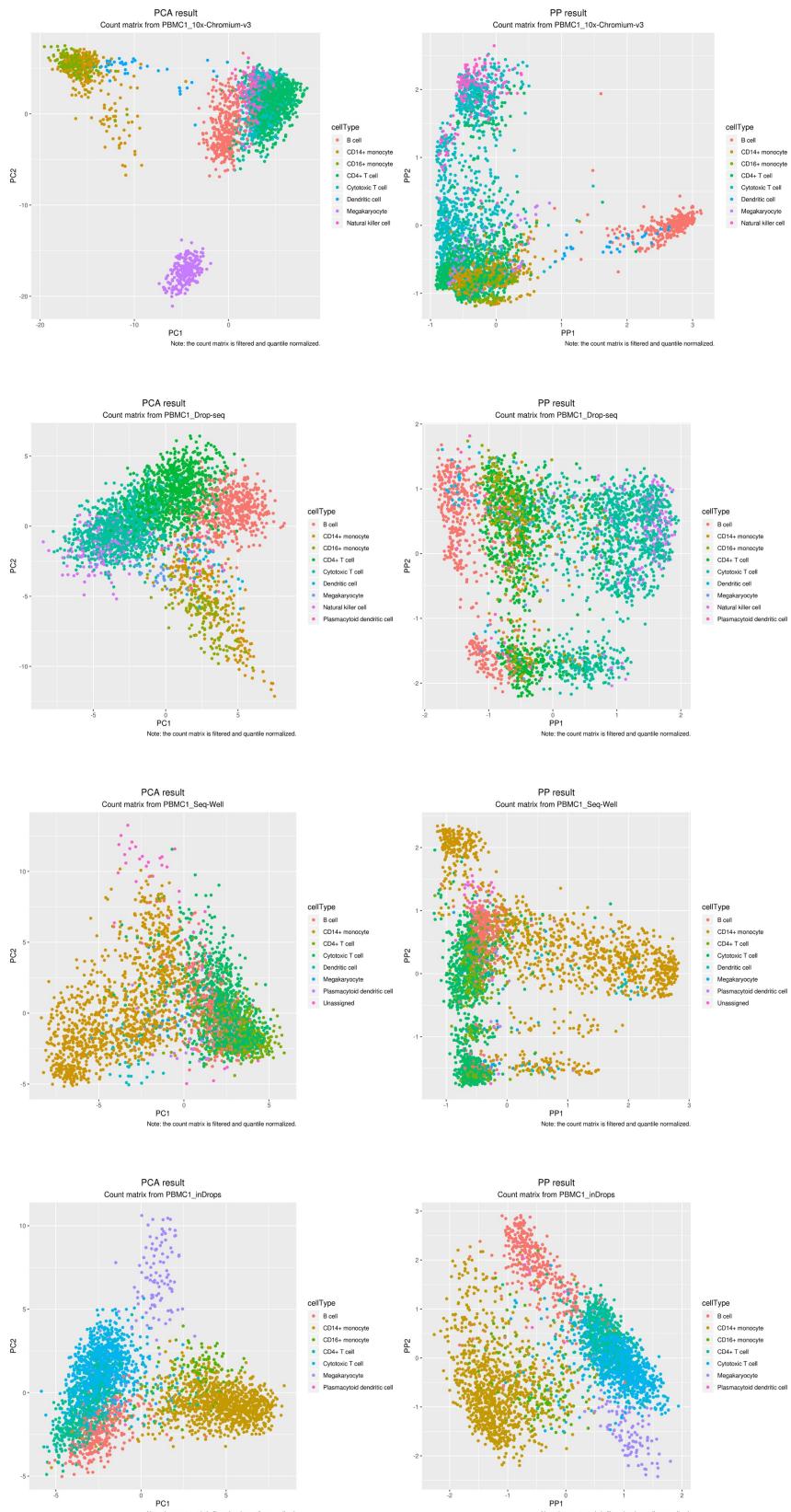


Figure 1: PCA (left) and PP (right) results on the count matrices from 7 different sequencing methods

## 5 Appendix

**Definition 5.1** (Entropy). The entropy of a probability measure  $\mu$  dominated by a  $\sigma$ -finite counting measure on  $\mathbb{R}^d$  is

$$H(\mu) = - \sum_{\omega \in \mathbb{R}^d} \mu(\{\omega\}) \log \mu(\{\omega\})$$

**Definition 5.2** (Differential entropy [20]). The differential entropy of a probability measure  $\nu$  dominated by Lebesgue measure on  $\mathbb{R}^d$  is defined as

$$H(\nu) = - \int \frac{d\nu}{d\lambda_d}(\omega) \log \frac{d\nu}{d\lambda_d}(\omega) \lambda_d(d\omega)$$

where  $\lambda_d$  is the d-dim Lebesgue measure and  $\frac{d\nu}{d\lambda_d}$  is the Radon-Nikodym derivative (i.e., see [1]) of  $\nu$  w.r.t.  $\lambda_d$ .

**Definition 5.3** (Kullback–Leibler divergence). Let  $\mathbb{P}$  and  $\mathbb{Q}$  be equivalent probability measures dominated by a common measure  $\mu$  (either counting or Lebesgue), then the Kullback–Leibler divergence (KL divergence) of  $\mathbb{P}$  and  $\mathbb{Q}$  is

$$KL(\mathbb{P} || \mathbb{Q}) = - \int \frac{d\mathbb{P}}{d\mu} \log \left( \frac{\frac{d\mathbb{Q}}{d\mu}}{\frac{d\mathbb{P}}{d\mu}} \right) d\mu$$

We have the following lemma of entropy.

**Lemma 5.1** (Entropy of affine transformation). Let  $X$  be a random vector, then the entropy of  $X$  is defined to be  $H(X) = H(\mathbb{P}_X)$  where  $\mathbb{P}_X$  is the induced probability measure. Let  $Y = AX + b$ , a non-singular affine transformation of  $X$  ( $A$  is invertible), then

$$H(Y) = \begin{cases} H(X) & \text{if } \mathbb{P}_X \text{ is discrete} \\ H(X) + \log |\det A| & \text{if } \mathbb{P}_X \ll \lambda_d \end{cases}$$

The proof is based on change of variable formula in measure theory (i.e. theorem 6.3.1 in [1]) and we have an important consequence:

**Lemma 5.2** (KL divergence is affine invariant). Suppose  $\mathbb{P}$  and  $\mathbb{Q}$  are two probability measures dominated by counting or Lebesgue measure on  $\mathbb{R}^d$  and  $A \in \mathbb{R}^{d \times d}$  is invertible, then

$$KL(\mathbb{P} || \mathbb{Q}) = KL(\mathbb{P}_A || \mathbb{Q}_A)$$

where  $\mathbb{P}_A$  and  $\mathbb{Q}_A$  are probability measures induced by the linear transformation  $A$ .

## References

- [1] D. M. Dabrowska. Advanced Probability with elements of real analysis and statistics. Lecture Notes. UCLA Biostatistics Department. 2019.
- [2] A. Z. Izenman. Modern Multivariate Statistical Techniques. *Springer*, 2008.
- [3] J. Li. Statistical Methods in Computational Biology. Lecture Notes. UCLA Statistics Department. 2019.
- [4] J. Li. Large Sample Theory with resampling methods. Lecture Notes. UCLA Statistics Department. 2019.
- [5] M. Wainwright. High Dimensional Statistics. *Cambridge University Press*, 2019.
- [6] T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer*, 2nd edition, 2009.
- [7] P. Bickel and K. Doksum. Mathematical Statistics: Basic Ideas and Selected Topics, Volumn I. *CRC Press*, 2nd edition, 2001.
- [8] P. Bickel, G. Kur and B. Nadler. *Projection pursuit in high dimensions*. PNAS **37** (2018) 9151-9156.
- [9] P. Diaconis and F. Freedman. *Asymptotics of graphical projection pursuit*. Ann. Statist. **12** (1984) 793-815.
- [10] J. Friedman and J. Tukey. *A projection pursuit algorithm for exploratory data analysis*. IEEE Trans. Comput. **C-23** (1974) 881-889.
- [11] P. Huber. *Projection Pursuit*. Ann. Statist. **13** (2005) 435-475.
- [12] A. Hyvärinen, J. Karhunen, and Erkki Oja. Independent Component Analysis. *John Wiley and Sons, INC.*, 2001.
- [13] K.V. Mardia. *Measures of multivariate skewness and kurtosis with applications*, Biometrika **57** (1970) 519–530.
- [14] J. Friedman. *Exploratory Projection Pursuit*. JASA Vol. 82, No. 397, (1987) 249-266.
- [15] G. Hinton and S. Roweis. *Stochastic Neighbor Embedding*. <https://www.cs.toronto.edu/~fritz/absps/sne.pdf>, 2003.
- [16] A. M. Pires and J. A. Branco. *Projection-pursuit approach to robust linear discriminant analysis*, Journal of Multivariate Analysis **101** (2010) 2464–2485.
- [17] P. Comon. *Independent Component Analysis, a new concept?*. Signal Processing **36** (1994) 287-314.
- [18] R. Sundberg. Statistical Modelling by Exponential Families. *Cambridge University Press*, 2019.
- [19] T. M. Cover and J. A. Thomas. Elements of information theory. *John Wiley and Sons, INC.*, 2nd edition, 2006.
- [20] A. N. Kolmogorov, I. M. Gelfand and A. M. Yaglom. *Amount of information and entropy for continuous distributions*. Selected works of A. N. Komogorov. Volume III: Information theory and the theory of algorithms. Springer. 1987.
- [21] J. Ding, X. Adiconis, et al. *Systematic comparative analysis of single cell RNA-sequencing methods*. bioRxiv. 2019.