Truncated Robust Principal Component Analysis and Noise Reduction for Single Cell RNA-seq Data

Krzysztof Gogolewski¹(✉), Maciej Sykulski²,³, Neo Christopher Chung¹, and Anna Gambin¹

¹ Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland
{kgogolewski,aniag}@mimuw.edu.pl, nchchung@gmail.com
² Department of Medical Genetics, Warsaw Medical University, Warsaw, Poland
macieksk@gmail.com
³ genXone Ltd., Poznań, Poland

Abstract. The development of single cell RNA sequencing (scRNA-seq) has enabled innovative approaches to investigating mRNA abundances. In our study, we are interested in extracting the systematic patterns of scRNA-seq data in an unsupervised manner, thus we have developed two extensions of robust principal component analysis (RPCA). First, we present a truncated version of RPCA (tRPCA), that is much faster and memory efficient. Second, we introduce a noise reduction in tRPCA with $L_2$ regularization (tRPCAL2). Unlike RPCA that only considers a low-rank $L$ and sparse $S$ matrices, the proposed method can also extract a noise $E$ matrix inherent in modern genomic data. We demonstrate its usefulness by applying our methods on the peripheral blood mononuclear cell (PBMC) scRNA-seq data. Particularly, the clustering of a low-rank $L$ matrix showcases better classification of unlabeled single cells. Overall, the proposed variants are well-suited for high-dimensional and noisy data that are routinely generated in genomics.

Keywords: Principal component analysis · Robust PCA
Truncated singular value decomposition · Matrix decomposition
Unsupervised learning · Single cell RNA-seq

1 Introduction

Single cell RNA sequencing (scRNA-seq) present new opportunities to elucidate systematic patterns of variation underlying biological processes and complex phenotypes. Conventionally, bulk RNA-seq data provide mean gene expression values from a large number of cells in that sample. However, a mixture of multiple cells that often have different functions or origins may hide relevant information, carry high variance related to their cellular composition, and might not be

K. Gogolewski and M. Sykulski—These authors equally contributed to this work.
reproducible in separate studies [1–3]. With scRNA-seq, we can overcome these challenges by measuring gene expression at a single cell resolution [4,5]. However, scRNA-seq data present new challenges for unsupervised learning methods because of unlabeled samples, higher dimensionality, dropouts, and sparsity.

Unsupervised learning techniques have become increasingly popular and useful for exploring and analyzing scRNA-seq data. In particular, principal component analysis (PCA) is most frequently used to reduce dimensions that enable a number of downstream statistical and machine learning [6]. Furthermore, closely related to factor analysis and latent variable models, principal components (PCs) help us to identify hidden and unmeasured structure that arise from biological and technical sources of variation [7–9]. Some of biological applications include tracking Definitive Endoderm Cells (DfE) to explain their lineage from Embryonic Stem Cells [10], classifying sensory neuron types [11], and identifying potentially damaged cells [12]. To account for an underlying sparse component (e.g., sparsely corrupted data or sparse latent structure), [13] proposed robust PCA (RPCA) that can decompose the observed data into low-rank and sparse components.

We build on the strength of RPCA [13] to introduce an computationally efficient truncated version and a noise reduction using $L_2$ regularization. In high dimensional genomic data, the systematic variation is likely contained in a small number of PCs, whereas lower-ranked PCs only contain noise. Therefore, our truncated RPCA (tRPCA) uses the top $k$ singular vectors to estimate low-rank and sparse components. Noise reduction of scRNA-seq data was possible by introducing an error component, in addition to low-rank and sparse components that were originally introduced in [13]. Advancements of matrix decomposition have a long history, including non-negative matrix factorization [14], sparse PCA [15], penalized matrix decomposition [16], and more. Inspired by these methods, our innovation enables separation of low-rank and sparse components, while imposing a $L_2$ penalty on a noise term inherent in large scale genomic data.

The paper is organized as follows. In Sect. 2 Methods, we present two proposed methods based on RPCA, namely its truncated version and noise reduction with $L_2$ regularization. We provide the algorithms and their characteristics. Section 3 contains the description and processing procedures for the scRNA-seq datasets used as the case study. In Sect. 4 Results and Discussion, we present the main results of our analysis, as well as provide some interpretations of low-rank, sparse, and noise components. Finally, in Sect. 5. Conclusions and Further Research, we summarize our work and discuss the future steps concerning the proposed methods.

2 Robust PCA and Extensions

**Robust PCA.** In our work we start from the decomposition algorithm proposed by Candes et al. [13] called robust PCA (RPCA). The aim of the RPCA is to decompose the input matrix $M$, into low-rank matrix $L$ and sparse matrix $S$
components. Simultaneously, the algorithm should minimize the following optimization problem:

$$\min_{L, S} \|L\|_* + \lambda_1 \|S\|_1, \text{ where } M = L + S$$

Here we denote $\|A\|_*$ as the nuclear norm of matrix $A$ and $\|A\|_1$ as the first norm of a vectorized $A$ matrix which are given by the following formulas:

$$\|A\|_* = \sum \sigma_i = \text{tr} \left( \sqrt{AA^T} \right), \text{ and } \|A\|_1 = \sum_{i,j} |a_{ij}|$$

In their work, authors discuss the assumptions that matrix $M$ should follow for the decomposition to exist. Moreover they prove that the parameter $\lambda_1$ can be set to $1/\sqrt{\min (m,n)}$, where $m, n$ are dimensions of the input matrix $M$, which guarantees proper decomposition into low-rank and sparse components as $m, n \to \infty$ under weak probabilistic assumptions, however, the spectrum of feasible values of $\lambda_1$ parameters is broad.

In order to solve the above problem, as proposed in [17], we use an implementation of a special case of the Alternating Directions method, which belongs to a more general class of augmented Lagrangian (AGL) multiplier algorithms. The approach is based on minimizing the following AGL operator with respect to $L$ and $S$ matrix alternately:

$$l(L, S, Y) = \|L\|_* + \lambda_1 \|S\|_1 + \langle Y, M - L - S \rangle + \frac{\mu}{2} \|M - L - S\|_F^2$$

where $Y$ is the Lagrange multiplier matrix, the inner product of matrices $\langle \cdot, \cdot \rangle$ is defined as the trace of their product, i.e. $\langle A, B \rangle = \text{tr}(AB^T)$, $\|A\|_F$ is the Forbenius norm of the form $\|A\|_F = \sqrt{\sum_{i,j} a_{i,j}^2}$ and $\mu$ is the penalty coefficient.

The outline of the solution is presented in the Algorithm 1, in which we use two shrinkage operators $S_\tau(x) = \text{sgn}(x) \cdot \max(|x| - \tau, 0)$ and $D_\tau = US_\tau(\Sigma)V^*$, where $USV^*$ is a SVD of $X$ and $\tau$ is the shrinkage threshold value. In case of initialization of the $\mu$ parameter and convergence condition, we set $\mu = \frac{m \cdot n}{4 \|M\|_1}$, as suggested in [17] and terminate the algorithm when $\|M - L - S\|_F \leq \delta \|M\|_F$ and $\delta = 10^{-7}$. The base implementation of the algorithm that we have extended in this work is publicly available as a R package [18] (https://cran.r-project.org/web/packages/rpca/).

**Truncated Version of RPCA.** First, we consider a truncated version of the algorithm, which calculates the $L$ matrix in the $L + S$ decomposition in such a way that it is of a given rank $k$ or the lowest possible rank greater than $k_0$, for which the problem has a solution that meet all its criteria. In order to achieve that behavior we use the truncated version of SVD (implementation from the irlba R package [19]) instead of a full SVD and iteratively modify the $\mu$ parameter with

$$\mu_{i+1}^{-1} = \max(c \cdot \mu_i^{-1}, \sigma_{k+1})$$
Algorithm 1. RPCA by Alternating Directions

1: procedure RPCA($\lambda$)
2:    $S_0, Y_0 \leftarrow 0; \mu > 0$
3: while not converged do
4:    compute $L_i^{+1} = D_{\mu^{-1}}(M - S_i + \mu^{-1} Y_i)$
5:    compute $S_{i+1} = S_{\lambda_1 \mu^{-1}}(M - L_{i+1} + \mu^{-1} Y_i)$
6:    compute $Y_{i+1} = Y_i + \mu \cdot (M - L_{i+1} - S_{i+1})$

Algorithm 2. truncated-RPCA

1: procedure tRPCA($\lambda$, $k$, $c$)
2:    $S_0, Y_0 = 0; \mu_0 > 0; k = k_0$
3: while not converged do
4:    compute $L_i^{+1} = D_{\mu_i^{-1}}(M - S_i + \mu_i^{-1} Y_i)$
5:    compute $S_{i+1} = S_{\lambda_i \mu_i^{-1}}(M - L_{i+1} + \mu_i^{-1} Y_i)$
6:    compute $Y_{i+1} = Y_i + \mu_i \cdot (M - L_{i+1} - S_{i+1})$
7:    compute $\mu_i^{-1} = \max(c \cdot \mu_i^{-1}, \sigma_{k+1})$
8: if $\mu_i^{-1} = \sigma_{k+1}$ then increase $k$

where $\sigma_k$ is the $k$-th singular value from the truncated SVD and $c < 1$ is the AGL constraints penalty growth rate.

As $\mu_i^{-1}$ decreases the penalty coefficient for $M = L + S$ constraints increases, which also speeds up the algorithm’s convergence, however, in theory, AGL algorithm converges to the constraint problem even when $\mu_i \not\to \infty$. Simultaneously, when $\mu_i^{-1}$ is set to the value of $\sigma_{k+1}$ we increase $k$, i.e. the number of computed SVD vectors, which is at same time the approximate expected rank of $L$ matrix.

The above modification (see Algorithm 2) reduces significantly the computation time of the algorithm compared to the original RPCA preserving its accuracy. However, in the case of real data (e.g. biomedical) the decomposition into low-rank and sparse matrices is not always feasible or easily obtainable. The data matrix usually has significant singular values (below the first $k$ low-rank important ones) that may come from biological activity, technical reasons, or other unknown sources. These prevent the recovery of low-rank component as when subtracted from $M$ they do not constitute a sparse matrix. We interpret these perturbations in the $L$ matrix as a noise or low-importance information. Since it does not have a sparse nature we extend the decomposition into $L + S + E$, where the matrix $E$ contains a dense noise controlled for using the $L_2$ norm on vectorized matrix $M$ (i.e. Frobenius norm).

**Noise Reduction.** In order to relax the assumptions on the input matrix we introduce the $E$ matrix to the decomposition. Now, the decomposition problem can be reformulated as follows:

$$M = L + S + E$$

$$\min_{L,S,E} ||L||_* + \lambda_1 ||S||_1 + \lambda_2 ||E||_F$$
The $E$ matrix is meant to contain the information of low importance or noise, which is carried by the lowest singular values in the SVD of $L$ matrix. To solve this problem we extend the Alternating Directions approach and we minimize the following AGL operator with respect to the $E$ matrix:

$$l(L, S, E, Y) = ||L||_* + \lambda_1||S||_1 + \lambda_2 ||E||_F$$

$$+ \langle Y, M - L - S - E \rangle + \frac{\mu}{2} ||M - L - S - E||_F^2$$

Solving $\partial l / \partial E = 0$ results in:

$$E \left( \frac{\lambda_2}{||E||_F} + \mu \right) = Y + \mu(M - L - S)$$

Let $C = Y + \mu(M - L - S)$, then $\exists d \in \mathbb{R} E = d \cdot C$. Assuming that $C \neq 0$ we determine the value of $d$. Since $d < 0$ results in a contradiction, we assume that $d \geq 0$ we have:

$$d = \frac{||C||_F - \lambda_2}{\mu ||C||_F} = \frac{1}{\mu} \left( 1 - \frac{\lambda_2}{||C||_F} \right) \geq 0$$

which holds for $||C||_F \geq \lambda_2$. We define the operator:

$$\mathcal{E}_\tau(X) = \max \left( 0, 1 - \frac{\tau}{||X||_F} \right) \cdot X$$

which describes how to determine the matrix $E$ which minimizes $l$.

Finally, we extend the algorithm of truncated-PCA by applying the defined operator $\mathcal{E}_\tau$. In our approach we apply the operator twice, both, after minimization with respect to the $L$, and $S$ matrix. It is worth to emphasize, that in the case of large $\lambda_2 > ||C||_F$ we end up with the previously introduced truncated-RPCA procedure. Moreover, in every iteration we adjust $k$ parameter to be of minimal value such that $D_{\mu^{-1}}$ operator can be properly applied. Algorithm 3 presents the pseudo-code of the whole decomposition procedure.

### 3 Single Cell Transcriptomic Data

In this study we use the publicly available scRNA-seq datasets provided by the 10x Genomics company (https://www.10xgenomics.com/solutions/single-cell/). Specifically, our results, that are presented in the next section, were obtained using the scRNA-seq datasets experiments performed on peripheral blood mononuclear cells (PBMCs) from a healthy donor. PBMCs are primary cells with relatively small amounts of RNA (1pg RNA/cell). The final dataset contains 2.7k individual single cells, sequenced on Illumina NextSeq 500 with approx. 69k reads per cell. Amplification was performed on 98bp read1 (transcript), 8bp I5 sample barcode, 14bp I7 GemCode barcode and 10bp read2 (UMI).
Algorithm 3. truncated-RPCA with L2 regularization

1: procedure tRPCAL2($\lambda_1, \lambda_2, k_0, c$)
2: $S_0, Y_0, E_0 = 0; \mu_0 > 0; k = k_0$
3: while not converged do
4: compute $L_{i+1} = D_{\mu-1}(M - S_i - E_i + \mu^{-1}Y_i)$
5: compute $E_{i+1} = E_{\lambda_2 \mu^{-1}}(M - S_i - L_{i+1} + \mu^{-1}Y_i)$
6: compute $S_{i+1} = S_{\lambda_1 \mu^{-1}}(M - E_{i+1} - L_{i+1} + \mu^{-1}Y_i)$
7: compute $E_{i+1} = E_{\lambda_2 \mu^{-1}}(M - S_{i+1} - L_{i+1} + \mu^{-1}Y_i)$
8: compute $Y_{i+1} = Y_i + \mu \cdot (M - E_{i+1} - L_{i+1} - S_{i+1})$
9: compute $\mu^{-1}_{i+1} = \max(c \cdot \mu^{-1}_i, \sigma_{k+1})$
10: if $\mu^{-1}_{i+1} == \sigma_{k+1}$ then increase $k$
11: else $k = 1 + \arg\max_j (\sigma_j > \mu^{-1}_{i+1})$

Along with the 2.7k PBMCs dataset, we have used the scRNA-seq data retrieved from homogeneous samples of specific cell types that constitute the PBMC sample. Each type-specific dataset has over 90% of purity for each subtype by Fluorescence Activated Cell Sorting (FACS) [20]. The transcriptomes were used in [21] and described the following types and subtypes: CD14$^+$ Monocytes, CD56$^+$ Natural Killer cells, CD19$^+$ B cells, CD34$^+$ cells and subfamilies of T cells: CD8$^+$ Cytotoxic T cells, CD8$^+$/CD45RO$^+$ Naive Cytotoxic T cells, CD4$^+$/CD45RA$^+$ Memory T cells, CD4$^+$ Helper T cells.

Each of the above datasets is given in the form of a counts matrix $A$ i.e. $i$-th row represents a gene and $j$-th column represents an individual cell. The value of $a_{ij}$ is the number of counts of the $i$-th gene for the $j$-th cell. Since our method is meant to filter out the sparse signal in $S$ and the dense noise in $E$ we do not apply the typical quality control step. All cells are used in the analysis and we expect all perturbations that break the linear behavior (e.g. biological or technical outliers or fluctuations) to remain in $S + E$ component of the decomposition. Additionally, for each dataset we filter out genes that had zero number of counts for all cells in a given set. Finally, the number of counts for each cell was normalized by its total number of counts and log-scaled. Further on we denote the processed 2.7k PBMCs data matrix by $M$.

Test Set Construction. In order to test our method we first set the labeling of cells from the PBMCs dataset. For each available type-specific dataset we calculate its average transcriptome. However, since the correlation between averaged subtype-specific transcriptomes within T-cell family is relatively high, for the purpose of this work, we label the cells with one of the five possible types: (i) Monocytes, (ii) Natural Killers, (iii) B cells, (iv) T cells, (v) Unknown. T cells family transcriptome is designated as an average among all T cells subtypes transcriptomes.

The criterion for labeling consists of two conditions. First, a cell is assumed to be of an unknown type if it does not correlate with any of the given profiles at least at the level of 0.5. Next, the cell is assumed to be of a specific type
if the separation between its correlation and correlations with other types is statistically significant (p-value < 0.05) otherwise it is assumed to be unknown.

Even though there are no transcriptomes available for other cell types, e.g. Megakaryocytes we are aware that they may also exist in our dataset and thus expect to find them using our decomposition method.

4 Results and Discussion

By definition our final extension of RPCA explains the input matrix data \( (M) \) in terms of compressed, low-rank information \( (L) \), sparse signal \( (S) \) and noise \( (E) \).

In order to validate our method on real data and evaluate its suitability for biomedical data analysis, we investigate the scRNA-seq 2.7k PBMCs data. We report, that the trRPCAL2 algorithm converged after 49 iterations, exec. time: 97 s (compared to 20 s PCA from R prcomp). As expected, due to the high background variance trRPCA and RPCA did not converge before 1000 iterations.

Clustering via Low-Rank Matrix. First we validate the quality of the dimension reduction by clustering cells basing on their low-rank representation kept in the \( L \) matrix. Using the unsupervised, hierarchical clustering algorithm we determined 5 clusters, which we visualized using t-SNE approach [22] (see Fig. 1). In contrast with expected cell types (derived from correlation with

![Fig. 1. Clustering of 2.7k PBMCs. In both panels, cells are visualized using t-SNE dimension reduction algorithm (perplexity = 35) run on the 10-dimensional representation of the original data \( (M) \) derived from \( L \) matrix. (a) Colors correspond to the cell types inferred from the correlation of each cell original transcriptome (columns of \( M \)) with type-specific PBMCs transcriptomes. We have determined: 630 Monocytes (orange), 251 B-cells (pink), 437 Natural Killer cells (blue) and 700 T-cells (yellow). Remaining 682 (gray) we assume to be of an unknown or tentative type. (b) Colors correspond to 5 clusters determined by hierarchical clustering method. Colors of the clusters are set such that the correspondence between predicted and original clusters is noticeable. Additionally, increased activity of CD8 in the gray cluster suggests that it can be mostly composed of cytotoxic T-cells. (Color figure online)
type-characteristic transcriptomes) we observe that the obtained clustering well determines all 4 main families of cells from PBMCs. Additionally, one more cluster separating NK and T cell family clusters was discovered. The cluster is described by increased activity of CD8A and CD8B (Bonferroni adjusted p-value $< 10^{-3}$) and regular activity of CD4, CD45 and CD25 genes in contrast to other cells, which suggests a cluster of mostly CD8$^+$ T cytotoxic cells and explains its similarity to NK cells [21,23].

Next, we compared our way of dimension reduction with the method analogous to the one used in [21]. With SVD we calculate top 10 singular values (in pursuance of the $L$ matrix rank) of the PBMC data matrix (M) using R irlba package. Then, we approximate the original data through the reduced 10-dimensional space. We perform the hierarchical clustering of all cells on the most characteristic marker genes per cell type (selected from the literature) from the described and the $L$ matrices. The aim is to verify how well the dimensionality reduction preserved the most reliable, biological information related to type-specific marker genes. It appeared that not only the $L$ matrix guarantees more accurate clustering, but also it contains more pronounced differences of the signal between clusters of both cells and genes, c.f. Fig. 2.

![Fig. 2. Marker gene based clustering comparison.](image-url)
**Monocyte Subtypes and Co-expression Detection.** The literature suggests existance of at least three subtypes of monocytes in PBMCs [24]. Their characterization can be based on the presence of CD14 (coded by CD14 gene) and CD16 (coded by FCGR3A, FCGR3B genes) clusters of differentiation: (i) the classical monocyte with high activity of CD14 ($CD14^{++}FCGR3A^{-}$); (ii) the intermediate monocyte with high activity of CD14 and low activity of FCGR3A ($CD14^{++}FCGR3A^{+}$); (iii) the non-classical monocyte with low activity of CD14 and co-expressed FCGR3A ($CD14^{+}FCGR3A^{++}$).

Interestingly, such classification of subtypes can be found using the low rank signal from the $L$ matrix (see Fig. 3). The activity of CD14 is almost uniquely distributed among cluster of monocyte cells and, simultaneously, the activity of FCGR3A changes with the gradient defining the cell subtype progression among all monocytes. Moreover, the Fig. 3c shows how the original expression values are distributed among decomposition matrices. The sparse peaks of activity are stored in $S$ and the linear part in $L$. Finally, $E$ contains remaining noise of mean 0 and the standard deviation of order $10^{-4}$ for both CD14 and FCGR3A.

---

**Fig. 3.** CD14 and FCGR3A activity levels. Panels present the activity of monocytes marker genes. (a) and (b) Figures present the activity of CD14 and FCGR3A genes among all cells, respectively. The level of activity (low to high) is spanned on the red to green color scale. (c) Consecutive panels present: (i) the log-transformed data from $M$; (ii) low-rank signal in $L$ matrix; (iii) sparse signal in $S$ matrix. In each panel cells (x-axis) are sorted by the activity level of CD14 (y-axis). (Color figure online)
The other interesting property of the $L$ matrix is the fact that it recovers co-expression patterns between genes. Namely, the activity of B cells can be detected e.g. by the presence of CD79 heterodimer composed of CD79A and CD79B proteins [25]. Their co-expression measured in terms of correlation was at the level of 0.227, while after the decomposition their low-rank signal had correlation of level 0.995. Similarly, the correlation between FCGR3A and GNLY characterizing Natural Killer cells increased from 0.400 to 0.949. Naturally, these observations result from filtering out the sparse and noise signals. Nonetheless, it is worth to emphasize that this type of information is retrieved by the method, because it can help in suggesting new co-expression patterns.

Sparse Signal Interpretation. The presence of megakaryocytes in our PBMC dataset, that was reported in the population of PBMCs sample from [21], was not evident using the low-rank $L$ matrix. Even though, a small cluster of cells of unknown type was separated by t-SNE method (see Fig. 1) it was not straightforward conclusion from clustering results. However, with the hierarchical clustering performed on the subset of unknown type cells and genes that had at least one non-zero entry form the sparse $S$ matrix we recover well-separated cluster of 9 cells. Further analysis confirmed that the cluster is characterized by high over-expression of PF4 gene, which is a well known marker for mature megakaryocytes [26], in comparison to other unknown cell types.

Noise Reduction Level. Finally, we want to discuss briefly the importance of the noise matrix $E$ and setting of both $\lambda_1$ and $\lambda_2$ parameters. The final decomposition quality in terms of information distribution among three matrices is mainly based on the choice of these crucial parameters. For the purpose of this study, we performed a grid-based search through the parameter space which resulted in $\lambda_1 = 0.016$, $\lambda_2 = 10.0$ and consequently $L + S + E$ decomposition with the following norms of the (vectorized) matrices: $||·||_*: 5.753, 60.289, 57.881$; $||·||_1: 3398.162, 60.289, 2670.012$; $||·||_2: 4.265, 2.826, 1.440$ ($L, S, E$ respectively).

To determine the order of magnitude for both parameters we have made use of the theory described in [13] as well as estimations based on the properties of the $MM^T$ matrix trace operator. Because tRPCAL2 algorithm mixes $L_1$ and $L_2$ norms, and because of their nature, final decomposition depends not only on relative or absolute values of lambdas, but also on distributions of elements in the decomposed matrices. We investigated the influence of $\lambda_1$ and $\lambda_2$ on decomposition properties such as: the rank of the resulting $L$ matrix, relative and absolute sparsity of the $S$ matrix and the variance of the noise level in $E$ matrix. We report few observations on PBMCs data: (i) relative sparsity of $S$ (i.e. proportion between sparsity of $S$ and original matrices) decreases exponentially with respect to $\lambda_1/\lambda_2$; the rank of matrix $L$ increases; (ii) sub-linearly as a function of $\lambda_1$ and fixed $\lambda_2$; (iii) polynomially as a function of $\lambda_2$ and fixed $\lambda_1$.

Finally, in theory it seems intuitive to expect elements of the $E$ matrix to be normally distributed with 0 mean. However, the computational experiments showed, that this distribution is a mixture of the zero-centered Gaussian and another low-variance Gaussian concentrated around non-negative,
\( \lambda_2 \)-dependent value. We assume that this is a linear dependency, nonetheless its precise description is still unexplored.

More formal investigation of theoretical properties of the \texttt{tRPCAL2} decomposition with respect to lambdas and \( L, S, E \) matrices could be of high interest in terms of future research.

5 Conclusions and Further Research

Concluding, in this paper we introduce an extension of the robust PCA matrix decomposition method. We propose a \( L + S + E \) decomposition of the matrix into: low-rank \( L \), sparse \( S \) and noise \( E \) components. Thanks to the reduction of \( L_2 \) moderated noise we restore the inner structure of the matrix, which approximates the original data with high accuracy, as well as recognize the sparse perturbation signal of the data. We present the case study based on the scRNA-seq data from 2.7k PBMCs. The method provides relatively fast and accurate dimension reduction and clustering of the high-dimensional data detecting different subtypes within a given cell type, co-expression patterns and novel subtypes.

One possible direction for the further research is to derive precise formulas for \( \lambda_1 \) and \( \lambda_2 \) parameters that guarantee optimal solutions of the decomposition problem. So far, simulation-based selection of the parameters is time consuming. Ideally, a lambda parameter selection method would result with the most natural \( L + S + E \) decomposition, taking into account user’s expectations in terms of, for example, Bayesian priors to relative magnitudes, and to other components’ statistics. This and the applicability of our method to other types of data we see as a promising direction of research. Preliminary results of video and image analysis, not described in this paper, suggest that the method can be successfully harnessed in the field of video-surveillance and image analysis. The most recent implementation of the \texttt{tRPCAL2} algorithm as an R package is available under https://github.com/macieksk/rpca.

Acknowledgments. This work was supported by the Polish National Science Centre grant no. 2016/21/N/ST6/01507 and no. 2016/23/D/ST6/03613. The authors thank B. Miasojedow, Ph.D. for comments and suggestions.

References