

# Regressão

Sparse reduced-rank regression for exploratory visualisation of  
paires multivariate data

Daniel Koiti Oshiro

Unicamp

12/06/22

- 1 Resumo
- 2 Introdução
- 3 Regressão
- 4 Resultados



## Resumo

“Dentro da genética, há o mundo das “ômicas”, um conjunto de técnicas moleculares que auxiliam na compreensão das diferentes moléculas biológicas que dão funcionalidades a um organismo.” (Biazon 2017).

Em particular, no estudo dos neurônios, podemos obter vários conjuntos de resultados dado as mesmas réplicas biológicas, com a existência de vários métodos, como o sequenciamento genético e a análise das propriedades eletrofisiológicas. No artigo selecionado é apresentado uma metodologia baseada em sparse reduced-rank regression (sRRR) para obter uma representação gráfica **interpretável** entre a relação dos genes e das propriedades eletrofisiológicas através de um *bibiplot*.

## Resumo

É utilizado o *elastic net*, pois fornece resultados esparsos e é computacionalmente eficiente.

Por fim, através de vários conjuntos de dados **Path-seq** é mostrado que o método tem melhor desempenho que a *full-rank* RRR e a RRR não esparsa, em relação à previsão. Surpreendentemente (ou não) conclui-se que o método empregado é uma boa ferramenta para análise exploratória de pares de dados multivariados junto com sua respectiva visualização.

## O problema

### Article

## Phenotypic variation of transcriptomic cell types in mouse motor cortex

<https://doi.org/10.1038/s41586-020-2907-3>

Received: 5 February 2020

Accepted: 16 October 2020

Published online: 12 November 2020

Open access

 Check for updates

Federico Scala<sup>1,2,3</sup>, Dmitry Kobak<sup>2,3</sup>, Matteo Bernabucci<sup>1,2</sup>, Yves Bernaerts<sup>4,5</sup>, Cathryn René Cadwell<sup>6</sup>, Jesus Ramon Castro<sup>7</sup>, Leonard Martmanis<sup>8</sup>, Xiaolong Jiang<sup>1,2</sup>, Sophie Laturneau<sup>9</sup>, Elanire Miranda<sup>10</sup>, Shalaka Mulherkar<sup>1</sup>, Zheng Huan Tan<sup>1</sup>, Zichen Yao<sup>9</sup>, Hongkui Zeng<sup>9</sup>, Rickard Sandberg<sup>9</sup>, Philipp Berens<sup>1,8,9,10,11</sup> & Andreas S. Tolias<sup>2,12</sup>

Cortical neurons exhibit extreme diversity in gene expression as well as in morphological and electrophysiological properties<sup>1,2</sup>. Most existing neural taxonomies are based on either transcriptomic<sup>3,4</sup> or morpho-electric<sup>5,6</sup> criteria, as it has been technically challenging to study both aspects of neuronal diversity in the same set of cells<sup>7</sup>. Here we used Patch-seq<sup>8</sup> to combine patch-clamp recording, biocytin staining, and single-cell RNA sequencing of more than 1,300 neurons in adult mouse primary motor cortex, providing a morpho-electric annotation of almost all transcriptomically defined neural cell types. We found that, although broad families of transcriptomic types (those expressing *Vip*, *Parvalb*, *Sst* and so on) had distinct and essentially non-overlapping morpho-electric phenotypes, individual transcriptomic types within the same family were not well separated in the morpho-electric space. Instead, there was a continuum of variability in morphology and electrophysiology, with neighbouring transcriptomic cell types showing similar morpho-electric features, often without clear boundaries between them. Our results suggest that neuronal types in the neocortex do not always form discrete entities. Instead, neurons form a hierarchy that consists of distinct non-overlapping branches at the level of families, but can form continuous and correlated transcriptomic and morpho-electrical landscapes within families.

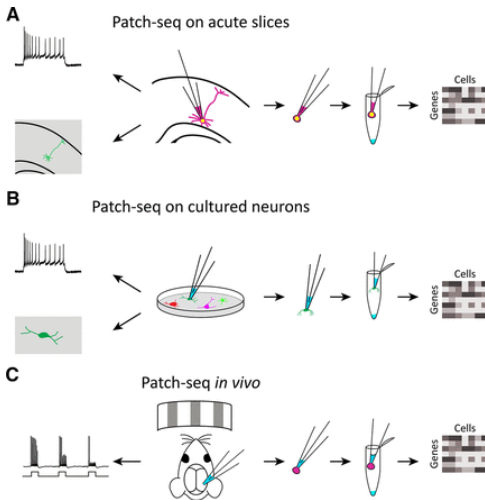
Figura: Problema de classificação.

## Path-seq

A análise de célula única (*Single-cell*) tem sido a base da neurociência, começando com as explicações incomuns da morfologia do neuronal. Como exemplo temos o estudo das propriedades eletrofisiológicas de neurônios individualmente. Recentemente, estes métodos foram introduzidos na ciência transcriptômica e resultaram em uma “resolução sem igual” na neurociência.<sup>a</sup>

---

<sup>a</sup>Lipovsek et al. 2021.



### Figura: Patch-seq analysis

## Path-seq

Resumindo: *Patch-seq* perfila as propriedades eletrofisiológicas e o transcriptoma dos mesmos neurônios individualmente, com o objetivo de identificar as relações subjacentes entre a expressão gênica e a função neuronal.<sup>a</sup>

Porém, *Patch-seq* é um experimento complicado e tem baixo rendimento, poucas células são caracterizadas no nível de expressão de milhares de genes e medidas eletrofisiológicas.<sup>b</sup> Ou seja, temos um problema de alta dimensão.

---

<sup>a</sup>Lipovsek et al. 2021.

<sup>b</sup>Kobak et al. 2021.

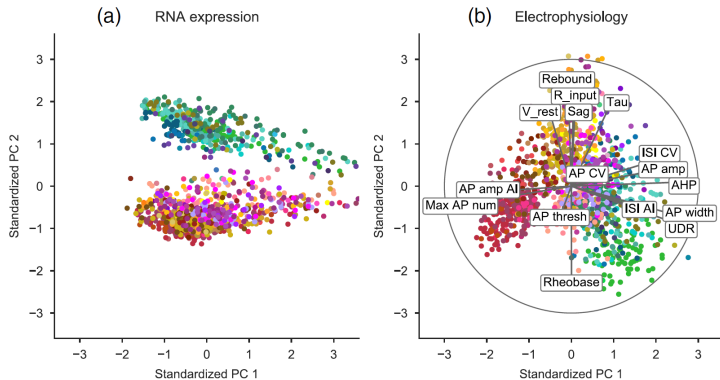


Name	Citation	Cells ( $n$ )	Genes ( $p$ )	Features ( $q$ )	Description
M1	Scala et al. (2020)	1213	1000	16	Motor cortex, all layers/types
V1	Gouwens et al. (2020)	3395	1252	55	Visual cortex, interneurons (L1–L6)
L4	Scala et al. (2019)	102	1000	13	Layer 4 <i>Sst</i> interneurons
L1	Cadwell et al. (2016)	44	3000	11	Layer 1 interneurons
S1	Fuzik et al. (2016)	80	1384	80	Layer 1/2 neurons

Figura: *Patch-seq*- cortex de um rato.

## Principal Component Analysis

Para ilustrar a estrutura do conjunto de dados e **motivar** o uso do *sRRR* para uma análise exploratória dos dados, os autores primeiramente usam o *PCA* em *M1*.



## *Principal Component Analysis<sup>a</sup>*

<sup>a</sup>Jolliffe 1986.

Relembrando o *PCA*:

Tome  $\mathbf{x}$  um vetor com  $p$  variáveis aleatórias. No caso em que a estrutura é complexa (ou as  $p$  é relativamente grande), estudar toda a estrutura das covariâncias destas variáveis não é vantajoso. Então procuramos pelas variáveis que preservam a maior parte da informação. Isto é, procuraremos por  $\ll p$  variáveis que descrevem a maior parte das variâncias e correlações existentes.

## Principal Component Analysis<sup>a</sup>

<sup>a</sup>Jolliffe 1986.

Ou seja, queremos vetores  $\alpha_i^t$  tal que maximizar  $\text{Var}[\alpha_i^t \mathbf{x}]$  sujeito que  $\alpha_i^t \mathbf{x}$  não é correlacionado com  $\alpha_j^t \mathbf{x}$ , para  $i \neq j$ .

Chamaremos a variável  $\alpha_i^t \mathbf{x}$  com a  $k$ -ésima maior variância de  $k$ -ésima PCA. Observe que:

$$\text{var}[\alpha_1^t \mathbf{x}] = \alpha_1^t \Sigma \alpha_1$$

Ou seja, maximizar esta variância implica em um  $\alpha_1$  não finito.

## Principal Component Analysis<sup>a</sup>

<sup>a</sup>Jolliffe 1986.

Imponhamos a condição  $\|\alpha_i^t\|_F^2 = 1$ . Então, aplicando os multiplicadores de Lagrange:

$$\max \alpha_1^t \Sigma \alpha_1 - \lambda(\alpha_1^t \alpha_1 - 1)$$

Derivando em relação à  $\alpha_1$  temos:

$$(\Sigma - \lambda I)\alpha_1 = 0$$

Ora, esta é a equação característica, logo,  $\lambda$  é um autovalor da matriz  $\Sigma$ . Se é um autovalor de então  $\Sigma \alpha_1 = \lambda \alpha_1$ . Dado que queremos maximizar  $\alpha_1^t \Sigma \alpha_1$ , então  $\lambda$  deve ser o maior valor possível.

## Principal Component Analysis

Com um pouco mais de trabalho derivamos resultados análogos para cada *PC*. Como exemplo, é **quase** fácil implementar o problema no *r*.

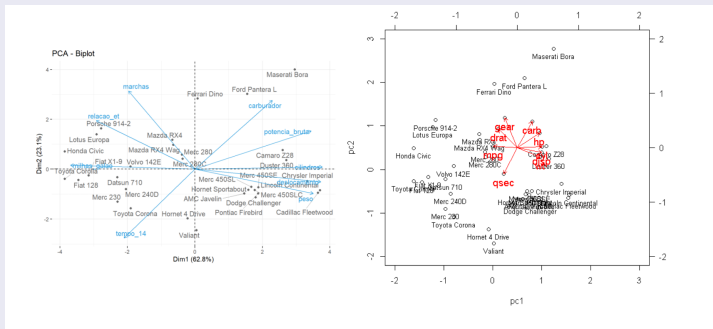


Figura: Comparação entre *biplot*.

## Sparse reduced rank regression (sRRR)

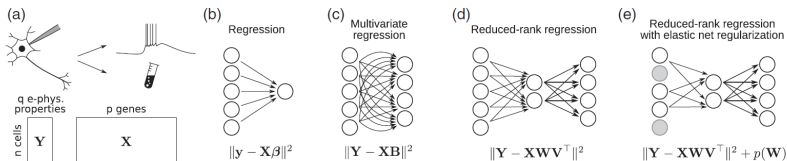


Figura: Metodologia

## Sparse reduced rank regression (sRRR)<sup>a</sup>

<sup>a</sup>Chen e Huang 2012.

Considere o problema de fazer uma regressão (tanto com o objetivo explicativo ou preditivo) de diversas variáveis respostas através de um conjunto comum de preditoras. Isto é, tome  $\mathbf{Y} = n \times q$  como a matriz resposta,  $\mathbf{X} = n \times p$  como a matriz de preditoras,  $\mathbf{B} = p \times q$  a matriz de coeficientes e  $\mathbf{E} = n \times q$  a matriz de erros, ou seja, assumindo linearidade temos:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$



## Sparse reduced rank regression (sRRR)<sup>a</sup>

<sup>a</sup>Chen e Huang 2012.

Minimizando o erro quadrático médio, temos que a matriz de coeficientes é dada por:

$$\hat{\mathbf{B}}_{OLS} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$$

Ora, então a regressão multivariada é equivalente a regressão linear realizada separadamente para cada variável de interesse. Logo, desconsideramos as possíveis correlações entre as respostas. De fato, abrindo as matrizes é fácil verificar o resultado empiricamente.

## Sparse reduced rank regression (sRRR)

	Estimate		Estimate		Estimate
cyl	0.700774	cyl	11.9261	cyl	0.720198
disp	-0.003038	disp	0.3255	disp	-0.011429
drat	5.369948	drat	-25.9269	drat	2.495166
wt	-1.972655	wt	-9.7578	wt	2.121838
vs	4.891396	vs	-17.7465	vs	4.087911
gear	0.334853	gear	35.5201	gear	-0.441540

Figura: Patch-seq analysis

	mpg	hp	qsec
cyl	0.700774335	11.926075	0.72019843
disp	-0.003037682	0.325524	-0.01142917
drat	5.369947538	-25.926876	2.49516594
wt	-1.972655073	-9.757776	2.12183811
vs	4.891395708	-17.746476	4.08791083
gear	0.334852565	35.520074	-0.44154020

Figura: Patch-seq analysis

## Sparse reduced rank regression (sRRR)<sup>a</sup>

<sup>a</sup>Chen e Huang 2012.

Uma maneira de utilizar da possibilidade que as variáveis respostas estão correlacionadas é limitar o *rank* da matriz **B** para um natural  $r \leq \min(p, q)$ . Logo, estamos assumindo que os coeficientes de regressão são múltiplos escalares dos outros, implicando que há uma “redução” no número de parâmetros. Como consequência, temos um aumento na eficiência da estimação. Matematicamente, podemos expressar **B** por:

$$\mathbf{B} = \mathbf{WV}^t$$

Onde o número de colunas de cada matriz é  $r$ . *WLOG* escolhemos  $\mathbf{VV}^t = \mathbf{I}$ .

## Sparse reduced rank regression (sRRR)<sup>a</sup>

<sup>a</sup>Chen e Huang 2012.

A solução é determinada pela decomposição em valores singulares tal que:

$$\hat{\mathbf{V}} = \mathbf{A}; \hat{\mathbf{W}} = \mathbf{B}_{OLS} \hat{\mathbf{V}}$$

Com  $\mathbf{A}$  sendo uma matriz diagonal dos autovalores (crescentes) de  $\mathbf{XB}_{OLS}$

## Sparse reduced rank regression (sRRR)<sup>a</sup>

<sup>a</sup>Chen e Huang 2012.

Porém, retornando ao artigo, no caso em que  $q \ll p$  os autores fazem um comentário:

“some of them [predictors] might not be useful for prediction. Thus, we would like to perform variable selection, or to exclude the redundant predictors”

Name	Citation	Cells ( $n$ )	Genes ( $p$ )	Features ( $q$ )	Description
M1	Scala et al. (2020)	1213	1000	16	Motor cortex, all layers/types
V1	Gouwens et al. (2020)	3395	1252	55	Visual cortex, interneurons (L1–L6)
L4	Scala et al. (2019)	102	1000	13	Layer 4 <i>Sst</i> interneurons
L1	Cadwell et al. (2016)	44	3000	11	Layer 1 interneurons
S1	Fuzik et al. (2016)	80	1384	80	Layer 1/2 neurons

Figura: *Patch-seq*- cortex de um rato.

## Sparse reduced rank regression (sRRR)

Então temos uma outra particularidade do problema, com exceção do conjunto de dados  $V1$ ,  $n \ll p$ . A regressão *sRRR* proposta utiliza da penalização do *lasso*, porém a regularização com o *elastic net* tem melhor desempenho quando as preditoras são altamente correlacionadas (Zou e Hastie 2005). Outra vantagem, é que a regularização do *elastic net* permite selecionar mais de  $n$  preditoras.

Mas, novamente temos um problema. Zou e Hastie 2005 apontam para um problema de *over* — *shrinkage* na penalidade do *elastic net*.

## *Relaxed Elastic Net*

Kobak et al. 2021 propõem o que eles chamam de *Relaxed Elastic Net*, baseado na nomenclatura do *Relaxed Lasso* e em abordagens similares, como a de De Mol et al. 2009.

Basicamente, o processo é:

- realizar o *sRRR* com a penalidade *elastic net*.
- com os coeficientes que não foram zerados, realiza-se a penalização do *ridge*.



## Relaxed Elastic Net

Neste caso, a regularização tomada implica uma perda:

$$\mathcal{L}_{sRRR} = \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{V}^t\|^2 + \lambda \left( \alpha \sum_{i=1}^p \|\mathbf{w}_i\|_2 + \frac{(1-\alpha)}{2} \|\mathbf{W}\|^2 \right)$$

E o a regressão tomada implica:

- realizar o *sRRR* com os parâmetros  $\lambda$  e  $\alpha$ .
- realizar novamente o *sRRR* com o mesmo  $\lambda$  e  $\alpha = 0$ .

## Relaxed Elastic Net

$$\mathcal{L}_{sRRR} = \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{V}^t\|^2 + \lambda \left( \alpha \sum_{i=1}^p \|\mathbf{w}_i\|_2 + \frac{(1-\alpha)}{2} \|\mathbf{W}\|^2 \right)$$

*We followed Friedman et al. (2010), Chen and Huang (2012) and others in imposing sparsity only in the predictor space: the lasso penalty is applied only to  $W$  but not to  $V$  and so feature selection only happens on the columns of  $X$  but not of  $Y$ . This arguably makes sense for the Patch-seq data considered in this manuscript but may be different for other kinds of datasets. See Chen et al. (2012a) for a discussion of sRRR with sparsity in both  $X$  and  $Y$ .*

### *Relaxed Elastic Net*

Então, segundo os autores, o problema é biconvexo e é resolvido fixando  $\mathbf{V}$  e encontrando  $\mathbf{W}_{opt}$  e então fixando  $\mathbf{W}$  e encontrando  $\mathbf{V}_{opt}$  até a convergência.

Uma particularidade é que não há garantias que a função perda é conjuntamente convexa para  $\mathbf{V}$  e  $\mathbf{W}$ , então a solução inicial deve ser tomada com cuidado. Perceba:

## Relaxed Elastic Net

É possível demonstrar que, fixado  $\mathbf{V}$ , minimizar a perda é equivalente à realizar o *elastic net* de  $\mathbf{YV}$  sobre  $\mathbf{X}$ . Então, basta aplicar os pacotes que já existem. De fato, veja que a perda é dada por:

$$\mathcal{L}_{sRRR} = \frac{1}{2n} \|\mathbf{YV} - \mathbf{XW}^t\|^2 + \lambda \left( \alpha \sum_{i=1}^p \|\mathbf{w}_i\|_2 + \frac{(1-\alpha)}{2} \|\mathbf{W}\|^2 \right)$$

Por outro lado, fixado  $\mathbf{W}$  basta realizarmos a decomposição em valores singulares de  $\mathbf{Y}^t \mathbf{XW}$  e tomarmos  $\mathbf{V}_{opt} = \mathbf{LR}^t$ . De fato, pois a perda é dada apenas por:

$$\|\mathbf{Y} - \mathbf{XWV}^t\|^2 = \kappa - 2tr(\mathbf{Y}^t \mathbf{XWV}^t)$$

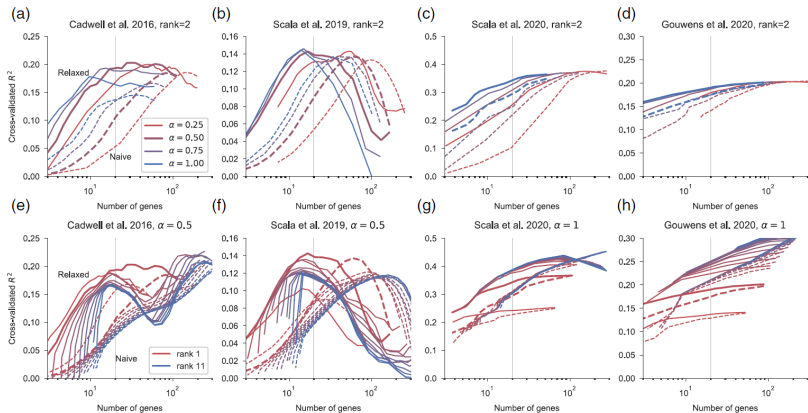


Figura: Resultados do *cross-validation*

## *Bibiplot*

Com o *sRRR* proposto são feitos pares de *biplots* em todos os conjuntos de dados, para  $r = 2$  e  $\lambda$  escolhido de maneira a escolher 20 genes.

Com os dados dos transcritos genéticos, obviamente foi usado **XW** para a representação do diagrama de dispersão. O *biplot* das propriedades eletrofisiológicas foi feito com **YV**. Vejamos os resultados:

(a) Cadwell et al. 2016

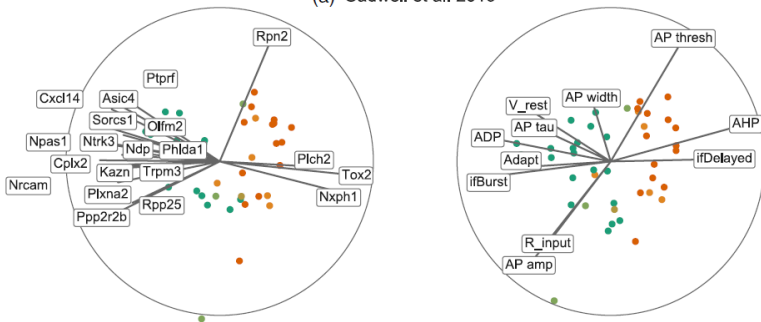


Figura: *Bibiplot de L1.*

(c) Scala et al. 2020

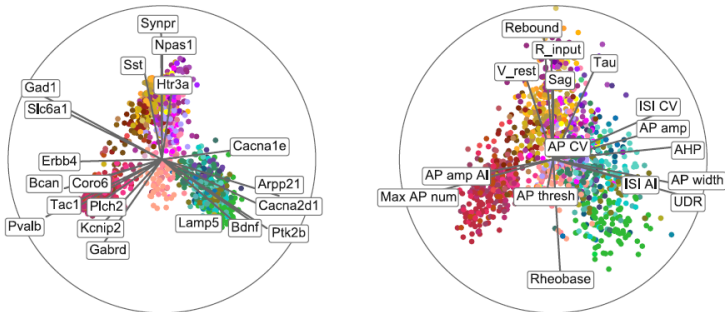


Figura: Bibiplot de M1.



(d) Gouwens et al. 2020

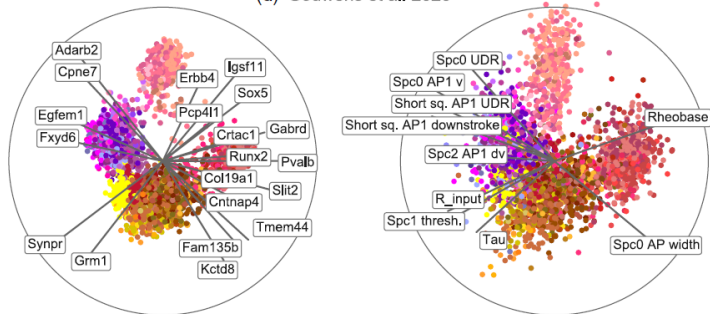


Figura: Bibiplot de V1.

(b) Scala et al. 2019

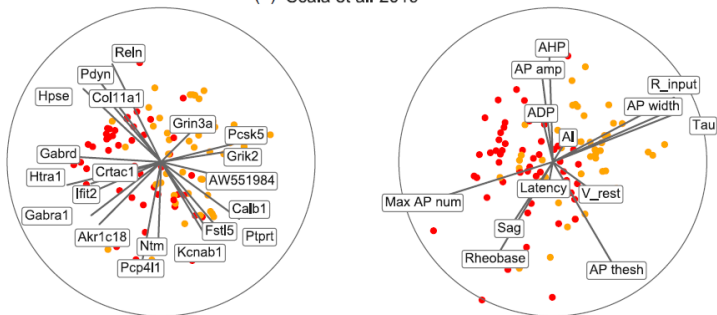









Figura: Bibiplot de L4.

-  Biazon, T. O. (2017). *Laboratório da USP traz metodologias ômicas para pesquisas e empresas*. Piracicaba; São Paulo: Jornal da USP.
-  Chen, L. e J. Huang (dez. de 2012). “Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection”. Em: *JASA. Journal of the American Statistical Association* 107.
-  De Mol, C. et al. (2009). “A Regularized Method for Selecting Nested Groups of Relevant Genes from Microarray Data”. Em: *Journal of Computational Biology* 16.5. PMID: 19432538, pp. 677–690. eprint: <https://doi.org/10.1089/cmb.2008.0171>.
-  Jolliffe, I. T. (1986). *Principal Component Analysis*. Berlin; New York: Springer-Verlag.

-  Kobak, D. et al. (2021). “Sparse reduced-rank regression for exploratory visualisation of paired multivariate data”. Em: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 70.4, pp. 980–1000. eprint: <https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssc.12494>.
-  Lipovsek, M. et al. (2021). “Patch-seq: Past, Present, and Future”. Em: *Journal of Neuroscience* 41.5, pp. 937–946. eprint: <https://www.jneurosci.org/content/41/5/937.full.pdf>.
-  Zou, H. e T. Hastie (2005). “Regularization and variable selection via the elastic net”. Em: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320. eprint: <https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2005.00503.x>.

## Discussão

É importante notar que o *biplot* pretendido está aparentemente impreciso, pois as considerações foram tomadas com o *scale* tanto para as componentes como na entrada da matriz, como o artigo faz ao introduzir o problema. Comparar o *biplot* com o *mtcars* deixa evidente.

## Discussão

Por exemplo, ao comparar o biplot de  $M1$ :

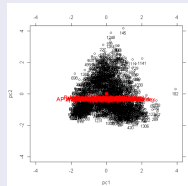


Figura:  $M1$

Porém os autores escalam as matrizes por um valor fator de 3, mais importante, eles selecionam apenas os dados informativos.

## Discussão

O  $R^2$  para cada *fold* foi calculado por:

$$R^2 = 1 - \frac{||\mathbf{Y}_{test} - \mathbf{X}_{test} \hat{\mathbf{W}} \hat{\mathbf{V}}^t||^2}{||\mathbf{Y}_{test}||^2}$$

## Discussão

Uma pergunta natural a se fazer é: existem outros métodos que podem promover outras visualizações? Elas são mais interpretáveis?

A resposta da primeira pergunta é um claro sim, porém em questão de interpretabilidade, depende. Ao tangenciar qualquer questão de qualidade não encontrei nenhuma resposta simples para garantir uma solução. Como “exercício”:



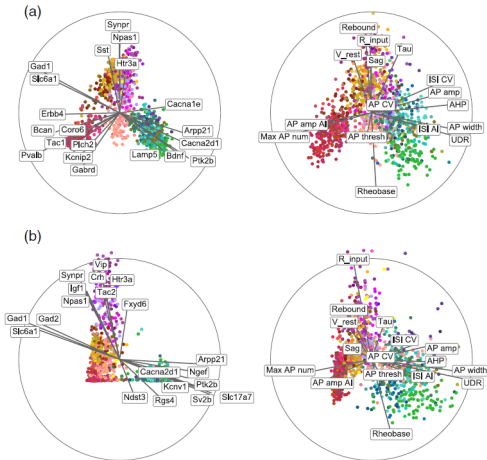


Figura: Comparação entre *biplots* de *M1*. Adaptado de Kobak et al. 2021.

## Discussão

Os autores concluem que a ferramenta/método proposto é eficiente para exploração e visualização de conjunto de dados deste tipo, porém quais as limitações fora de *Patch – seq*? Quais as características do *relaxed elastic net*? A estabilidade depende do tratamento anterior dos dados?