Sparse reduced-rank regression for exploratory visualization of multimodal data sets

Dmitry Kobak[1,✉], Yves Bernaerts[1,2], Marissa A. Weis[1], Federico Scala[3], Andreas Tolias[3], and Philipp Berens[1,4,✉]

[1] *Institute for Ophthalmic Research, University of Tübingen, Germany*
[2] *International Max Planck Research School for Intelligent Systems, Germany*
[3] *Department of Neuroscience, Baylor College of Medicine, Houston, Texas, USA*
[4] *Department of Computer Science, University of Tübingen, Germany*

name.surname@uni-tuebingen.de

August 9, 2019

Abstract

In genomics, transcriptomics, and related biological fields (collectively known as 'omics'), it is common to work with $n \ll p$ data sets with the dimensionality much larger than the sample size. In recent years, combinations of experimental techniques began to yield multiple sets of features for the same set of biological replicates. One example is Patch-seq, a method combining single-cell RNA sequencing with electrophysiological recordings from the same cells. Here we present a framework based on sparse reduced-rank regression for obtaining an interpretable visualization of the relationship between the transcriptomic and the electrophysiological data. We use an elastic net regularization penalty that yields sparse solutions and allows for an efficient computational implementation. Using several publicly available Patch-seq data sets, we show that sparse reduced-rank regression outperforms both sparse full-rank regression and non-sparse reduced-rank regression in terms of predictive performance, and can outperform existing methods for sparse partial least squares and sparse canonical correlation analysis in terms of out-of-sample correlations. We introduce a 'bibiplot' visualization in order to display the dominant factors determining the relationship between transcriptomic and electrophysiological properties of neurons. We believe that sparse reduced-rank regression can provide a valuable tool for the exploration and visualization of multimodal data sets, including Patch-seq.

1 Introduction

Since the days of Ramón y Cajal, neuroscientists have classified neurons into cell types, which are often considered the fundamental building blocks of neural circuits (Masland, 2004). Classically, these types have been defined based on their electrophysiology or anatomy, but due to the recent rise of single-cell transcriptomics, a definition of cell types based on genetics is becoming increasingly popular (Poulin et al., 2016). For example, single-cell RNA sequencing has been used to establish a census of neurons in the retina (Shekhar et al., 2016; Macosko et al., 2015), the cortex (Zeisel et al., 2015; Tasic et al., 2016, 2018), the whole brain (Saunders et al., 2018), and the entire nervous system (Zeisel et al., 2018) of mice. Despite this success, it has proven difficult to integrate the obtained cell type taxonomy based on the transcriptome with information about physiology and anatomy (Tripathy et al., 2017; Zeng and Sanes, 2017) and it remains unclear to what extent neural types are discrete or show continuous variation (Zeng and Sanes, 2017; Harris et al., 2018).

A recently developed technique called Patch-seq (Cadwell et al., 2016, 2017; Fuzik et al., 2016; Földy et al., 2016) allows to isolate and sequence RNA content of cells characterized electrophysiologically and/or morphologically (Figure 1a), opening the way to relate gene expression patterns to physiological characteristics on the single-cell level. Patch-seq experiments are laborious and low throughput, resulting

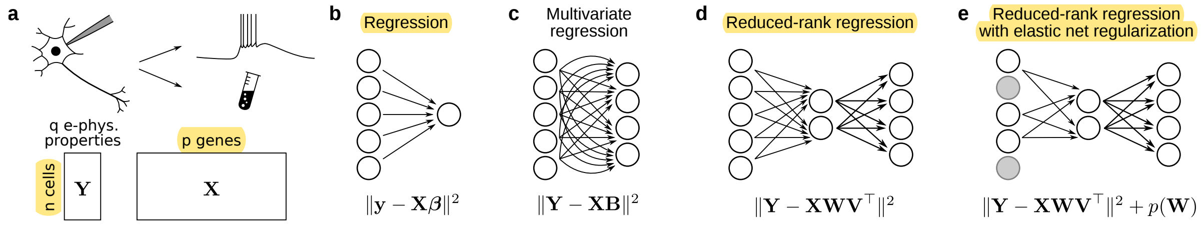
a
q e-phys. properties
p genes
n cells
Y
X
b
Regression
$\|\mathbf{y} - \mathbf{X}\beta\|^2$
c
Multivariate regression
$\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2$
d
Reduced-rank regression
$\|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{V}^\top\|^2$
e
Reduced-rank regression with elastic net regularization
$\|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{V}^\top\|^2 + p(\mathbf{W})$

Figure 1: a. Schematic illustration of a Patch-seq experiment: electrophysiological activity is recorded by patch-clamping, followed by RNA extraction and sequencing. Below: data matrices after computational characterization of electrophysiological properties ($\mathbf{Y}$) and estimation of gene counts ($\mathbf{X}$). **b–e.** Schematic illustrations and loss functions for several regression methods. **b.** Simple regression. **c.** Multivariate regression. **d.** Reduced-rank regression. **e.** Regularized reduced-rank regression. Gray circles denote predictors that are left out of the sparse model.

in multimodal data sets with a particular statistical structure: a few dozen or hundreds of cells are characterized with expression levels of many thousands of genes as well as dozens of electrophysiological measurements (Figure 1a). Integrating and properly visualizing genetic and physiological information in this $n \ll p$ regime requires specialized statistical techniques that could isolate a subset of relevant genes and exploit information about the relationships within both data modalities to increase statistical power.

Here we developed sparse reduced-rank regression based on the elastic net penalty to obtain an interpretable and intuitive visualization of the relationship between high-dimensional single-cell transcriptomes and electrophysiological information obtained using techniques like Patch-seq. We used three existing Patch-seq data sets (Fuzik et al., 2016; Cadwell et al., 2016; Scala et al., 2018) to demonstrate and validate our method. Our sparse RRR method extends sparse RRR of Chen and Huang (2012) and, as we show, outperforms it on our data.

Our code in Python is available at `https://github.com/berenslab/patch-seq-rrr`.

2 Results

2.1 Patch-seq data

A Patch-seq experiment yields two paired data matrices (Figure 1a): an $n \times p$ matrix $\mathbf{X}$ containing expression levels of p genes for each of the n cells, and an $n \times q$ matrix $\mathbf{Y}$ containing q electrophysiological properties of the same n cells. We assume that both matrices are centered, i.e. column means have been subtracted.

To illustrate the structure of such data sets and motivate the development of sparse RRR for exploratory visualization, we use principal component analysis (PCA) on the largest available Patch-seq data set (Scala et al., 2018). It contains $n = 102$ somatostatin-positive interneurons from layer 4 and layer 5 of primary visual and somatosensory cortex in mice (Figure 2). Each cell was described with $q = 13$ electrophysiological properties and we used $p = 1000$ 'most variable' genes that were selected in the original publication. Note that there are ∼25 thousand genes in the mouse genome, and ∼18 thousand were detected in at least one cell in this particular experiment; it is, however, a common practice to select a smaller set of genes for downstream analysis (Luecken and Theis, 2019), as most detected genes have zero counts for most of the cells and are likely not informative.

PCA in the transcriptomic space (Figure 2a) reveals that PC1, in this case, is an experimental artefact mostly driven by the variability in the expression depth between cells (correlation between PC1 and row sums of $\mathbf{X}$ was 0.73), whereas PC2 captures a biologically meaningful difference between cells in layer 4 and in layer 5. In contrast, PCA in the electrophysiological space (Figure 2b) separates cells primarily by the cortical area: cells from the somatosensory cortex tend to have narrower action potentials and higher firing rates than cells from the visual cortex. Thus, there appears to be no relationship between the leading PCs of the two modalities. The aim of reduced-rank regression is to uncover such relationships.

The visualisation in Figure 2b is known as a 'biplot' (Gabriel, 1971). Lines represent correlations

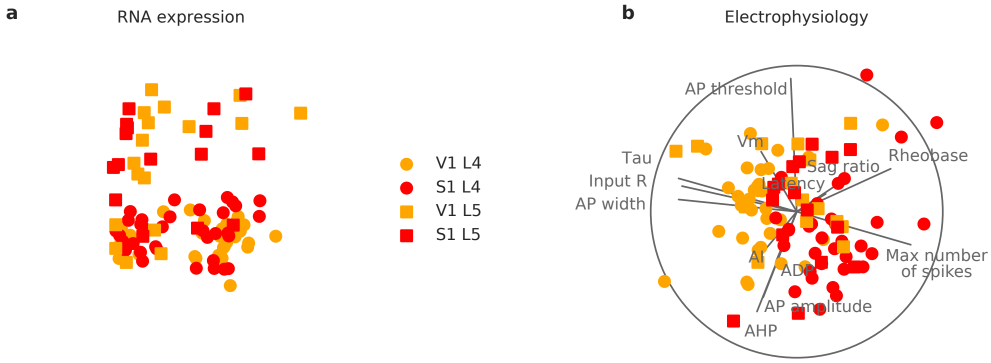

Figure 2: a. Principal component analysis (PCA) of the transcriptomic data in the Scala et al. data set (Scala et al., 2018). Color denotes cortical area (orange: visual cortex; red: somatosensory cortex), marker shape denotes cortical layer (circles: layer 4; squares: layer 5). Both PCs were standardized. **b.** PCA biplot of the electrophysiological data in the same data set. Grey lines show correlations of individual electrophysiological features with PC1 and PC2. The circle shows maximal possible correlations (sometimes called 'correlation circle'). The relative scaling of the scatter plot and the lines/circle is arbitrary.

between each electrophysiological property and PC1/PC2: the horizontal coordinate of each line's tip shows correlation with PC1 and the vertical coordinate shows correlation with PC2. The circle, sometimes called 'correlation circle'(citation ?), shows the maximum attainable correlation. The scaling between the scatter plot and the lines/circle is arbitrary. Following Gabriel (1971), we standardize both PCs and scale the lines/circle by an arbitrary factor of 3 (so that most points in the scatter plot are contained within the circle). We do not show a biplot in the transcriptomic space (Figure 2a) because the PCA in the gene space is not sparse, making the biplot practically impossible to display and interpret as it would have to show all 1000 genes from $\mathbf{X}$. This motivates the sparsity constraint that we impose on RRR.

2.2 Reduced-rank regression

To relate gene expression patterns to electrophysiological properties, one could use the transcriptomic data to predict any given electrophysiological property, e.g. action potential threshold. This is a *regression* problem: each gene is a predictor and action potential threshold is the response variable (Figure 1b). To predict multiple electrophysiological properties at the same time, one can combine individual regressions into a *multivariate regression* problem where the response is a multivariate vector (Figure 1c). The loss function of multivariate linear regression (known as ordinary least squares, OLS) is

$$\mathcal{L}_{\mathrm{OLS}} = \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 \tag{1}$$

and its well-known solution is given by

$$\hat{\mathbf{B}}_{\mathrm{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}. \tag{2}$$

Here and below all matrix norms are Frobenius norms.

Different electrophysiological properties tend to be strongly correlated and so one can construct a more parsimonious model where gene expression is predicting latent factors that in turn predict all the electrophysiological properties together (Figure 1d). These latent factors form a 'bottleneck' in the linear mapping and allow exploiting correlations between the predicted elecrophysiological properties to increase statistical power and decrease overfitting. This approach is called *reduced-rank regression* (RRR) (Izenman, 1975; Velu and Reinsel, 2013). Its loss function is

$$\mathcal{L}_{\mathrm{RRR}} = \|\underbrace{\mathbf{Y}}_{n*m} - \underbrace{\mathbf{X}}_{n*p}\underbrace{\mathbf{W}}_{p*r}\underbrace{\mathbf{V}^\top}_{r*m}\|^2, \tag{3}$$

m*r, m=1, then V is a vector of length r.

V is the reg coeff, like B.
XW is the latent factors.

where $\mathbf{W}$ and $\mathbf{V}$ each have r columns. Without loss of generality it is convenient to require that $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$. The product $\mathbf{W}\mathbf{V}^\top$ forms the matrix of regression coefficients that has rank r.

This decomposition allows to interpret $\mathbf{W}$ as a mapping that transforms $\mathbf{X}$ into r latent variables and $\mathbf{V}$ as a mapping that transforms the latent variables into $\mathbf{Y}$ (Figure 1e). As a result, RRR can be viewed not only as a prediction method, but also as a dimensionality reduction method, allowing visualization and exploration of the multimodal data set. Latent factors $\mathbf{XW}$ can be interpreted as capturing low-dimensional genetic variability that is predictive of electrophysiological variability, while $\mathbf{YV}$ can be interpreted as low-dimensional electrophysiological variability that can be predicted from the genetic variability.

RRR can be directly solved by applying singular vector decomposition (SVD) to the results of multivariate regression. Indeed, the RRR loss can be decomposed into the OLS loss and the low-rank loss:

$$\mathcal{L}_{\text{RRR}} = \|\mathbf{Y} - \mathbf{XWV}^\top\|^2 = \|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{\text{OLS}}\|^2 + \|\mathbf{X}\hat{\mathbf{B}}_{\text{OLS}} - \mathbf{XWV}^\top\|^2, \tag{4}$$

The first term corresponds to the variance of $\mathbf{Y}$ that is unexplainable by any linear model. The minimum of the second term can be obtained by computing the SVD of $\mathbf{X}\hat{\mathbf{B}}_{\text{OLS}}$. The right singular vectors corresponding to the r largest singular values give $\hat{\mathbf{V}}$, and $\hat{\mathbf{W}} = \hat{\mathbf{B}}_{\text{OLS}}\hat{\mathbf{V}}^\top$.

2.3 Reduced-rank regression with elastic net penalty

SRRR

As there are over 20 thousand genes in a mouse genome (with 1000–5000 typically retained for analysis) while the typical sample size of a Patch-seq data set is on the order of $n \approx 100$, the regression problems discussed above are in the $n \ll p$ regime and need to be regularized. Here we use elastic net regularization, which combines ℓ_1 (lasso) and ℓ_2 (ridge) penalties (Zou and Hastie, 2005). Elastic net enforces sparsity and performs feature selection: only a small subset of genes are selected into the model while all other genes get zero regression coefficients (Figure 1e). Our elastic net RRR extends a previously suggested sparse RRR (Chen and Huang, 2012) that used the lasso penalty on its own. The elastic net penalty has well-known advantages compared to the pure lasso penalty, e.g. it allows to select more than n predictors and can outperform lasso when predictors are strongly correlated (Zou and Hastie, 2005).

The loss function of our regularized RRR is:

$$\mathcal{L}_{\text{enRRR}} = \|\mathbf{Y} - \underset{\text{n*p}}{\mathbf{X}}\underset{\text{p*r}}{\mathbf{W}}\mathbf{V}^\top\|^2 + \lambda\Big(\alpha\sum_{i=1}^{p}\|\mathbf{W}_{i\cdot}\|_2 + (1-\alpha)\|\mathbf{W}\|^2\Big) \quad \text{s.t. } \mathbf{V}^\top\mathbf{V} = \mathbf{I}. \tag{5}$$

The penalties are only applied to the matrix $\mathbf{W}$ because $\mathbf{V}$ is constrained to have a fixed ℓ_2 norm, and applying ℓ_1 penalty to it would be inappropriate because we do not wish to make $\mathbf{V}$ sparse. We used the same parametrization of the penalty as in the popular `glmnet` library (Friedman et al., 2010): α controls the trade-off between the lasso ($\alpha = 1$) and the ridge ($\alpha = 0$) while λ controls the overall regularization strength. Following (Chen and Huang, 2012), the lasso penalty term $\sum_{i=1}^{p}\|\mathbf{W}_{i\cdot}\|_2$ group: r>1 $= \sum_{i=1}^{p}\sqrt{\sum_{j=1}^{r}W_{ij}^2}$ computes the sum of ℓ_2 norms of each row of $\mathbf{W}$. This is known as *group lasso* (Yuan and Lin, 2006) because it is the ℓ_1 norm of the vector of row ℓ_2 norms; it encourages the entire rows of $\mathbf{W}$, and not just its individual elements, to be zeroed out, corresponding to some of the genes being left out of the model entirely. See Discussion about this choice.

X = (x1, x2, ..., xp)
W=(w1, w2,..., wp)T
XW = x1*w1+...+xp*wp

x_i: n-by-1
w_i: 1-by-r

This optimization problem is biconvex and can be solved with an iterative 'alternating' approach: in turn, we fix $\mathbf{V}$ and find the optimal $\mathbf{W}_{\text{opt}}$ and then fix $\mathbf{W}$ and find the optimal $\mathbf{V}_{\text{opt}}$ until convergence. For a fixed $\mathbf{V}$, the least-squares term can be re-written as

$$\begin{aligned}\|\mathbf{Y} - \mathbf{XWV}^\top\|^2 &= \operatorname{tr}(\mathbf{Y}^\top\mathbf{Y}) + \operatorname{tr}(\mathbf{VW}^\top\mathbf{X}^\top\mathbf{XWV}^\top) - 2\operatorname{tr}(\mathbf{VW}^\top\mathbf{X}^\top\mathbf{Y}) \\ &= \text{const} + \operatorname{tr}(\mathbf{V}^\top\mathbf{Y}^\top\mathbf{YV}) + \operatorname{tr}(\mathbf{W}^\top\mathbf{X}^\top\mathbf{XW}) - 2\operatorname{tr}(\mathbf{W}^\top\mathbf{X}^\top\mathbf{YV}) \\ &= \text{const} + \|\mathbf{YV} - \mathbf{XW}\|^2,\end{aligned} \tag{6}$$

meaning that for a fixed $\mathbf{V}$, the loss is equivalent to

$$\mathcal{L}_{\text{enRRR}} \,|\, \mathbf{V} \sim \|\underset{\text{n*m m*r}}{\mathbf{YV}} - \underset{\text{n*p p*r}}{\mathbf{XW}}\|^2 + \lambda\Big(\alpha\sum_{i=1}^{p}\|\mathbf{W}_{i\cdot}\|_2 + (1-\alpha)\|\mathbf{W}\|^2\Big). \tag{7}$$

This is the loss of multivariate elastic net regression of $\mathbf{YV}$ on $\mathbf{X}$, and so the optimal $\mathbf{W}_{\text{opt}}$ can be obtained using the `glmnet` library (Friedman et al., 2010) (using `family="mgaussian"` option for row-wise lasso penalty) which has readily available interfaces for Matlab, Python, and R.

For a fixed $\mathbf{W}$, the loss does not depend on the penalty terms and the least-squares term can be written as

$$\begin{aligned} \|\mathbf{Y}-\mathbf{X}\mathbf{W}\mathbf{V}^\top\|^2 &= \|\mathbf{Y}\|^2 + \operatorname{tr}(\mathbf{V}\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{V}^\top) - 2\operatorname{tr}(\mathbf{Y}^\top\mathbf{X}\mathbf{W}\mathbf{V}^\top) \\ &= \text{const} - 2\operatorname{tr}(\mathbf{Y}^\top\mathbf{X}\mathbf{W}\mathbf{V}^\top). \end{aligned} \tag{8}$$

This is an example of the orthogonal Procrustes problem (Gower and Dijksterhuis, 2004). Maximizing $\operatorname{tr}(\mathbf{Y}^\top\mathbf{X}\mathbf{W}\mathbf{V}^\top)$ is achieved by the 'thin' SVD of $\mathbf{Y}^\top\mathbf{X}\mathbf{W}$. If the r left and right singular vectors are stacked in columns of $\mathbf{L}$ and $\mathbf{R}$ respectively (we order them by singular values, in decreasing order), then $\mathbf{V}_{\text{opt}} = \mathbf{L}\mathbf{R}^\top$. We provide a short proof in the Appendix.

Given that the loss function is biconvex but possibly not jointly convex in $\mathbf{V}$ and $\mathbf{W}$, it can be important to choose a reasonable initialization. We initialized $\mathbf{V}$ by the r leading right singular vectors of $\mathbf{X}^\top\mathbf{Y}$ and found this strategy to work well.

2.4 Relaxed elastic net

It has been argued that elastic net or even the lasso penalty on its own can lead to an over-shrinkage with non-zero coefficients shrinking 'too much' (Zou and Hastie, 2005). There have been several suggestions in the literature on how to mitigate this effect (Efron et al., 2004; Zou and Hastie, 2005; Meinshausen, 2007). *Relaxed lasso* (Meinshausen, 2007) performs lasso (setting $\alpha = 1$ and $\lambda = \lambda_1$) and then, using only the terms with non-zero coefficients, performs another lasso with a different penalty ($\alpha = 1$, $\lambda = \lambda_2$; usually $\lambda_2 < \lambda_1$). If $\lambda_2 = 0$, then this has also been called *LARS-OLS hybrid* (Efron et al., 2004).

Similar two-stage procedures for the elastic net penalty are not as established. We found that we obtain an improvement in predictive performance if after RRR with elastic net penalty with coefficients λ and α, we take the genes with non-zero coefficients and run RRR again using $\alpha = 0$ (i.e. pure ridge) and the same value of λ. This procedure does not introduce any additional tuning parameters but substantially outperforms pure elastic net RRR on our data, as we show below. We called it 'relaxed elastic net', following the 'relaxed lasso' terminology Meinshausen (2007). The solution of the first round of RRR we call 'naïve', following Zou and Hastie (2005).

A similar approach was suggested by De Mol et al. (2009) who performed elastic net using $\lambda = \lambda_1$ and some small fixed value of α, followed by pure ridge ($\alpha = 0$) regression using $\lambda = \lambda_2$ and only genes selected in the first stage. This approach also has two hyperparameters that need to be selected using cross-validation, but requires a manual choice of α for the first elastic net. If α is also treated as an adjustable hyperparameter, then it becomes a more flexible generalization of our approach. We found that the procedure of De Mol et al. (2009) performed similarly well to our relaxed elastic net for our data sets.

2.5 Cross-validation

We used cross-validation (CV) to select the values of r, λ, and α that maximize the predictive performance of the RRR model. The cross-validation estimates of R^2 are shown in Figure 3 for the Cadwell et al. and the Scala et al. data sets. We used 10 times repeated 11-fold CV for Cadwell et al. ($n = 44$) and 10 times repeated 10-fold CV for Scala et al. ($n = 102$). See Methods for the pre-processing details. The test-set R^2 for each fold was computed as

$$R^2 = 1 - \frac{\|\mathbf{Y}_{\text{test}} - \mathbf{X}_{\text{test}}\hat{\mathbf{W}}\hat{\mathbf{V}}^\top\|^2}{\|\mathbf{Y}_{\text{test}}\|^2}, \tag{9}$$

where $\mathbf{X}_{\text{test}}$ and $\mathbf{Y}_{\text{test}}$ were centered using the corresponding training-set means. We averaged the resulting R^2 across all folds and repetitions.

We found that $\alpha = 0.5$ outperformed other values on the Cadwell et al. data set (Figure 3a), suggesting that adding an additional ridge penalty to the sparse RRR model of Chen and Huang (2012) can be helpful. At the same time, $\alpha = 0.5$ and $\alpha = 1$ performed equally well for the Scala et al. data set. For simplicity and consistency, we always chose $\alpha = 0.5$ for the downstream analysis, and recommend it as a default setting.

The optimal λ corresponded to ~30 selected genes for the Cadwell et al. data set (Figure 3a) and to ~15 selected genes for the Scala et al. data set (Figure 3c), but the performance was comparably good

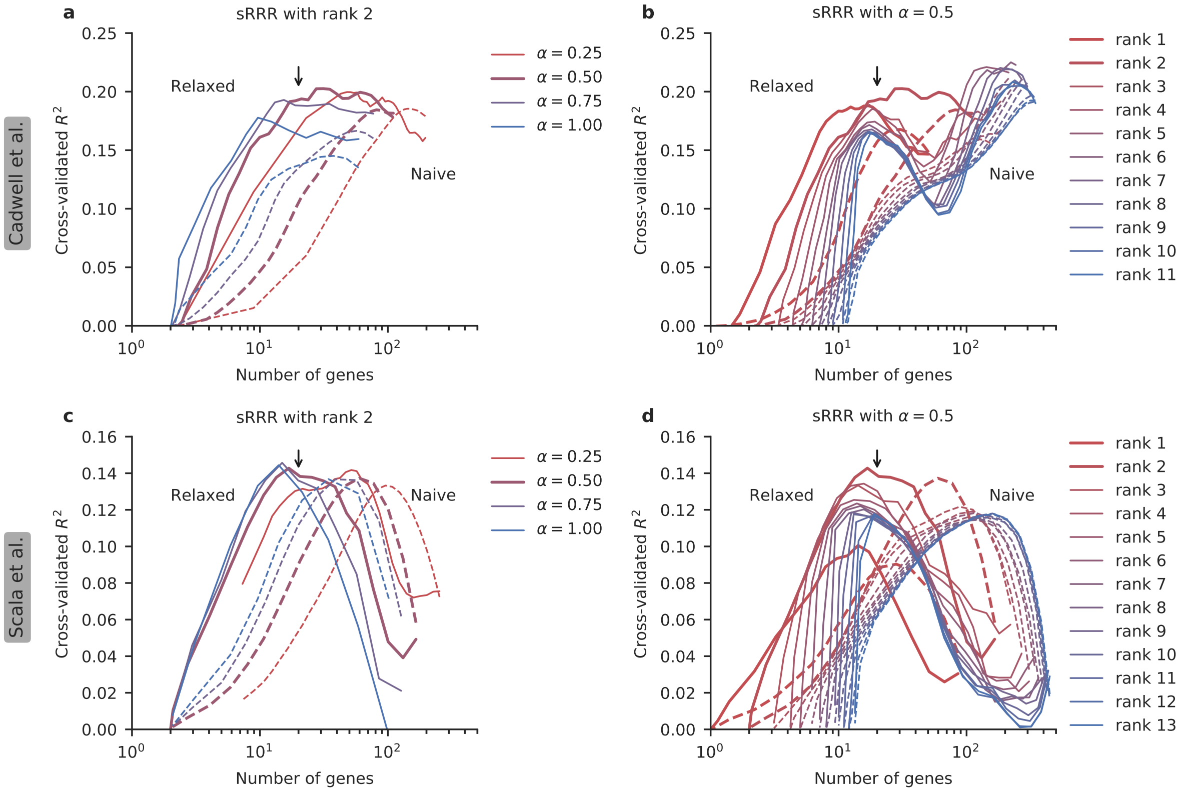

Figure 3: a. Cross-validation performance of sparse RRR with $r = 2$ in the Cadwell et al. data set, depending on α (color-coded, see legend) and λ. Horizontal axis shows the average number of selected genes obtained for each λ. Dashed lines: naive sparse RRR. Solid lines: relaxed sparse RRR. Black arrow points at our parameter choice ($\alpha = 0.5$ and λ yielding 20 selected genes). Thick lines highlight $\alpha = 0.5$. **b.** Cross-validatation performance with $\alpha = 0.5$ depending on the rank (color-coded, see legend). Thick lines highlight $r = 1$ and $r = 2$. **c–d.** The same for the Scala et al. data set.

in the range of ∼10–50 genes. For the downstream analysis, we chose the value of λ yielding 20 selected genes. Selecting many more genes than that makes visualisation difficult (see below).

The optimal value of rank was $r = 2$ in both data sets (Figure 3b,d). A lower rank $r = 1$ had worse performance due to under-fitting, especially for the Scala et al. data set, whereas higher ranks $r > 2$ led to a drop in performance due to over-fitting. Note that the full rank ($r = 11$ and $r = 13$ for the two data sets respectively) corresponds to the standard multivariate elastic net regression and we verified that our algorithm yields the same solution as `glmnet` does on its own. The much better performance of $r = 2$ shows that r can act as a regularization parameter, making sparse reduced-rank regression outperform sparse full-rank regression.

Finally, in both data sets the relaxed version of sparse RRR strongly outperformed the naive version. In particular, the same or even superior performance could be reached with many fewer genes (Figure 3a–d). If a high number of genes was selected in the model, the relaxed version performed worse than the naive version, suggesting that our 'relaxed' approach might be too simplistic; but for the low number of selected genes yielding optimal performance (10–50 genes) the relaxed version clearly had superior performance. Note that sparse RRR of Chen and Huang (2012) corresponds to the 'naive' version with $\alpha = 1$.

2.6 Bibiplot visualisation

We applied our sparse RRR approach with $r = 2$, $\alpha = 0.5$, and λ chosen to yield 20 selected genes to the Cadwell et al. (Figure 4a,b) and the Scala et al. (Figure 4c,d) data sets. For each of the data sets,

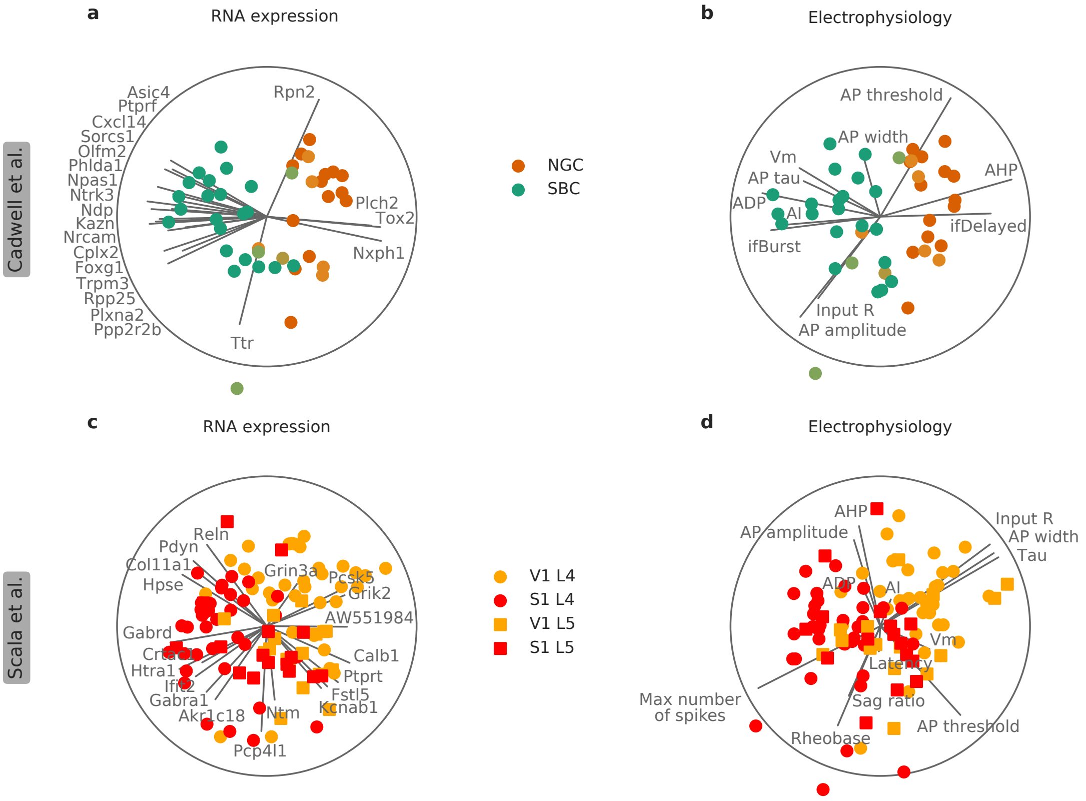

Figure 4: a. RRR biplot of the transcriptomic space in the Cadwell et al.(add citation since there are two paper 16' and 17) data set. Color codes cell type (orange: neurogliaform cells, NGC; green: single bouqet cells, SBC). Only the genes selected by the model are shown. See Figure 2b for the details of biplot visualization. **b.** RRR biplot of the electrophysiological space in the Cadwell et al. data set. **c–d.** The same for the Scala et al. data set. Color denotes cortical area (orange: visual cortex; red: somatosensory cortex), marker shape denotes cortical layer (circles: layer 4; squares: layer 5).

we visualized the results with a pair of biplots, a graphical technique that we suggest to call a 'bibiplot'.

To construct a biplot in the transcriptomic space, we use the bottleneck representation $\mathbf{XW}$ for the scatter plot, and show lines for all genes that are selected by the model (even though other genes can also have non-zero correlations with $\mathbf{XW}$). The biplot in the electrophysiological space is constructed using $\mathbf{YV}$ and shows all available electrophysiological properties. If R^2 of the model is high, then the two scatter plots will be similar to each other. Comparing the directions of variables between the two biplots can suggest which electrophysiological variables are associated with which genes.

The Cadwell et al. data set encompasses two types of interneurons from layer 1 of mouse cortex: neurogliaform cells (NGC) and single bouqet cells (SBC). Accordingly, the first RRR component captured the difference between the two cell types (Figure 4a,b). The second RRR component had only two genes associated with it (Figure 4a) and contributed only a very small increase in cross-validated R^2, as one can see comparing the cross-validation curves for $r = 1$ and $r = 2$ (Figure 3b). We conclude that the second RRR component in this data set is only weakly detectable.

In the Scala et al. data set (Figure 4c,d), the most salient feature in the bibiplot is the separation between the cells recorded in the visual and the somatosensory cortices. The selected genes here are pointing in all directions, and indeed the second component contributed a substantial increase in R^2 (Figure 3d). This suggests that both components are biologically meaningful. It is worth noting that the RRR biplot in the electrophysiological space (Figure 4d) was very similar to the PCA biplot (Figure 2b). This indicates that the sparse RRR model explained the dominant modes of variation among the dependent variables.

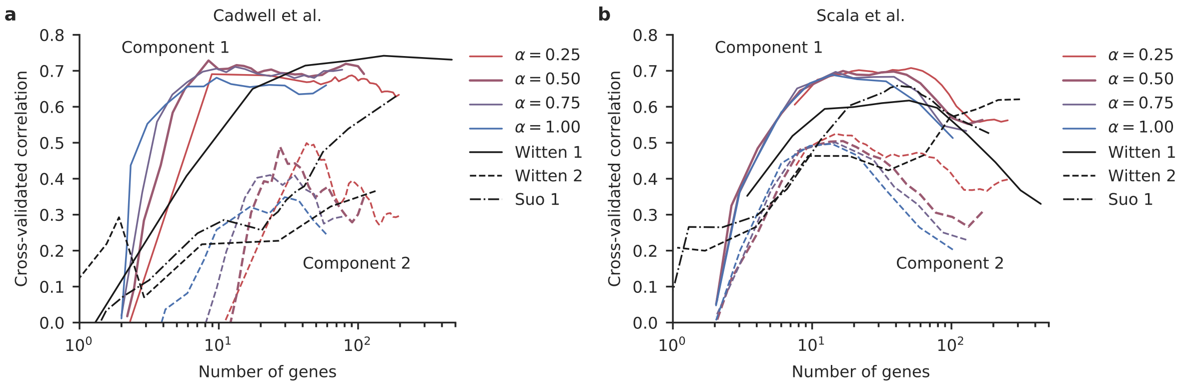

Figure 5: a. Cross-validation estimates of correlations between transcriptomic and electrophyiological RRR components with $r = 2$ in the Cadwell et al. data set, depending on α (color-coded, see legend) and λ. Horizontal axis shows the average number of selected genes obtained for each λ. Solid colored lines: RRR component 1. Dashed colored lines: RRR component 2. Solid and dashed black lines: sparse CCA method of Witten et al. (2009), components 1 and 2. Dash-dotted black line: sparse CCA method of Suo et al. (2017). **b.** The same for the Scala et al. data set.

2.7 Comparison to sparse CCA and PLS

Reduced-rank regression does not directly aim to maximize the correlation between $\mathbf{Xw}$ and $\mathbf{Yv}$, where $\mathbf{w}$ and $\mathbf{v}$ are corresponding columns of $\mathbf{W}$ and $\mathbf{V}$, even though high correlation is needed to achieve high R^2. Nevertheless, one can ask what is the cross-validated estimate of this correlation in the first and in the second pair of RRR components. We used the same cross-validation scheme to measure these out-of-sample correlations (Figure 5)[1]. With the hyper-parameters used above ($r = 2$, $\alpha = 0.5$, λ chosen to yield 20 genes), the correlations in the Cadwell et al. data set were 0.7 for component 1 and 0.4 for component 2 (Figure 3a). In the Scala et al. data set, they were 0.7 and 0.5, respectively (Figure 3b).

A statistical method that directly maximizes correlation between $\mathbf{Xw}$ and $\mathbf{Yv}$ is called 'canonical correlation analysis' (CCA). A number of different methods for sparse CCA have been suggested in the last decade (Waaijenborg et al., 2008; Wiesel et al., 2008; Parkhomenko et al., 2009; Witten et al., 2009; Witten and Tibshirani, 2009; Lykou and Whittaker, 2010; Hardoon and Shawe-Taylor, 2011; Chen et al., 2012; Chu et al., 2013; Wilms and Croux, 2015; Gao et al., 2017; Suo et al., 2017), of which the sparse CCA of Witten et al. (2009) is arguably the most well-known (judging by the number of citations in Google Scholar). We reimplemented the algorithm of Witten et al. and used the cross-validation procedure described above to measure its out-of-sample performance (Figure 5, solid and dashed black lines). We found that it performed noticeably worse than our sparse RRR: correlations for both data sets and both components (1st and 2nd) were either similar or lower than with sparse RRR, at least in the regime of 10–50 selected genes. We also implemented a recently suggested sparse CCA algorithm of Suo et al. (2017) that directly builds up on the Witten et al. approach. We found that the sparse CCA of Suo et al. performed worse than sparse CCA of Witten et al., in particular for the Cadwell et al. data set (Figure 5, dash-dotted black lines).

It remains beyond the scope of this paper to investigate why our sparse RRR is competitive with or even outperforms these sparse CCA methods in terms of out-of-sample correlations. One possible explanation is that Witten et al. and Suo et al. use regularization approaches that are suboptimal for our data sets. To understand this under-/over-regularization, note that RRR maximizes explained variance in $\mathbf{Y}$, i.e. correlation between $\mathbf{Xw}$ and $\mathbf{Yv}$, times the standard deviation of $\mathbf{Yv}$. Another related method is called 'partial least squares' (PLS): it maximizes the covariance between $\mathbf{Xw}$ and $\mathbf{Yv}$, i.e. correlation, times the standard deviation of $\mathbf{Yv}$, times the standard deviation of $\mathbf{Xw}$. In some sense, both RRR and PLS can be seen as particular regularized versions of CCA, because they bias $\mathbf{w}$ and $\mathbf{v}$ towards the high-variance directions in $\mathbf{X}$ and $\mathbf{Y}$, somewhat similar to the ridge penalty. The method of

[1] In some related previous work (González et al., 2008, 2009), cross-validated correlations were computed by pooling test set points across all cross-validation splits. We observed that this procedure can sometimes yield biased results; we compute test-set correlation within each test set, and then average across CV splits.

Witten et al. maximizes covariance (and so could in fact be called 'sparse PLS' and not 'sparse CCA'), which might provide too strong ℓ_2 regularization. The method of Suo et al. maximizes correlation and does not use any ridge penalty, which might cause too weak ℓ_2 regularization.

3 Discussion

We proposed sparse reduced-rank regression (RRR) as a tool for interpretable data exploration and visualization of Patch-seq recordings. It allows to visualize the variability across cells in transcriptomic and electrophysiological modalities in a consistent way, and to find a sparse set of genes explaining electrophysiological variability. We used cross-validation to tune the hyper-parameters and to estimate the out-of-sample performance of the model.

3.1 Comparison to other regression methods

Our method directly builds up on the sparse RRR of Chen and Huang (2012) who added the lasso penalty to the RRR loss function. We extended this approach by using an elastic net penalty that combines lasso and ridge regularization. This adds flexibility to the method and indeed we showed that in some cases non-zero ridge penalty is beneficial for the predictive performance (Figure 3). In addition, we introduced a 'relaxed elastic net' approach to mitigate the over-shrinkage bias associated with 'naive elastic net' or lasso solutions (Efron et al., 2004; Zou and Hastie, 2005; Meinshausen, 2007; De Mol et al., 2009). The sparse RRR method of Chen et al. (2012) corresponds to our 'naive' RRR with $\alpha = 1$, and in our experiments performed much worse than our 'relaxed' version (Figure 3).

On the other hand, sparse reduced-rank regression outperformed sparse full-rank regression (which is directly available e.g. in the popular `glmnet` library), suggesting that reduced-rank constraint is not only useful for visualisation but also provides additional regularization and reduces overfitting.

Elastic net regularization has two parameters, α and λ, and cross-validation sometimes indicates that α can be varied in some range without affecting the model performance (Figure 3). This allows the researcher to control the trade-off between a sparser solution and a more comprehensive gene selection. If there is a set of genes that are highly correlated among each other, then large α will tend to select only one of them, whereas small α will tend to assign similar weights to all of them. In the data sets analyzed here, we found that $\alpha = 0.5$ yielded a good compromise.

3.2 Comparison to other dimensionality reduction methods

Reduced rank-regression is closely related to two other classical dimensionality reduction methods analyzing two paired data matrices ('two views'): canonical correlation analysis (CCA) and partial least squares (PLS). They can be understood as looking for projections with maximal correlation (CCA) or maximal covariance (PLS) between $\mathbf{X}$ and $\mathbf{Y}$, whereas RRR looks for projections with maximal explained variance in $\mathbf{Y}$. In recent years, multiple approaches to sparse CCA (Waaijenborg et al., 2008; Wiesel et al., 2008; Parkhomenko et al., 2009; Witten et al., 2009; Witten and Tibshirani, 2009; Lykou and Whittaker, 2010; Hardoon and Shawe-Taylor, 2011; Chen et al., 2012; Chu et al., 2013; Wilms and Croux, 2015; Gao et al., 2017; Suo et al., 2017) and sparse PLS (Lê Cao et al., 2008, 2011; Chun and Keleş, 2010) have been suggested in the literature. Here, we chose sparse RRR at the core of our framework, because for the Patch-seq data it seems more meaningful to predict electrophyiological properties from transcriptomic information instead of treating them symmetrically, as genes give rise to physiological function. In addition, sparse RRR allows a mathematically simple formulation for rank $r > 1$ (using group lasso), whereas all sparse PLS/CCA methods cited above are iterative: after extracting the k-th component, matrices $\mathbf{X}$ and $\mathbf{Y}$ are 'deflated' and the algorithm is repeated to extract the $(k + 1)$-th component, resulting in a cumbersome procedure that is often difficult to analyze mathematically.

That said, by comparing our sparse RRR method with sparse PLS/CCA methods of Witten et al. (2009) and Suo et al. (2017), we unexpectedly found that our method is competitive as a CCA variant. This suggests that there is room for new sparse CCA methods that would be more appropriate at least for the kind of data sets studied here.

Sparse (and non-sparse) CCA and sparse PLS have been applied to biological data sets in order to integrate multi-omics data (Lê Cao et al., 2008, 2009; González et al., 2008, 2009, 2012), with the recent

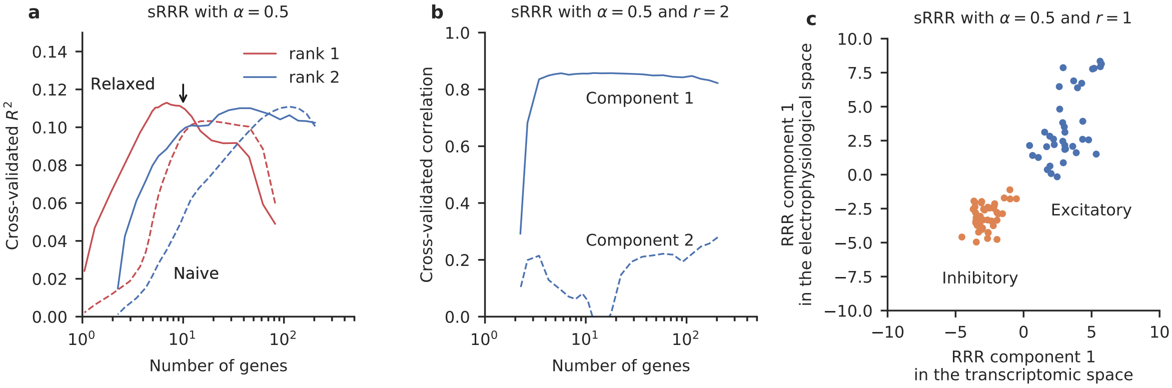

Figure 6: **a.** Cross-validation estimates of R^2 in sparse RRR with $r = 1$ and $r = 2$ and $\alpha = 0.5$ in the Fuzik et al. data set. Horizontal axis shows the average number of selected genes obtained for each λ. **b.** Cross-validation estimates of correlations between transcriptomic and electrophyiological RRR components with $r = 2$ and $\alpha = 0.5$. **c.** The RRR component using $r = 1$ and $\alpha = 0.5$ in the transcriptomic space (horizontal axis) and in the electrophysiological space (vertical axis). The value of λ was chosen to yield 10 select genes.

`mixOmics` package for R providing a convenient implementation for several of these methods (Rohart et al., 2017). We believe that our sparse RRR can be a useful addition to this array of multi-omics statistical techniques.

This series of multi-omics papers always used two separate plots for visualizing the CCA/PLS results within each modality: a ‘sample plot’, also called a ‘units plot’ (in our case this would be a scatter plot of Patch-seq cells), and a ‘variable plot’, also called a ‘correlation circle plot’ (in our case this would be a scatter plot of selected genes or electrophysiological properties, together with the correlation circle). We found it convenient to combine these two plots into a single biplot (Gabriel, 1971). This allows to use two biplots (what we called a ‘bibiplot’) instead of four separate plots.

3.3 RRR with $r \neq 2$

For both Cadwell et al. and Scala et al. data sets, cross-validation suggested $r = 2$ as the optimal rank, conveniently allowing us to use two-dimensional scatter plots for visualization. When higher-dimensional Patch-seq data sets become available, an optimal RRR model will likely need more than two components. In this case, one will have to resort to showing several biplots for different pairs of components, or alternatively to perform separate RRR analyses on the subsets of the data.

We observed an opposite case when we applied sparse RRR to the Patch-seq data set from Fuzik et al. (2016) with $n = 80$ inhibitory and excitatory neurons from mouse somatosensory cortex (Figure 6). The first RRR component strongly separated excitatory from inhibitory neurons, which is not surprising given the large differences in gene expression and in firing patterns between these two classes of neurons. However, subsequent RRR components did not seem to carry much signal in this data set. The RRR model with $r = 1$ and $\alpha = 0.5$ outperformed the model with $r = 2$ in terms of R^2 (Figure 6a), while the correlation for the 2nd component when using $r = 2$ was around zero (Figure 6b). This suggests effectively a one-dimensional shared subspace.

3.4 Limitations and outlook

Following Friedman et al. (2010) and Chen and Huang (2012), we used group lasso that induces row-wise sparsity in $\mathbf{W}$. This means that the same set of genes is selected for all RRR components. For $r = 2$, as used in this manuscript, the same set of genes influences the 1st and the 2nd component, which has both advantages and disadvantages. Our sparse RRR algorithm is easy to modify for the standard lasso case: using $\sum \|W_{i\cdot}\|_1$ in Eq. (5) instead of $\sum \|W_{i\cdot}\|_2$ would induce element-wise sparsity. In this case the loss in Eq. (7) can be minimized separately for each column of $\mathbf{W}$ (e.g. also using `glmnet`). Using this approach, different sets of selected can be selected for different RRR components and the same value of

λ can yield different number of selected genes for different components. However, when using 'relaxed' elastic net and performing RRR again, all components will get non-zero contributions from all selected genes. Further work would be needed to formulate a 'relaxed' version of the element-wise sparse RRR that would preserve element-wise sparsity. Empirically, we found that for the data sets considered here, the performance of the element-wise sparse RRR without relaxation was similar to the performance of the non-relaxed row-wise sparse RRR but worse than the performance of the relaxed row-wise sparse RRR.

One important caveat is that the list of selected genes should not be interpreted as definite. There are two reasons for that. First, the model performance (Figure 3) was unaffected in some range of parameters corresponding to selecting from $\sim$10 to $\sim$50 genes, meaning that the choice of regularization strength in this interval remains an analyst's call. Second, even for fixed regularization parameters, a somewhat different set of genes may be selected every time the experiment is repeated. This is a general feature of all sparse models, especially when $n \ll p$, known as 'instability' (Xu et al., 2011). Gene selection stability can be estimated with bootstrapping, and indeed we found that even the most reliably selected genes in the Cadwell et al. and Scala et al. datas sets were selected only $\sim$85% of times. There is an interplay between these two factors. Stronger ℓ_1 regularization leads to a sparser model with less selection reliability. Weaker ℓ_1 regularization leads to a less sparse model with higher selection reliability. We stress that such instability is an inherent feature of *all* sparse methods (Xu et al., 2011). We used bootstrapping to confirm that sparse CCA of Witten et al. had similar levels of selection instability.

In principle, it would be possible to generalize our regression framework to nonlinear mappings, using e.g. a neural network with a bottleneck instead of the low-rank linear mapping shown in Figure 1e. This can be an interesting direction for future research, but fitting such models would require much larger sample sizes than currently available for Patch-seq data.

In conclusion, we believe that sparse RRR can be a valuable tool for exploration and visualization of multimodal data sets. We expect that our method can be relevant beyond the scope of Patch-seq data. For example, spatial transcriptomics (Lein et al., 2017) combined with two-photon imaging may allow characterizing the transcriptome and physiology of individual cells in the intact tissue, yielding large multi-modal data sets. Similarly, other types of 'multi-omics' data where single-cell or bulk transcriptomic data are combined with some other type of measurements (e.g. chemical, medical, or even behavioural), may benefit from interpretable visualization techniques such as the one introduced here.

4 Methods

4.1 Data preprocessing

Cadwell et al. We used read counts table from the original publication (Cadwell et al., 2016). In this data set there are $n = 51$ interneurons (from 53 sequenced interneurons, 2 were excluded in the original publication as 'contaminated'), $p = 15\,074$ genes identified by the authors as 'detected', and $q = 11$ electrophysiological properties. We excluded all cells for which at least one electrophysiological property was not estimated, resulting in $n = 44$. We restricted the gene pool to the $p = 3000$ most variable genes, the same ones identified in the original publication. We used the expert classification of cells into two classes performed in the original publication for annotating cell types. Out of $n = 44$ cells, only 35 cells were classified unambiguously (score 1 or score 5 on the scale from 1 to 5); the remaining 9 cells received intermediate scores.

Fuzik et al. We used UMI counts table from the original publication (Fuzik et al., 2016). In this data set there are $n = 83$ cells, $p = 24\,378$ genes after excluding ERCC spike-ins, and $q = 89$ electrophysiological properties. Out of 83 sequenced cells, we were only able to match $n = 80$ to the electrophysiological data. We used only $q = 80$ electrophysiological properties for which the data were available for all these cells (the fact that $n = q = 80$ is coincidental). We selected $p = 1\,384$ genes with average expression above 0.5 (before standardization) for the RRR analysis.

Scala et al. We used read counts table from (Scala et al., 2018). This manuscript was prepared in parallel; the data are going to be publicly released after the publication of the original paper. In this data set there are $n = 110$ Patch-seq neurons; we used the same $n = 102$ as in the original publication (after excluding low quality cells). We use the same $p = 1000$ genes selected in the original publication, and the same $q = 13$ electrophysiological properties.

Preprocessing For the full-length data sets (Cadwell et al. and Scala et al.) we performed sequencing depth normalization by converting the counts to counter per million (CPM). For the UMI-based data set (Fuzik et al.) we divided the values for each cell by the cell sum over all genes ('sequencing depth') and multiplying the result by the median sequencing depth size across all cells. We then log-transformed the data using $\log_2(x + 1)$ transformation. Finally, we standardized all gene expression values and all electrophysiological properties to zero mean and unit variance.

Data availability All data sets were either downloaded from the original publications or provided by the authors and can be found at https://github.com/berenslab/patch-seq-rrr or following the links therein. Our full analysis code in Python is also available there.

5 Appendix. Procrustes problem

Given $\mathbf{A}$, the Procrustes problem is to maximize $\mathrm{tr}(\mathbf{A}\mathbf{V}^\top)$ subject to $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$ Gower and Dijksterhuis (2004). Let us denote by $\mathbf{A} = \mathbf{L}\mathbf{Q}\mathbf{R}^\top = \tilde{\mathbf{L}}\tilde{\mathbf{Q}}\mathbf{R}^\top$ the 'thin' and the 'full' SVD of $\mathbf{A}$. Now we have:

$$\mathrm{tr}(\mathbf{A}\mathbf{V}^\top) = \mathrm{tr}(\tilde{\mathbf{L}}\tilde{\mathbf{Q}}\mathbf{R}^\top\mathbf{V}^\top) = \mathrm{tr}(\tilde{\mathbf{Q}}\mathbf{R}^\top\mathbf{V}^\top\tilde{\mathbf{L}}) = \mathrm{tr}(\tilde{\mathbf{Q}}\mathbf{H}) = \sum q_i H_{ii} \leq \sum q_i = \mathrm{tr}(\mathbf{Q}).$$

Here $\mathbf{H} = \mathbf{R}^\top\mathbf{V}^\top\tilde{\mathbf{L}}$ is a matrix with orthonormal rows as can be verified directly, and so it must have all its elements not larger than one. It follows that the whole trace is not larger than the sum of singular values of $\mathbf{A}$. Using $\mathbf{V} = \mathbf{L}\mathbf{R}^\top$ yields exactly this value of the trace, hence it is the optimum.

Authors' contributions

PB and DK conceptualized the project, DK developed the statistical method and wrote the software, DK, YB, and MW performed computational experiments, FS performed Patch-seq experiments under the supervision of AT and helped analyzing the data, PB supervised the project. DK and PB wrote the paper with input from all authors.

Acknowledgements

We thank Rickard Sandberg, Cathryn Cadwell, and Jiaolong Xiang for discussions, Cathryn Cadwell, Janos Fuzik and their co-authors for making their data available and Shreejoy Tripathy for comments and help with data processing. This work was funded by the German Ministry of Education and Research (FKZ 01GQ1601), the German Research Foundation (EXC 2064 project number 390727645, BE5601/4-1) and the National Institute Of Mental Health of the National Institutes of Health under Award Number U19MH114830. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Cathryn R Cadwell, Athanasia Palasantza, Xiaolong Jiang, Philipp Berens, Qiaolin Deng, Marlene Yilmaz, Jacob Reimer, Shan Shen, Matthias Bethge, Kimberley F Tolias, et al. Electrophysiological, transcriptomic and morphologic profiling of single neurons using patch-seq. *Nature Biotechnology*, 34 (2):199, 2016.

Cathryn R Cadwell, Federico Scala, Shuang Li, Giulia Livrizzi, Shan Shen, Rickard Sandberg, Xiaolong Jiang, and Andreas S Tolias. Multimodal profiling of single-cell morphology, electrophysiology, and gene expression using Patch-seq. *Nature Protocols*, 12(12):2531, 2017.

Lisha Chen and Jianhua Z Huang. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545, 2012.

Xi Chen, Liu Han, and Jaime Carbonell. Structured sparse canonical correlation analysis. In *Artificial Intelligence and Statistics*, pages 199–207, 2012.

Delin Chu, Li-Zhi Liao, Michael K Ng, and Xiaowei Zhang. Sparse canonical correlation analysis: New formulation and algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12): 3050–3065, 2013.

Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

Christine De Mol, Sofia Mosci, Magali Traskine, and Alessandro Verri. A regularized method for selecting nested groups of relevant genes from microarray data. *Journal of Computational Biology*, 16(5):677–690, 2009.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Csaba Földy, Spyros Darmanis, Jason Aoto, Robert C Malenka, Stephen R Quake, and Thomas C Südhof. Single-cell rnaseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. *Proceedings of the National Academy of Sciences*, 113(35):E5222–E5231, 2016.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

János Fuzik, Amit Zeisel, Zoltán Máté, Daniela Calvigioni, Yuchio Yanagawa, Gábor Szabó, Sten Linnarsson, and Tibor Harkany. Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nature Biotechnology*, 34(2):175, 2016.

Karl Ruben Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

Chao Gao, Zongming Ma, Harrison H Zhou, et al. Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

Ignacio González, Sébastien Déjean, Pascal GP Martin, Alain Baccini, et al. CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12):1–14, 2008.

Ignacio González, Sébastien Déjean, Pascal GP Martin, Olivier Gonçalves, Philippe Besse, and Alain Baccini. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, 17(02):173–199, 2009.

Ignacio González, Kim-Anh Lê Cao, Melissa J Davis, and Sébastien Déjean. Visualising associations between paired 'omics' data sets. *BioData Mining*, 5(1):19, 2012.

John C Gower and Garmt B Dijksterhuis. *Procrustes problems*, volume 30. Oxford University Press on Demand, 2004.

David R Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83 (3):331–353, 2011.

Kenneth D Harris, Hannah Hochgerner, Nathan G Skene, Lorenza Magno, Linda Katona, Carolina Bengtsson Gonzales, Peter Somogyi, Nicoletta Kessaris, Sten Linnarsson, and Jens Hjerling-Leffler. Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics. *PLoS Biology*, 16(6):e2006387, 2018.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

Kim-Anh Lê Cao, Debra Rossouw, Christele Robert-Granié, and Philippe Besse. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

Kim-Anh Lê Cao, Pascal GP Martin, Christèle Robert-Granié, and Philippe Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10(1):34, 2009.

Kim-Anh Lê Cao, Simon Boitard, and Philippe Besse. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12(1): 253, 2011.

Ed Lein, Lars E Borm, and Sten Linnarsson. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*, 358(6359):64–69, 2017.

Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6), 2019.

Anastasia Lykou and Joe Whittaker. Sparse CCA using a lasso with positivity constraints. *Computational Statistics & Data Analysis*, 54(12):3144–3157, 2010.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Richard H Masland. Neuronal cell types. *Current Biology*, 14(13):R497–R500, 2004.

Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8 (1):1–34, 2009.

Jean-Francois Poulin, Bosiljka Tasic, Jens Hjerling-Leffler, Jeffrey M Trimarchi, and Rajeshwar Awatramani. Disentangling neural cell diversity using single-cell transcriptomics. *Nature Neuroscience*, 19 (9):1131, 2016.

Florian Rohart, Benoit Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11):e1005752, 2017.

Arpiar Saunders, Evan Z Macosko, Alec Wysoker, Melissa Goldman, Fenna M Krienen, Heather de Rivera, Elizabeth Bien, Matthew Baum, Laura Bortolin, Shuyu Wang, et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*, 174(4):1015–1030, 2018.

Federico Scala, Dmitry Kobak, Shen Shan, Yves Bernaerts, Sophie Laturnus, Cathryn Rene Cadwell, Leonard Hartmanis, Jesus Castro, Zheng Huan Tan, Rickard Sandberg, et al. Neocortical layer 4 in adult mouse differs in major cell types and circuit organization between primary sensory areas. *bioRxiv*, page 507293, 2018.

Karthik Shekhar, Sylvain W Lapan, Irene E Whitney, Nicholas M Tran, Evan Z Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z Levin, James Nemesh, Melissa Goldman, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5):1308–1323, 2016.

Xiaotong Suo, Victor Minden, Bradley Nelson, Robert Tibshirani, and Michael Saunders. Sparse canonical correlation analysis. *arXiv*, 2017.

Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335, 2016.

Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72, 2018.

Shreejoy J Tripathy, Lilah Toker, Brenna Li, Cindy-Lee Crichlow, Dmitry Tebaykin, B Ogan Mancarci, and Paul Pavlidis. Transcriptomic correlates of neuron electrophysiological diversity. *PLoS Computational Biology*, 13(10):e1005814, 2017.

Raja Velu and Gregory C Reinsel. *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media, 2013.

Sandra Waaijenborg, Philip C Verselewel de Witt Hamer, and Aeilko H Zwinderman. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

Ami Wiesel, Mark Kliger, and Alfred O Hero III. A greedy approach to sparse canonical correlation analysis. *arXiv*, 2008.

Ines Wilms and Christophe Croux. Sparse canonical correlation analysis from a predictive point of view. *Biometrical Journal*, 57(5):834–851, 2015.

Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–27, 2009.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3): 515–534, 2009.

Huan Xu, Constantine Caramanis, and Shie Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):187–193, 2011.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015.

Amit Zeisel, Hannah Hochgerner, Peter Lonnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Haring, Emelie Braun, Lars Borm, Gioele La Manno, et al. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014, 2018.

Hongkui Zeng and Joshua R Sanes. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature Reviews Neuroscience*, 18(9):530, 2017.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.