Choosing the number of principal components for denoising single-cell RNA sequencing data

Aaron Lun[1,*]

1 Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom

1 Background

Principal components analysis (PCA) [1, 2] is widely used for dimensionality reduction in a variety of scientific fields including single-cell RNA sequencing (scRNA-seq) data analysis. PCA constructs new variables – i.e., principal components (PCs) – that are linear functions of the original features, successively maximize variance, and are uncorrelated with each other [3]. The first few PCs explain the most variance in the data and are used as a proxy for the original data set in exploratory analysis. By compressing the input data into a low-dimensional space, PCA removes much of the random high-dimensional noise and reduces the size of the data. This improves the resolution of interesting signal (e.g., subpopulations or trajectories in scRNA-seq data) as well as the efficiency of downstream procedures like clustering and visualization.

More specifically: consider an input scRNA-seq matrix $\mathbf{Y}$ where the rows are the samples (i.e., cells) and the columns are the features (i.e., genes). Assume that the matrix is column-centered, i.e., the mean of each column is zero. We perform an SVD to obtain the usual $\mathbf{Y} = \mathbf{UDV}'$. Each column of $\mathbf{UD}$ is a PC containing scores for all cells. From the Eckart-Young theorem, the best k-rank approximation of $\mathbf{Y}$ is

$$\tilde{\mathbf{Y}}_k = \mathbf{U}_{(,k)}\mathbf{D}_{(k,k)}\mathbf{V}'_{(,k)} \ ,$$

where $\mathbf{M}_{(i,j)}$ contains only the first i rows and j columns of the matrix $\mathbf{M}$. We also note that the Euclidean distances between cells using their PC coordinates are the same as the distances computed from $\tilde{\mathbf{Y}}_k$. This motivates the direct use of first k PCs in downstream procedures involving distance calculations (e.g., hierarchical clustering, nearest-neighbour searches), which reduces computational work compared to using $\tilde{\mathbf{Y}}_k$.

The question now becomes: what is the best choice of k? If too few PCs are used, we may discard some relevant biological signal that is only captured in later PCs. On the other hand, if too many PCs are used, we introduce unnecessary noise and offset the efficiency benefits of having a low-dimensional data set. Deciding how many PCs to retain is a long-standing topic of study [4, 5], and depends on whether the aim is to recover the true rank of the signal matrix [5]; to minimize error compared to the true signal [6]; or to obtain components with a straightforward scientific interpretation [7].

In this report, we explore a number of computational methods for choosing the number of PCs to retain in scRNA-seq data. We use a variety of simulations to evaluate different methods in terms of their ability to remove noise and recover the true biological signal. We demonstrate that the strategies used in existing scRNA-seq analysis software are suboptimal but robust and largely effective.

2 A brief review of some PC selection methods

2.1 Elbow detection

The scree plot displays the percentage of variance explained by successive PCs. (We will denote the variance explained by PC k as σ_k^2, where $\sigma_k^2 > \sigma_{k+1}^2$.) Here, the aim is to visually detect the difference between early PCs that capture structure and later PCs that capture random noise. A sharp drop from the k^{th} to $(k+1)^{\text{th}}$ PC suggests that most of the structure is captured in the first k PCs [8]. This manifests as an "elbow" in the scree plot at k, after which there is a plateau in the percentage of variance explained. To detect the elbow, we consider a line connecting the first and n^{th} PCs, using $n = 50$ by default. The elbow is defined as the point on the curve in the scree plot that has the largest perpendicular distance to the line. We then retain all PCs up to the elbow point (but not including) the elbow point. This algorithm is relatively robust compared to derivative-based methods that are sensitive to unstable numerical differentiation.

2.2 Parallel analysis

Horn's parallel analysis [9] involves permuting all observations for each feature of the input matrix and performing PCA on this permuted matrix. This yields a variance explained ω_k^2 for the k^{th} PC. Any PC with a σ_k^2 (from the original input matrix) comparable to ω_k^2 is considered to be uninteresting, as the PC explains no more variance than expected under a random model containing no structure. One can visualize this strategy by considering a scree plot and discarding all PCs past the first intersection of the curves for the variances explained from the original and permuted PCAs.

Several definitions of "comparable" can be used to define the first uninteresting PC. The simplest is to remove all PCs past and including the first PC where $\sigma_k^2 < \omega_k^2$. However, this retains too many PCs in noisy datasets where the original and permuted variance-explained curves are similar and intersect slowly. Another solution is to repeat the permutations many times, and define the threshold as an upper quantile of ω_k^2 for each k [10]. A PC with σ_k^2 below this quantile is considered to be feasibly sampled from the distribution of ω_k^2 under a random model, and is subsequently discarded. We use the 95^{th} percentile, which yields a more conservative estimate of the number of PCs.

2.3 The Marchenko-Pastur law

The Marchenko-Pastur law [11] specifies the asymptotic distribution of singular values for a large random matrix with independent identically distributed (i.i.d.) entries. This distribution has a strict upper bound that could be used as a threshold on the number of PCs to retain [12]. To justify this, we assume that our input matrix can be written as $\mathbf{Y} = \mathbf{T} + \mathbf{X}$, where $\mathbf{T}$ is a r-rank matrix of true signal and $\mathbf{X}$ is a matrix of i.i.d. noise. We further assume that the first r rotation vectors are exactly linear combinations of the r basis vectors of $\mathbf{T}$, i.e., the additional noise due to $\mathbf{X}$ does not affect the identification of the true basis vectors. This means that the difference between $\mathbf{Y}$ and the low-rank approximation derived from the first r PCs will be equal to $\mathbf{X}$. Thus, we should take all PCs with singular values greater than the Marchenko-Pastur limit for $\mathbf{X}$.

In practice, the limit needs to be adjusted to account for the magnitude of the noise. This is most simply achieved by scaling the limit by the standard deviation of the noise, which is itself estimated by modelling technical variation in the space of the input data [13]. An alternative approach is to fit the Marchenko-Pastur distribution to the observed set of singular values for the input matrix [12], adjusting for scaling differences between the observed and theoretical distributions. However, this requires a complete

singular value decomposition (SVD) of the input matrix. This is time-consuming as it precludes the use of approximate SVD algorithms, so we will not consider it here.

2.4 The Gavish-Donoho method

Gavish and Donoho [6] describe a method for determining the optimal threshold for the singular values. They consider a matrix containing both true signal and i.i.d. noise with a constant standard deviation. By using only the first k PCs that have singular values above the optimal threshold, we can minimize the mean-squared error (MSE) of the low-rank approximation of the input matrix to the true signal. The Gavish-Donoho threshold needs to be scaled by an estimate of the standard deviation of the noise, which can again be obtained by modelling technical noise in the input matrix.

2.5 Jackstraw

The *Seurat* package for single-cell RNA-seq data analysis [14] uses the jackstraw procedure [15] for determining whether features are significantly associated with individual PCs. Briefly, observations for each feature in a randomly chosen subset of features are permuted. The PCA is performed on the modified matrix and a measure of association is computed for each PC with the permuted observations for the features in the chosen subset. This is repeated over several iterations to obtain a null distribution of association measures. The association measure is also computed for each feature and PC using the original data. A p-value is subsequently obtained for each gene and PC by comparing its original association measure with that of the null distribution.

In practice, this approach yields a feature-by-PC matrix of p-values that needs to be consolidated into a decision regarding the number of PCs to retain. We do so by combining the p-values for each PC using Simes' method [16]. This yields a combined p-value representing the evidence against the global null hypothesis for that PC, i.e., that it is not significantly associated with any features. We then retain PCs up to the first PC that accepts the global null at a error threshold of 0.05. We note that a variety of other strategies can be used to consolidate multiple p-values into a single decision. However, Simes' method is appealing as it is robust to correlations between features [17].

2.6 Summation of technical components

The `denoisePCA` function in the *scran* package [13] defines the number of PCs to retain based on the total technical noise in the input data. Technical noise is expected to be random and uncorrelated across genes, and thus should be mostly present in the later PCs. The number of PCs to retain is chosen as the smallest value l such that

$$\sum_{k=l+1}^{N} \sigma_k^2 \leq \omega_t^2 \,,$$

where N is the total number of PCs and ω_t^2 is the sum of technical components for all features, estimated as previously described. This discards later PCs until the discarded variance is equal to the total technical variance in the data. The value of l is a lower bound on the number of PCs required to capture the biological signal. Equality to this lower bound occurs when the signal is wholly captured by the first few PCs, though in practice, l will be lower than the rank of the matrix containing the signal. Furthermore, we only use features where the total variance is greater than the estimated technical component. This ensures that there is some $l \in [1, N]$ that satisfies the above inequality.

3 Assessing accuracy with the MSE

We assume that $\mathbf{Y}$ is the sum of a signal matrix $\mathbf{T}$ of the same dimensions, containing the true signal (i.e., gene expression) in each entry of the matrix; and a noise matrix of the same dimensions, where each entry has an expectation of zero and is independently distributed from the other entries. Our aim is to choose k such that the Frobenius norm of $\tilde{\mathbf{Y}}_k - \mathbf{T}$ is minimized. In other words, we want to minimize the MSE of the low-rank approximation $\tilde{\mathbf{Y}}_k$ from the true signal $\mathbf{T}$. This is arguably the only relevant measure of performance when PCA is applied to remove noise. We do not intend to interpret the individual PCs, nor do we care about the true rank of $\mathbf{T}$. Moreover, recall that the distances between cells computed in the PC space are the same as the distances computed from $\tilde{\mathbf{Y}}_k$. Thus, minimizing the MSE of $\tilde{\mathbf{Y}}_k$ will also improve the accuracy of the PC-based distances with respect to the true distances computed from $\mathbf{T}$.

We can also view the choice of k as a compromise between bias and precision. At the largest value of k, there is no bias as $E(\tilde{\mathbf{Y}}_k) = E(\mathbf{Y}) = \mathbf{T}$, but precision is poor due to the presence of high-dimensional noise. As k decreases, noise is removed and precision improves. However, this comes at the cost of introducing more bias when PCs capturing aspects of $\mathbf{T}$ are discarded. The MSE is the sum of the bias $||E(\tilde{\mathbf{Y}}_k) - \mathbf{T}||_F^2$ and the variance $||\tilde{\mathbf{Y}}_k - E(\tilde{\mathbf{Y}}_k)||_F^2$, and provides a single overall measure of accuracy.

4 Evaluations with simple simulations

4.1 Simulation design

We considered a simple simulation involving S subpopulations of C cells with G genes. A proportion of genes P were chosen to drive biological heterogeneity. For each gene g in the chosen set, the mean expression μ_{gp} in subpopulation p was randomly sampled from a Normal$(0, s^2)$ distribution. For the remaining genes, the mean expression was set to zero for all subpopulations. Each cell was randomly assigned to one subpopulation. The observed expression for gene g in cell c assigned to p was defined as $\mu_{gp} + \epsilon_{gc}$ where $\epsilon_{gc} \sim$ Normal$(0, w_g^2)$ and represents the effect of technical noise. We tested all combinations of parameters for S from 5 to 20; s^2 from 0.2 to 1; P from 0.2 to 1; C from 200 to 5000; G from 1000 to 5000; and w_g^2 of 1 (i.e., constant variance) or sampled from a Gamma$(2, 2)$ or Gamma$(0.2, 0.2)$ distribution. We also repeated the simulations where, instead of assigning each cell to a single subpopulation to create clusters, we created trajectories by treating each cell as a linear combination of two randomly chosen subpopulations with mixing proportion sampled from a Uniform$(0, 1)$ distribution.

We evaluated each choice of the number of PCs by computing the MSE of the low-rank approximation $\tilde{\mathbf{Y}}_k$ from the $\mathbf{T}$ matrix containing the known signal for each cell. For each cell, the corresponding column of $\mathbf{T}$ was defined as the mean expression vector of the assigned subpopulation (i.e., μ_{gp}) for the cluster simulations. For trajectories, each column of $\mathbf{T}$ was defined as the known linear combination of the two chosen subpopulations for each cell. In each simulation scenario, we determined the optimal (i.e., lowest) MSE by testing all possible values of k. We then reported the results for each method in terms of fold-increases from the optimal MSE. For the technical summation method, we estimated ω_t^2 as $||\mathbf{Y} - \mathbf{T}||_F^2$. (In practice, $\mathbf{T}$ is neither known nor necessary for estimating the technical components, but we have used it here for simplicity.) For the Gavish-Donoho and Marchenko-Pastur methods, the average variance of the noise was defined as ω_t^2/G, assuming i.i.d. noise across genes.

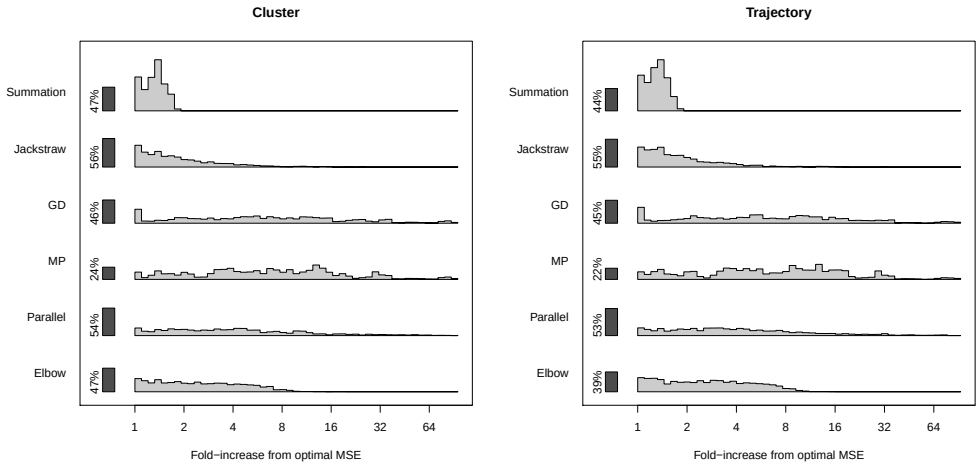


Figure 1. Distributions of MSEs for the number of PCs chosen by each method in the cluster (left) or trajectory simulations (right). MSEs were obtained from 10 simulation iterations for each combination of parameters, and are shown as fold increases from the optimal MSE for each iteration. For each method, the dark grey bar on the left represents the percentage of iterations in which the observed MSE was equal to the optimal value, while the light grey histogram shows the distribution of all MSEs that were greater than the optimum. GV, Gavish-Donoho; MP, Marchenko-Pastur.

4.2 Simulation results

Across all simulation scenarios, technical summation performs consistently well with MSEs close to or at the optimum (Figure 1). In fact, we see that the MSE from summation is never greater than twice the optimum in any scenario. This motivates its general use for scRNA-seq data as part of the *scran* analysis pipeline. On the other hand, technical summation returns the optimum MSE less frequently than some of the other methods. This reflects the fact that summation consistently yields a choice of k below the true rank of $\mathbf{T}$ (denoted r), which likely contributes to a systematic shift from the optimal MSE in some scenarios where the true rank is the optimal choice. (Of course, there is no guarantee that r is the optimal choice in all scenarios. Nonetheless, it is instructive to consider the ability of methods to recover r when it is optimal.)

The performance of the Gavish-Donoho method warrants some closer inspection. When the variance of the noise w_g^2 is constant, this method is almost perfect with respect to minimizing the MSE (Figure 2). This is consistent with the theoretical guarantees on its performance [6]. When w_g^2 varies across genes, the MSE increases above that of other methods. This reflects the sensitivity of the Gavish-Donoho method to violations of the assumption of a constant variance for the noise, which is unfortunate as (sc)RNA-seq data often exhibits strong mean-variance relationships [13, 18]. It is possible to mitigate violations of this assumption by scaling the input matrix so that the technical components are the same across all genes. However, scaling would also distort the magnitudes of the biological components, e.g., by upweighting stably expressed high-abundance genes with small technical components. This changes the nature of the problem and obviously precludes the accurate recovery of $\mathbf{T}$.

The use of the Marchenko-Pastur limit also exhibits poor performance, which can be traced back to its underying assumptions. Like the Gavish-Donoho method, the application of the Marchenko-Pastur law requires the assumption of i.i.d. noise. Violations of this assumption lead to increased MSE when the limit is used to choose number of PCs (Figure 2). However, even when the i.i.d. assumption holds, we observe

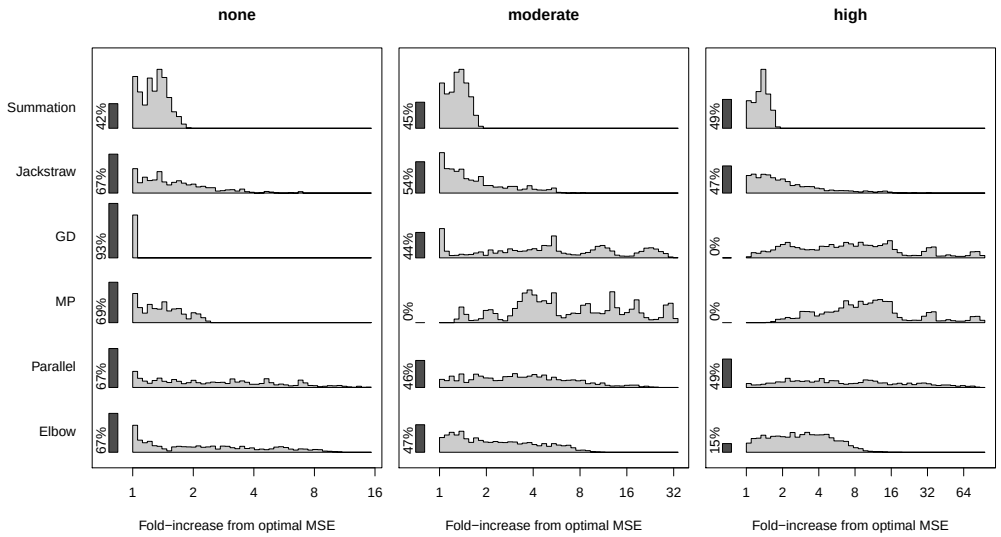
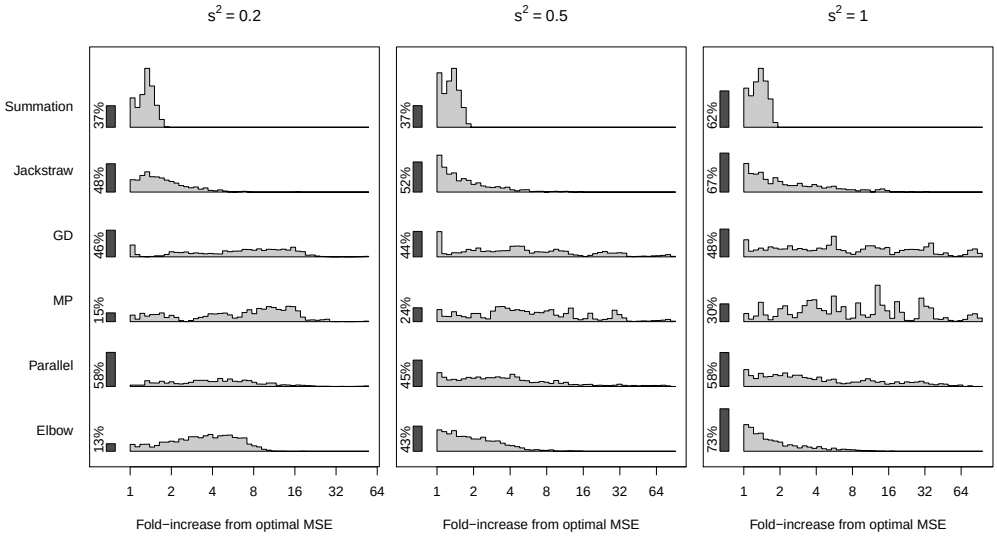


Figure 2. Distributions of MSEs for the number of PCs chosen by each method in simulations with different levels of heteroskedasticity. MSEs were obtained from 10 simulation iterations for each combination of parameters with no variability in gene-wise variances, i.e., $w_g^2 = 1$; moderate variability, i.e., $w_g^2 \sim \text{Gamma}(2, 2)$; or high variability, i.e., $w_g^2 \sim \text{Gamma}(0.2, 0.2)$. Bars and abbreviations are as described in Figure 1.

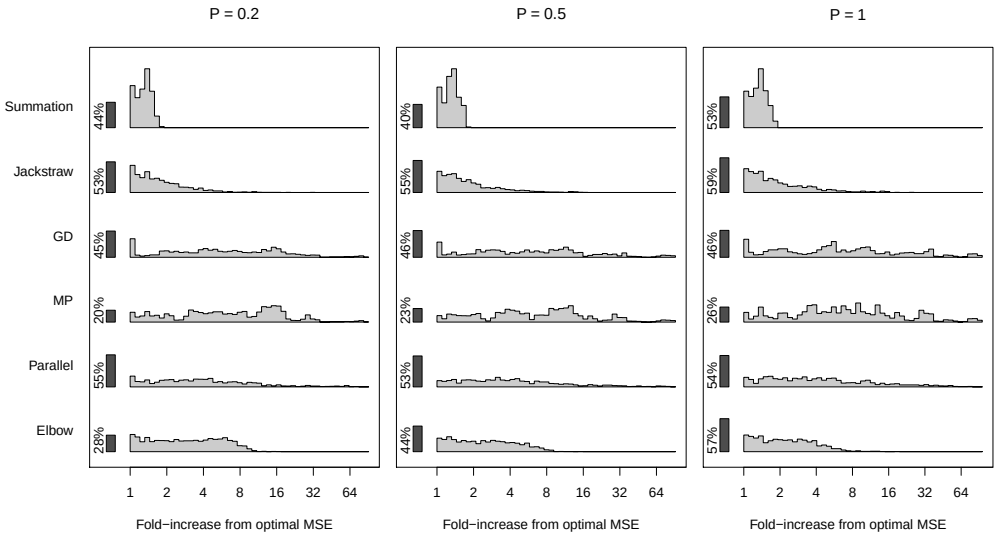
inflated MSEs compared to the optimum. This is driven by the fact that the r basis vectors of $\mathbf{T}$ do not define the same subspace as the first r rotation vectors in noisy datasets. Instead, some of the variance from the true signal will be captured by the $> r$ rotation vectors. This increases the later singular values above the Marchenko-Pastur limit and generally results in an overstatement of the number of PCs to retain.

Of the methods that do not require an estimate of the technical noise, the jackstraw performs slightly better than parallel analysis. However, both have a long tail of large MSEs, mostly caused by the retention of more PCs than the optimal choice. This is attributable to the same effect described for the Marchenko-Pastur method. Recall that the jackstraw tests for associations between gene expression and each PC. In noisy data sets, genuine signal is captured by later PCs and results in significant associations for more than r PCs. The same effect applies to parallel analysis where the variance explained by later PCs is increased above the permutations. The jackstraw is also subject to random type I errors that introduce further variation in k across iterations.

Elbow detection from the scree plot has large MSEs when the true signal is weak relative to the noise (Figure 3). This occurs when the subpopulation means are weakly separated, either due to small differences in μ_{gp} or low P. In such cases, the scree plot will not exhibit any sharp curvature as noise will contribute strongly to the variance explained by all PCs. Thus, any detected elbow is unlikely to have a strong relation to the true rank of the signal. As the signal increases in strength, the curvature in the scree plot becomes more pronounced and the performance of elbow detection improves. By comparison, the behaviour of each of the other methods is mostly stable with respect to the strength of the biological signal. This is because they are less reliant on making distinctions between PCs based on the proportions of variance explained.

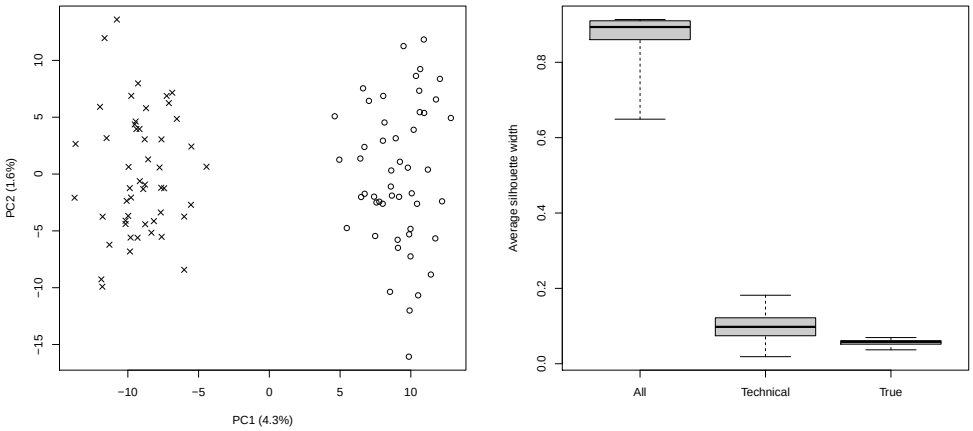


(a)



(b)

Figure 3. Distributions of MSEs for the number of PCs chosen by each method in simulations with different levels of separation between subpoplations. MSEs for each method were obtained from each of 10 simulation iterations for each combination of parameters with (a) weak separation ($s^2 = 0.2$), moderate separation ($s^2 = 0.5$) or or strong separation ($s^2 = 1$) between subpopulation means; or (b) a low proportion of genes ($P = 0.2$), a moderate proportion ($P = 0.5$) or all genes ($P = 1$) involved in separating subpopulations. Bars and abbreviations are as described in Figure 1.



(a) **(b)**

Figure 4. Effect of removing biological noise in a simulation with two subpopulations. (a) PCA plot of the simulation design involving 1000 genes and 100 cells. Cells were split into two subpopulations, denoted by the point shape. Subpopulation means were set to 0 or 0.5 for all genes in each population. Biological noise was introduced by adding an independent standard Normal variate to each observation. Technical noise was also introduced by adding another independent standard Normal variate to each observation. (b) Average silhouette widths for each cell, computed using the known subpopulation labels. Distances were calculated using the true expression for each cell (subpopulation mean with added biological noise); or using the first few PCs, with k chosen by the Gavish-Donoho method to remove all random noise or only the technical noise.

4.3 Comments on technical noise

In these simulations, we have only considered the removal of technical noise in ϵ_{gc}. This is because the technical component of variation is relatively easy to estimate in real scRNA-seq data. We can add a constant quantity of spike-in transcripts to each cell, and estimate the technical noise from the variation in the coverage of the spike-ins across cells. Alternatively, for data involving unique molecular identifiers [19], we could assume that the noise is driven by Poisson sampling. This is justified by the removal of PCR amplification biases and associated overdispersion in the counts. However, there also exists biological noise caused by transcriptional kinetics [20] and cellular processes like the cell cycle [21]. Choosing k based on ω_t^2 will not remove such biological noise.

We argue that biological noise is part of the true signal $\mathbf{T}$ and should not be removed. We are only interested in removing the technical noise introduced by the scRNA-seq protocol as this is always irrelevant to the underlying biology. The remaining heterogeneity should be preserved as it reflects the true state of the cell population. Even random biological noise can be critical to the interpretation of the results, e.g., when studying fate commitment decisions [22]. Its removal may exaggerate the separation between subpopulations (Figure 4), possibly promoting the identification of spurious cell types. It is also difficult to estimate the magnitude of biological noise, and to disentangle random noise from structured variation related to cell cycle or metabolic activity. Deciding which aspects of biological noise are "uninteresting" depends heavily on domain expertise and is beyond the scope of this report.

Table 1. Number of PCs retained by each method on real scRNA-seq data, set to a maximum of 50 for computational efficiency (capped choices denoted by *).

Method	mESC	PBMC
Summation	24	10
Jackstraw	4	8
GD	6	50*
MP	50*	50*
Parallel	9	10
Elbow	4	4

5 Testing with real data-based simulations

We generated log-normalized expression values from real scRNA-seq data by following a published workflow [13]. We first removed low-quality cells with low total counts, low total numbers of detected features and high spike-in proportions [23]. We computed size factors for each gene using the deconvolution method [24], and separate size factors for each spike-in transcript [25]. We divided the counts by the relevant size factors to obtain log-transformed normalized expression values. We estimated the technical component of variance of the transformed data for each gene by fitting a mean-dependent trend to the spike-in variances. We performed feature selection by only retaining endogenous genes with total variance greater than the technical component. If spike-in transcripts were not available, we instead used mitochondrial proportions for quality control, and we estimated the technical component for each gene by assuming Poisson noise.

Given a log-normalized expression matrix $\mathbf{Z}$, we generated a s-rank approximation $\tilde{\mathbf{Z}}_s$ as described above. For the purpose of simulating data, we treated $\tilde{\mathbf{Z}}_s$ as the matrix of true signal. We then created the observed matrix $\mathbf{Y}'$ by computing $\tilde{\mathbf{Z}}_s + \mathbf{S}$, where $\mathbf{S}$ is a matrix of the same dimensions as $\mathbf{Z}$. Each row of $\mathbf{S}$ consists of values independently sampled from a Normal distribution with mean zero and variance equal to the estimated technical component for the corresponding gene. This strategy allows us to incorporate aspects of real biological structure and technical noise into our simulations. We then evaluated each method by computing the MSE for the chosen k with respect to $\tilde{\mathbf{Z}}_s$. As the true rank of $E(\mathbf{Z})$ is not known, we repeated the simulations using s from 10 to 30.

We used our framework to generate simulated data from a mouse embryonic stem cell (mESC) data set [26] and a droplet-based data set using peripheral blood mononuclear cells (PBMCs) [27]. For the mESC data set, we only considered the single batch in which spike-ins were added, involving approximately 200 cells. For the PBMC data set, we considered approximately 4000 cells after quality control. Methods for choosing k were then evaluated as described for the simple simulations. Parallel analysis and technical summation performed consistently well in both simulation scenarios (Figure 5). In comparison, other methods exhibited much larger MSEs in one or both simulations. These results are roughly consistent with those from the simple simulations. The summation approach is often suboptimal but consistently achieves low error, while other methods can be more frequently optimal but yield larger maximum MSEs.

We also examined the behaviour of each method on the original data $\mathbf{Z}$. It is difficult to perceive any consistent pattern, though in general, methods that require the magnitude of technical noise (i.e., summation, Gavish-Donoho, Marchenko-Pastur) retain more PCs than the others (Table 1). This is attributable to the fact that random biological noise is retained, requiring more PCs to account for the greater intrinsic dimensionality. The other methods do not distinguish between technical and biological noise and aim to remove both, resulting in retention of fewer PCs upon denoising.

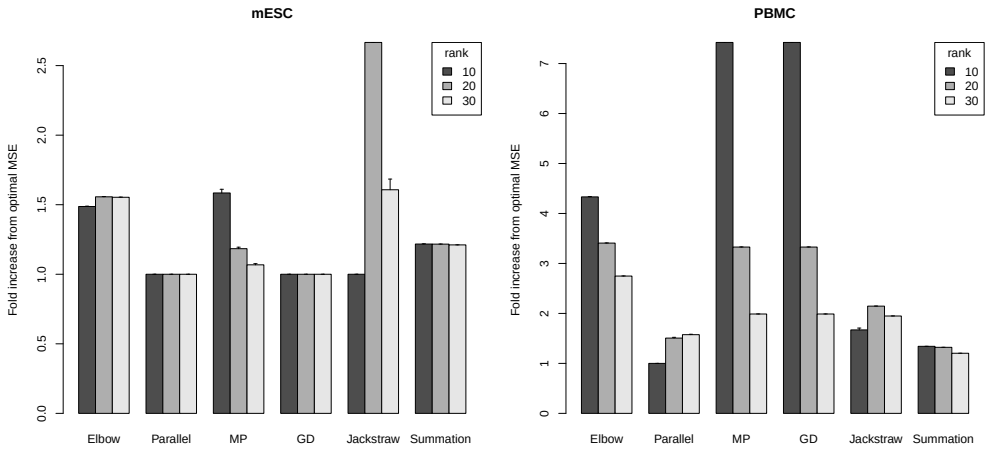


Figure 5. Average MSE for the number of PCs chosen by each method in the simulations based on the mESC and PBMC data sets, shown as fold-increases from the optimal MSE. The height of each bar represents the average fold-increase in the MSE from 10 simulation iterations at each rank s. Error bars represent one standard error.

6 Discussion

Here, we have performed a variety of simulations to assess different strategies for choosing the number of PCs to retain in scRNA-seq data analysis. The technical summation method from *scran* consistently yields low MSEs in all scenarios. While it is not always optimal, summation does not yield large MSEs either, with no MSE being more than 2-fold greater than the optimum. This is despite the fact that it does not guarantee minimization of the MSE or even recovery of the rank of $\mathbf{T}$. Other methods such as the jackstraw approach from *Seurat* achieve the optimal MSE more frequently than summation but also exhibit larger MSEs. Our results suggest that the summation approach is both satisfactory and robust for general use in a wide range of scenarios.

It is also worth considering the practicalities involved in the use of each method. The Gavish-Donoho, Marchenko-Pastur and technical summation methods require an estimate of the standard deviation of the noise. In scRNA-seq experiments, this is most conventionally achieved by adding spike-in transcripts and estimating the technical component of the variation in the expression values. However, this may not always be possible (e.g., droplet-based experiments) in which case assumptions need to be made about the nature of the noise. For example, we might assume that noise is driven by Poisson sampling in data sets with unique molecular identifiers, where overdispersion due to amplification noise is not an issue. The jackstraw method and parallel analysis do not require the magnitude of the noise but instead repeat the PCA on permuted data. This avoids the need for spike-ins or assumptions but involves more computational work, which is unappealing when dealing with large data sets. Elbow detection in the scree plot is the simplest method as it does not use the noise magnitude or permutations.

In this report, we have only discussed the retention of first k PCs. This is because the first k components provide the best low-rank approximation $\tilde{\mathbf{Y}}_{(k)}$ to $\mathbf{Y}$. However, we are primarily interested in the approximation to the true signal matrix $\mathbf{T}$, for which an arbitrary subset of PCs may be better. For example, the first PC could be driven by one very noisy gene with equal expected expression in all samples – ignoring this PC would improve the approximation of the remaining PCs to $\mathbf{T}$. One could imagine formulating a method that achieves even lower MSEs than the optimum in our simulations by retaining specific PCs based on per-gene estimates of the noise.

7 Additional information

Simulation code is available at `https://github.com/LTLA/PCSelection2018`. All simulations were performed using R version 3.5.0 and *scran* version 1.9.18, *Seurat* version 2.3.4 and *RMTstat* version 0.3. Count tables for real scRNA-seq data were obtained from `https://www.ebi.ac.uk/teichmann-srv/espresso/` for the mESC data set and from `https://support.10xgenomics.com/single-cell-gene-expression/datasets` for the PBMC data set. This work was supported by Cancer Research UK (award no. A17197 to Dr. John Marioni).

References

1. K. Pearson. 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.

2. H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24(6):417, 1933.

3. I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*, 374(2065):20150202, Apr 2016.

4. K. I. Howard and R. A. Gordon. Empirical note on the "number of factors" problem in factor analysis. *Psychol. Rep.*, 12(1):247–250E, 1963.

5. W. R. Zwick and W. F. Velicer. Comparison of five rules for determining the number of components to retain. *Psychol. Bull.*, 99(3):432, 1986.

6. M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inf. Theory*, 60(8):5040–5053, 2014.

7. S. B. Franklin, D. J. Gibson, P. A. Robertson, J. T. Pohlmann, and J. S. Fralish. Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, 6(1):99–106, 1995.

8. R. B. Cattell. The scree test for the number of factors. *Multivar. Behav. Res.*, 1(2):245–276, 1966.

9. J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, Jun 1965.

10. A. Buja and N. Eyuboglu. Remarks on Parallel Analysis. *Multivariate Behav Res*, 27(4):509–540, Oct 1992.

11. V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

12. K. Shekhar, S. W. Lapan, I. E. Whitney, N. M. Tran, E. Z. Macosko, M. Kowalczyk, X. Adiconis, J. Z. Levin, J. Nemesh, M. Goldman, S. A. McCarroll, C. L. Cepko, A. Regev, and J. R. Sanes. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5):1308–1323, Aug 2016.

13. A. T. Lun, D. J. McCarthy, and J. C. Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res*, 5:2122, 2016.

14. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36(5):411–420, Jun 2018.

15. N. C. Chung and J. D. Storey. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 31(4):545–554, Feb 2015.

16. R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

17. S. K. Sarkar and C.-K. Chang. The simes method for multiple hypothesis testing with positively dependent test statistics. *J. Am. Stat. Assoc.*, 92(440):1601–1608, 1997.

18. C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, 15(2):R29, Feb 2014.

19. S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lonnerberg, and S. Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2):163–166, Feb 2014.

20. J. K. Kim and J. C. Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.*, 14(1):R7, Jan 2013.

21. A. Scialdone, K. N. Natarajan, L. R. Saraiva, V. Proserpio, S. A. Teichmann, O. Stegle, J. C. Marioni, and F. Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, Sep 2015.

22. G. Balazsi, A. van Oudenaarden, and J. J. Collins. Cellular decision making and biological noise: from microbes to mammals. *Cell*, 144(6):910–925, Mar 2011.

23. D. J. McCarthy, K. R. Campbell, A. T. Lun, and Q. F. Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, 04 2017.

24. A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17:75, Apr 2016.

25. A. T. L. Lun, F. J. Calero-Nieto, L. Haim-Vilmovsky, B. Gottgens, and J. C. Marioni. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.*, 27(11):1795–1806, 11 2017.

26. A. A. Kolodziejczyk, J. K. Kim, J. C. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Buhler, P. Liu, J. C. Marioni, and S. A. Teichmann. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485, Oct 2015.

27. G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, 8:14049, Jan 2017.