DropLasso: A robust variant of Lasso for single cell RNA-seq data

Beyrem Khalfaoui, Jean-Philippe Vert

DropLasso: A robust variant of Lasso for single cell RNA-seq data

Beyrem Khalfaoui [1,2] and Jean-Philippe Vert [1,2,3,]

[1] MINES ParisTech, PSL Research University,
CBIO - Centre for Computational Biology, F-75006 Paris, France
[2] Institut Curie, PSL Research University, INSERM, U900, F-75005 Paris, France.
[3] Ecole Normale Supérieure, Department of Mathematics and Applications,
CNRS, PSL Research University, F-75005 Paris, France.
firstname.lastname@mines-paristech.fr

Abstract

Single-cell RNA sequencing (scRNA-seq) is a fast growing approach to measure the genome-wide transcriptome of many individual cells in parallel, but results in noisy data with many dropout events. Existing methods to learn molecular signatures from bulk transcriptomic data may therefore not be adapted to scRNA-seq data, in order to automatically classify individual cells into predefined classes.
We propose a new method called DropLasso to learn a molecular signature from scRNA-seq data. DropLasso extends the dropout regularisation technique, popular in neural network training, to estimate sparse linear models. It is well adapted to data corrupted by dropout noise, such as scRNA-seq data, and we clarify how it relates to elastic net regularisation. We provide promising results on simulated and real scRNA-seq data, suggesting that DropLasso may be better adapted than standard regularisations to infer molecular signatures from scRNA-seq data.
DropLasso is freely available as an R package at `https://github.com/jpvert/droplasso`

1 Introduction

The fast paced development of massively parallel sequencing technologies and protocols has made it possible to measure gene expression with more precision and less cost in recent years. Single-cell RNA sequencing (scRNA-seq), in particular, is a fast growing approach to measure the genome-wide transcriptome of many individual cells in parallel (Kolodziejczyk *et al.*, 2015). By giving access to cell-to-cell variability, it represents a major advance compared to standard "bulk" RNA sequencing to investigate complex heterogeneous tissues (Macosko *et al.*, 2015; Tasic *et al.*, 2016; Zeisel *et al.*, 2015; Villani *et al.*, 2017) and study dynamic biological processes such as embryo development (Deng *et al.*, 2014) and cancer (Patel *et al.*, 2014).

The analysis of scRNA-seq data is however challenging and raises a number of specific modelling and computational issues (Ozsolak and Milos, 2011; Bacher and Kendziorski, 2016). In particular, since a tiny amount of RNA is present in each cell, a large fraction of polyadenylated RNA can be stochastically lost during sample preparation steps including cell lysis, reverse transcription or amplification. As a result, many genes fail to be detected even though they are expressed, a type of errors usually referred to as *dropouts*. In a standard scRNA-seq experiment it is common to observe more than 80% of genes with no apparent expression in each single cell, an important proportion of which are in fact dropout errors (Kharchenko *et al.*, 2014). The presence of so many zeros in the raw data can have significant impact on the downstream analysis and biological conclusions, and has given rise to new statistical models for data normalisation and visualisation (Pierson and Yau, 2015; Risso *et al.*, 2018) or gene differential analysis (Kharchenko *et al.*, 2014).

Besides exploratory analysis and gene-per-gene differential analysis, a promising use of scRNA-seq technology is to automatically classify individual cells into pre-specified classes, such as particular cell types in a cancer tissue. This requires to establish cell type specific "molecular signatures" that could be shared and used consistently across laboratories, just like standard molecular signatures are commonly used to classify

tumour samples into subtypes from bulk transcriptomic data (Ramaswamy *et al.*, 2001; Sørlie *et al.*, 2001, 2003). From a methodological point of view, molecular signatures are based on a *supervised analysis*, where a model is trained to associate each genome-wide transcriptomic profile to a particular class, using a set of profiles with class annotation to select the genes in the signature and fit the parameters of the models. While the classes themselves may be the result of an unsupervised analysis, just like breast cancer subtypes which were initially defined from a first unsupervised clustering analysis of a set of tumours (Perou *et al.*, 2000), the development of a signature to classify any new sample into one of the classes is generally based on a method for supervised classification or regression.

Signatures based on a few selected genes, such as the 70-gene signature for breast cancer prognosis of van de Vijver *et al.* (2002), are particularly useful both for interpretability of the signature, and to limit the risk of overfitting the training set. Many techniques exist to train molecular signatures on bulk transcriptomic data (Haury and Vert, 2010), however, they may not be adapted to scRNA-seq data due to the inflation of zeros resulting from dropout events.

Interestingly and independently, the term "dropout" has also gained popularity in the machine learning community in recent years, as a powerful technique to regularise deep neural networks (Srivastava *et al.*, 2014). Dropout regularisation works by randomly removing connexions or nodes during parameter optimisation of a neural network. On a simple linear model (a.k.a. single-layer neural network), this is equivalent to randomly creating some dropout noise to the training examples, i.e., to randomly set some features to zeros in the training examples (Wager *et al.*, 2013; Baldi and Sadowski, 2013). Several explanations have been proposed for the empirical success of dropout regularisation. Srivastava *et al.* (2014) motivated the technique as a way to perform an ensemble average of many neural networks, likely to reduce the generalisation error by reducing the variance of the estimator, similar to other ensemble averaging techniques like bagging (Breiman, 1996) or random forests (Breiman, 2001). Another justification for the relevance of dropout regularisation, particularly in the linear model case, is that it performs an intrinsic data-dependent regularisation of the estimator (Wager *et al.*, 2013; Baldi and Sadowski, 2013) which is particularly interesting in the presence of rare but important features. Yet another justification for dropout regularisation, particularly relevant for us, is that it can be interpreted as a *data augmentation* technique, a general method that amounts to adding virtual training examples by applying some transformation to the actual training examples, such as rotations of images or corruption by some Gaussian noise; the hypothesis being that the class should not change after transformation. Data augmentation has a long history in machine learning (e.g., Schölkopf *et al.*, 1996), and is a key ingredient of many modern successful applications of machine learning such as image classification (Krizhevsky *et al.*, 2012). As shown by van der Maaten *et al.* (2013), dropout regularisation in the linear model case can be interpreted as a data augmentation technique, where corruption by dropout noise enforces the model to be robust to dropout events in the test data, e.g., to blanking of some pixels on images or to removal of some words in a document. Wager *et al.* (2014) show that in some cases, data augmentation with dropout noise allows to train model that should be insensitive to such noise more efficiently than without.

Since scRNA-seq data are inherently corrupted by dropout noise, we therefore propose that dropout regularisation may be a sound approach to make the predictive model robust to this form of noise, and consequently to improve their generalisation performance on scRNA-seq supervised classification. Since plain dropout regularisation does not lead to feature selection and to the identification of a limited number of genes to form a molecular signature, we furthermore propose an extension of dropout regularisation, which we call *DropLasso* regularisation, obtained by adding a sparsity-inducing ℓ_1 regularisation to the objective function of the dropout regularisation, just like *lasso* regression adds an ℓ_1 penalty to a mean squared error criterion in order to estimate a sparse model (Tibshirani, 1996). We show that the ℓ_1 penalty can be integrated in the standard stochastic gradient algorithm used to implement dropout regularisation, resulting in a scalable stochastic *proximal* gradient descent formulation of DropLasso. We also clarify the regularisation property of DropLasso, and show that it is to elastic net regularisation what plain dropout regularisation is to the plain ridge regularisation. Finally, we provide promising results on simulated and real scRNA-seq data, suggesting that specific regularisations like DropLasso may be better adapted than standard regularisations to infer molecular signatures from scRNA-seq data.

2 Methods

2.1 Setting and notations

We consider the supervised machine learning setting, where we observe a series of n pairs of the form $(x_i, y_i)_{i=1,\dots,n}$. For each $i \in [1, n]$, $x_i \in \mathbb{R}^d$ represents the gene expression levels for d genes measured in the i-th cell by scRNA-seq, and $y_i \in \mathbb{R}$ or $\{-1, 1\}$ is a label to represent a discrete category or a real number associated to the i-th cell, e.g., a phenotype of interest such as normal vs tumour cell, or an index of progression in the cell cycle. For $i \in [1, n]$ and $j \in [1, d]$, we denote by $x_{i,j} \in \mathbb{R}$ the expression level of gene j in cell i. From this training set of n annotated cells, the goal of supervised learning is to estimate a function to predict the label of any new, unseen cell from its transcriptomic profile. We restrict ourselves to linear models $f_w : \mathbb{R}^d \to \mathbb{R}$, for any $w \in \mathbb{R}^d$, of the form

$$\forall u \in \mathbb{R}^d, \quad f_w(u) = \sum_{i=1}^{d} w_i u_i \,.$$

To estimate a model on the training set, a popular approach is to follow a penalised maximum likelihood or empirical risk minimisation principle and to solve an objective function of the form

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(w, x_i, y_i) + \lambda \Omega(w) \right\}, \tag{1}$$

where $L(w, x_i, y_i)$ is a loss function to assess how well f_w predicts y_i from x_i, Ω is an (optional) penalty to control overfitting in high dimensions, and $\lambda > 0$ is a regularisation parameter to control the balance between under- and overfitting. Examples of classical loss functions include the square loss:

$$L_{\text{square}}(w, x_i, y_i) = \left(y_i - \sum_{j=1}^{d} w_j x_{i,j} \right)^2 ,$$

and the logistic loss:

$$L_{\text{logistic}}(w, x_i, y_i) = \log \left(1 + \exp(-y_i \sum_{j=1}^{d} w_j x_{i,j}) \right) ,$$

which are popular losses when y_i is respectively a continuous ($y_i \in \mathbb{R}$) or discrete ($y_i \in \{-1, 1\}$) label. As for the regularisation term $\Omega(w)$ in (1), popular choices include the ridge penalty (Hoerl and Kennard, 1970):

$$\Omega_{\text{ridge}}(w) = \|w\|_2^2 = \sum_{i=1}^{d} w_i^2 \,,$$

and the lasso penalty (Tibshirani, 1996):

$$\Omega_{\text{lasso}}(w) = \|w\|_1 = \sum_{i=1}^{d} |w_i| \,.$$

The properties, advantages and drawbacks of ridge and lasso penalties have been theoretically studied under different assumptions and regimes. The lasso penalty additionally allows feature selection by producing sparse solutions, i.e., vectors w with many zeros; this is useful to in many bioinformatics applications to select "molecular signatures", i.e., predictive models based on the expression of a limited number of genes only. It is known however that lasso can be unstable in particular when there are several highly correlated features in the data. It also cannot select more features than the number of observations and its accuracy is often dominated by that of ridge. For these reasons, another popular penalty is elastic net, which encompasses the advantages of both penalties Zou and Hastie (2005) :

$$\Omega_{\text{ridge}}(w) = \alpha \|w\|_2^2 + (1 - \alpha)\|w\|_1 \,,$$

where $\alpha \in [0, 1]$ allows to interpolate between the lasso ($\alpha = 0$) and the ridge ($\alpha = 1$) penalties.

2.2 DropLasso

For scRNA-seq data subject to dropout noise, we propose a new model to train a sparse linear model robust to the noise by artificially augmenting the training set with new examples corrupted by dropout. Formally, given a vector $u \in \mathbb{R}^d$ and a dropout mask $\delta \in \{0,1\}^d$, we consider the corrupted pattern $\delta \odot u \in \mathbb{R}^d$ obtained by entry-wise multiplication $(\delta \odot u)_i = \delta_i u_i$. In order to consider all possible dropout masks, we make δ a random variable with independent entries following a Bernoulli distribution of parameter $p \in [0,1]$, i.e., $P(\delta_i = 1) = p$, and consider the following DropLasso regularisation for any $\lambda > 0$, $p \in [0,1]$ and loss function L:

$$\min_{w \in \mathbb{R}^d} \left(\frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\delta_i \sim B(p)^d} L(w, \delta_i \odot \frac{x_{i,}}{p}, y_i) + \lambda \left\| w \right\|_1 \right) . \tag{2}$$

In this equation, the expectation over the dropout mask corresponds to an average of 2^d terms. The division by p in the term x_i/p is here to ensure that, on average, the inner product between w and $\delta_i \odot \frac{x_{i,}}{p}$ is independent of p, because:

$$\begin{aligned} \mathop{\mathbb{E}}_{\delta_i \sim B(p)^d} \sum_{j=1}^{d} w_j \left(\delta_i \odot \frac{x_{i,}}{p} \right)_j &= \sum_{j=1}^{d} \mathop{\mathbb{E}}_{\delta_{i,j} \sim B(p)} w_j \delta_{i,j} \frac{x_{i,j}}{p} \\ &= \sum_{j=1}^{d} w_j x_{i,j} \, . \end{aligned}$$

When $p = 1$ and $\lambda > 0$, the only mask with positive probability is the constant mask with all entries equal to 1, which performs no dropout corruption. In that case, DropLasso (2) therefore boils down to standard lasso. When $\lambda = 0$ and $p < 1$, on the other hand, DropLasso boils down to the standard dropout regularisation proposed by Srivastava *et al.* (2014) and studied, among others, by Wager *et al.* (2013); Baldi and Sadowski (2013); van der Maaten *et al.* (2013). In general, DropLasso interpolates between lasso and dropout. For $\lambda > 0$, it inherits from lasso regularisation the ability to select features associated with ℓ_1 regularisation (Bach *et al.*, 2011). We therefore propose DropLasso as a good candidate to select molecular signatures (thanks to the sparsity-inducing ℓ_1 regularisation) for data corrupted with dropout noise, in particular scRNA-seq data (thanks to the dropout data augmentation).

2.3 Algorithm

For any convex loss function L such as the square or logistic losses, DropLasso (2) is a non-smooth convex optimisation problem whose global minimum can be found by generic solvers for convex programs. Due to the dropout corruption, the total number of terms in the sum in (2) is $n \times 2^d$. This is usually prohibitive as soon as d is more than a few, e.g., in practical applications when d is easily of order 10^4 (number of genes). Hence the objective function (2) can simply not be computed exactly for a single candidate model w, and even less optimised by methods like gradient descent.

To solve (2), we instead propose to follow a stochastic gradient approach to exploit the particular structure of the model, in particular the fact that it is fast and easy to generate a sample randomly corrupted by dropout noise. A similar approach is used for standard dropout regularisation when L is differentiable Srivastava *et al.* (2014), however in our case we additionally need to take care of the non-differentiable ℓ_1 norm; this can be handled by a forward-backward algorithm which, plugged in the stochastic gradient loop, leads to the proximal stochastic gradient algorithm presented in Algorithm 1. The fact that Algorithm 1 is correct, i.e., converges to the solution of (2), follows from general results on stochastic approximations (Robbins and Siegmund, 1971).

2.4 DropLasso and elastic net

As we already mentioned, DropLasso interpolates between lasso ($p = 1, \lambda > 0$) and dropout ($p \in [0,1]$, $\lambda = 0$). On the other hand, dropout regularisation is known to be related to ridge regularisation (Wager *et al.*, 2013; Baldi and Sadowski, 2013); in particular, for the square loss, dropout regularisation boils down to ridge regression after proper normalisation of the data, while for more general losses it can be approximated

Algorithm 1 Solving DropLasso

Require: Training set $(x_i, y_i)_{i=1,\dots,n}$, initialisation $w_0 \in \mathbb{R}^d$, learning rate $\gamma_0 > 0$, number of passes $n_{passes} \in \mathbb{N}$, $\lambda \geq 0$, $p \in [0, 1]$

procedure DropLasso
  $w^0 \leftarrow w_0$
  $t \leftarrow 0$
  for $iter = 1$ `to` n_{passes} **do**
    $\pi \leftarrow$ random permutation of $[1, n]$ ▷ Shuffle training set
    for $i = 1$ `to` n **do** ▷ (Mini-)batch also possible
      $t \leftarrow t + 1$
      $\gamma_t \leftarrow \gamma_0/(1 + \gamma_0 \lambda t)$
      Sample $\delta \sim Bernoulli(p)^d$
      $z \leftarrow \delta \odot x_{\pi(i)}/p$
      $w^{t+1} \leftarrow S_{\gamma_t \lambda}(w^t - \gamma_t \nabla_w L(w^t, z, y_{\pi(i)}))$
      ▷ $S_{\gamma_t \lambda}$is the soft-thresholding operator
    end for
  end for
return (w^{t+1})
end procedure

by reweighted version of ridge regression. Here we show that DropLasso largely inherits these properties, and in a sense is to elastic net what dropout is to ridge.

Let us start with the square loss. In that case we have the following:

Property 1. *For the square loss, DropLasso is equivalent to an elastic net regression if the data are normalised so that all features have the same norm. If data are not normalised, DropLasso is equivalent to an elastic net regression with a weighted ridge penalty.*

Proof: By developing the error function and marginalising over the Bernoulli variables we get :

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\delta_i \sim B(p)^d} L(w, \delta_i \odot \frac{x_{i,}}{p}, y_i) + \lambda \left\|w\right\|_1 \\
=&\frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\delta_i \sim B(p)^d} \left(y_i - \sum_{j=1}^{d} w_j \delta_{i,j} \frac{x_{i,j}}{p} \right)^2 + \lambda \left\|w\right\|_1 \\
=&\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} w_j x_{i,j} \right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{d} w_j^2 x_{i,j}^2 \mathrm{Var}\left(\frac{\delta_{i,j}}{p} \right) + \lambda \left\|w\right\|_1 \\
=&\frac{1}{n}\sum_{i=1}^{n} L(w, x_i, y_i) + \frac{1-p}{p}\sum_{j=1}^{d} \left\|x_{.,j}\right\|_2^2 w_j^2 + \lambda \left\|w\right\|_1 . \qquad \square
\end{aligned}
$$

In the case of the logistic loss, we can also adapt a result of Wager *et al.* (2013) which relates dropout to an adaptive version of ridge regression:

Property 2. *: For the logistic loss, DropLasso can be approximated when the dropout probability p is close to 1 by an adaptive version of elastic net that automatically scales the data but also that encourages more confident predictions.*

Proof: Writing the Taylor expansion for the logistic loss up to the second order when the dropout is small (p close to 1), we get:

$$\begin{aligned} L(w, \delta_{\mathbf{i}} \odot \frac{x_{i,}}{p}, y_i) \simeq & L(w, x_{i,}, y_i) \\ & + \sum_{j=1}^{d} \frac{\partial L(w, x_{i,}, y)}{\partial x_{i,j}} \left(\frac{\delta_{i,j}}{p} - 1 \right) x_{i,j} \\ & + \frac{1}{2} \sum_{j=1}^{d} \frac{\partial^2 L(w, x_{i,}, y)}{\partial^2 x_{i,j}} \left(\frac{\delta_{i,j}}{p} - 1 \right)^2 x_{i,j}^2 \,. \end{aligned}$$

Taking the expectation with respect to δ_i, the first order term cancels out since

$$\mathop{\mathbb{E}}_{\delta_i \sim B(p)^d} \left[\delta_i \odot \frac{x_{i,}}{p} \right] = x_i \,.$$

We then get:

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\delta_i \sim B(p)^d} L(w, \delta_i \odot \frac{x_{i,}}{p}, y_i) + \lambda \left\| w \right\|_1 \\ & \simeq \sum_{i=1}^{n} L(w, x_{i,}, y_i) + \frac{1-p}{p} \sum_{j=1}^{d} \alpha_j w_j^2 + \lambda \left\| w \right\|_1 \,, \end{aligned}$$

where, for the logistic loss,

$$\alpha_j = \sum_{i=1}^{n} P(Y = 1 \mid X = x_i, w) P(Y = 0 \mid X = x_i, w) x_{i,j}^2 \,. \qquad \square$$

In words, this shows that the dropout penalty can be approximated by a weighted data-dependent version of ridge regression, where the ridge penalty is controlled both by the size of the features $x_{i,j}^2$, but also by the fact that the prediction for each sample is confident or not.

3 Results

3.1 Simulation results

We first investigate the performance of DropLasso on simulated data, and compare it to standard dropout and elastic net regularisation. We design a toy simulation to illustrate in particular how corruption by dropout noise impacts the performances of the different methods. The simulation goes as follow :

- We set the dimension to $d = 100$.
- Each sample is a random vector $z \in \mathbf{N}^d$ with entries following a Poisson distribution with parameter 1. We introduce correlations between entries by first sampling a Gaussian copula with covariance $\Sigma_d = \mathbf{I}_d + \mathbf{1}_d^\top \mathbf{1}_d$, then transforming each entry in $[0, 1]$ into an integer using the Poisson quantile function.
- The "true" model is a logistic model with sparse weight vector $w \in \mathbb{R}^d$ satisfying $w_i = 0.05$ for $i = 1, \ldots, 10$ and $w_i = 0$ for $i = 1, \ldots, d$.
- Using w as the true underlying model and z as the true observations, we simulate a label $y \sim \text{Bernoulli}(1/(1 + \exp(-\sum_{j=1}^{d} w_j z_j)))$
- We introduce corruption by dropout events by multiplying entry-wise z with an i.i.d Bernoulli variables δ with probability q.

We simulate $n = 100$ (z, x, y) samples to train different models, evaluate their performance on $m = 400$ independent samples, and repeat the whole process 10 times. Each method estimates a model using the (x, y) pairs in the training set only, i.e., does only see the corrupted samples. Elastic net and DropLasso both have two parameters. In order to make a fair comparison, we fixed the α parameter of elastic net to $\alpha = 0.5$, and the dropout probability in DropLasso to $p = 0.5$. In both cases, we vary the remaining λ parameter over a large grid of 100 values, estimate the classification performance on the test set in terms of area under the receiving operator curve (AUC), and report the best average AUC over the grid.

Table 1: Average test AUC of different methods on simulations with different amount of dropout noise. The $*$ indicates that the performance of DropLasso is significantly higher than that of elastic net ($P < 0.05$)

Method / noise	*no noise*	*q=0.8*	*q=0.6*	*q=0.4*
Elastic net	0.641	0.612	0.557	0.528
Dropout	0.626	0.613	0.551	0.525
DropLasso	**0.642**	**0.625**	**0.565**	**0.542 ***

Table 1 shows the classification performance in terms of best average AUC of elastic net, dropout and DropLasso, when we vary the amount of dropout corruption in the data. We first observe that for all methods, the performance drastically decreases when dropout noise increases, confirming the difficulty induced by dropout events to learn predictive models. Second, we note that, whatever the amount of noise, DropLasso outperforms dropout. This illustrates the benefit of incorporating the ℓ_1 lasso penalty in the objective function of dropout when the true model is sparse. Third, and more importantly, we observe that DropLasso outperforms elastic net in all settings, and that the difference in performance increases when the amount of dropout noise increases. This confirms the intuition that DropLasso is more efficient that elastic net in situations where data are corrupted by dropout noise.

Besides classification accuracy, it is also of interest to investigate to what extent the different methods select the correct variables, which are known in our simulations. For each λ value in the grid, we compute the number of true and false positives among the features selected by elastic net and DropLasso, and plot these values averaged over the 10 repeats in Figure 1. Interestingly, we observe that elastic net seems to outperform DropLasso when there is no noise in the observed data ($q = 1$), but seems to lose its ability to recover the correct features as the amount of dropout noise increases quicker than DropLasso, and in the high noise regime ($q = 0.4$) DropLasso eventually outperforms elastic net. Although the differences are limited in this simulation setting, this illustrates again that DropLasso is more robust than elastic net in the presence of dropout noise.

3.2 Classification on Single Cell RNA-seq

We now turn on to real scRNA-seq data. To evaluate the performance of methods for supervised classification, we collected 7 publicly available scRNA-seq datasets amenable to this setting, as summarised in Table 2. The first 6 datasets were processed Soneson and Robinson (2017), and we obtained them from the *conquer* website[1], a collection of consistently processed, analysis-ready and well documented publicly available scRNA-seq data sets. We used the already preprocessed length-scaled transcripts per million mapped reads (see Soneson and Robinson, 2017, for details about data processing). These datasets were used by Soneson and Robinson (2017) to assess the performance of methods for gene differential analysis between classes of cells, and we follow the same splits of cells into classes for our experiments of supervised classification. The last dataset was collected from Li *et al.* (2017) and was originally used for the analysis of transcriptional heterogeneity in colorectal tumours. Gene expression levels were quantified as fragments per kilo-base per million reads (FPKM) which also allowed to compare the classifiers after different normalisations. We used the available sample annotations to create a binary classification problem where we want to discriminate tumour from normal epithelial cells, as described in Table 2. For all datasets, we filtered out genes that were not expressed in any sample. In order to also study the performance of the different methods for a varying number of variables, we also created for each dataset 3 reduced datasets by first selecting respectively the top 100, 1,000 and 10,000 genes with highest variance among the samples, regardless of their labels. On

[1] http://imlspenticton.uzh.ch:3838/conquer/

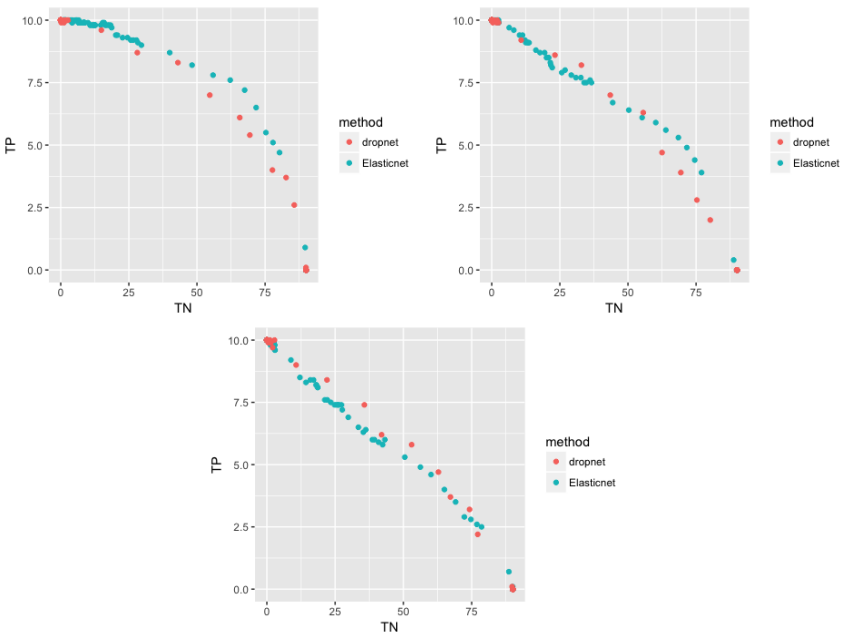

Figure 1: Performance on feature selection for elastic net and DropLasso on simulated data. From left to right, the amount of dropout noise in the data increases from no noise (left, $q = 1$), to $q = 0.8$ (centre) and $q = 0.4$ (right)

each of the $7 \times 4 = 28$ resulting dataset, we compare the performance of 4 regularisation methods for logistic regression: lasso, dropout, elastic net and DropLasso. As in the simulation study, we fix the probability of dropout to $p = 0.5$ for dropout and DropLasso regularisation. For elastic net, we fix $\alpha = 0.5$ to balance the ℓ_1 and ℓ_2 norms. Finally, for lasso, elastic net and DropLasso, we run the models with 100 values for λ, regularly spaced after log transform between $\lambda_{\min} = 10^{-5}$ and $\lambda_{\max} = 10^5$, on 20% of the data chosen in such way that labels are balanced, and evaluate the performance of the resulting models on the 80% remaining data. We report in Table 2 the maximal test AUC, averaged over the 5 repeats, taken over the grid of λ.

Table 2: Experimental datasets

Dataset	**Classification Task**	**Variables**	**Samples**	**Organism**
EMTAB2805	G1 vs G2M	20 614	96 ; 96	mouse
GSE74596	NKT0 vs NKT17	14 172	44 ; 45	mouse
GSE45719	16-cell stage blastomere vs midblastocyst	21 605	50 ; 60	mouse
GSE63818-GPL16791	Primordial Germ Cells vs Somatic Cells	30 022	26 ; 40	human
GSE48968-GPL13112	BMDC: 1h LPS vs 4h LPS Stimulation	17 947	95 ; 96	mouse
GSE60749-GPL13112	Culture condition: 2i+LIF vs Serum+LIF	39 351	90 ; 94	mouse
GSE81861	Epithelial cells: Tumour (colorectal) vs Normal	36 400	160 ; 160	human

On each of the $7 \times 4 = 28$ resulting dataset, we compare the performance of 4 regularisation methods for logistic regression: lasso, dropout, elastic net and DropLasso. As in the simulation study, we fix the probability of dropout to $p = 0.5$ for dropout and DropLasso regularisation. For elastic net, we fix $\alpha = 0.5$ to balance the ℓ_1 and ℓ_2 norms. Finally, for lasso, elastic net and DropLasso, we run the models with 100 values for λ, regularly spaced after log transform between $\lambda_{\min} = 10^{-5}$ and $\lambda_{\max} = 10^5$, on 20% of the data chosen in such way that labels are balanced, and evaluate the performance of the resulting models on the 80% remaining data. We report in Table 2 the maximal test AUC, averaged over the 5 repeats, taken over the grid of λ.

The first observation is that the performances reached by all methods on all datasets are generally very high, and can reach an AUC above 0.9 on each of the 7 datasets. This suggests that the labels chosen in these datasets are sufficiently different in terms of transcriptomic profiles that they can be easily recognised most of the time. We still notice some differences in performance between datasets, with GSE60749-GPL13112 being

Table 3: Best average test AUC score for lasso, dropout, elastic net and DropLasso regularised logistic regression on 7 datasets and their 3 reduced datasets. $*$ and $**$ indicate when the performance of DropLasso is significantly better than that of elastic net ($P < 0.05$ and $P < 0.01$, respectively)

Dataset	Number of variables	LASSO	Dropout	Elastic net	DropLasso
EMTAB2805	100	0.95	0.94	**0.966**	0.964
	1 000	0.956	0.989	0.980	**0.990** *
	10 000	0.764	0.961	0.817	**0.961** *
	All (20 614)	0.72	0.928	0.796	**0.946** **
GSE74596	100	0.997	0.996	0.994	**0.998**
	1 000	0.988	0.997	0.994	**0.999**
	10 000	0.769	0.960	0.909	**0.990***
	All (14 172)	0.844	0.915	0.943	**0.966**
GSE45719	100	0.999	0.990	0.999	**0.999**
	1 000	0.997	0.999	0.999	**1**
	10 000	0.995	0.998	0.998	**1** *
	All	0.990	0.999	0.999	**1**
GSE63818-GPL16791	100	0.94	0.977	0.984	**0.998** *
	1 000	0.945	0.998	0.985	**1** *
	10 000	0.951	0.995	0.987	**0.998** *
	All	0.932	0.970	0.976	**0.989**
GSE48968-GPL13112	100	0.995	0.992	0.996	**0.997**
	1 000	0.962	0.992	0.996	**0.997**
	10 000	0.939	0.97	0.978	**0.992** *
	All	0.948	0.962	0.96	**0.987** *
GSE60749-GPL13112	100	**1**	**1**	**1**	**1**
	1 000	**1**	0.998	**1**	**1**
	10 000	**1**	0.999	**1**	**1**
	All	**1**	0.997	**1**	**1**
GSE81861	100	0.852	0.808	0.854	**0.887** *
	1 000	0.89	0.915	0.898	**0.933** *
	10 000	0.827	0.9	0.881	**0.927** **
	All	0.8	0.851	0.84	**0.9** **

the easiest one while EMTAB2805 is the most challenging, for all methods. Soneson and Robinson (2017) also noticed a difference in signal-to-noise ratio between these datasets, in the context of gene differential analysis. Second, we observe that the best performance is obtained by DropLasso on 27 out of the 28 datasets. The difference between the 4 regularisers is visualised in the box plots on Figure 2, which summarise the AUC values over the 7 experiments with all genes for each method. In 14 of the 28 experiments, DropLasso significantly outperforms elastic net at significance level $P < 0.05$, and in 3 of these cases the significance level is $P < 0.01$. This confirms that on real scRNA-seq data, DropLasso also brings a consistent benefit over dropout of elastic net regularisation. Finally, regarding the impact of the number of features selected, we observe that in most datasets and for most methods the best performance is reached for 100 or 1,000 genes. Remembering that genes are only selected based on their variance, independently of any class label information, this suggests that all method suffer in the high-dimensional regime and that a simple pre-filtering of genes can help by reducing the dimension of the problem. This also indirectly confirms the importance of regularisation in high dimension, and the relevance of developing adequate regularisers incorporating prior knowledge about the data and their noise.

To conclude this section, we now compare the lists of genes in the molecular signatures estimated by elastic net and DropLasso regularisation. We illustrate this comparison on the first dataset, EMTAB2805, where the goal is to discriminate mouse cells at the G1 from the G2M cell cycle stages. To this end, we retrain the different methods with the tuning parameter λ corresponding to the best accuracy on the 1000-filtered datasets with all the samples, and then we perform a bioinformatic analysis of Gene Ontology annotations using DAVID (Huang *et al.*, 2009) on the subset of genes with non-zero coefficients for each method.

For this dataset and the best tuning parameters, DropLasso selects 186 variables while elastic net only selects 48 variables. The analysis of the selected genes shows enrichment in the functional terms "cell division", "cell cycle" and "mitosis" for both methods. For DropLasso, 21 genes are related to the functional

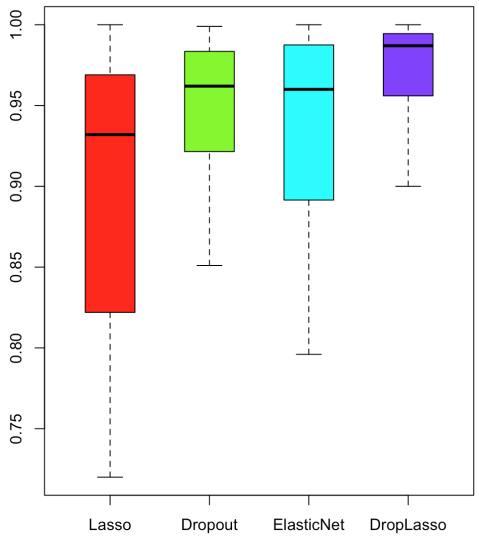

Figure 2: Average CV-AUC for all 7 datasets

term "cell division". Among those, one can find CDC28, CDC23 and CDca8 which were not selected by elastic net, that only selected 10 genes related to cell division. Similarly 30 genes that have previously been annotated with cellular processes consistent with the "cell cycle" term are selected in DropLasso versus 9 in elastic net, and 22 genes related to "mitosis" versus 10 in elastic net. Therefore, DropLasso could potentially allow for the discovery of more functionally related genes in its signature than elastic net.

4 Discussion

ScRNA-seq is changing the way we study cellular heterogeneity and investigate a number of biological process such as differentiation or tumourigenesis. Yet, as the throughput of scRNA-seq technologies increases and allows to process more and more cells simultaneously, it is likely that the amount of information captured in each individual cell will remain limited in the future and that dropout noise will continue to affect scRNA-seq (and other single-cell technologies).

Several techniques have been proposed to handle dropout noise in the context of data normalisation or gene differential expression analysis, and shown to outperform standard techniques widely used for bulk RNA-seq data analysis. In this paper we investigate a new setting which, we believe, will play an important role in the future: supervised classification of cell populations into pre-specified classes, and selection of molecular signatures for that purpose. Molecular signatures for the classification of tissues from bulk RNA-seq data has already had a tremendous impact in cancer research, and as more and more cell types are investigated and discovered with scRNA-seq it is likely that specific molecular signatures will be useful in the future to automatically sort cells into their classes.

DropLasso, the new technique we propose, borrows the recent idea of dropout regularisation from machine learning, and extends it to allow feature selection. While a parallel between dropout regularisation and (data-dependent) ridge regression has already been shown by Wager *et al.* (2013) and Baldi and Sadowski (2013), it is reassuring that we are able to extend this parallel to DropLasso and elastic net regularisation.

More interesting is the fact that, on both simulated and real data, we obtained promising results with DropLasso. They suggest that, again, specific models tailored to the data and noise give an edge over generic models developed under different assumptions.

The intuition behind why dropout (and DropLasso) perform well on scRNA-seq data, however, remains a bit unclear. Our main motivation to use them in this context was to see them as data augmentation techniques, where training data are corrupted according to the noise we assume in the data. While we believe this is fundamentally the reason why we obtained promising results, alternative explanations for the

success of dropout have been proposed, and may also play a role in the context of scRNA-seq. They include for example the interpretation of dropout as a regulariser similar to a data-dependent weighted version of ridge regularisation, which works well in the presence of rare but important features (Wager *et al.*, 2013); it would be interesting to clarify if the regularisation induced by DropLasso on scRNA-seq data exploits some fundamental property of these data, and may be replaced by a more direct approach to model this.

Finally, this first study of dropout and DropLasso regularisation on biological data paves the way to many future direction. For example, it is known that the probability of dropout in scRNA-seq data depends on the gene expression level (Kharchenko *et al.*, 2014; Risso *et al.*, 2018). It would therefore be interesting to study both theoretically and empirically if a dropout regularisation following a similar pattern may be useful. Second, instead of independently perturbing the different features one may create a correlation between the dropout events in different genes. Creating a correlation may be a way to create new regularisation by generating a *structured* dropout noise. It may for example be possible to derive a correlation structure for dropout noise from prior knowledge about gene annotations or gene networks in order to enforce a structure in the molecular signature, just like structured ridge and lasso penalties have been used to promote structure molecular signatures with bulk transcriptomes (Rapaport *et al.*, 2007; Jacob *et al.*, 2009).

Funding

This work has been supported by the European Research Council (grant ERC-SMAC-280032).

References

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2011). Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, **4**(1), 1–106.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, **17**(63).

Baldi, P. and Sadowski, P. J. (2013). Understanding dropout. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Adv. Neural. Inform. Process Syst.*, pages 2814–2822. Curran Associates, Inc.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, **24**(2), 123–140.

Breiman, L. (2001). Random forests. *Mach. Learn.*, **45**(1), 5–32.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**(6167), 193–6.

Haury, A.-C. and Vert, J.-P. (2010). On the stability and interpretability of prognosis signatures in breast cancer. In *Proceedings of the Fourth International Workshop on Machine Learning in Systems Biology (MLSB10)*. To appear.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression : biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucl. Acids Res.*, **37**, 1–13.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, New York, NY, USA. ACM.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**(7), 740–742.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Molecular Cell*, **58**(4), 610–620.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Adv. Neural. Inform. Process Syst.*, volume 25, pages 1097–1105. Curran Associates, Inc.

Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., Kong, S. L., Chua, C., Hon, L. K., and Tan, W. S. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.*, **49**(5), 708.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**(5), 1202–1214.

Ozsolak, F. and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., and Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**(6190), 1396–1401.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, **406**(6797), 747–752.

Pierson, E. and Yau, C. (2015). Dimensionality reduction for zero-inflated single cell gene expression analysis. *Genome Biol.*, **16**(241).

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E., and Golub, T. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA*, **98**(26), 15149–15154.

Rapaport, F., Zynoviev, A., Dutreix, M., Barillot, E., and Vert, J.-P. (2007). Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Comm.*, **9**(1), 284.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier.

Schölkopf, B., Burges, C., and Vapnik, V. (1996). Incorporating invariances in support vector learning machines. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *ICANN 96: Proceedings of the 1996 International Conference on Artificial Neural Networks*, pages 47–52, London, UK. Springer-Verlag.

Soneson, C. and Robinson, M. D. (2017). Bias, robustness and scalability in differential expression analysis of single-cell RNA-seq data. Technical Report 143289, bioRxiv.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lønning, P., and Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA*, **98**(19), 10869–10874.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C., Lnning, P., Brown, P., Brresen-Dale, A., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA*, **100**(14), 8418–8423.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**(1), 1929–1958.

Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., Levi, B., Gray, L. T., Sorensen, S. A., Dolbeare, T., Bertagnolli, D., Goldy, J., Shapovalova, N., Parry, S., Lee, C., Smith, K., Bernard, A., Madisen, L., Sunkin, S. M., Hawrylycz, M., Koch, C., and Zeng, H. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.*, **19**(2), 335–346.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**(1), 267–288.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**(25), 1999–2009.

van der Maaten, L., Chen, M., Tyree, S., and Weinberger, K. Q. (2013). Learning with marginalized corrupted features. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, number 28 in JMLR Proceedings, pages 410–418. JMLR.org.

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., *et al.* (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, **356**(6335), eaah4573.

Wager, S., Wang, S., and Liang, P. S. (2013). Dropout training as adaptive regularization. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Adv. Neural. Inform. Process Syst.*, volume 26, pages 351–359. Curran Associates, Inc.

Wager, S., Fithian, W., Wang, S., and Liang, P. S. (2014). Altitude training: Strong bounds for single-layer dropout. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Adv. Neural. Inform. Process Syst.*, pages 100–108. Curran Associates, Inc.

Zeisel, A., Manchado, A. B. M., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., and Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**(6226), 1138–42.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.