



**HAL**  
open science

# Differentiable Feature Selection, a Reparameterization Approach

Jérémie Dona, Patrick Gallinari

► **To cite this version:**

Jérémie Dona, Patrick Gallinari. Differentiable Feature Selection, a Reparameterization Approach. ECML PDKK 2021, Sep 2021, Virtual, France. hal-03291389

**HAL Id: hal-03291389**

**<https://hal.science/hal-03291389>**

Submitted on 20 Jul 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Differentiable Feature Selection, a Reparameterization Approach

Jérémie Donà <sup>1</sup> and Patrick Gallinari<sup>1,2</sup>

<sup>1</sup> Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

<sup>2</sup> Criteo AI Labs, Paris, France

`firstname.lastname@lip6.fr`

**Abstract.** We consider the task of feature selection for reconstruction which consists in choosing a small subset of features from which whole data instances can be reconstructed. This is of particular importance in several contexts involving for example costly physical measurements, sensor placement or information compression. To break the intrinsic combinatorial nature of this problem, we formulate the task as optimizing a binary mask distribution enabling an accurate reconstruction. We then face two main challenges. One concerns differentiability issues due to the binary distribution. The second one corresponds to the elimination of redundant information by selecting variables in a correlated fashion which requires modeling the covariance of the binary distribution. We address both issues by introducing a relaxation of the problem via a novel reparameterization of the logitNormal distribution. We demonstrate that the proposed method provides an effective exploration scheme and leads to efficient feature selection for reconstruction through evaluation on several high dimensional image benchmarks. We show that the method leverages the intrinsic geometry of the data, facilitating reconstruction.

**Keywords:** Representation Learning · Sparse Methods

## 1 Introduction

Learning sparse representations of data finds essential real-world applications as in budget learning where the problem is limited by the number of features available or in embedded systems where the hardware imposes computational limitations. Feature selection serves similar objectives giving insights about variable dependencies and reducing over-fitting [7]. Combined with a reconstruction objective, feature selection is a sensible problem when collecting data is expensive which is often the case with physical processes. For example, consider optimal sensor placement. This task consists in optimizing the location of sensors measuring a scalar field over an area of interest (e.g pressure, temperature) to enable truthful reconstruction of the signal on the whole area. It finds applications in climate science [13, 12], where key locations are monitored to evaluate the impact of climate change on snow melt and Monsoon. These examples illustrate

how feature selection for reconstruction may be critically enabling for large scale problems where measurements are costly.

Common practices for feature selection involves a  $\ell_1$ -regularization over the parameters of a linear model to promote sparsity [17]. Initiated by [17], several refinements have been developed for feature selection. For example, [16] employs a  $\ell_{2,1}$ -norm in a linear auto-encoder. [8] impose a  $\ell_1$ -penalty on the first layer of a deep auto-encoder to select features from the original signal. Finally, Group-Lasso methods extended lasso by applying the sparse  $\ell_1$ -penalty over precomputed chunks of variables to take prior knowledge into account while selecting features. These approaches suffer from two main limitations: the design of the groups for Group-Lasso methods and the loss of the intrinsic structure of the data as both [16, 8] treat the input signal as a vector. Moreover, non-linear  $\ell_1$  based methods for feature selection and reconstruction are intrinsically ill posed, see section 6.5. Like Group-Lasso methods, our proposition aims at selecting variables in a correlated fashion, to eliminate redundant information, while leveraging the structure of the data. We illustrate its efficiency on images but it can be adapted to exploit patterns in other types of structured data as graphs.

We propose a novel sparse embedding method that can tackle feature selection through an end-to-end-approach. To do so, we investigate the learning of binary masks sampled from a distribution over binary matrices of the size of the image, with 1 indicating a selected pixel. We alleviate differentiability issues of learning categorical variables by relying on a continuous relaxation of the problem. The learned latent binary distribution is optimized via a stochastic exploration scheme. We consider the dependency between the selected pixels and we propose to sample the pixels in the mask in a correlated fashion to perform feature selection efficiently. Accordingly, we learn a correlated logitNormal distribution via the reparameterization trick allowing for an efficient exploration of the masks space while preserving structural information for reconstruction. Finally, sparsity in the embedding is enforced via a relaxation of the  $\ell_0$ -norm. To summarize, we aim at learning a binary mask for selecting pixels from a distribution of input signals  $x$ , with  $x \in \mathbb{R}^{n \times n}$  for images, enabling an accurate reconstruction. We formulate our problem as learning jointly a parametric sampling operator  $S$  which takes as input a random variable  $z \in \mathcal{Z} \subseteq \mathbb{R}^d$  and outputs binary masks, i.e.  $S : \mathcal{Z} \rightarrow \{0, 1\}^{n \times n}$ . We introduce two ways to learn the sampling operator  $S$ . For reconstruction, an additional operator denoted  $G$  learns to reconstruct the data  $x$  from the sparse measurements  $s \odot x$ . Our proposed approach is fully differentiable and can be optimized directly via back-propagation. Our main contributions are:

- We introduce a correlated logitNormal law to learn sparse binary masks, optimized thanks to the reparameterization trick. This reparameterization is motivated statistically. Sparsity is enforced via a relaxed  $\ell_0$ -norm.
- We formulate the feature selection task for 2-D data as the joint learning of a binary mask and a reconstruction operator and propose a novel approach to learn the parameters of the considered logitNormal law.

- We evidence the efficiency of our approach on several datasets: Mnist, CelebA and a complex geophysical dataset.

## 2 Related Work

Our objective of learning binary mask lies in between a few major domains: density modeling, feature selection and compressed sensing.

**Density Modeling via Reparameterization** Sampling being not differentiable, different solutions have been developed in order to estimate the gradients of the parameters of a sampling operator. Gradient estimates through score functions [9, 4] usually suffer from high variance or bias. Reparameterization [24] provides an elegant way to solve the problem. It consists in sampling from a fixed distribution serving as input to a parametric transformation in order to obtain both the desired distribution and the gradient with respect to the parameters of interest. However, the learning of categorical variables remains tricky as optimizing on a discrete set lacks differentiability. Continuous relaxation of discrete variables enables parameters optimization through the reparameterization trick. Exploiting this idea, [37, 44] developed the concrete law as a reparameterization of the Gumbel max variable for sampling categorical variables [32]. Alternative distributions, defining relaxations of categorical variables can be learned by reparameterization such as the Dirichlet or logitNormal distribution [29, 26]. Nonetheless, most previous approaches learn factorized distribution, thus selecting variables independently when applied to a feature selection task. In contrast, we rely on the logitNormal distribution to propose a reparameterization scheme enabling us to sample the mask pixels jointly, taking into account dependencies between them and exploiting the patterns present in 2-D data.

**Feature Selection** Wrapper methods, [7, 10, 19] select features for a downstream task whereas filter methods [21, 20, 27] rank the features according to tailored statistics. Our work belongs to the category of *embedded* methods, that address selection as part of the modeling process.  $\ell_1$ -penalization over parameters, as for instance in Lasso and in Group Lasso variants [34, 11, 25], is a prototypical embedded method.  $\ell_1$ -penalty was used for feature selection for example in [36, 16] learning a linear encoding with a  $\ell_{2,1}$ -constraint for a reconstruction objective. Auto-encoders [22] robustness to noise and sparsity is also exploited for feature selection [15, 38, 28]. For example, AEFS [8] extends Lasso with non linear auto-encoders, generalizing [36]. Another line of work learns embeddings preserving local properties of the data and then find the best variables in the original space to explain the learned embedding, using either  $\ell_1$  or  $\ell_{2,1}$  constraints [5, 14]. Closer to our work, [1] learn a matrix of weights  $m$ , where each row follow a concrete distribution [37]. That way each row of matrix  $m$  samples one feature in  $x$ . The obtained linear projection  $m.x$  is decoded by a neural network, and  $m$  is trained to minimize the  $\ell_2$ -loss between reconstructions and

targets. Because  $x$  is treated as a vector, here too, the structure of the data is ignored and lost in the encoding process. Compared to these works, we leverage the dependencies between variables in the 2-D pixel distribution, by sampling binary masks via an adaptation of the logitNormal distribution.

**Compressed Sensing** Our work is also related to compressed sensing (CS) where the objective is to reconstruct a signal from limited (linear) measurements [41]. Deep learning based compressed sensing algorithms have been developed recently: [40] use a pre-trained generative model and optimize the latent code to match generated measurements to the true ones; The measurement process can be optimized along with the reconstruction network as in [42]. Finally, [35] use a CS inspired method based on the pivots of a QR decomposition over the principal components matrix to optimize the placement of sensors for reconstruction, but scales poorly for large datasets. Our approach differs from CS. Indeed, for CS, measurements are linear combinations of the signal sources, whereas we consider pixels from the original image. Thus, when CS aims at reconstructing from linear measurements, our goal is to preserve the data structural information to select a minimum number of variables for reconstruction.

### 3 Method

We now detail our framework to learn correlated sparse binary masks for feature selection and reconstruction through an end-to-end approach. The choice of the logitNormal distribution, instead of the concrete distribution [37], is motivated by the simplicity to obtain correlated variables thanks to the stability of independent Gaussian law by addition as detailed below. We experimentally show in section 4 that taking into account such correlations helps the feature selection task. This section is organised as follows : we first introduce in section 3.1 some properties of the logitNormal distribution and sampling method for this distribution. We detail in section 3.2 our parameterization for the learning of the masks distribution. Finally, in section 3.3 we show how to enforce sparsity in our learned distribution before detailing our reconstruction objective in section 3.4.

#### 3.1 Preliminaries: logitNormal Law on $[0, 1]$

Our goal is to sample a categorical variable in a differentiable way. We propose to parameterize the sampling on the simplex by the logitNormal law, introduced in [3]. We detail this reparameterization scheme for the unidimensional case since we aim at learning binary encodings. It can be generalized to learn k-dimensional one-hot vector, see supplementary materials section 6.1. Let  $z \sim \mathcal{N}(\mu, \sigma)$ , and  $Y$  defined as:

$$Y = \text{sigmoid}(z) \tag{1}$$

Then  $Y$  is said to follow a logitNormal law. This distribution defines a probability over  $[0, 1]$ , admits a density and its cumulative distribution function has an analytical expression used to enforce sparsity in section 3.3.

This distribution can take various forms as shown in fig. 1 and be flat as well as bi-modal. By introducing a temperature in the sigmoid so that we have,  $\text{sigmoid}_\lambda(z) = \frac{1}{1+\exp^{-z/\lambda}}$ , we can polarize the logitNormal distribution. In Proposition 1 we evidence the link between the 0-temperature logitNormal distribution and Bernoulli distribution:

**Proposition 1 (Limit Distribution).** *Let  $W \in \mathbb{R}^n$  be a vector and  $b \in \mathbb{R}$  a scalar. Let  $Y = \text{sigmoid}_\lambda(W \cdot z^T + b)$ , where  $z \sim \mathcal{N}(0, I_n)$ , when  $\lambda$  decrease towards 0,  $Y$  converges in law towards a Bernoulli distribution and we have:*

$$\lim_{\lambda \rightarrow 0} \mathbb{P}(Y = 1) = 1 - \Phi\left(\frac{-b}{\sqrt{\sum_i w_i^2}}\right) \quad (2)$$

$$\lim_{\lambda \rightarrow 0} \mathbb{P}(Y = 0) = \Phi\left(\frac{-b}{\sqrt{\sum_i w_i^2}}\right) \quad (3)$$

Where  $\Phi$  is the cumulative distribution function of the Normal law  $\mathcal{N}(0, 1)$ ,

The proof is available in supplementary, section 6.3. Proposition 1 characterizes the limit distribution as the temperature goes down to 0, and  $Y$  defines a differentiable relaxation of a Bernoulli variable. This proposition is used to remove randomness in our learned mask distribution, see section 4.

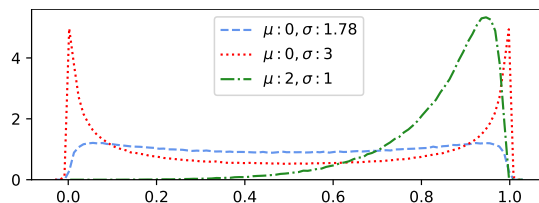


Fig. 1: Density of the logitNormal law for various couple  $(\mu, \sigma)$ :  $(\mu = 0, \sigma = 1.78)$  (dashed line),  $(\mu = 0, \sigma = 3)$  (dotted line)  $(\mu = 2, \sigma = 1)$  (dotted and dashed).

We relax the objective of learning of a binary mask in  $\{0, 1\}$  by learning in  $[0, 1]$  using the logitNormal law. Let  $m \in \mathbb{N}$ , be the dimension of the desired logitNormal variable  $Y$ . A simple solution for learning the logitNormal distribution of the masks is via independent sampling.

**Independent Sampling** A common assumption is that the logitNormal samples originate from a factorized Normal distribution [26]. Thus, the learned parameters of the distribution are: the average  $\mu \in \mathbb{R}^m$  and the diagonal coefficients of the covariance matrix  $\sigma \in \mathbb{R}^m$ , according to:

$$Y = \text{sigmoid}_\lambda(\mu + z \odot \sigma) \quad (4)$$

where  $\odot$  is the element-wise product and  $Y \in \mathbb{R}^m$ . Note that, for feature selection on images, one aims at learning a binary mask and thus the latent space has the same dimension as the images, i.e.  $m = n \times n$ , then  $z \in \mathbb{R}^{n \times n}$ .

This sampling method has two main drawbacks. First, the coordinates of  $z$  are independent and so are the coordinates of  $Y$ , therefore such sampling scheme does not take correlations into account. Also, the dimension of the sampling space  $\mathcal{Z}$  is the same as  $Y$  which might be prohibitive for large images.

We address both limitations in the following section, by considering the relations between the pixel values. In that perspective, Group-Lasso selects variables among previously designed group of variables [34], reflecting different aspects of the data. Similarly, we want to select variables evidencing different facets of the signal to be observed. Indeed, finding the best subset of variables for the reconstruction implies to eliminate the redundancy in the signal and to explore the space of possible masks. We propose to do so by selecting the variables in a correlated fashion, avoiding the selection of redundant information.

**Correlated Sampling:** To palliate the limitations of independent sampling, we model the covariance between latent variables by learning linear combinations between the variables in the prior space  $\mathcal{Z}$ . Besides, considering dependencies between latent variables, this mechanism reduces the dimension of the sampling space  $\mathcal{Z}$ , allowing for a better exploration of the latent space. In order to generate correlated variables from a lower dimensional space, we investigate the following transformation: let  $z \sim \mathcal{N}_d(0, I_d) \in \mathcal{Z} = \mathbb{R}^d$  with  $d \ll m$ ,  $W \in \mathcal{M}_{m,d}(\mathbb{R})$  a weight matrix of size  $m \times d$  and  $b \in \mathbb{R}^m$  a real vector, then

$$Y = \text{sigmoid}_\lambda(Wz + b) \quad (5)$$

represents  $m$ -one dimension logitNormal laws due to the stability of independent Gaussian laws by addition. However, the Normal law induced by  $Wz + b$  has now a full covariance matrix and not only diagonal coefficient as in eq. (4). This reparameterization provides a simple way to sample correlated (quasi)-binary variables, even for high dimension latent space, i.e with  $m$  large.

Compared to [1], our proposition offers a significant advantage for feature selection in images. Indeed, let  $G$  be the neural network aiming to reconstruct data  $x$  from the selected variable. With our proposition  $G$  can access a sparse version of the original signal  $Y \odot x$  and can thus leverage both the pixel values and their position in the image for reconstruction. In [1] only the selected feature values without structural information are available for the reconstruction.

### 3.2 Parameterizing logitNormal Variables for Feature Selection

Now we have established how to compute correlated logitNormal variables following eq. (5), we detail our parameterization for learning. Let  $S : \mathcal{Z} \rightarrow [0, 1]^{n \times n}$  be our sampling operator that generates a binary mask from a random sample  $z$ . We consider two approaches to parameterize  $S$  so that it follows a logitNormal law. Our first proposition denoted vanilla parameterization directly optimizes  $W$  and  $b$  from eq. (5), while our second approach proposes to explore and optimize the spaces of linear combinations  $W$  and biases  $b$ .

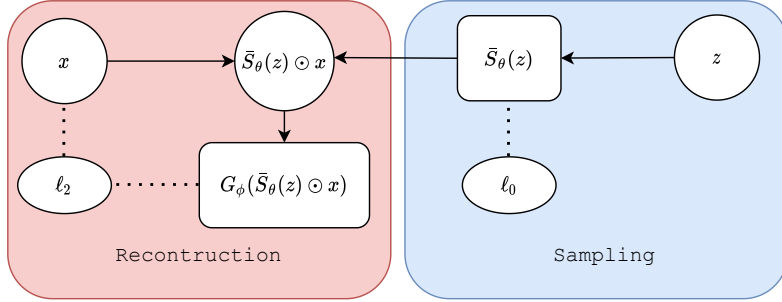


Fig. 2: Algorithmic flow of our framework for feature selection for reconstruction.  $S_\theta(z)$  has a correlated logitNormal distribution. We sample  $z \sim \mathcal{N}(0, 1)$ .  $S_\theta(z)$  defines the binary masks and  $G_\phi$  estimate  $x$  from  $x^{obs} = \bar{S}_\theta(z) \odot x$ .

**Vanilla Parameterization:** A simple approach is to parameterize  $S$  as  $S_\theta$  according to eq. (5). Then, the optimized parameters are:  $\theta = (W, b)$  with  $W \in \mathcal{M}_{n \times n, d}(\mathbb{R})$  and  $b \in \mathbb{R}^{n \times n}$ . This sampling process can be summarized by eq. (6):

$$\begin{cases} \text{Initialize } W \in \mathcal{M}_{n \times n, d}(\mathbb{R}), b \in \mathbb{R}^{n \times n} \\ z \sim \mathcal{N}(0, I_d) \end{cases} \quad (6a)$$

$$S_\theta(z) = \text{sigmoid}(W \cdot z + b) \quad (6b)$$

In that case each variable in  $S_\theta(z)$  follows a logitNormal law. The selected variables are indicated for  $S_\theta = 1$ . The optimization process allows two degrees of freedom ( $b$  and  $W$ ) for the control of the variance, of the covariance and of the average of the variables of the masks. Note that, this parameterization corresponds to a linear layer followed by a sigmoid activation so that besides tractability for the distribution of  $Y$ , it presents the advantage of a simple implementation. Unlike [1], our proposition preserves the structure of the data.

**HyperNetworks Parameterization:** Aiming to learn a matrix  $W$  and a bias vector  $b$  that fully characterizes our logitNormal law as eq. (5), we leveraged in eq. (6) the stability of independent Gaussian law by addition. However, the space of the linear combinations to be learned is high dimensional and structured, hence hard to learn. Also, the optimization of the parameterization as eq. (6) is highly dependent on the initialization, as we optimize  $W$  and  $b$  from a (randomly) chosen start point. Therefore, we want to be able to reach a wider space of parameters  $W, b$ . To do so, we build on [45] that successfully leverages latent code pre-processing with neural network in the context of adversarial learning for image generation, and [46] where a neural network generates the weights of another neural network to facilitate learning. Therefore, instead of learning directly  $W, b$  as in the vanilla approach we propose to learn to sample on the space of linear combination  $W$  and biases  $b$ . The core idea is to leverage neural networks expressivity to enrich the space of reachable matrices  $W$  and vectors  $b$  compared to the vanilla approach. To do so we use the random sample  $z$  to



extract a representation vector  $r \in \mathbb{R}^k$ . This representation  $r$  serves as input to neural networks  $F_b, F_W$  providing estimates of  $W$  and  $b$ . To sum up, in the HyperNetwork approach we learn a logitNormal law according to:

$$\begin{cases} z \sim \mathcal{N}(0, I_d), r = F_{rep}(z) \in \mathbb{R}^k, & (7a) \\ W = F_W(r) \in \mathcal{M}_{n \times n, d}(\mathbb{R}), & (7b) \\ b = F_b(r) \in \mathbb{R}^{n \times n}, \text{ and finally:} & (7c) \\ S_\theta(z) = \text{sigmoid}(F_b(r) + F_W(r).z), & (7d) \end{cases}$$

Note that as desired  $S_\theta(z)$  follows a logitNormal law  $\mathcal{LN}(F_b(r), F_W(r)^T.F_W(r))$ .

This proposition presents several advantages. First, in eq. (6)  $W$  is a randomly initialized weight matrix, then we only explore one trajectory of optimization from this (randomly chosen) starting point. Also, instead of learning a distribution of masks, this parameterization learns a distribution of transport matrices and biases. Therefore, both  $F_W$  and  $F_b$  stochastically explore a direction for each sample of  $z$ , providing more feedback with respect to the objective of feature selection for reconstruction. This parameterization of  $W$  and  $b$  offers a way to explore efficiently the space of biases and linear combinations. Also, because it rely on matrix multiplication, this procedure is computationally barely less efficient than the naive one when  $F_W$  and  $F_b$  are small neural networks.

We show experimentally the superiority of this approach in section 4.

### 3.3 Sparsity Constraint: $\ell_0$ -Relaxation

We detail our approach promoting sparsity. Frequently, sparsity in regression settings is enforced thanks to a  $\ell_1$  penalty on the parameters. However,  $\ell_1$  approaches may suffer from a shrinking effect due to ill-posedness as detailed in section 6.5. Consequently, we introduce an alternative approximation of the  $\ell_0$ -formulation better suited to our feature selection application: we minimize the expected  $\ell_0$ -norm, i.e the probability of each variable in our binary mask to be greater than 0. Thus, we need a non zero probability of sampling 0 which is not the case with the current scheme. Accordingly, we introduce a stretching scheme to obtain a non-zero mass at points 0 and 1 while maintaining differentiability.

**Stretched Distribution** To create a mass at 0, we proceed as in [31]. Let  $Y \in [0, 1]^m$  be a logitNormal variable,  $\gamma < 0$  and  $\eta > 1$  and  $HT$  be the hard-threshold function defined by  $HT(Y) = \min(\max(Y, 0), 1)$ , the stretching is defined as:

$$\bar{Y} = HT\{(\eta - \gamma)Y + \gamma\} \quad (8)$$

Thanks to this stretching of our distribution, we have a non zero probability to be zero, i.e  $\mathbb{P}(\bar{Y} = 0) > 0$  and also  $\mathbb{P}(\bar{Y} = 1) > 0$ . Further details are available in supplementary section 6.6. We can now derive a relaxed version of the  $\ell_0$ -norm penalizing the probability of the coordinates of  $\bar{Y}$  to be greater than 0.

**Sparsity Constraint:** Let  $L_0(\bar{Y})$  the expected  $\ell_0$ -norm of our stretched output  $\bar{Y}$ . Using the notation in eq. (8), we have:

$$L_0(\bar{Y}) = \mathbb{E}[\ell_0(\bar{Y})] = \sum_{i=1}^m \mathbb{P}(\bar{Y}_i > 0) = \sum_{i=1}^m 1 - F_Y\left(\frac{-\gamma}{\eta - \gamma}\right), \quad (9)$$

where  $F_Y$  denotes the cumulative distribution function (CDF) of  $Y$ . This loss constrains the random variable  $Y$  to provide sparse outputs as long as we can estimate  $F_Y$  in a differentiable way. In the case of the logitNormal law, we maintain tractability as  $Y$  satisfies eq. (5) or eq. (4). Thus, for our  $m$ -dimensional logitNormal law defined as in eq. (5), we have:

$$L_0(\bar{Y}) = \sum_i 1 - \Phi\left(\frac{\log\left(\frac{-\gamma}{\eta}\right) - b}{\sqrt{\sum_j W_{j,i}^2}}\right), \quad (10)$$

where  $\Phi$  is the CDF of the unitary Normal law. Detailed computations are available in supplementary materials section 6.4. Minimizing eq. (10) promotes sparsity in the law of  $Y$  by minimizing the expected true  $\ell_0$ -norm of the realisation of the random variable  $Y$ . We have developed a constraint that promotes sparsity in a differentiable way. Now we focus on how to learn efficiently the parameters of our correlated logitNormal law.

### 3.4 Reconstruction for Feature Selection

We have designed a sparsity cost function and detailed our parameterization to learn our sampling operator, we focus on the downstream task. Consider data  $(x_i, y_i)_{i \in [1..N]}$ , consisting in paired input  $x$  and output  $y$ . Feature selection consists in selecting variables in  $x$  with a mask  $s$ , so that the considered variables:  $s \odot x$  explain at best  $y$ . Let  $G$  be a prediction function and  $\mathcal{L}$  a generic cost functional, feature selection writes as:

$$\min_{s, f} \mathbb{E}_{x, y} \mathcal{L}(G(s \odot x), y) \quad \text{s.t. } \|s\|_0 < \lambda, \quad (11)$$

In this work we focus on a reconstruction as final task, i.e  $y = x$ . Besides the immediate application of such formulation to optimal sensors placement and data compression, reconstruction as downstream task requires no other source of data to perform feature selection. Naturally, this framework is adaptable to classification tasks. As a sparse auto-encoding technique, feature selection with a reconstruction objective aims at minimizing the reconstruction error while controlling the sparsity. In this case  $G_\phi : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$  is our reconstruction network (of parameter  $\phi$ ) taking as inputs the sparse image. The feature selection task with an  $\ell_2$ -auto-encoding objective writes as:

$$\min_{\theta, \phi} \mathbb{E}_x \|G_\phi(\bar{S}_\theta(z) \odot x) - x\|_2 + \lambda_{sparse} L_0(\bar{S}_\theta(z)) \quad (12)$$

A schematic view of our proposition, illustrating the sampling and the reconstruction component is available in Figure 2. An algorithmic description in the vanilla case (eq. (6)) is available in supplementary section 6.7.

## 4 Experiments

We provide experimental results on 3 datasets: MNIST, CelebA and a geophysical dataset resulting from complex climate simulations [43, 33]. We use the traditional train-test split for MNIST and a 80-20 train-test split for the other datasets. The geophysical dataset is composed of surface temperatures anomalies (deviations between average temperature at each pixel for a reference period and observations) and contains 21000 samples (17000 for train). The data have both high (Gulf stream, circum-polar current ...) and low frequencies (higher temperature in the equatorial zone, difference between northern and southern hemispheres ...) that need to be treated accurately due to their influence on the Earth climate. Accuracy in the values of reconstructed pixel is then essential for the physical interpretation. These dense images represent complex dynamics and allow us to explore our method on data with crucial applications and characteristics very different from the digits and faces.

### 4.1 Experimental and Implementation Details

**Baselines** Besides our models Vanilla logitNormal, denoted *VLN*, and its hypernetworks counterpart denoted *HNet-LN*, we consider as competing methods the following approaches:

1. Concrete-Autoencoder [1] denoted *CAE*.
2. To assess the relevance of our correlated proposition, we investigate a binary mask approach based on the independent logitNormal mask that corresponds to equation eq. (4) denoted *ILN*,
3. Another independent binary mask method based on the concrete law [37], see supplementary materials section 6.10, denoted *SCT*.

**Implementation Details** For all binary mask based methods, we use a Resnet for  $G_\phi$ , [30] following the implementation of [23].  $F_{rep}$ ,  $F_W$  and  $F_b$  are two layers MLP with leaky relu activation. For CAE, because the structure of the data is lost in the encoding process, we train  $G_\phi$  as a MLP for MNIST and a DcGAN for geophysical data and CelebA. Thorough experimental details are available in section 6.11. The code is available at: [https://github.com/JeremDona/feature\\_selection\\_public](https://github.com/JeremDona/feature_selection_public)

**Removing Randomness:** All masked based algorithms learn distributions of masks. To evaluate the feature selection capabilities, we evaluate the different algorithms using fixed masks. We rely on proposition 1 to remove the randomness during test time. Let  $S_\theta^0$  be the 0-temperature distribution of the estimated  $S_\theta$ . We first estimate the expected  $\ell_0$ -norm of the 0-temperature distribution:  $L_0(S_\theta^0)$ . We then estimate two masks selecting respectively the  $10 \times \lfloor \frac{L_0(S_\theta^0)}{10} \rfloor$  and  $10 \times \lceil \frac{L_0(S_\theta^0)}{10} \rceil$  most likely features (rounding  $L_0(S_\theta^0)$  up and down to the nearest ten). This method has the advantage of implicitly fixing a threshold in

the learned mask distribution to select or reject features. More details on the method are available in section 6.9.

We now illustrate the advantage of selecting features in a correlated fashion.

#### 4.2 Independent vs Correlated Sampling Scheme:

**Is a Covariance Matrix Learned ?** Because we model the local dependencies in the sampling by learning linear mixing of latent variables  $z$ , we first verify the structure of the covariance matrix. Figure 3 reports the learned covariance matrix of the sampling for MNIST dataset using eq. (6) method. Besides the diagonal, extra-diagonal structures emerge, revealing that local correlations are taken into account to learn the sampling distribution.

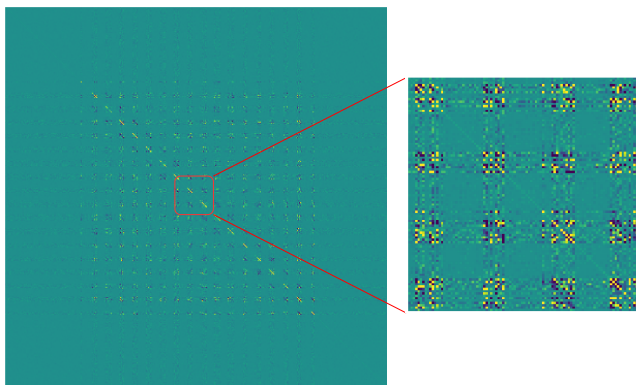


Fig. 3: Covariance matrix learned with eq. (6), with  $\approx 30$  pixels selected. Yellow values indicates high positive covariance, blue ones low negative covariance

**Independent Sampling Does not Choose** We show in fig. 4 the empirical average of the sampled masks for each masked base competing algorithm where all algorithms were trained so that at  $L_0(S_\theta^0) \approx 30$ . Figure 4 clearly shows that concrete base algorithm (SCT) and in a lesser sense (ILN) do not select features, but rather put a uniformly low probability to sample pixels in the center of the image. This means that both algorithms struggle at discriminating important features from less relevant ones. On the other hand, our correlated propositions, Vanilla logitNormal (V-LN, eq. (6)) and particularly the hyper-network approach (HNetL, eq. (7)) manage to sparsify the distribution prioritizing the selection of important pixels for reconstruction.

#### 4.3 Feature Selection and Reconstruction

We now quantitatively estimate the impact of our choices on the reconstruction error on the various datasets. First, the mean squared error reconstruction

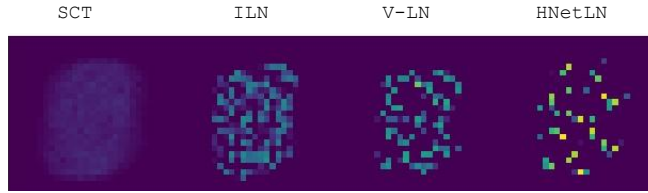


Fig. 4: Masks empirical distribution for competing binary masks algorithms on the MNIST datasets for about 30 features in the sampled mask

results from table 1 tells us that considering the spatial structure of the data enhances reconstruction performance. Indeed, mask based methods consistently over-perform CAE where the data structure is linearized in the encoding process. Furthermore for mask based method, correlated sampling (row V-LN and HNet-LN) also consistently improves over independent sampling based method (row ILN and SCT). Finally, our hyper-network HNet-Ln proposition also improves over the vanilla approach validating our proposition. Samples for all datasets are available in supplementary section 6.12

Table 1: Average Reconstruction Error (MSE) on MNIST, Climate and CelebA datasets for all considered baselines

# Features	MNIST				Climat			CelebA		
	20	30	50	100	200	300	100	200	300	
CAE	3.60	3.05	2.40	2.07	1.98	1.96	7.65	6.42	5.7	
ILN	3.67	2.41	1.41	1.44	1.05	0.83	7.1	2.56	1.87	
SCT	3.72	3.61	2.60	2.20	1.89	1.51	7.99	3.31	2.44	
VLN (Ours)	3.22	2.19	1.33	<b>1.11</b>	<b>0.93</b>	0.79	3.11	1.96	1.50	
HNet-Ln (Ours)	<b>2.15</b>	<b>1.53</b>	<b>1.06</b>	1.78	0.96	<b>0.60</b>	<b>2.81</b>	<b>1.7</b>	<b>1.46</b>	

#### 4.4 Quality of the Selected Features: MNIST Classification

We now assess the relevance of the selected features of our learned masks on another task. To do so, for each learned distribution we train a convolutional neural network, with a DeGAN architecture on MNIST classification task. Here also, the randomness in test set is removed. For each mask we run 5 experiments to account for the variability in the training. Classification results reported in table 2 indicate that both our correlated logitNormal propositions consistently beat all considered baselines, validating our choices to learn a sampling scheme in a correlated fashion. Indeed, our propositions systematically reach the lowest minimum and average classification error.

Table 2: Classification error in percent for MNIST on test set for all considered baselines. Minimum and average are taken over 5 runs.

# Features	20		30		50	
Metric	Min	Mean	Min	Mean	Min	Mean
CAE	24.4	31.64	8.89	19.60	5.45	6.65
ILN	21.58	28.26	7.96	16.63	4.17	5.33
SCT	20.88	32.79	9.49	18.22	4.11	6.77
VLN (Ours)	<b>12.15</b>	<b>24.74</b>	<b>6.38</b>	<b>15.07</b>	3.32	<b>4.67</b>
HNet-LN (Ours)	19.23	25.07	7.24	17.80	<b>2.84</b>	6.45

#### 4.5 Extension: cGAN

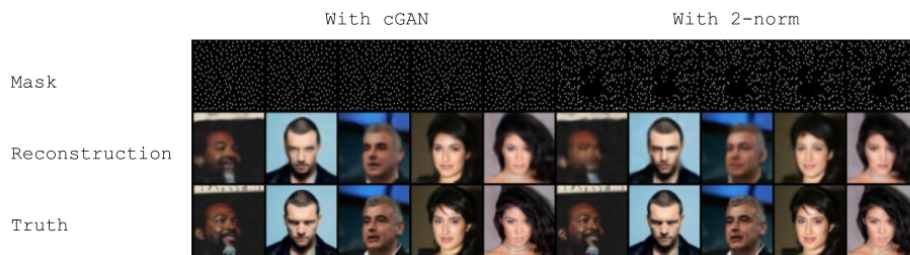


Fig. 5: Examples of masks (first row), reconstructions (second row) and true data (last row) for CelebA dataset using either a cGAN (4 first columns) or simple auto-encoding (4 last columns) for 200 selected features. Best viewed in color.

We detailed in the previous experiments feature selection results obtained thanks to an  $\ell_2$ -auto-encoding approach. This choice was motivated because in physical measurement all points are equals: we don't want to favor the reconstruction of some part of the image while neglecting another. However, for images such as CelebA all points are not equal: the face part of the image being more interesting than the background. Indeed, a realistic reconstruction can be preferred to a well reconstructed background. Moreover,  $\ell_2$ -auto-encoding suffers from blur in the reconstruction. In that perspective, we can leverage conditional generative adversarial networks (cGAN) approaches [39, 23] that solves the blurriness occurring in  $\ell_2$ -decoding. We implement the cGAN approach of [23]. Figure 5 illustrates that despite both method show good reconstruction, the cGAN approach on CelebA enables a stronger focus on faces facilitating realistic reconstruction. We refer to section 6.13 for more details and samples.

## 5 Conclusion

In this work, we formulate the feature selection task as the learning of a binary mask. Aiming to select features in images for reconstruction, we developed a novel way to sample and learn a correlated discrete variable thanks to a reparameterization of the logitNormal distribution. The proposed learning framework also preserves the spatial structure of the data, enhancing reconstruction performance. We experimentally show that our proposition to explore the space of covariance matrices and average vectors as in eq. (7) is efficient providing us with a sampling with lower variance. Finally, we experimentally evidenced the advantage of learning a correlated sampling scheme instead of independent ones.

## References

1. Balm, M.F., Abid, A., Zou, J.: Concrete Autoencoders: Differentiable Feature Selection and Reconstruction. In: 36th International Conference on Machine Learning, in Proceedings of Machine Learning Research 97:444-453 (2019)
2. Aitchison, J.: The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*. 44, 2, 139–177 (1982).
3. Aitchison, J., and S. M. Shen.: Logistic-Normal Distributions: Some Properties and Uses. *Biometrika* 67, no. 2 (1980)
4. Bengio Y., Nicholas L., Courville A: Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. ArXiv abs/1308.3432 (2013)
5. Deng C., Chiyuan Z., Xiaofei H.: Unsupervised feature selection for multi-cluster data. In: 16th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining. (2010) <https://doi.org/10.1145/1835804.1835848>
6. Ian J. Goodfellow, Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.: Generative adversarial nets. In: 27th International Conference on Neural Information Processing Systems - Volume 2 (2014).
7. Guyon I. Elisseeff A.: An Introduction of Variable and Feature Selection. In: *Journal of Machine Learning Research Special Issue on Variable and Feature Selection*. 3. 1157 - 1182. (2003) <https://doi.org/10.1162/153244303322753616>.
8. K. Han, Y. Wang, C. Zhang, C. Li and C. Xu: Autoencoder Inspired Unsupervised Feature Selection. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2941-2945. (2018). <https://doi.org/10.1109/ICASSP.2018.8462261>.
9. Williams, R.J: Simple statistical gradient-following algorithms for connectionist reinforcement learning. In: *Machine Learning* 8, 229–256 (1992). <https://doi.org/https://doi.org/10.1007/BF00992696>
10. Xing E., Jordan M., Karp R. : Feature selection for high-dimensional genomic microarray data. In: *Proceedings of the 8th International Conference on Machine Learning* (2001).
11. Simon, N., Friedman, J., Hastie, T., Tibshirani, R. . A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2), 231-245. (2013)
12. McPhaden, M. J., G. Meyers, K. Ando, Y. Masumoto, V. S. N. Murty, M. Ravichandran, F. Syamsudin, J. Vialard, L. Yu, and W. Yu: RAMA: The Research Moored Array for African–Asian–Australian Monsoon Analysis and Prediction, *Bulletin of the American Meteorological Society* 90, 4 (2009)

13. Haeberli, W. Hoelzle, M. Paul, F. Zemp, M.: Integrated monitoring of mountain glaciers as key indicators of global climate change: the European Alps. *Annals of Glaciology*, 46, 150–160. Cambridge University Press. (2007)
14. Hou C., Nie F., Li X., Yi D., Wu Y.: Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection, in *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 793-804, (2014), <https://doi.org/10.1109/TCYB.2013.2272642>.
15. Vincent P., Larochelle H, Bengio, Y., Manzagol P. Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*. 1096-1103. (2008) <https://doi.org/10.1145/1390156.1390294>.
16. S. Yang, R. Zhang, F. Nie and X. Li, Unsupervised Feature Selection Based on Reconstruction Error Minimization, *ICASSP - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2019), pp. 2107-2111, <https://doi.org/10.1109/ICASSP.2019.8682731>.
17. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267-288 (1996)
18. Han, K., Wang, Y., Zhang, C., Li, C.: Autoencoder Inspired Unsupervised Feature Selection. 2941-2945. (2018) <https://doi.org/10.1109/ICASSP.2018.8462261>
19. Maldonado, S., Weber, R.: A wrapper method for feature selection using support vector machines. *Inf. Sci.* 179(13), 2208-2217. (2009) <https://doi.org/10.1016/j.ins.2009.02.014>
20. Yu, L., Liu, H.: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: *20th International Conference on Machine Learning*. 2. 856-863. (2003)
21. Braunstein, S.L., Ghosh, S., Severini, S. : The Laplacian of a Graph as a Density Matrix: A Basic Combinatorial Approach to Separability of Mixed States. *Ann. Comb.* 10, 291–317 (2006). <https://doi.org/10.1007/s00026-006-0289-3>
22. Hinton, G.E., Salakhutdinov, R.: Reducing the Dimensionality of Data with Neural Networks. *Science*, 313, 504 - 507. (2006)
23. Isola, P., Zhu, J., Zhou, T., Efros, A.: Image-to-Image Translation with Conditional Adversarial Networks. 5967-5976. (2017) <https://doi.org/10.1109/CVPR.2017.632>
24. Kingma, D. P., Welling, M.: Auto-Encoding Variational Bayes. *2nd International Conference on Learning Representations, ICLR , Banff, AB, Canada*, (2014)
25. Zhou, Y., Jin, R., Hoi, S.C.: Exclusive Lasso for Multi-task Feature Selection. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, in *Proceedings of Machine Learning Research* 9:988-995 (2010)
26. Kovcsiký, T. Melis, G. Grefenstette, E., Dyer, C., Ling, W., Blunsom, P., Hermann, K.: Semantic Parsing with Semi-Supervised Sequential Autoencoders. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2016)
27. Koller D., Sahami M.: Toward optimal feature selection. In: *13th International Conference on International Conference on Machine Learning (ICML)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1996)
28. Ng A.: Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning (ICML)*. Association for Computing Machinery, New York, NY, USA, (2004) <https://doi.org/10.1145/1015330.1015435>
29. Figurnov, M., Mohamed S., Mnih. A.: Implicit reparameterization gradients. In: *32nd International Conference on Neural Information Processing Systems (NIPS)*, (2018). Curran Associates Inc., Red Hook, NY, USA, 439–450.



30. K. He, X. Zhang, S. Ren and J. Sun: Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, <https://doi.org/10.1109/CVPR.2016.90>
31. Louizos, C., Welling, M., Kingma, D.: Learning Sparse Neural Networks through  $L_0$  Regularization. In: International Conference on Learning Representations, ICLR, (2017)
32. Luce, R. D.: Individual choice behavior. John Wiley. (1959)
33. Sepulchre, P., Caubel, A., Ladant, J.-B., Bopp, L., Boucher, O., Braconnot, P., Brockmann, P., Cozic, A., Donnadiou, Y., Dufresne, J.-L., Estella-Perez, V., Ethé, C., Fluteau, F., Foujols, M.-A., Gastineau, G., Ghattas, J., Hauglustaine, D., Hourdin, F., Kageyama, M., Khodri, M., Marti, O., Meurdesoif, Y., Mignot, J., Sarr, A.-C., Servonnat, J., Swingedouw, D., Szopa, S., and Tardif, D.: IPSL-CM5A2 – an Earth system model designed for multi-millennial climate simulations, *Geosci. Model Dev.*, 13, 3011–3053, (2020) <https://doi.org/10.5194/gmd-13-3011-2020>,
34. Yuan, M., Lin, Y. : Model Selection and Estimation in Regression With Grouped Variables. In: *Journal of the Royal Statistical Society Series B.* 68. 49-67. (2006) <https://doi.org/10.1111/j.1467-9868.2005.00532.x>
35. Manohar, K., Brunton, B., Kutz, J., Brunton, S. : Data-Driven Sparse Sensor Placement for Reconstruction: Demonstrating the Benefits of Exploiting Known Patterns. In: *IEEE Control Systems*, 38, 63-86. (2018)
36. Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon C.K. Shiu. Un-supervised feature selection by regularized self-representation. In: *Pattern Recognition*. 48, 2, 438–446. (2015), <https://doi.org/10.1016/j.patcog.2014.08.006>
37. Maddison, C., Mnih, A., Teh, Y.: The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations, ICLR*, (2017)
38. Makhzani, A., Frey, B.: k-Sparse Autoencoders. In *International Conference on Learning Representations, ICLR*, (2014)
39. Mirza, M., Osindero, S.: Conditional Generative Adversarial Nets. (2014)
40. Bora, A., Jalal, A., Price, E., Dimakis, A.G.: Compressed Sensing using Generative Models. In: *34th International Conference on Machine Learning*, (2017)
41. Donoho D.: Compressed sensing, in *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289-1306, (2006), <https://doi.org/10.1109/TIT.2006.871582>
42. Wu, Y., Rosca, M., Lillicrap, T.: Deep Compressed Sensing. *Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 97:6850-6860 (2019)
43. IPSL: IPSL CMA5.2 simulation, [http://forge.ipsl.jussieu.fr/igcmg\\\_doc/wiki/DocHconfigAips1cm5a2](http://forge.ipsl.jussieu.fr/igcmg\_doc/wiki/DocHconfigAips1cm5a2), (2018)
44. Jang, E., Gu, S., Poole, B.: Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations, ICLR*, (2017)
45. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396-4405. (2019)
46. Ha, D., Dai, A, Le, Q.: HyperNetworks. In *International Conference on Learning Representations, ICLR*, (2017)

## 6 Supplementary Material

### 6.1 The logitNormal Distribution:

The logitNormal distribution defines a probability distribution over the simplex, see [3] Initially introduced to describe compositional data [2], it is defined as:

**Definition 2.** *LogitNormal* Let  $X$  be random variable defined over  $\mathbb{R}^n$  such that  $X \sim \mathcal{N}(\mu, \sigma)$ . Then, consider the following transformation:

$$Y_{-n} = e^X / (1 + \sum_{j=1}^n e^{X_j}), \text{ and } Y_{n+1} = 1 - \sum_{j=1}^n Y_j$$

Then the vector  $Y = (Y_1, \dots, Y_{n+1})$  follows a logitNormal distribution denoted  $\mathcal{LN}(\mu, \sigma)$  and is defined over the  $\mathbb{R}^{n+1}$  simplex. Moreover,  $Y$  admits a density and can be found in [3].

If  $Y \sim \mathcal{LN}(\mu, \sigma)$ , it defines a probability distribution over the simplex which makes it practical to model compositional data, i.e. “data where the involved data forms some sort of proportion of a whole” [2].

### 6.2 Reparameterizing the logitNormal Distribution:

Using Definition 2 of the logitNormal distribution, we can use the reparameterization trick in order to learn the parameters of a logitNormal law from samples of Normal law.

**Theorem 3.** *Reparameterization:* Let  $X = (X_i)_{i \leq n}$  such that  $X_i \sim \mathcal{N}(0, 1)$  and all  $X_i$  are iid ( $X \in \mathbb{R}^n$ ),  $W \in \mathcal{M}_{m \times n}(\mathbb{R})$ , and  $b \in \mathbb{R}^m$ , then :

$$Y_{-n} = \exp(WX + b) / (1 + \sum_i \exp(W_i \cdot X + b_i))$$

$$Y = (Y_{-n}, 1 - \sum_{j=1}^n Y_j) \tag{13}$$

$$Y \sim \mathcal{LN}(b, \Sigma) \tag{14}$$

This comes from the simple fact that an affine transformation of i.i.d.  $\mathcal{N}(0, 1)$  follows also a Normal law, which co-variance matrix can be expressed through the matrix of linear weights. Moreover, this advantageously correspond to a neural network layer with an extended sigmoidal function.

### 6.3 Proof For 0-Temperature

Here we prove the convergence of the reparameterization of the logitNormal law for the zero temperature.

*Proof.* Let  $(\lambda_n)_{n \geq 0}$  be a positive sequence decreasing towards 0. We prove the 0-temperature convergence for  $z \sim \mathcal{N}(\mu, \sigma)$ . Let  $Y_n = \text{sigmoid}_{\lambda_n}(z)$ . We investigate the convergence in distribution of  $Y_n$  towards a Bernoulli distribution. Let  $f$  be a continuous bounded function. We have:

$$\begin{aligned} \mathbb{E}(f(Y_n)) &= \int_0^1 f(Y_n) dP_{Y_n} = \int_{\mathbb{R}} f(\text{sigmoid}_{\lambda_n}(z)) dP_z \\ &= \int_{\mathbb{R}} f(\text{sigmoid}_{\lambda_n}(z)) \frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} dz \end{aligned}$$

We first have point-wise convergence of the sequence of function inside the integral. Indeed,

$$\text{If } z > 0, \lim_{n \rightarrow \infty} \text{sigmoid}_{\lambda_n}(z) = 1.$$

$$\text{If } z < 0, \lim_{n \rightarrow \infty} \text{sigmoid}_{\lambda_n}(z) = 0. \text{ We have:}$$

$$\lim_{n \rightarrow \infty} f(\text{sigmoid}_{\lambda_n}(z)) \frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi\sigma}} f(\delta_{z>0}) \exp^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$$

The domination is verified using the function:

$$g(z) = \frac{1}{\sqrt{2\pi\sigma}} \|f\|_{\infty} \times \exp^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$$

We can finally apply the theorem of dominated convergence:

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E}(f(Y_n)) &= \mathbb{E}(\lim_{n \rightarrow \infty} f(Y_n)) \\ &= \int \frac{1}{\sqrt{2\pi\sigma}} f(\delta_{z>0}) \exp^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} dz \\ &= \frac{1}{\sqrt{2\pi\sigma}} f(0) \int_{-\infty}^0 \exp^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} dz + \frac{1}{\sqrt{2\pi\sigma}} f(1) \int_0^{+\infty} \exp^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} dz \\ &= f(0)\Phi\left(-\frac{\mu}{\sigma}\right) + f(1)(1 - \Phi\left(-\frac{\mu}{\sigma}\right)) \\ &= \mathbb{E}_{b \sim \mathcal{B}(1 - \Phi\left(-\frac{\mu}{\sigma}\right))} f(b), \end{aligned}$$

where  $\mathcal{B}$  denotes Bernoulli distribution. Finally, we can conclude that  $Y_n$  converges in law towards a Bernoulli distribution such that:  $Y_n \rightarrow \mathcal{B}(1 - \Phi\left(-\frac{\mu}{\sigma}\right))$

#### 6.4 Proof $L_0$ -logitNormal:

$$\begin{aligned}
 L_0(S_\theta(z)) &= \sum_i 1 - \mathbb{P}(\bar{z} \leq 0) \\
 &= \sum_i 1 - \mathbb{P}(\text{sigmoid}(Wz + b) \leq -\gamma/(\eta - \gamma)) \\
 &= \sum_i 1 - \mathbb{P}(W_i \cdot z \leq \log\left(\frac{-\gamma}{\eta}\right) - b) \\
 &\text{as } W_i \cdot z \text{ has a normal law } \mathcal{N}\left(0, \sqrt{\sum_j w_{j,i}^2}\right) \\
 &= \sum_i 1 - \Phi\left(\frac{\log\left(\frac{-\gamma}{\eta}\right) - b}{\sqrt{\sum_j W_{j,i}^2}}\right)
 \end{aligned}$$

#### 6.5 Ill posedness of the $\ell_1$ -formulation:

Consider the auto encoding setting with a  $\ell_1$ -norm instead of the derived  $L_0$ . The optimization problem is:

$$\mathcal{L}_{\ell_2} = \lambda_{\ell_2} \mathbb{E}_{x \sim p_x} \|x - G_\phi(S_\theta(z) \odot x)\|_2 + \lambda_s \cdot L_1(S_\theta) \quad (15)$$

Let  $(G_\phi^*, S_\theta^*)$  be an optimal solution, i.e that realizes the minimum of the above optimization cost function. Then, consider:  $S_2 = S_\theta^*/2$  and  $G_2$  defined as  $G_2(x) = G_\theta^*(2 * x)$ . Then the MSE term of eq. (15) for the couple  $(G_2, S_2)$  is equivalent as the one with  $(G_\phi^*, S_\theta^*)$ , however the  $\ell_1$ -norm of  $(S_2)$  is lower. Therefore  $(G_\phi^*, S_\theta^*)$  is not optimal and the problem of eq. (15) is ill-posed. However, note that, in the case of binary vectors,  $\ell_0$ -norm and  $\ell_1$ -norm are equals.

#### 6.6 On the Stretching Scheme:

We initially start from a distribution  $p$  that lives in  $[0, 1]$  and need to transform it in order to obtain a non zero probability of sampling 0 while maintaining both tractability and differentiability. We denote this function  $f$ . We need  $f^{-1}(0)$  to be a non-zero measure set of the original support. In other words, we need  $f$  to be a surjection, and  $f^{-1}(0)$  to be Lebesgue measurable with a non zero mass. Instead of the *HT* function we could have used a stretched *relu* function. One significant advantage of the chosen function is that it also creates a non-zero probability of sampling 1 therefore enforcing the binary behaviour of our masks. Unbalanced binary scheme can also be investigated in future works. Indeed one can think of creating a higher portion of the stretched distribution above one, enforcing the binary behaviour of the mask.

#### 6.7 Algorithm

We present here the algorithm for the proposed logitNormal based feature selection algorithm.

**Algorithm 1** Differentiable Feature Selection**Result:** Converged  $S$   $G_\phi$ Initialize  $\theta = (W, b)$  and  $G_\phi$ **while** *Convergence not reached* **do**    sample batch  $x = (x_1, \dots, x_n)$  and  $z = (z_1, \dots, z_n)$ , such that  $z_i \sim \mathcal{N}(0, I_d)$     Compute  $S_\theta(z) = \text{sigmoid}_\lambda(Wz + b)$  and the observations  $x^{obs} = \bar{S}_\theta(z) \odot x$     Estimate reconstruction  $\hat{x} = G_\phi(x^{obs})$      $L = \|x - \hat{x}\|_2 + \lambda_{sparse} L_0(\bar{S}_\theta(z))$     Update  $\phi$  and  $\theta$ :

$$\phi \leftarrow \phi - \frac{\partial L}{\partial \phi}$$

$$\theta \leftarrow \theta - \frac{\partial L}{\partial \theta}$$

**end****6.8 Practical Consideration on the Temperature:**

As duely noted by [37], the temperature in sigmoid activation plays a crucial role in the training. This remark holds for our work. Indeed, in our work decreasing the temperature in the sigmoid, amounts to increase the variance and the absolute value of the average of the initial Gaussian distribution.

Also aiming at approximating binary distribution, we don't want any interior mode as in the green curve depicted in fig. 1:  $\mathcal{LN}(0, 1)$  has an interior maximum point. This case is not acceptable for the approximation of Bernoulli random variable as, it could allow a leakage of information, i.e the distribution is not approximating a binary distribution anymore. Therefore, during training one should ensure that the learned distribution has no interior maxima. Fortunately, it suffices to sufficiently decrease the temperature  $\lambda$  of the sigmoid in order to recover two modes at 0 and 1. Indeed, decreasing sufficiently the temperature in the sigmoid pushes the interior maximum towards the edges. In practice, we observe that initializing our  $W$  so that  $W.z$  with a variance higher than 0.5 with a temperature of  $\lambda = 0.3$  suffices.

**6.9 Removing the Randomness**

Both our propositions of eq. (6) or eq. (7) estimates distribution in the spaces of binary variables. To collapse the distribution, one can take advantage of proposition 1 and select the  $K$  desired number of features. One can also, empirically select the  $K$  features the mask with the highest probability to be selected. Both approaches lead to similar results in practice. Note that in both cases, if  $K$  is far from the observed number of pixel, the selected features may not be the best subset of the learned distribution.

In practice, we chose to collapse the distribution using Proposition 1: We first estimate the expected  $\ell_0$ -norm of the distribution, which equals to  $\sum (1 - \phi(-\frac{\mu_i}{\sigma_i}))$ . Let  $L_0$  be the value of the expected  $\ell_0$ -norm of our learned distribution.

We then select two masks made of the most likely features to be selected: the first one has  $L_0$  rounded *down* to the nearest ten pixels. The other one has  $L_0$  rounded *up* to the nearest ten selected pixels. Note that, for SCT baseline, we use a property similar to Proposition 1 for the concrete distribution, available in [37].

### 6.10 Concrete Law

Introduced by [37] to approximate discrete variables, binary concrete random variable is defined as follows:

$$\begin{aligned} u &\sim \mathcal{U}([0, 1]) \\ G &= \log(u) - \log(1 - u) \\ X &= \text{sigmoid}\left(\frac{\log(\alpha) + G}{\lambda}\right), \end{aligned}$$

And  $X$  follows a relaxed binary concrete law.

### 6.11 Experimental Details

All experiments were trained on Titan XP GPU via using Pytorch framework and mixed precision training. For all experiments the expected  $\ell_0$ -norm is normalized by the number of pixels in the signal. Also for all algorithms trained using correlated logitNormal approach, the dimension of  $z$  is 16, i.e.  $z \sim \mathcal{N}(0, I_{16})$ .

For all mask based methods,  $G_\phi$  is a resnet following the implementation of [23] with 2 residual blocks and 16 filters.

**Mnist** All masked based algorithms were trained using ADAM optimizer with  $\beta = (0.9, 0.99)$  and a learning rate of  $2.10^{-4}$  for 550 epochs with batch size 256. CAE method was trained for 1400 epochs with a temperature decreasing form 10 to 0.01 following recommendation of the authors.

**Climate Data** All masked based algorithms were trained using ADAM optimizer with  $\beta = (0.9, 0.99)$  and a learning rate of  $2.10^{-4}$  for 550 epochs with batch size 128. CAE method was trained for 1400 epochs with a temperature decreasing form 10 to 0.01 following recommendation of the authors.

**CelebA** All masked based algorithms were trained using ADAM optimizer with  $\beta = (0.9, 0.99)$  and a learning rate of  $2.10^{-4}$  for 140 epochs with batch size 128. CAE method was trained for 400 epochs with a temperature decreasing form 10 to 0.01 following recommendation of the authors.

**Hyperparameters Search** Except for CAE where the number of selected features is a structural constraint, we search the hyperparameter space by sampling from the interval  $[10^{-2}; 1]$  discretized by steps of  $3 \cdot 10^{-2}$ . For all dataset, the CAE method was trained with a decreasing temperature from 10 to 0.01 following the guidelines of the authors [1]. For the mask method based on the concrete distribution the temperature of the sigmoid was set to  $\lambda = 2/3$  following the recommendation of [37]. For logitNormal based algorithm, the temperature was fixed to  $\lambda = 0.3$ .

**Initialization** For all mask based methods, we chose the initialization parameters so that the resulting distribution of the each variable in the mask is symmetrical, with as many chances to be sampled than to be rejected, i.e. for all variable  $i$  in the masks:  $\mathbb{P}(S_\theta(z)_i < \epsilon) \approx \mathbb{P}(S_\theta(z)_i > 1 - \epsilon) \approx 0.2$ . That way, all distribution can explore the space of binary masks. Also, in order to verify whether a covariance matrix is learned during training for the logitNormal sampling method of eq. (6),  $W$  is initialized with using an uniform law.

### 6.12 Additional Samples:

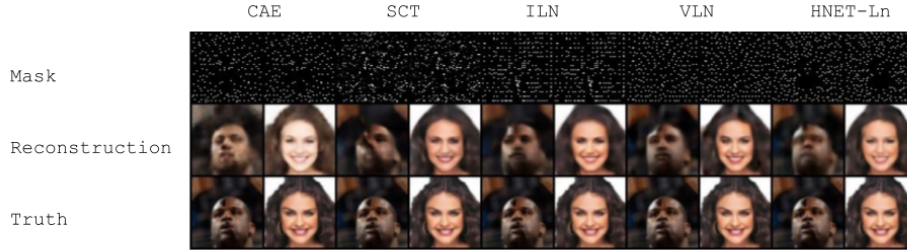


Fig. 6: Sample of masks (first row), Reconstruction (second row) and True Data (Last row) for CelebA dataset on all considered algorithms for 200 features with  $\ell_2$ -encoding

### 6.13 cGAN Details and Samples

Simply speaking, a cGAN has two main learnable functions: a discriminator network with parameters  $\psi$  named  $D_\psi$  trained to differentiate "true" data labeled as 1 from data generated by  $G_\phi$  labeled as 0. A generative network with parameter  $\phi$  denoted  $G_\phi$ .  $G_\phi : \mathbb{R}^p \times \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^n$  takes as input a random variable  $\gamma \in \mathbb{R}^p$  and our conditional information  $x^{obs} = \bar{S}_\theta(z) \odot x \in \mathbb{R}^{n \times n}$ , and aims at fooling  $D_\psi$ , making it classify the conditionally generated images as true. For our

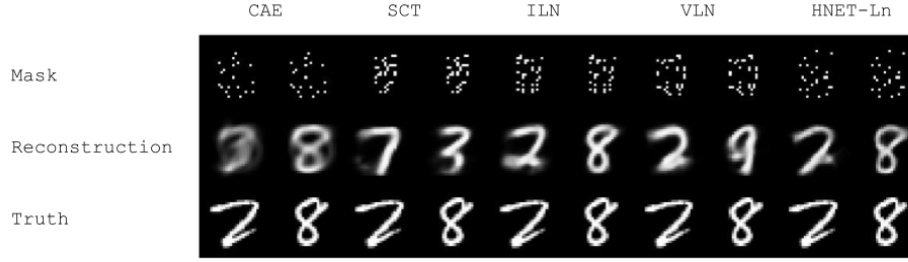


Fig. 7: Sample of masks (first row), Reconstruction (second row) and True Data (Last row) for Mnist dataset on all considered algorithms for 20 selected features with  $\ell_2$ -encoding

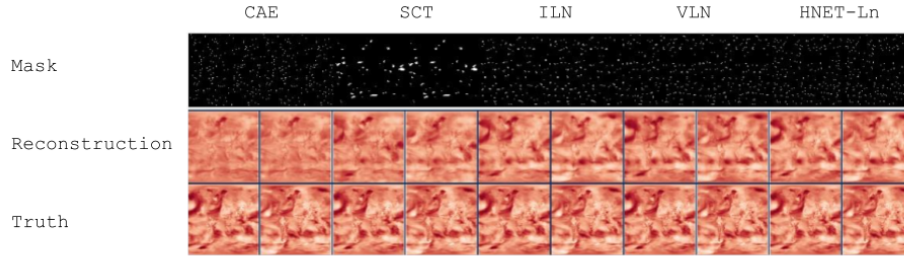


Fig. 8: Sample of masks (first row), Reconstruction (second row) and True Data (Last row) for the Geophysical Dataset on all considered algorithms for 200 features with  $\ell_2$ -encoding

experiments we used the cGAN implementation of [23] optimizing the following loss, with  $x^{obs} = x \odot \bar{S}_\theta(z)$ :

$$\begin{aligned} \min_{\phi, \theta} \max_{\psi} \mathbb{E}_{z, x} \log D_\psi(x, x^{obs}) + \mathbb{E}_{z, x} \log \{1 - D_\psi(G_\phi(x^{obs}), x^{obs})\} \\ + \lambda_{sparse} \times \ell_0(\bar{S}_\theta(z)) + \lambda_{rec} \times \ell_1(x - G_\phi(x^{obs})), \end{aligned} \quad (16)$$

Consider  $S_\theta$  fixed, one interesting advantage about the cGAN approach is that we can prove that the optimal distribution  $p_{G_\phi}$  for  $G_\phi$  is given  $x^{obs}$ :  $p_{G_\phi}(x, x^{obs}) = p_{x \sim data}(x|x^{obs})$  which means that  $G_\phi$  will sample according to the observed data distribution.

*Proof.* To lighten notation, we will use the notation  $y = x \odot \bar{S}_\theta(z)$  as conditioning variable, giving the following game value function:

$$V(G, D) = \mathbb{E}_{x, y} \log D(x, y) + \mathbb{E}_{z, y} \log \{1 - D(G(z, y), y)\}$$



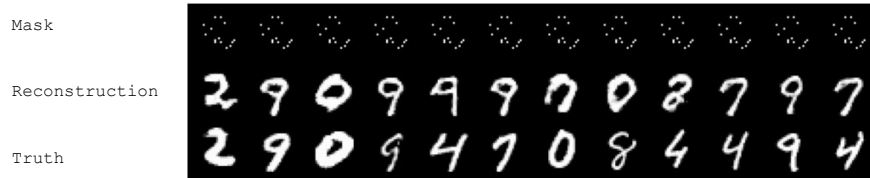


Fig. 9: Samples of masks (first row), reconstruction (second row) and true data (last row) for Mnist dataset obtained using a cGAN approach following [23], i.e including a  $\ell_1$ +Gan loss as reconstruction objective for approximately 15 sampled pixels ( $\lambda_s = 100$ )

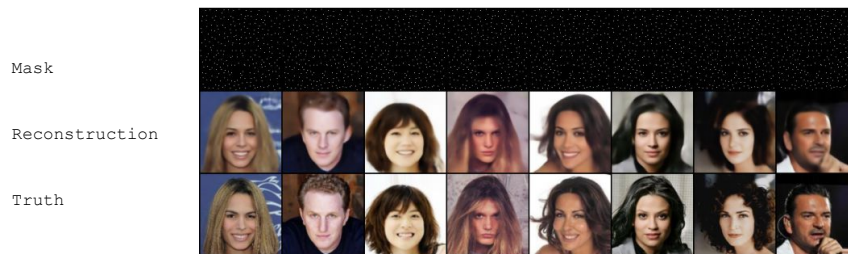


Fig. 10: Samples of masks (first row), reconstruction (second row) and true data (last row) for CelebA dataset obtained using a cGAN approach following [23], i.e including a  $\ell_1$ +Gan loss as reconstruction objective for approximately 1.7% sampled pixels ( $\lambda_s = 100$ )

Following [6], we can write:

$$\begin{aligned}
 V &= \int_{x,y} \log D(x,y) p_x(x,y) dx dy + \int_{z,y} \log\{1 - D(G(z,y), y)\} p_z(z) p_y(y) dz dy \\
 &\quad \text{if } G \text{ induce a distribution } p_g, \\
 V &= \int_{x,y} [\log D(x,y) p_x(x|y) p_y(y) + \log\{1 - D(x,y)\} p_g(x|y) p_y(y) dx dy] \\
 &= \int_y \left( \int_x \log D(x,y) p_x(x|y) + \log\{1 - D(x,y)\} p_g(x|y) \right) p_y(y) dy
 \end{aligned}$$

Then classically the maximal value of  $x \rightarrow a \log(x) + b \log(1 - x)$  is reached in  $\frac{a}{a+b}$ . Thus, given  $y$ , the optimal distribution followed by  $D$ :

$$p_D(x,y) = \frac{p_x(x|y)}{p_x(x|y) + p_g(x|y)}$$

The optimal distribution of  $G$  is completely doable at  $y$  fixed following the original reasoning of [6]