Deep Variational Sufficient Dimensionality Reduction

Ershad Banijamali[1,2] **, Amir-Hossein Karimi**[1] **, Ali Ghodsi**[3]
[1]School of Computer Science, University of Waterloo
[2]Vector Institute
[3]Department of Statistics and Actuarial Science, University of Waterloo
{sbanijam,a6karimi,aghodsib@uwaterloo.ca}

Abstract

We consider the problem of sufficient dimensionality reduction (SDR), where the high-dimensional observation is transformed to a low-dimensional sub-space in which the information of the observations regarding the label variable is preserved. We propose DVSDR, a deep variational approach for sufficient dimensionality reduction. The deep structure in our model has a bottleneck that represent the low-dimensional embedding of the data. We explain the SDR problem using graphical models and use the framework of variational autoencoders to maximize the lower bound of the log-likelihood of the joint distribution of the observation and label. We show that such a maximization problem can be interpreted as solving the SDR problem. DVSDR can be easily adopted to semi-supervised learning setting. In our experiment we show that DVSDR performs competitively on classification tasks while being able to generate novel data samples.

1 Introduction

Dimensionality reduction is a long-standing problem in machine learning. The initial motivation behind dimensionality reduction was to visualize data and many unsupervised, supervised, and semi-supervised algorithms have been designed for this purpose. Recent advancements in the area of deep learning has pushed the boundaries of the state-of-the-art across a variety of fields including object classification [16, 9, 15], speech recognition [4, 5], and sample synthesis [3, 11, 6], among others. Casting classical algorithms in statistical machine learning within the framework of deep learning has been transformative for the aforementioned applications, e.g. [1]. This inspired us to revisit the problem of sufficient dimensionality reduction and tackle it using the tools provided to us by deep learning.

Sufficient dimensionality reduction (SDR) [2] is a technique that aims to find a low-dimensional representation of data that retains predictive information regarding the label (response) variable. The original paper of SDR presents a way of quantifying this information using information theoretic notions and introduces an iterative algorithm for extracting the features that maximizes this information. Although SDR methods are typically applied to continuous target variables, there exist methods based on distance covariance that can estimate the central subspace, the intersection of all dimensionality reduction subspaces, for discrete target variables [13]. In this paper we consider the discrete target scenario.

2 Model description

In this section we first overview SDR and then present a graphical model interpretation of SDR and propose a deep model in the framework of variational autoencoder that approximates the posterior in the graphical model and learns the low-dimensional space efficiently.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

2.1 Sufficient Dimensionality Reduction

Consider the frequently encountered goal of regression: predicting a future value for a univariate label $\mathbf{y} \in \mathbb{R}^D$ for a given observation $\mathbf{x} \in \mathbb{R}^p$. In large scale domains, traditional regression methods may require large amounts of training data to avoid overfitting. Therefore, there is a pertinent need for dimensionality reduction methods that replace the original covariate $\mathbf{x}$ with another variable $\mathbf{z} \in \mathbb{R}^d$ that retains most or all of the information and variation in $\mathbf{x}$. When $\mathbf{z}$ retains all the relevant information about $\mathbf{y}$, the dimensionality reduction is said to be *sufficient*: $p(\mathbf{y}|\mathbf{z}(\mathbf{x})) = p(\mathbf{y}|\mathbf{x})$.

Alternatively, sufficient dimension reduction (SDR) techniques can be viewed as methods that find a low dimensional representation such that the remaining degrees of freedom become conditionally independent of the output values, i.e.:

$$\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{z} \tag{1}$$

In other words, $\mathbf{z}(\mathbf{x})$ carries all the information of $\mathbf{x}$ needed to predict a future value for $\mathbf{y}$.

2.2 SDR, a graphical model interpretation

The goal in SDR, as stated in above, is to discover a latent space that can be used for classification. Therefore, compared to the unsupervised dimensionality reduction algorithms, SDR aims to find a latent variable with high predictive power. SDR problem can be explained by either of the graphical models in Fig. 1. Although the objective of SDR might be derived from Fig. 1b more intuitively, we argue that the latent space that is learned in Fig. 1a carries more information about $\mathbf{x}$ and therefore is more generalizable. In fact, deriving $p(\mathbf{y}|\mathbf{x})$ from the joint distribution $p(\mathbf{x}, \mathbf{y})$ yields a better representation in the latent space that is more robust against overfitting [14]. This is especially true when we parameterize these conditional probabilities using neural networks with hundreds of thousands of parameters. Another benefit of the model in Fig. 1a is that it can leverage unlabeled samples to build the latent space and therefore may be used for semi-supervised learning. Whereas the model in Fig. 1a can only use the labeled samples.

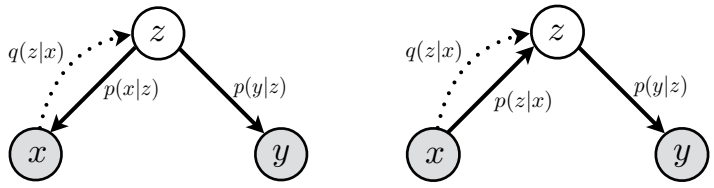


Figure 1: Candidate graphical models for sufficient dimensionality reduction

In fact, we assume that $\mathbf{z}$ can not only generate high-dimensional observations, but can also generate the target variable. Therefore, our objective is to find a latent variable that keeps the information of $\mathbf{X}$ and can be used for classification. In the graphical model of Fig. 1b, this is equivalent to maximizing the joint distribution of evidences, $p(\mathbf{X}, \mathbf{Y})$ for all pairs of $(\mathbf{x}_i, \mathbf{y}_i)$. One way to maximize this likelihood is using the variational inference. Based on this graphical model the variational lower bound of the log-likelihood of the joint distribution can be written as:

$$\log p(\mathbf{x}, \mathbf{y}) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\big[\log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{y}|\mathbf{z})\big] - \mathrm{KL}\big(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big). \tag{2}$$

We assume that the prior distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Our goal is to maximize this lower bound. In this model the posterior approximating distribution $q(\mathbf{z}|\mathbf{x})$ plays the role of a probabilistic dimensionality reduction function, $\mathbf{z}(\mathbf{x}) \sim q(\mathbf{z}|\mathbf{x})$.

2.3 Deep Variational Learning

Variational autoencoders (VAEs) [8, 12] are deep models that can maximize the lower bounds of type Eq. 2 efficiently in large scale. In these models, the conditional distributions are parameterized by neural networks. An *encoder* network parameter set ϕ models $q_\phi(\mathbf{z}|\mathbf{x})$, which is a transformation

function that maps high-dimensional observation to the latent space. A *decoder* with parameter set θ models $p_\theta(\mathbf{x}|\mathbf{z})$, which is a mapping from the latent space to observation space. And a *classifier* with parameter set ψ models $p_\theta(\mathbf{y}|\mathbf{z})$, which is a mapping from the latent space to the space of labels. Therefore the lower bound in 2, denoted by $\mathcal{L}(\mathbf{x}, \mathbf{y})$ can be rewritten as:

$$\mathcal{L}_{\phi,\theta,\psi}(\mathbf{x}, \mathbf{y}) \quad = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\psi(\mathbf{y}|\mathbf{z})\big] - \mathrm{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big). \quad (3)$$

Our approach to solve the SDR problem is to maximize this lower bound using deep variational learning, and we call it deep variational SDR (DVSDR). By maximizing $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\psi(\mathbf{y}|\mathbf{z})\big]$ for each labeled sample, we build a latent representation that has a high predictive power and this is exactly what SDR aims for.

2.4 Semi-supervised DVSDR

In many cases in practical problems, the label information is scarce and therefore a semi-supervised learning method should be exploited. DVSDR can be easily adopted to a semi-supervised learning setting. Suppose $\mathbf{X}_L \subset \mathbf{X}$ is the set of all labeled samples for which we have the label set $\mathbf{Y}$, and $\mathbf{X}_U = \mathbf{X} \backslash \mathbf{X}_L$ is the set of unlabeled samples. Since our latent variable is inferred purely based on our observation variable (and not label variable), we can split the objective function to two parts: $\mathcal{L}^\ell_{\phi,\theta,\psi}(\mathbf{x}, \mathbf{y})$, which is equal to 3, for the labeled points and $\mathcal{L}^u_{\phi,\theta}(\mathbf{x})$, for the unlabeled points, which has the following form:

$$\mathcal{L}^u_{\phi,\theta}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - \mathrm{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big). \quad (4)$$

$\mathcal{L}^u_{\phi,\theta}(\mathbf{x})$ has the same form as a regular VAE. Therefore the overall objective of the semi-supervised model is:

$$\max_{\phi,\theta,\psi} \sum_{\mathbf{x}\in\mathbf{X}^\ell} \mathcal{L}^\ell_{\phi,\theta,\psi}(\mathbf{x}, \mathbf{y}) + \sum_{\mathbf{x}\in\mathbf{X}^u} \mathcal{L}^u_{\phi,\theta}(\mathbf{x}). \quad (5)$$

In fact, in the semi-supervised setting, the low-dimensional latent variable for unlabeled samples is used to only reconstruct the observations, but for the labeled samples it is used to reconstruct the samples and predict the labels.

3 Experiment Results

We evaluate the performance of DVSDR in two different tasks: Classification and new sample generation.

3.1 Classification

We compare the classification performance of the proposed model with Adversarial Autoencoder (AAE) [10] and semi-supervised learning with deep generative models [7] that are both deep generative models with bottleneck.

	MNIST (All labeled samples)	MNIST (1000 labeled samples)
VAE (M1+M2)	0.96 ± 0.03	2.40 ± 0.02
AAE	0.86 ± 0.03	$\mathbf{1.60 \pm 0.08}$
DVSDR	$\mathbf{0.80 \pm 0.04}$	2.1 ± 0.06

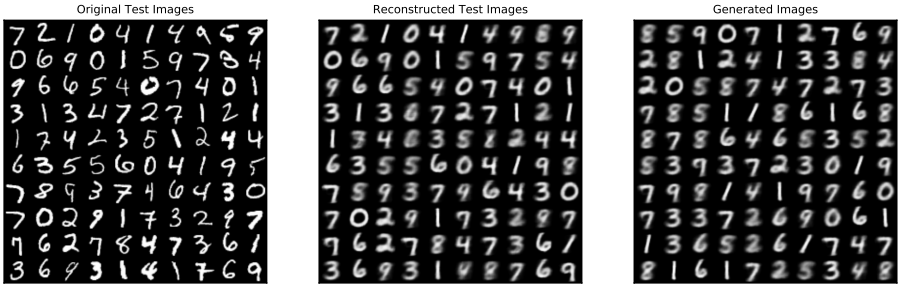
Table 1: Classification Error rate

As we can see our model outperforms the other two models when we use all the label information, but when only 1000 labeled samples are used, AAE performs better. This is because AAE directly impose a supervised loss on the latent space by matching the distribution of the latent space with a categorical distribution.

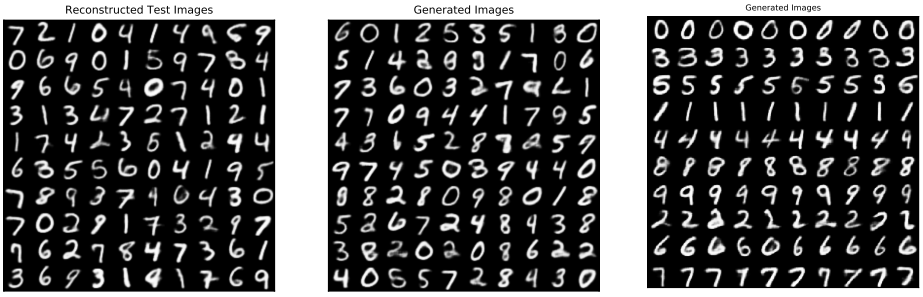
3.2 Generation

Sample generation is a bi-product of the DVSDR algorithm. Using the structure of DVSDR not only can we preserve the information in the observation set while reconstructing the input samples, but also we can generate novel data by sampling from the prior $p(\mathbf{z})$ and feeding it to the decoder.

Fig. 2 and 3 show the results of sample reconstruction/generation for MNIST and Fashion MNIST datasets. For MNIST we use DVSDR with a latent space of dimensionality 2 and 15, and for Fashion MNIST dimensionality of 10. We can generate high quality images using the proposed model. When we fit a mixture of Gaussian over the latent space of the model for the MNIST dataset, we can generate samples from different classes, this is because the representations of the points in each class in the latent space are very well separated.



(a) 2-dimensional latent space



(b) 15-dimensional latent space (c)

Figure 2: (a,b) Reconstructed and generated images using DVSDR with 2-D and 15-D latent space. (c) Fitting a mixture of Gaussian with 10 components on the latent space and sampling from each component

Figure 3: Fashion MNIST. 10-D latent space

References

[1] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 24(12):5017–5032, 2015.

[2] A. Globerson and N. Tishby. Sufficient dimensionality reduction. *Journal of Machine Learning Research*, 3(Mar):1307–1331, 2003.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[7] D. Kingma, D. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 27:3581–3589, 2014.

[8] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[11] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[12] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1278–1286, 2014.

[13] W. Sheng and X. Yin. Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 25(1):91–104, 2016.

[14] R. Shu, H. H. Bui, and M. Ghavamzadeh. Bottleneck conditional density estimation. In *International Conference on Machine Learning*, pages 3164–3172, 2017.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.