Dimension Reduction Using Projection Pursuit

Elvis Cui and Heather Zhou

Dept. of Biostatistics, UCLA
Dept. of Statistics, UCLA

January 7, 2020

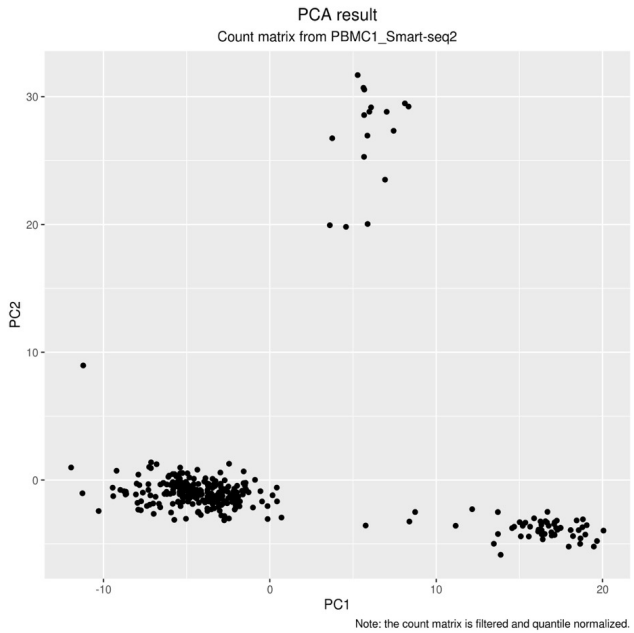
Overview

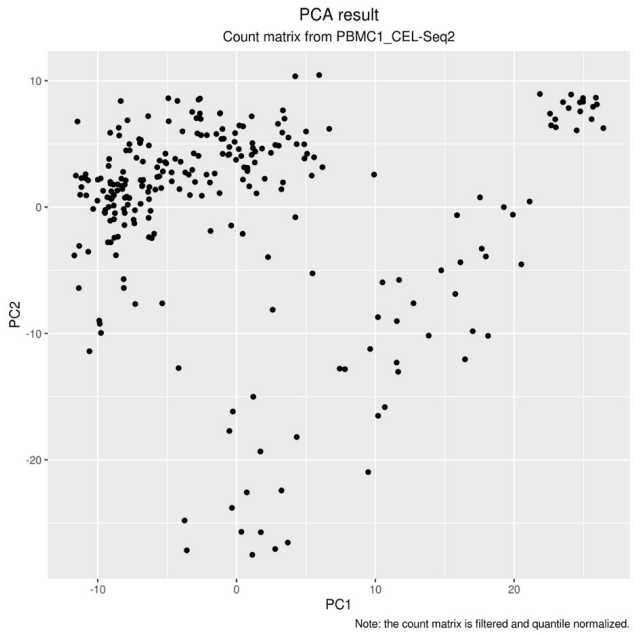
1. Data pre-processing
2. PCA is not consistent in high dimensions
   - Marcenko-Pastur Law
3. Projection pursuit
   - Definitions
   - Examples of projection indexes
   - PP for highly correlated data (genetic data)
   - Diaconis-Freedman theorem
   - **Bickel-Kur-Nadler approximation theorem**
4. Applications to RNA-sequencing data
5. References

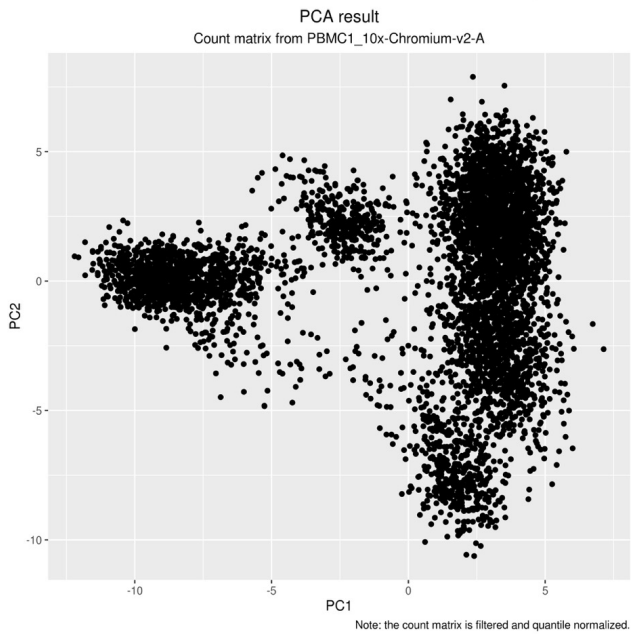
Data pre-processing

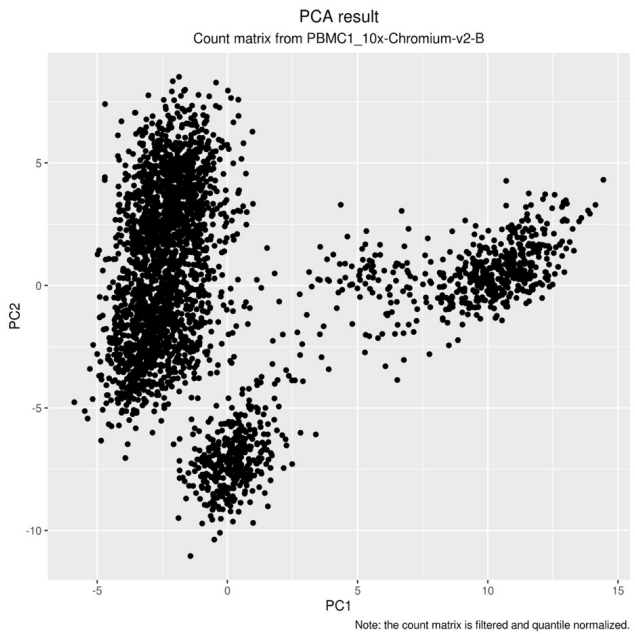
1. Start with the PBMC (frozen human peripheral blood mononuclear cells) count matrix
2. Get 8 sub count matrices (cell by gene):
    1. PBMC1_Smart-seq2 (311*33,694 $\rightarrow$ 311*3,105 after filtering)
    2. PBMC1_CEL-Seq2 (257*33,694 $\rightarrow$ 257*3,680)
    3. PBMC1_10x-Chromium-v2-A (5,172*33,694 $\rightarrow$ 5,172*674)
    4. PBMC1_10x-Chromium-v2-B (3,057*33,694 $\rightarrow$ 3,057*1,006)
    5. PBMC1_10x-Chromium-v3 (4,033*33,694 $\rightarrow$ 4,033*1,879)
    6. PBMC1_Drop-seq (4,683*33,694 $\rightarrow$ 4,683*560)
    7. PBMC1_Seq-Well (5,125*33,694 $\rightarrow$ 5,125*336)
    8. PBMC1_inDrops (6,184*33,694 $\rightarrow$ 6,184*210)
3. Filter genes
    1. Filter out genes with zero expression in more than 80% cells
    2. Filter out genes with genes with coefficient of variation in the bottom 20 percent
4. Quantile normalize the count matrices so that in each count matrix, the genes have the same empirical distribution in every cell

Dimension reduction by PCA (Part 1)

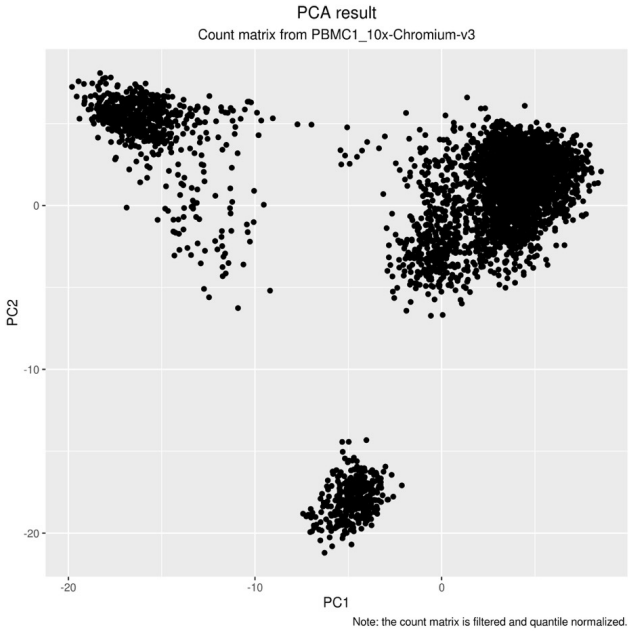


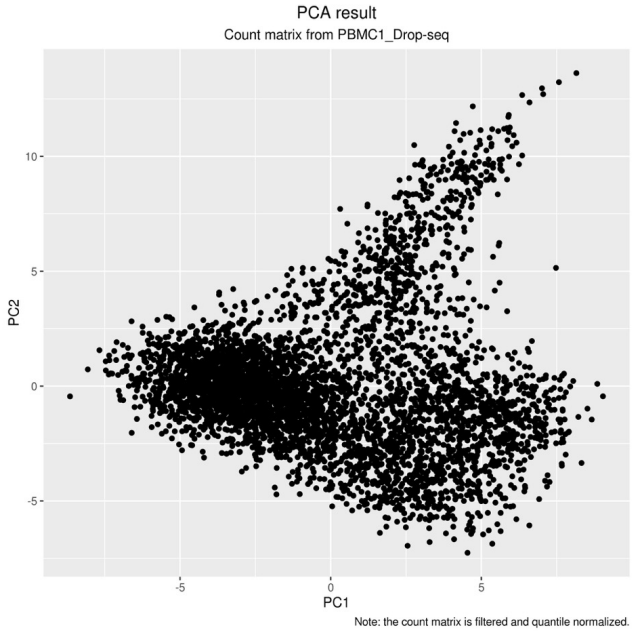


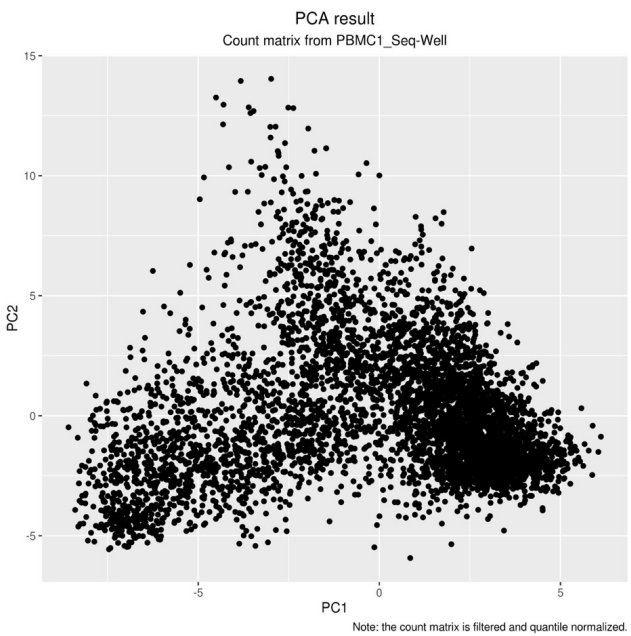


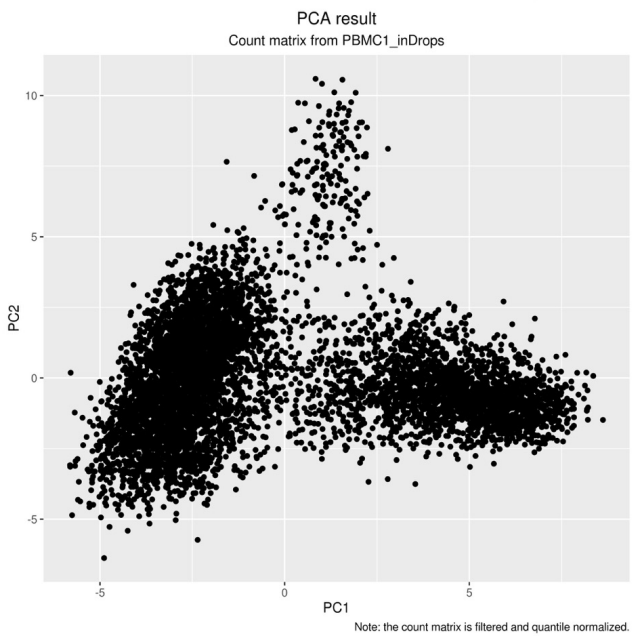


Dimension reduction by PCA (Part 2)









Empirical spectral distribution

Random matrix theory gives us the asymptotic results:

Theorem 1 (Marcenko-Pastur Law (1967))

Suppose $\mathbf{X}\mathbf{X}^T \sim \mathcal{W}_p(n, \mathbf{I}_p)$. *Define the* **empirical spectral distribution**

$$G_p(k) = \frac{1}{p}\#\{\hat{\lambda}_j \leq k\}$$

If $\frac{p}{n} \to \gamma \in (0, \infty)$, *then,* $G_r(k) \to G(k)$ *almost surely, where the limiting distribution* $G(k)$ *has density* $g(k) = G'(k)$*:*

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

This is the famous **Marcenko-Pastur Law** *in statistical physics.*

Therefore, PCA fails in high dimensional settings.

Projection Pursuit

- **Problems with PCA**: Get misleading conclusions based on such projections (or principal components) [has been well studied].
- **Goal**: Find an "interesting" projection in lower dimensions (perhaps 2 or 3).
- **Solution**: Projection Pursuit (PP)! (Friedman and Tukey, 1974)
- **1 A general framework**: Sample mean, CCA, PCA, ICA and liquid association can be derived as special cases of PP.
- **2 Curse of dimensionality**: PP is one of the very few methods able to bypass the "curse of dimensionality" (Huber, 1985).
- **3 Approximation ability**: PP is able to approximate any distributions provided $\frac{p}{n} \to \infty$ (**Bickel et al, 2018**)
- **4 Fourier inversion theorem**: Any d-dim distribution is uniquely determined by its 1-d projection due to uniqueness of characteristic functions (Biostats 255B; **Stats 203**).

Definition of PP

Definition 2 (Projection index (Freedman and Tukey, 1974))

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

Definition 3 (Projection pursuit (Huber, 1985))

Projection pursuit 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.

Sidenote:

- For visualization, $\binom{d}{2}$ or $\binom{d}{3}$ will be impossible.
- Tukey and Tukey (1981) provided an interesting fact about PP and random number generator (see Fig. 2.1 in Huber, 1985).

Examples of projection indexes

Example 4 (1-d location estimators)

This is class I projection index in Huber (1985). we choose $Q(\cdot)$ s.t.

$$Q(sZ + t) = sQ(Z) + t,\ Z = X^T a \in \mathbb{R}$$

Thus, $Q(a^T X) = \mathbb{E}(a^T X)$ is class I and maximized by $a_0 = \frac{\mu}{\|\mu\|}$.

Example 5 (PCA as location invariance, scale equivariance)

This is class II projection index in Huber (1985). We choose $Q(\cdot)$ s.t.

$$Q(sZ + t) = |s|Q(Z),\ Z = X^T a \in \mathbb{R}$$

Thus, PCA is a special case of class II projection index:

$$Q(a^T X) = [\mathbb{E}(a^T (X - \mu))^2]^{1/2} \text{ s.t. } \|a\| = 1$$

Examples of projection indexes cont'd

Example 6 (Canonical correlation analysis)

Set $Q(a^T X, b^T Y) = Corr(X^T a, Y^T b)$, then this is also a class II projection index (i.e. location invariance).

Example 7 (Shannon entropy as affine invariance)

This is class III in Huber (1985). We require that $Q(\cdot)$ is affine invariant:

$$Q(AX + b) = Q(X), \ A \in \mathbb{S}_+^{d \times d}$$

Thus, standardized negative Shannon entropy (Shannon, 1948) is a class III projection index:

$$\begin{aligned} Q(X) &= \int \log(f) f dx + \log((2\pi e)^{1/2} \sigma(X)) \\ &= -\int \log(\frac{\phi}{f}) f dx \end{aligned}$$

Examples of projection indexes cont'd

Example 8 (Johnson-Lindenstrauss embedding)

Choose the functional $Q(\cdot)$ to be

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

That is, we want to preserve L_2 distance between all data points. Besides MDS we learned in class, **Johnson-Lindenstrauss embedding** (for an introduction, see e.g. Stats 200C; Wainwright, 2019) will be a perfect alternative.

Let $A \in \mathbb{R}^{r \times d}$ filled with **sub-Gaussian random variables**, then with exponential-decaying probability, **the inner product and L_2 distance** is preserved:

$$\mathbb{P}(\exists (i,j) \text{ s.t. } \|Ax_i - Ax_j\|_2^2 > (1+\delta)\|x_i - x_j\|_2^2) < \epsilon$$

Examples of projection indexes cont'd

Example 9 (Fisher information as projection index)

Define the standardized Fisher information projection index:

$$Q(X) = \sigma^2(X) \int (\frac{f'}{f})^2 f dx - 1$$

This can be rewritten as

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

where $\phi(\cdot)$ the pdf of $\mathcal{N}(0, 1)$. Clearly, this is also affine invariant (thus, class III).

Examples of projection indexes cont'd

Example 10 (Polynomial-Based Indexes)

Recall $Z = a^T X$ with density function f_Z. **Polynomial-based projection indexes** take the general form of weighted integrated squared error,

$$Q(Z) = \int [\phi(z) - f_Z(z)]^2 w(z) dz$$

- **J. Friedman's index** (1987): $Q(Z) = \int_{-1}^{1} [p_U(u) - \frac{1}{2}]^2 du$, where $U = 2\Phi(Z) - 1$ and $\Phi(\cdot)$ is the cdf of $\mathcal{N}(0, 1)$. This index can be approximated by (Jones and Sibson, 1987)

$$Q(Z) \approx \frac{(\kappa_3(Z))^2}{12} + \frac{(\kappa_4(Z))^2}{48}$$

- **P. Hall's index** (1989): $Q(Z) \propto (\mathbb{E}(\phi(Z)) - \frac{1}{2\pi^{1/2}})$.

PP for highly correlated data

- Suppose $X = f * Z$ where

  X : a column of count matrix (population version)

  f : convolution kernel

  Z : hidden variable

  $*$: convolution operator

- Then we have

$$X_t = \sum_{s=0}^{t} f_s Z_{t-s}$$

- **Goal**: find a deconvolution kernel q s.t.

$$q * X = Z$$

PP for highly correlated data cont'd

- Note * forms a semi-group:

$$q * X = q * f * Z = (q * f) * Z$$

- Thus $(q * X)_t$ is a linear combination of Z.

Definition 11 (PP for high correlated data)

(Donoho, 1981; Huber, 1985)

1. Restrict $length(q) = d$.
2. Derive $(X_t, \cdots, X_{t+d-1}) \in \mathbb{R}^d$.
3. Find a **least normal** (or other indexes) 1-d projection:

$$q := f^{-1}$$

Diaconis-Freedman theorem

- Define **empirical projection distribution**:

$$\widehat{G}_z(t) = \frac{1}{n}\sum_{j=1}^{n} 1(x_j^T z \leq t)$$

Theorem 12 (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 $d, n \to \infty$, $\forall \epsilon > 0$

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

where ρ 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.

Bickel-Kur-Nadler Theorem

However, in a special case, we have the following remarkable result:

Theorem 13 (Bickel et al, 2018)

Suppose $\frac{d}{n} \to \infty$. *Let* $G(t)$ *be an* **arbitrary cumulative distribution function**. *There there* $\exists$ *a sequence of projections* $z = z_n \in \mathbb{S}^{d-1}$ *s.t. the following holds:*

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

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.

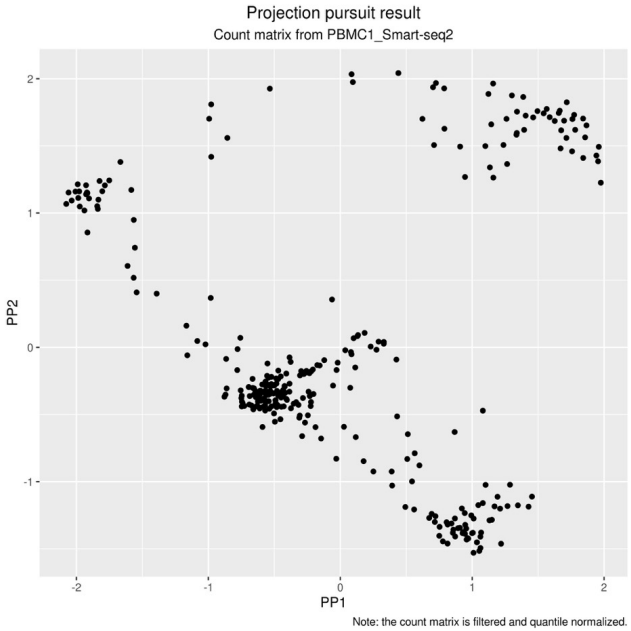
Applying projection pursuit to RNA-seq data

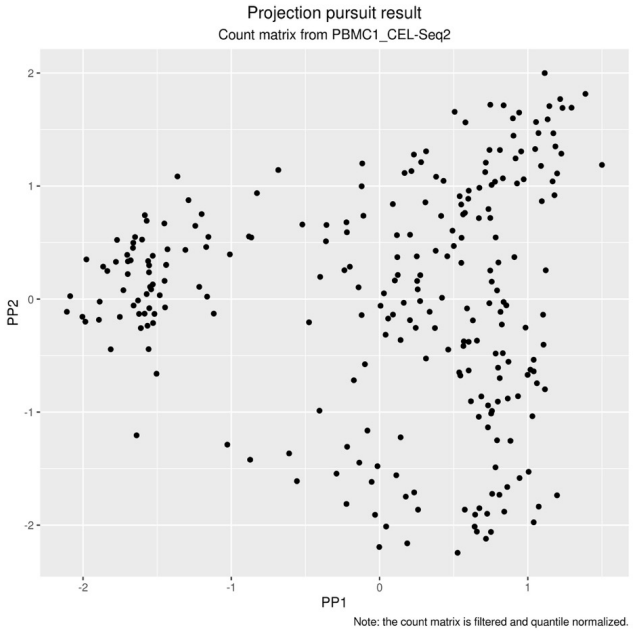
1. The projected data $\mathbf{X}\hat{\mathbf{b}} \in \mathbb{R}^{n \times 2}$ is found by

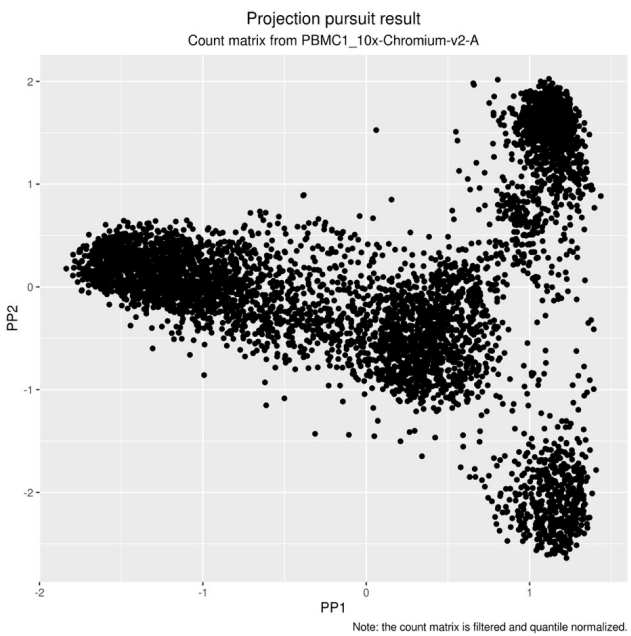
$$\hat{\mathbf{b}} := \arg\max Q(\mathbf{X}\mathbf{b})$$

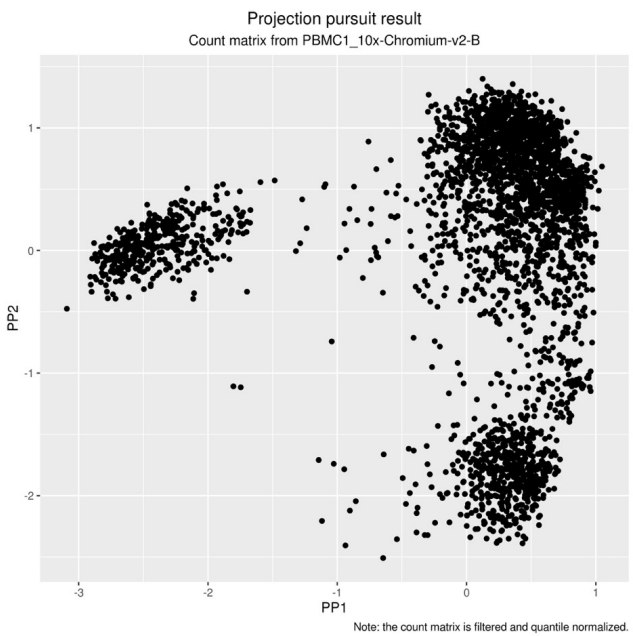
   where $Q(\cdot)$ denotes the negative standardized Shannon entropy.
2. In other words, we maximize the negative Shannon entropy of the projected data, which means we maximize the "distance" of the empirical distribution of the projected data from the standard multivariate Gaussian distribution.

Dimension reduction by projection pursuit (Part 1)

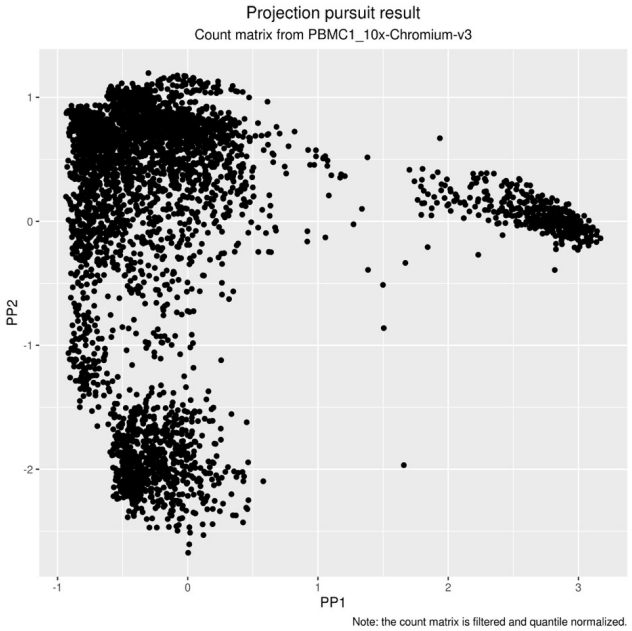


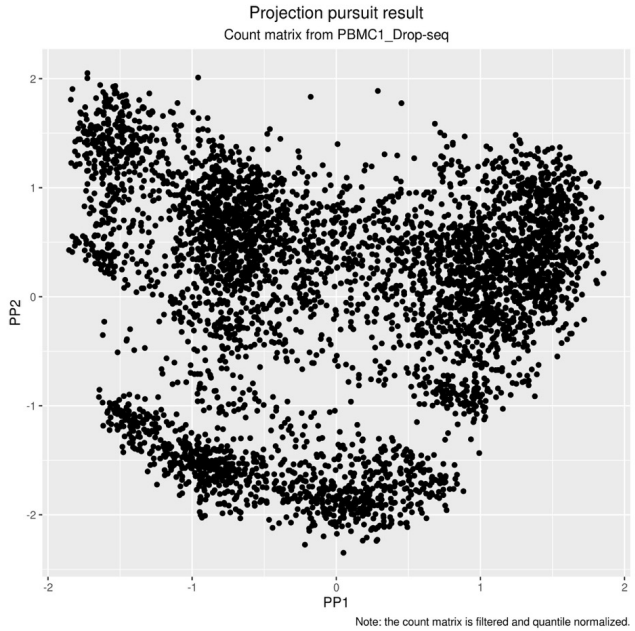


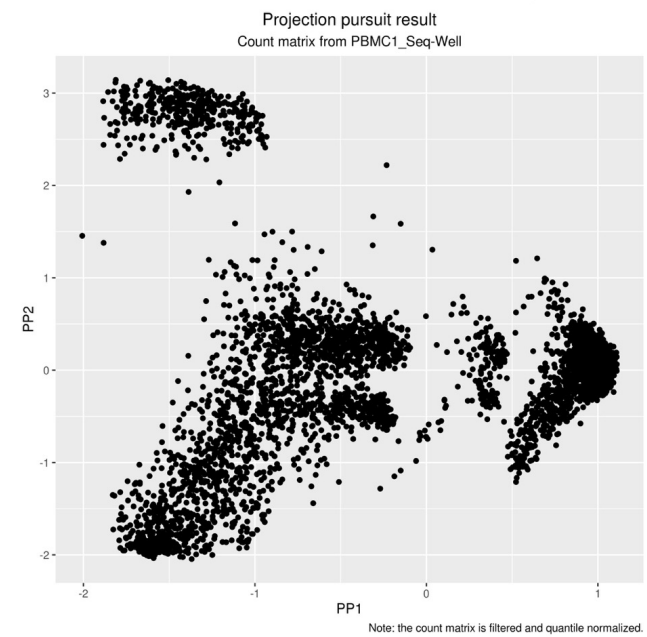


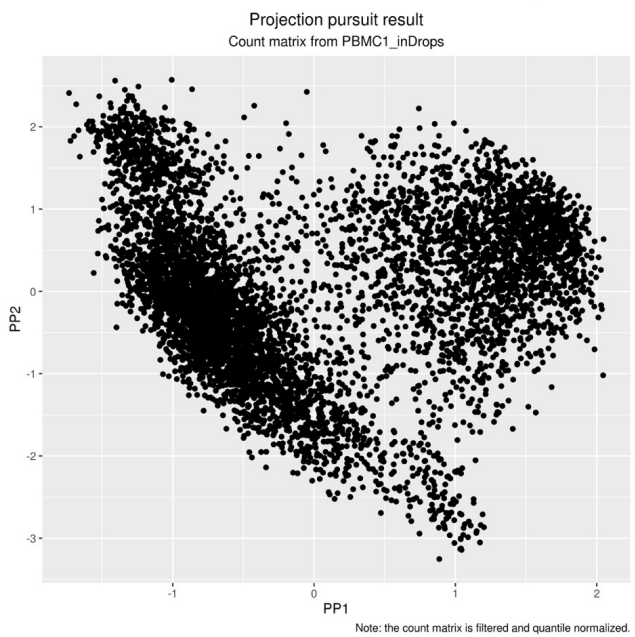


Dimension reduction by projection pursuit (Part 2)



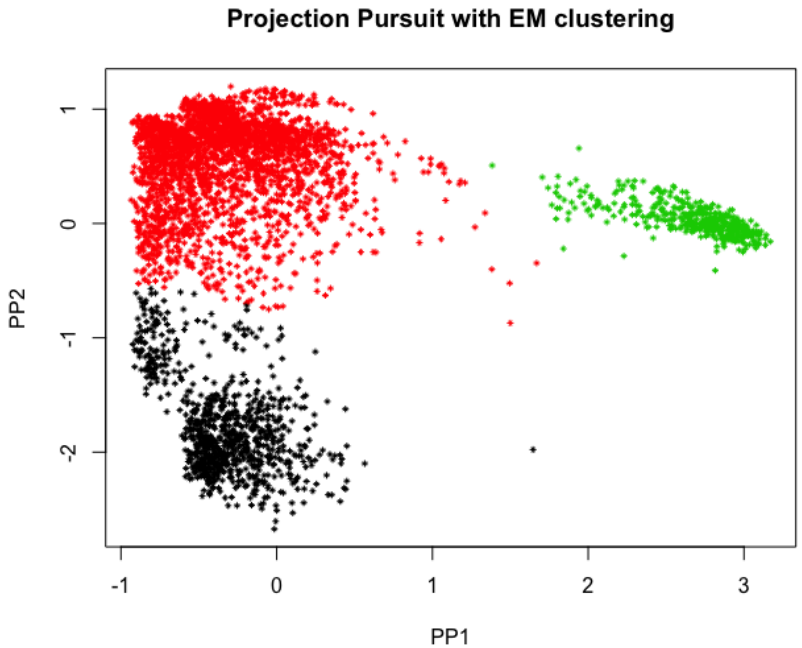


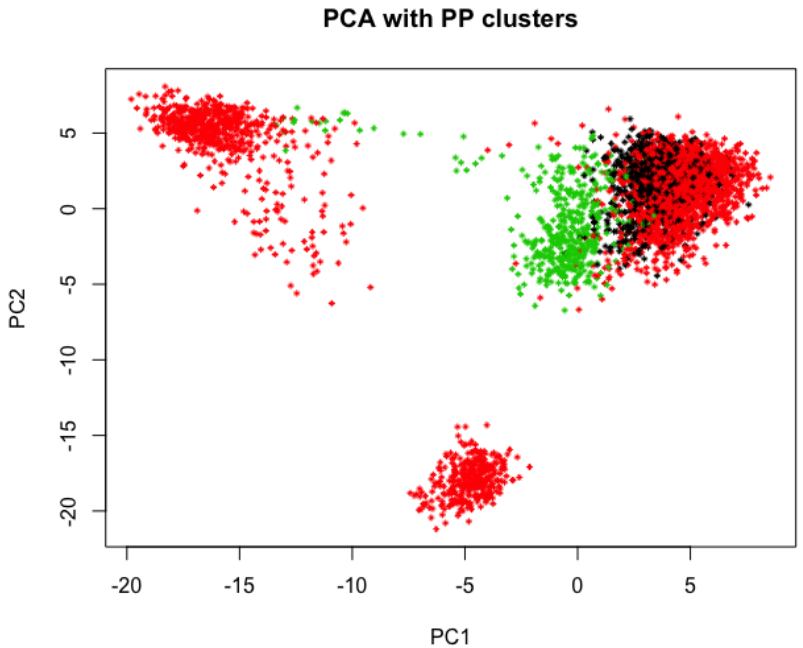




PP and PCA produce different results

Take PBMC1_10x-Chromium-v3 as an example:





References

1. Bickel, P. (2018). Projection pursuit in high dimensions. PNAS **37** 9151-9156.
2. Diaconis, P. and Freedman, F. (1984). Asymptotics of graphical projection pursuit. Ann. Statist. **12** 793-815.
3. Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Comput. **C-23** 881-889.
4. Huber, P (1985). Projection Pursuit. Ann. Statist. **13** 435-475.
5. Izemann, A. J. (2008). Modern multivariate statistical methods. Springer. Chapter 7, 12, 15 and 16.