Principal Genes Selection

Ian C. Covert
University of Washington
Seattle, WA, USA
icovert@uw.edu

Uygar Sümbül
Allen Institute
Seattle, WA, USA
uygars@alleninstitute.org

Su-In Lee
University of Washington
Seattle, WA, USA
suinlee@uw.edu

Abstract

Unsupervised feature selection involves finding a small number of highly informative features, in the absence of a specific supervised learning task. Here, we propose the *restricted autoencoder* (RAE) framework for selecting features that can accurately reconstruct the rest of the features. We justify our approach through a proof that the reconstruction ability of a set of features bounds its performance in downstream supervised learning tasks. Based on this theory, we present a learning algorithm for RAEs that iteratively eliminates features using learned per-feature corruption rates. We apply the RAE framework to two high-dimensional biological datasets—single cell RNA sequencing and microarray gene expression data, which pose important problems in cell biology and precision medicine—and demonstrate that RAEs outperform nine baseline methods: they select features with better reconstruction ability, and that perform better in downstream classification tasks.

1 Introduction

Many domains involve high-dimensional observations $X \in \mathbb{R}^d$, and it is often desired to select a small number of representative features *a priori* and observe only this subset. Unsupervised feature selection should select features that are informative in a general sense, and not just for a specific supervised learning task. As a motivating example, we may be restricted to measuring the expression levels of only a small number of genes, and then use these measurements in a variety of future prediction tasks, such as disease subtype prediction, cell type classification, and so on. We refer to these as the *principal genes*, because they capture as much information as possible.

In this work, we present an approach for selecting features based on their reconstruction ability. We provide a novel theoretical result which shows that the ability of a set of features to impute all the remaining features bounds its performance in downstream supervised learning tasks. Based on this theory, we introduce the framework of *restricted autoencoders* (RAEs) to learn a model that reconstructs the full observation vector while relying on a subset of inputs. In experiments on single-cell RNA sequencing data and gene microarray data, we demonstrate that RAEs outperform nine baseline methods, selecting genes with better reconstruction ability, and that perform better in downstream classification tasks.

2 Restricted autoencoders for feature selection

2.1 Imputation objective

For a random variable $X \in \mathbb{R}^d$, feature selection algorithms determine a set $\mathcal{S} \subset \{1, 2, \dots, d\}$ of selected indices, and a set $\mathcal{R} \equiv \{1, 2, \dots, d\} \setminus \mathcal{S}$ of rejected indices. We use the notation $X^{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ and $X^{\mathcal{R}} \in \mathbb{R}^{|\mathcal{R}|}$ to denote selected and rejected features, respectively.

Machine Learning in Computational Biology (MLCB) 2019, Vancouver, Canada.

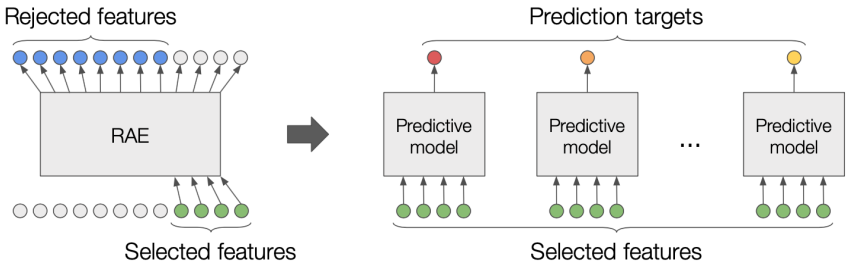

Figure 1: Features are selected by learning a restricted autoencoder (RAE), and can be applied in downstream prediction tasks.

The goal of unsupervised feature selection is to select features $X^{\mathcal{S}}$ that are most representative of the full observation vector X. An approach that has received some interest in prior work is to measure how well $X^{\mathcal{S}}$ can reconstruct $X^{\mathcal{R}}$ [15, 18, 7, 14, 17, 26, 10]. It is intuitive to consider reconstruction ability, because if the rejected features can be reconstructed perfectly, then no information is lost when selecting a subset of features. There are many other methods for unsupervised feature selection [5, 11, 25, 1, 24, 16], but none designed explicitly to ensure strong performance in prediction tasks.

To make our motivation precise, we derive a justification that has not been presented in prior work. First, we define the *imputation loss* to quantify how much information $X^{\mathcal{S}}$ contains about $X^{\mathcal{R}}$.

Definition 1 (Imputation loss)**.** *The imputation loss $\mathcal{L}(\mathcal{S})$ quantifies how well $X^{\mathcal{S}}$ can reconstruct $X^{\mathcal{R}}$ using an unrestricted function. It is defined as:*

$$\mathcal{L}(\mathcal{S}) = \min_{h} \mathbb{E}[\, ||X^{\mathcal{R}} - h(X^{\mathcal{S}})||^2 \,] \tag{1}$$

Using the notion of imputation loss, we attempt to describe how well $X^{\mathcal{S}}$ can predict a target variable $Y \in \mathbb{R}$. Specifically, we consider the degradation in performance (*performance loss*) when a model is fitted to $X^{\mathcal{S}}$ instead of X. The following result shows that the performance loss is bounded by $\mathcal{L}(\mathcal{S})$.

Theorem 1 (Performance loss)**.** *Assume a prediction target Y such that the conditional expectation $\mathbb{E}[Y \mid X = x]$ is (C, α)-Hölder continuous with exponent $0 < \alpha \leq 1$, so that the following holds almost everywhere in the distribution of X:*

$$\big|\, \mathbb{E}[Y \mid X = a] - \mathbb{E}[Y \mid X = b] \,\big| \leq C \cdot ||a - b||_2^{\alpha}. \tag{2}$$

Then, the performance loss for features $X^{\mathcal{S}}$ can be upper bounded:

$$\min_{f_1} \mathbb{E}\big[\big(Y - f_1(X^{\mathcal{S}})\big)^2\big] - \min_{f_2} \mathbb{E}\big[\big(Y - f_2(X)\big)^2\big] \leq C^2 \cdot \mathcal{L}(\mathcal{S})^{\alpha}. \tag{3}$$

The bound in Theorem 1 suggests that $X^{\mathcal{S}}$ should be selected to minimize $\mathcal{L}(\mathcal{S})$, because doing so reduces the upper bound on the performance loss. To conserve space, we omit a proof. Given this result, we proceed with an approach to select $X^{\mathcal{S}}$ by minimizing the imputation loss as follows:

$$\mathcal{S}^* = \operatorname*{arg\,min}_{|\mathcal{S}|=k} \big\{ \min_{f} \mathbb{E}[\, ||X^{\mathcal{R}} - f(X^{\mathcal{S}})||^2 \,] \big\} \tag{4}$$

2.2 Learning restricted autoencoders

We aim to select k features by solving the problem in Eq. 4, but the objective has combinatorial complexity. We therefore propose the framework of *restricted autoencoders* (RAEs) to jointly optimize for $\mathcal{S}$ and f, and train a model to reconstruct the full observation vector while relying on a subset of the inputs. The approach is depicted in Figure 1.

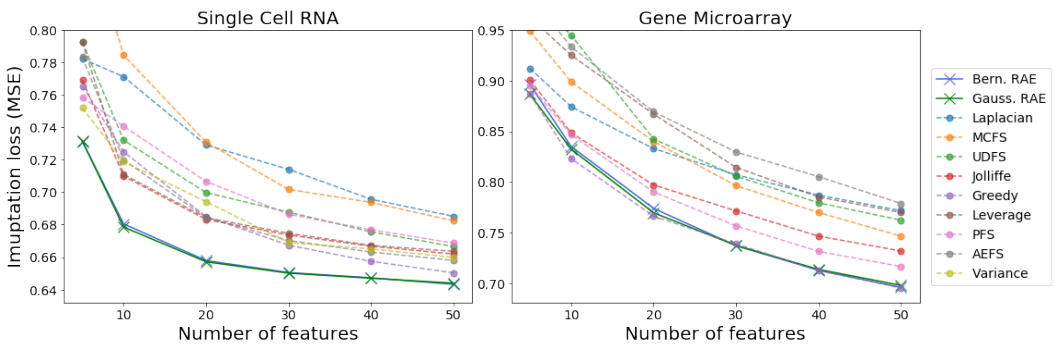

Figure 2: Imputation loss results. The MSE is normalized by the total variance of each dataset.

To train the model, we propose a learning algorithm based on backwards elimination. By leveraging feature ranking methods, it is possible to learn a RAE by iteratively training a reconstruction model, ranking features, and eliminating the lowest ranked features, in a procedure that is analogous to recursive feature elimination [9]. This proved to be more effective than simply selecting the top ranked features, and performed better than sparsity inducing penalties [8, 10, 21].

To rank features we consider two sensitivity measures, both of which are based on learning per-feature corruption rates. The first method stochastically sets inputs to zero using learned dropout rates p_j for each feature $j \in \mathcal{S}$ [2]. Similarly, the second method injects Gaussian noise using learned per-feature standard deviations σ_j. We refer to these methods as Bernoulli RAE and Gaussian RAE, due to the kind of noise they inject. Based on the logic that important features tolerate less corruption, we rank features according to p_j or σ_j.

During training, both methods require penalty terms to encourage non-zero corruption rates, and a hyperparameter λ to control the tradeoff between accurate reconstruction and the amount of noise. The objective functions for each iteration of the elimination algorithm are shown in Eqs. 5-6, and both are optimized using stochastic gradient methods and the reparameterization trick [13, 20].

$$\min_{\theta,p} \mathbb{E}_{m\sim B(p)}\big[\,\mathbb{E}_X[(X - h_\theta(X^{\mathcal{S}} \odot m))^2]\,\big] - \lambda \sum_{j\in\mathcal{S}} \log p_j \tag{5}$$

$$\min_{\theta,\sigma} \mathbb{E}_{z\sim N(0,\sigma^2)}\big[\,\mathbb{E}_X[(X - h_\theta(X^{\mathcal{S}} + z))^2]\,\big] + \lambda \sum_{j\in\mathcal{S}} \log(1 + \frac{1}{\sigma_j^2}) \tag{6}$$

3 Experiments

3.1 Datasets and baselines

We apply the RAE feature selection approach to two publicly available biological datasets: single-cell RNA sequencing data from the Allen Brain Atlas ($n = 24{,}411$, $d = 5{,}081$) [22], and microarray gene expression data from cancer patients, with labeled samples from the Gene Expression Omnibus [6] ($n = 11{,}963$, $d = 7{,}592$) and labeled samples from The Cancer Genome Atlas [23] ($n = 590$, $d = 7{,}592$). We performed standard pre-processing, using log1p of the expression counts for the single-cell data, and applying batch correction to the combined gene microarray datasets. For both of these data domains, determining a small subset of informative features is an important problem. In precision medicine, a key goal is to identify a small set of expression markers for subtype classifications. In cell biology, pre-selection of a small number of genes is required for fluorescent *in-situ* hybridization (FISH) methods [19, 3] that measure expression levels on intact tissue.

We compare RAEs with nine baseline methods. Jolliffe B4 [12], principal feature selection (PFS) [4], greedy feature selection (GFS) [7], the leverage score method [17], and autoencoder feature selection (AEFS) [10] either explicitly or implicitly relate to reconstruction ability, albeit primarily with a linear function. Max variance simply selects features with the largest variance; due to batch correction,

Table 1: Classification accuracy using subsets of features

	Cell type classification					Cancer subtype classification				
# Features	5	10	20	30	50	5	10	20	30	50
Laplacian	0.219	0.251	0.443	0.505	0.680	0.676	0.640	**0.748**	0.748	0.748
MCFS	0.111	0.278	0.532	0.622	0.713	0.532	0.514	0.613	0.685	0.685
UDFS	0.291	0.510	0.656	0.702	0.767	0.505	0.532	0.631	0.640	0.649
PFS	0.268	0.335	0.465	0.565	0.649	0.622	0.685	0.703	0.721	0.712
AEFS	0.320	0.574	0.759	0.806	0.847	0.523	0.486	0.550	0.640	0.604
Variance	0.447	0.541	0.741	0.793	0.822					
Leverage	0.463	0.634	0.780	0.816	**0.852**	0.523	0.568	0.613	0.658	0.649
Jolliffe	0.264	0.557	0.712	0.793	0.844	0.667	0.676	0.622	0.685	0.703
Greedy	0.203	0.367	0.580	0.691	0.820	0.657	0.673	0.684	**0.750**	**0.753**
B. RAE	0.484	**0.674**	**0.789**	**0.822**	0.845	**0.679**	**0.687**	0.701	0.721	**0.753**
G. RAE	**0.487**	**0.667**	0.771	**0.822**	0.846	0.645	0.678	0.686	0.694	0.740

it could not be applied to the microarray data. Laplacian scores [11] and multi-cluster feature selection (MCFS) [1] aim to preserve local structure through spectral information, and unsupervised discriminative feature selection (UDFS) [24] aims to retain local discriminative information.

3.2 Imputation performance

We first demonstrate that RAEs select features that achieve a low imputation loss. Both datasets were split into training, validation and test sets, and we used only the unlabeled samples for the gene microarray data. We selected features using each of the methods, and then trained separate imputation models to predict only the rejected features. Hyperparameter choices were made on validation data. We found that RAEs are robust to both shallow and deep architectures, and that iteratively eliminating features was critical (we eliminated 20% of the remaining features at each iteration).

Figure 2 displays the results, showing the imputation loss for different numbers of selected features. RAEs achieve the best performance on both datasets. The gap is larger on the single-cell RNA sequencing data, where the RAEs perform significantly better than all baselines. RAEs and GFS achieve similar results on the microarray data, outperforming all other methods by a large margin.

3.3 Downstream classification performance

Next, we assess the performance of selected features in downstream prediction tasks. Both datasets have associated classification problems that, in certain settings, would need to be performed using a subset of features: cell type classification (150 types) for the single-cell RNA data, and cancer subtype classification (4 types) for the microarray data. For the single-cell data, we used the same dataset split; for cancer classification we split the labeled TGCA samples. MLPs were trained for each task, and the reported accuracy is the average performance of 10 models on the test data.

Table 1 displays the results for both datasets. Features selected by RAEs perform very well in both tasks, particularly when using a smaller number of features. Overall, RAEs achieve the best performance most of the time ($7/10$), and when they do not, they are still among the best. We posit that RAEs perform well because they select *principal genes* that contain maximum information, and are therefore guaranteed to perform well in a variety of prediction problems (see Theorem 1).

4 Discussion

In this work we presented the framework of *restricted autoencoders* (RAEs) for selecting features based on their reconstruction ability, and a learning algorithm based on learning per-feature corruption rates. We showed theoretically and empirically that the reconstruction ability of a set of features has implications for their performance in downstream prediction tasks. Experiments on single-cell RNA sequencing and microarray gene data demonstrated the ability of RAEs to select highly informative *principal genes*, which could impact how biologists determine genes for FISH and precision medicine.

References

[1] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 333–342. ACM, 2010.

[2] Chun-Hao Chang, Ladislav Rampasek, and Anna Goldenberg. Dropout feature ranking for deep learning models. *arXiv preprint arXiv:1712.08645*, 2017.

[3] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

[4] Ying Cui and Jennifer G Dy. Orthogonal principal feature selection. 2008.

[5] Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *Journal of machine learning research*, 5(Aug):845–889, 2004.

[6] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[7] Ahmed K Farahat, Ali Ghodsi, and Mohamed S Kamel. An efficient greedy method for unsupervised feature selection. In *2011 IEEE 11th International Conference on Data Mining*, pages 161–170. IEEE, 2011.

[8] Jean Feng and Noah Simon. Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*, 2017.

[9] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[10] Kai Han, Yunhe Wang, Chao Zhang, Chao Li, and Chao Xu. Autoencoder inspired unsupervised feature selection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2941–2945. IEEE, 2018.

[11] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2006.

[12] Ian T Jolliffe. Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(2):160–173, 1972.

[13] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[14] Mahdokht Masaeli, Yan Yan, Ying Cui, Glenn Fung, and Jennifer G Dy. Convex principal feature selection. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 619–628. SIAM, 2010.

[15] George P McCabe. Principal variables. *Technometrics*, 26(2):137–144, 1984.

[16] Pabitra Mitra, CA Murthy, and Sankar K. Pal. Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):301–312, 2002.

[17] Dimitris Papailiopoulos, Anastasios Kyrillidis, and Christos Boutsidis. Provable deterministic leverage score sampling. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 997–1006. ACM, 2014.

[18] F Questier, R Put, D Coomans, B Walczak, and Y Vander Heyden. The use of cart and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems*, 76(1):45–54, 2005.

[19] Arjun Raj, Patrick Van Den Bogaard, Scott A Rifkin, Alexander Van Oudenaarden, and Sanjay Tyagi. Imaging individual mrna molecules using multiple singly labeled probes. *Nature methods*, 5(10):877, 2008.

[20] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[21] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. Neural granger causality for nonlinear time series. *arXiv preprint arXiv:1802.05842*, 2018.

[22] 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.

[23] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[24] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. L2, 1-norm regularized discriminative feature selection for unsupervised. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[25] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157. ACM, 2007.

[26] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon CK Shiu. Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446, 2015.