Performance and accessibility of statistical learning algorithms for applied data analysis

Cédric Beaulac

University of Toronto

December the 18th 2020

Introduction

- Performance: Accuracy, generalization, stability, etc...
- Accessibility: Hyper-parameter tuning , availability of packages, assumptions, computing power needed, etc...
- Different applications: higher education, healthcare and computer vision.

Plan of the presentation

Higher education data set: a random forest application

Branch-Exclusive Splits Tree (BESTree): A new decision tree algorithm

Survival analysis variational autoencoder (SAVAE)

Analysis of HWD+: a new handwritten digit data set

Project #1 Analysis of an higher education data set

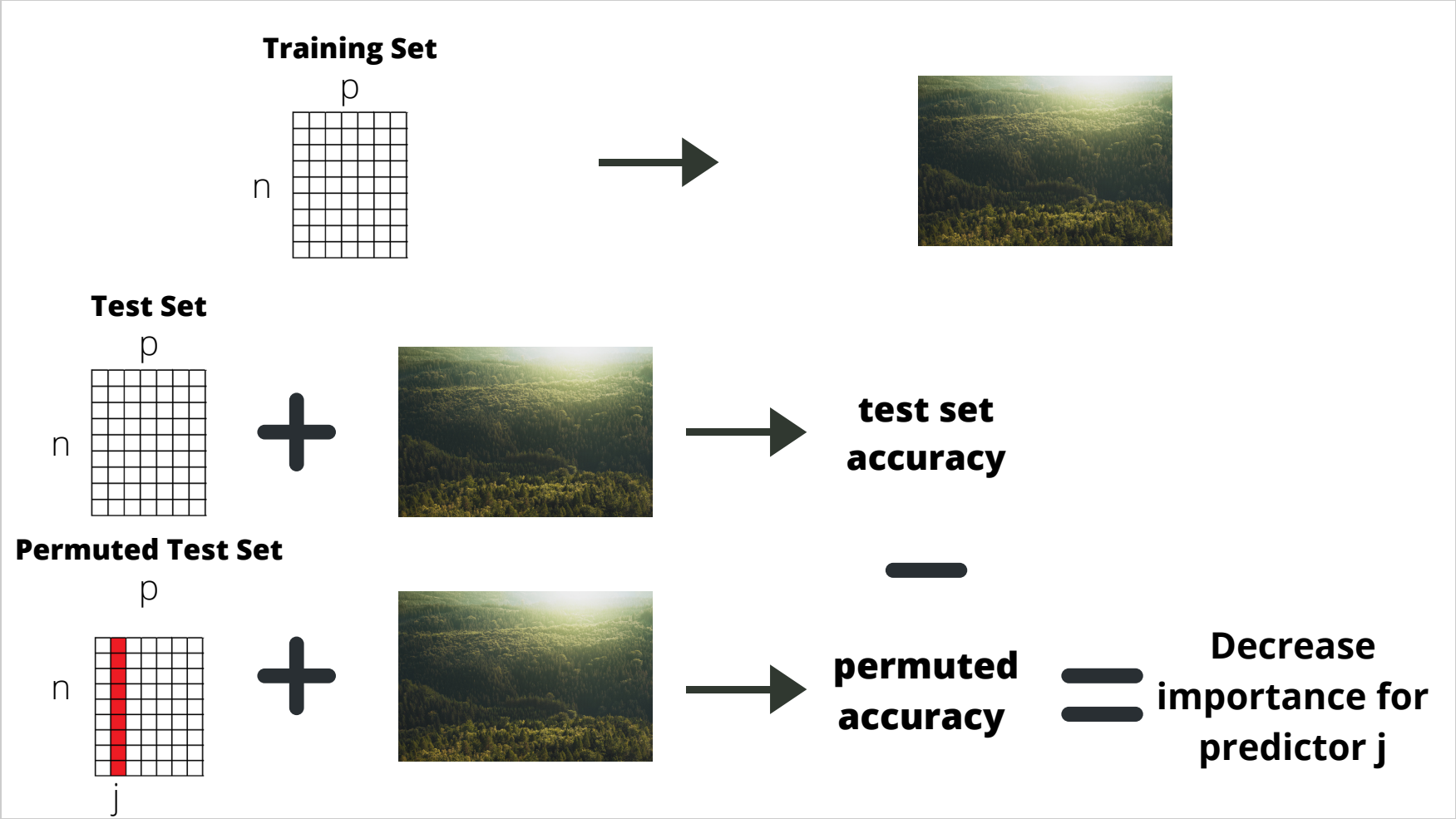
The data set

- First year data for 38 842 students.
- Number of credits and averaged grade in every department (71 departments).
- We decided to use random forests (and decision trees):
  1. Correlated predictors are problematic with logistic regression
  2. We assume lots of interactions between predictors
  3. We want to explore variable importance computations

Random forest : variable importance

- A variable importance assesses the *significance* of predictors in the model.
- Does so by evaluating the contribution of individual predictors to the accuracy of the model.
- Different from *statistical significance* which does not really address the *size* of the effect.
- Decrease Importance: *How significant is the reduction in accuracy when a predictor is randomized ?*

Random forest : variable importance



Research questions

- (1) We want to predict if students will complete their undergraduate program.
- For (1), we predict using strictly academic data.
- (2) We want to discuss *grade inflation*, a well-known issue well documented in research in higher education.
- For (2), our assumption is that grades in low-grading department will be considered *important* grades when predicting success.
- Grades in department where it is *harder* to get good grades should be a better predictor for program completion.

Results

- We predict program completion with 79% accuracy (73% with logistic regression).
- Grades in Mathematics, Chemistry, Finance, Biology and Economics are consistently among the most importance grades in prediction program completion.
- They are considered low-grading department in the literature.

Projet #2 BESTree: A new decision tree algorithm

Motivation

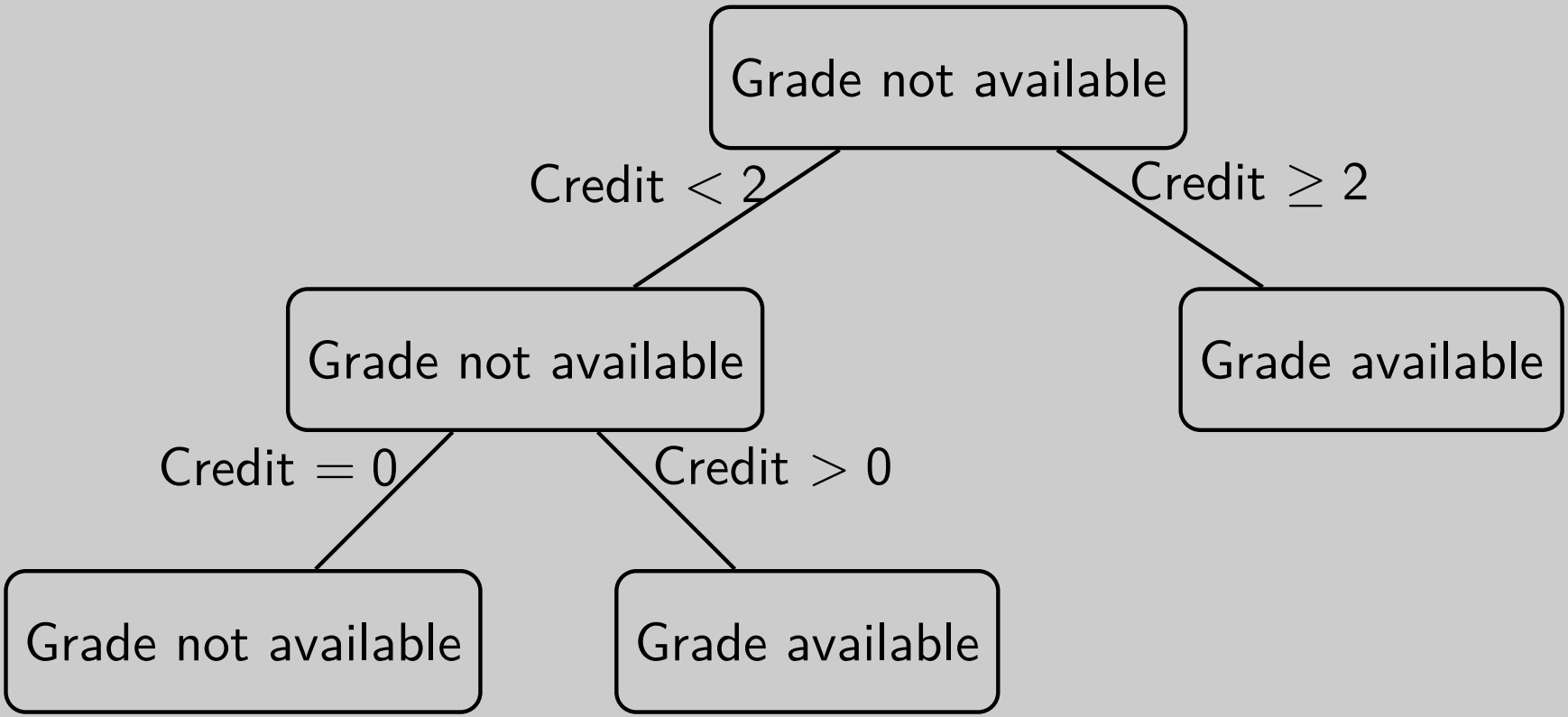
- There are lots of missing values in the previously discussed data set.
- No courses in a department means the average grade in that department is missing.

Student ID	Credits Math	Grade Math	Credits Econ	Grade Econ	Credits Hist	Grade Hist
101	2	72	3	88	0	NA
208	0	NA	0	NA	5	78

Table: An example of the set of predictors for 2 students and 3 departments

Concept

- We want to use the structure of decision trees to *naturally* account for missing value.
- We define when variables are available for the partitioning process using the tree branches.
- The *grade in mathematics* variable is available on branches such that *number of credits in mathematics* > 0.


Grade not available
Credit < 2
Credit ≥ 2
Grade not available
Grade available
Credit $= 0$
Credit > 0
Grade not available
Grade available

How does it works ?

- We consider a subset of all possible trees.
- We define the subset of trees in two steps.
- To begin, we determine the subset of predictors available for the partitioning process in the root node.

  Credit variables

- Then we define the *branch conditions* to access other variables.

  Grades are available on branches such that credit > 0.

Advantages

- Needs no data processing: no imputation nor omission.
- Keeps information about which variables were missing.
- Allows us to evaluate the importance of the missing pattern (which can be a good predictor itself).
- Uses the tree forming procedure to accomplish all of those.
  1. Can be extended to various forest forming procedure
  2. Does not slow the partitioning

Results

- When compared with decision tree alternatives, we show that BESTree:
  1. perfoms as good or better for different missing data structures(MNAR,MAR) on simulated data sets,
  2. produces trees that are more interpretable,
  3. and has higher accuracy on the higher education grade data set.
- We built a R-package publicly available on CRAN.

Project #3 : Adapting variational autoencoders (VAEs) for survival analysis

The data set: AHOD0031 trial

- We obtained a data set from the *Children's Oncology Group*.
- For the 1 712 patients symptoms, the treatment and the response were collected. (Largest randomized trial of pediatric HL ever conducted)
- The response is a time-to-event variable where the majority is right-censored.

Our model: SAVAE (Survival Analysis VAE)

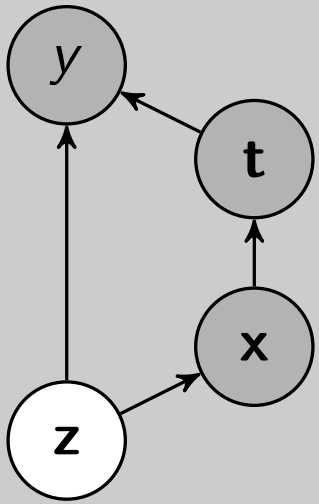


Figure: The response is identified by y, the treatment by $\mathbf{t}$, the symptoms by $\mathbf{x}$ and the latent variable by $\mathbf{z}$ (we assume to be true health status). The graphical representation illustrates
$p(x, y, t, z) = p(z)p(x|z)p(t|x)p(y|t, z)$

SAVAE: ELBO

$$
\begin{aligned}
\ln p_\theta(\mathbf{x}, y, \mathbf{t}) &\geq \text{ELBO} \\
&= \mathbf{E}_{q_\phi}\left[\ln \frac{p_\theta(\mathbf{x}, \mathbf{t}, y, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x}, y)}\right] \\
&= \mathbf{E}_{q_\phi}[\ln p_\theta(\mathbf{z}) + \ln p_\theta(\mathbf{x}|\mathbf{z}) + \ln p_\theta(\mathbf{t}|\mathbf{x}) + \ln p_\theta(y|\mathbf{t}, \mathbf{z}) \\
&- \ln q_\phi(\mathbf{z}|\mathbf{x}, y)].
\end{aligned} \tag{1}
$$

where

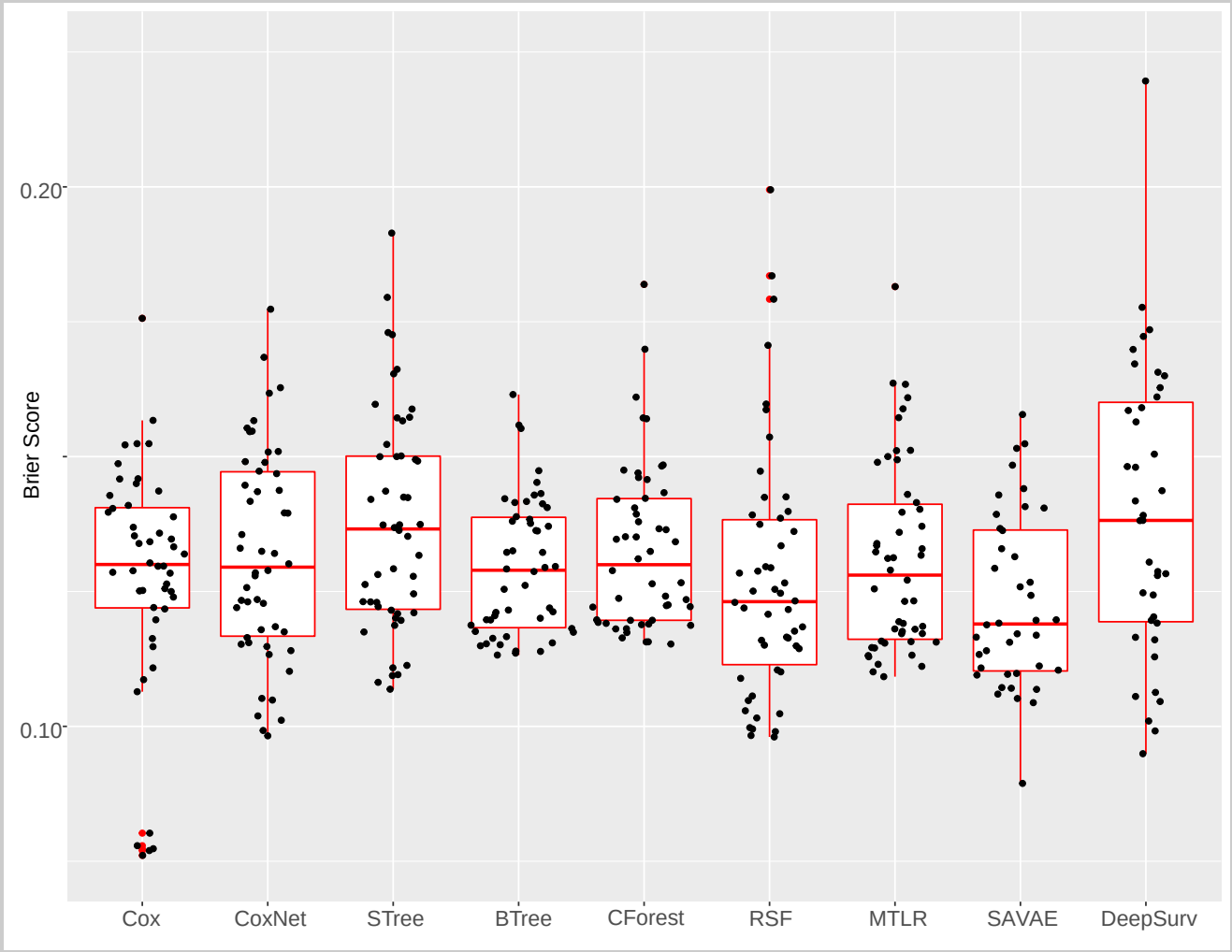
$$
\log p_\theta(y|\mathbf{t}, \mathbf{z}) = \delta \log f_\theta(y|t, \mathbf{z}) + (1 - \delta) \log S_\theta(y|\mathbf{t}, \mathbf{z}), \tag{2}
$$

with $\delta = 1$ if y is observed and 0 fi y is right-censored.

Results

- Our model outperforms the Cox proporitonal hazard benchmark according to the Brier Score and also outperforms a plethora of new ML models.
- When experimenting with a wide range of models:
  1. Modern models suffer from accessibility issues and
  2. provide no improvements over CoxPH.

Comparative results



Projet #4 : HWD+ a new handwritten digit data set

Motivations

Inspired by the now-famous *MNIST data set*.

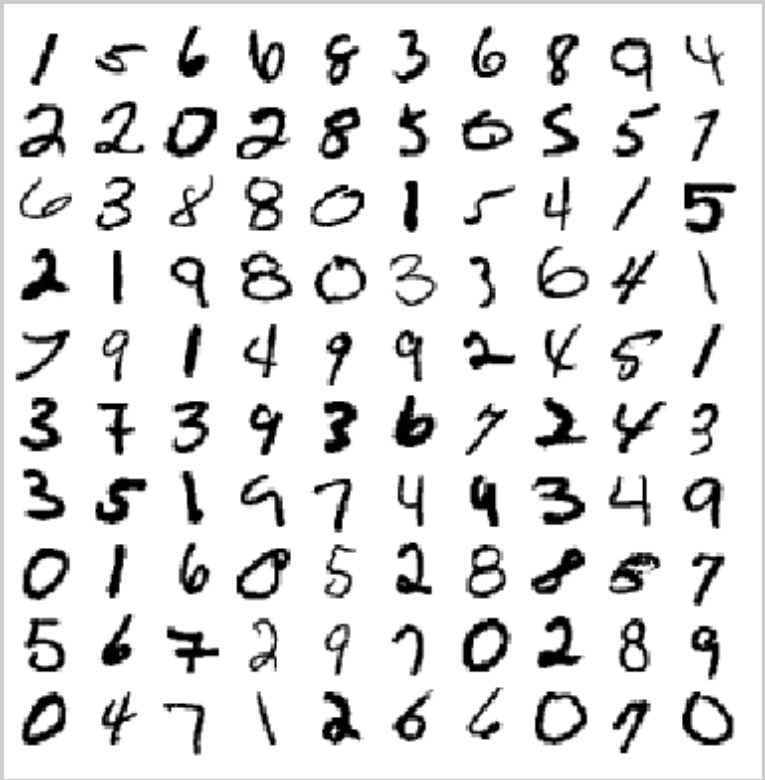


Figure: Sample from the *MNIST data set*.

Motivation

- Since writing styles depend on the individual, can we do writer identification solely using digits?
- We collect a data set.
- Different because MNIST:
  1. has small and standardize images,
  2. as only one attached variable,
  3. possible to achieve very high accuracy.
- Similar to MNIST because:
  1. we want to experiment with semi-supervised models,
  2. we want to provide an easy-to-use alternative to MNIST.

Data set

- The COVID strikes back.
- 97 writers, 14 replications of every digit for a total of 13 580 images in high resolution (500 x 500).
- Attached variables: digit, ID, age, biological gender, height, language, handiness, education level and main writing medium.
- Publicly available on my website in multiple formats.

Data set: a sample

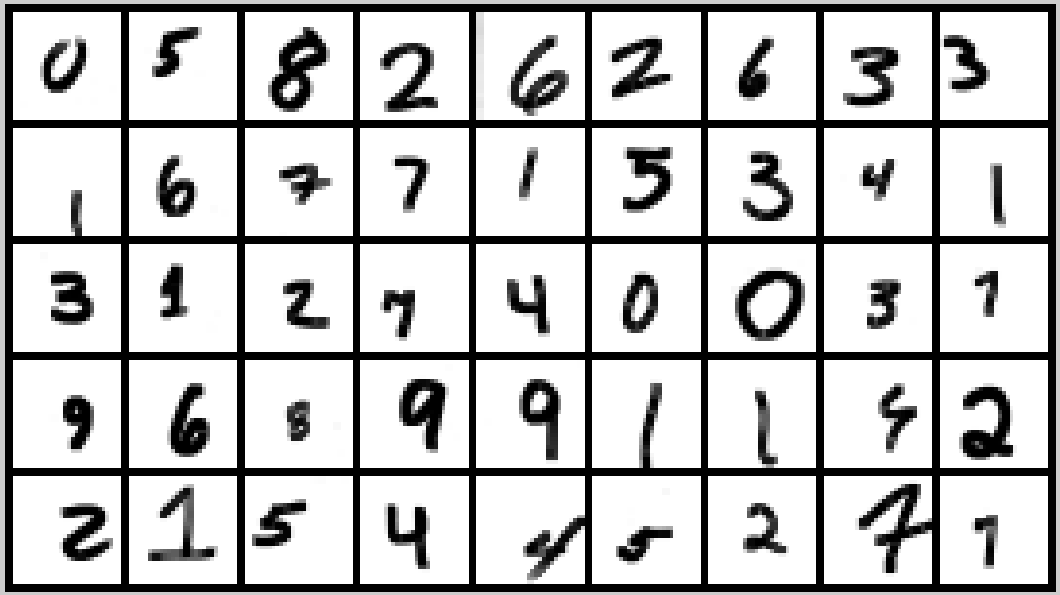


Figure: Sample of 45 images.

Questions

1. Predict writers and other characteristics (not directly the content of the image).
2. Evaluate the effect of the image resolution (performance vs accessibility).
3. Leverage the similarity with the MNSIT data set in semi-supervised application.
4. Generate images of new digits that mimics a writing style (ID).

Controlled image generation

- Let us discuss (4). Can we decide the digit or the style?
- Yes, using the generative component of VAEs made for classification.
- Controlled image generation is sampling on a conditional distribution.
- *Our assumption:* If there exist a signal between the image and the variable then we can use this variable for controlled image generation.

Results : controlled image generation

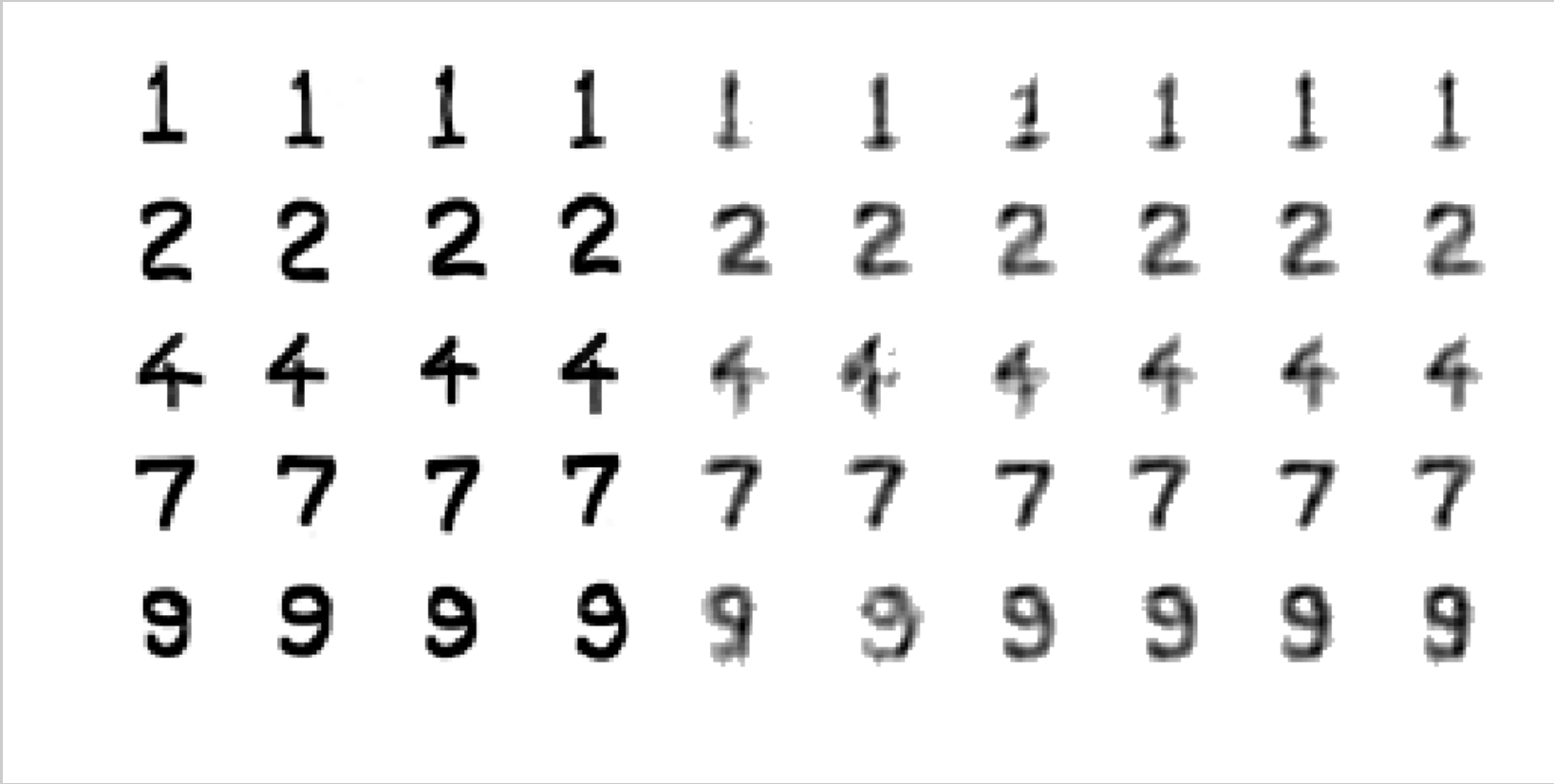


Figure: Example with ID12.

Results : controlled image generation

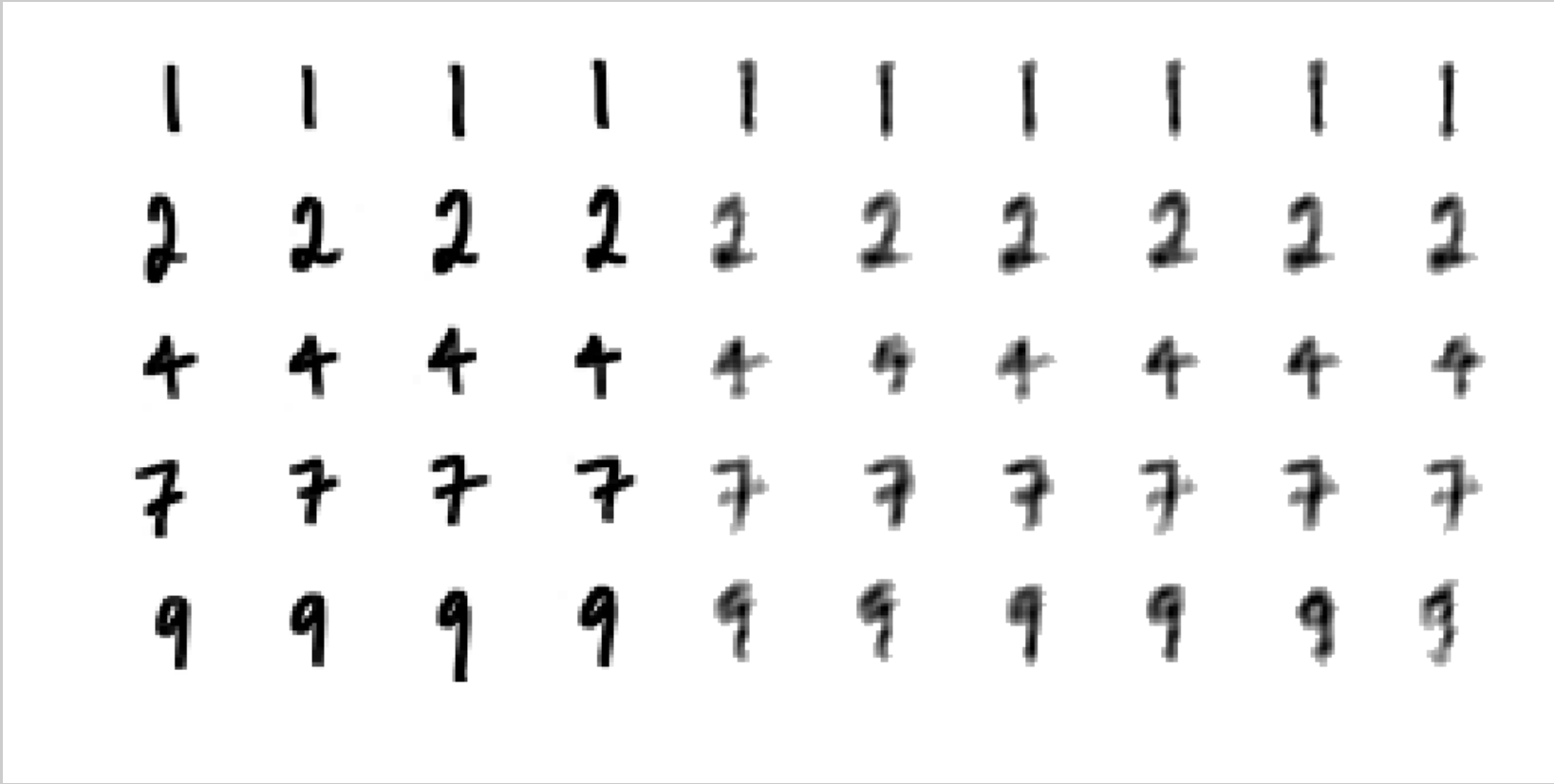


Figure: Example with ID14.

Results : controlled image generation

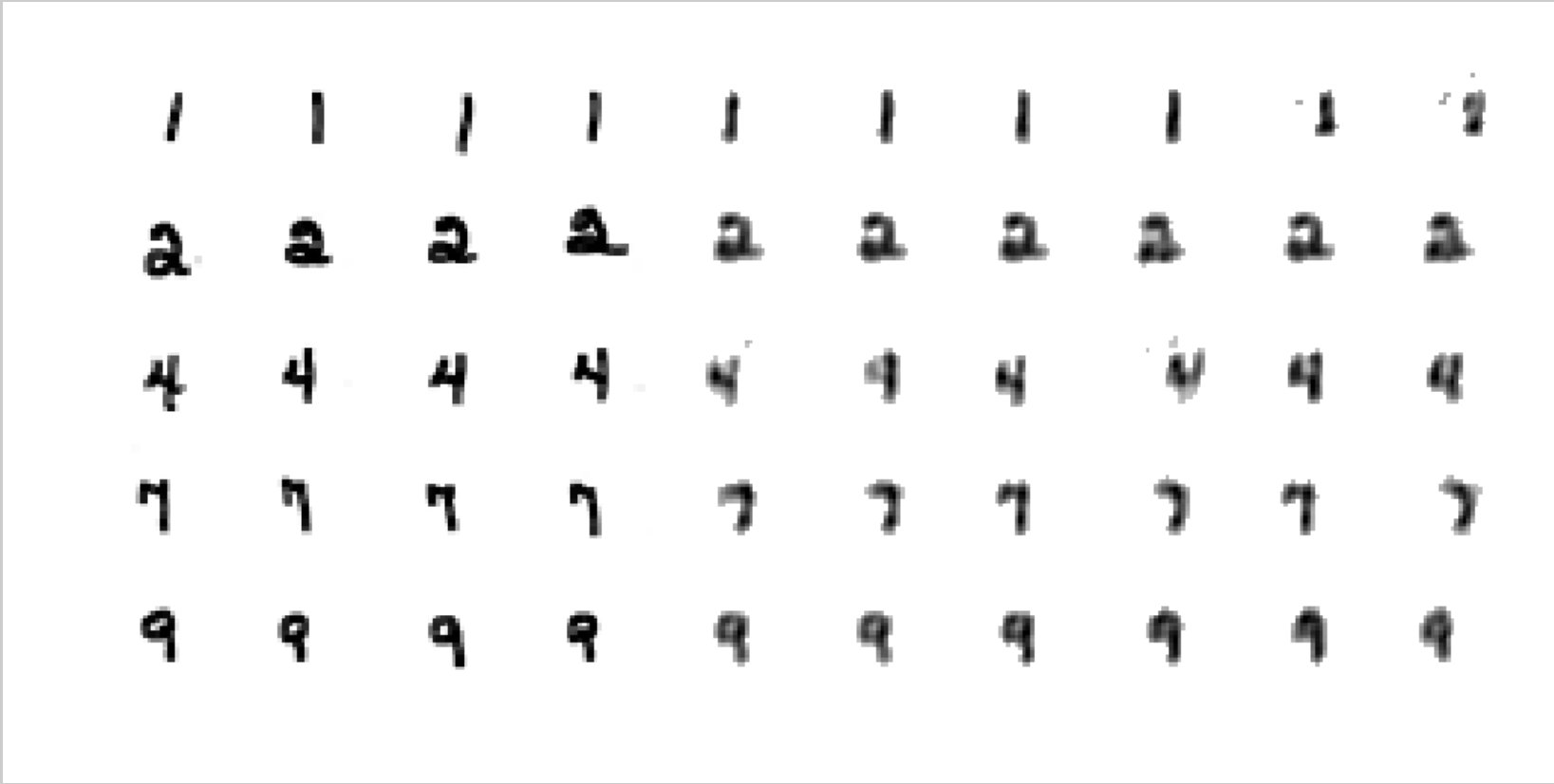


Figure: Example with ID29.

Results : controlled image generation

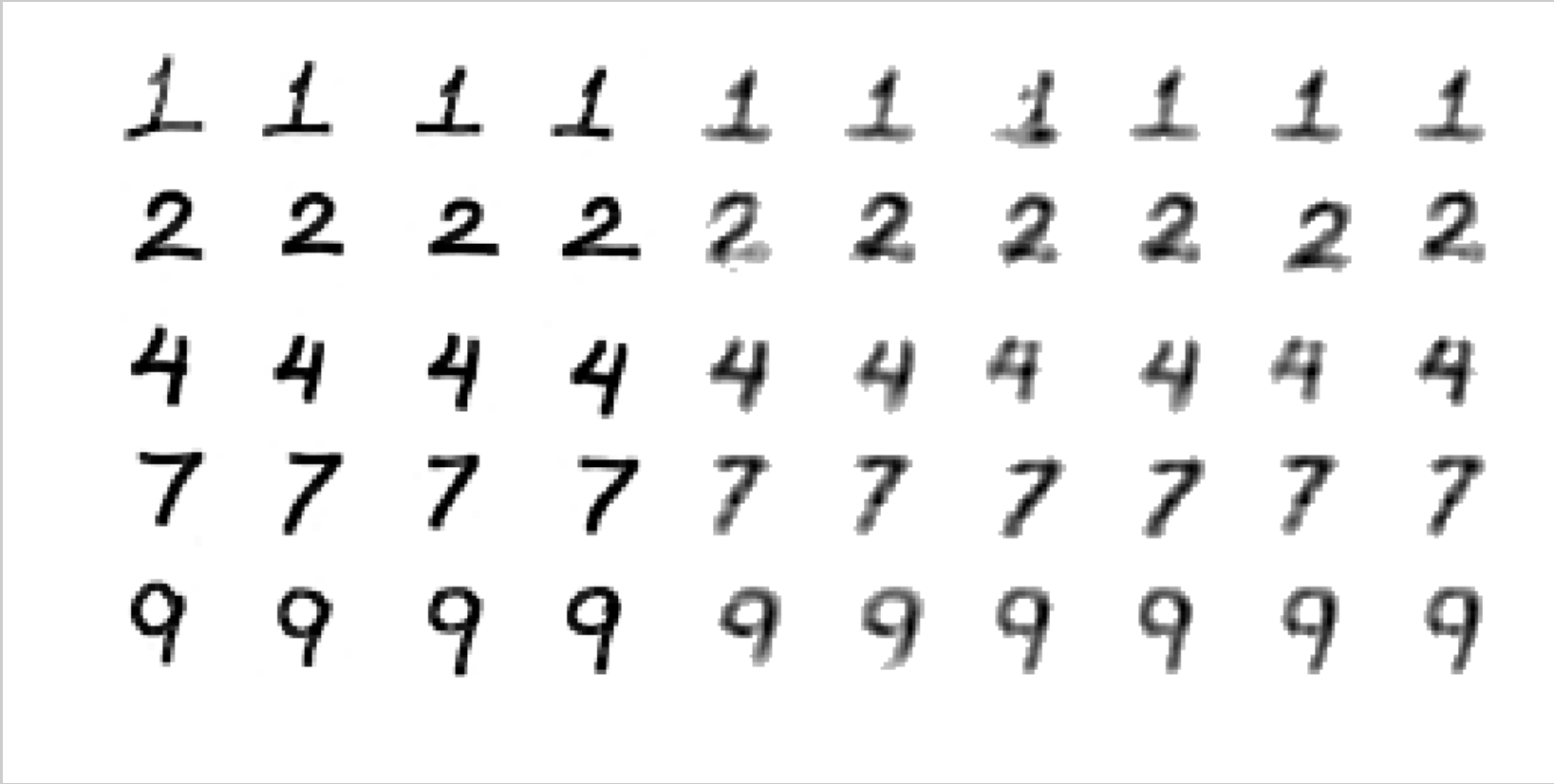
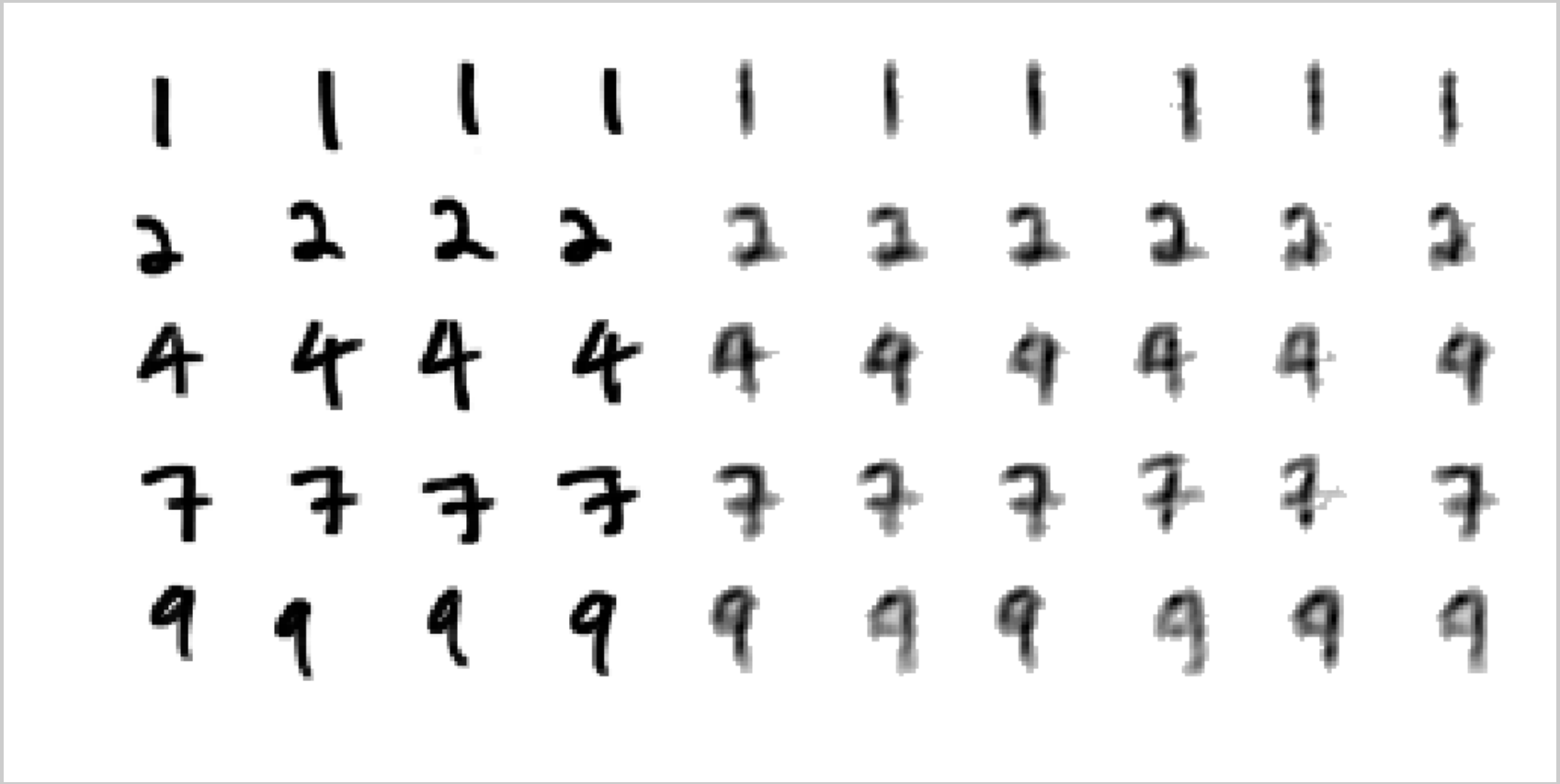


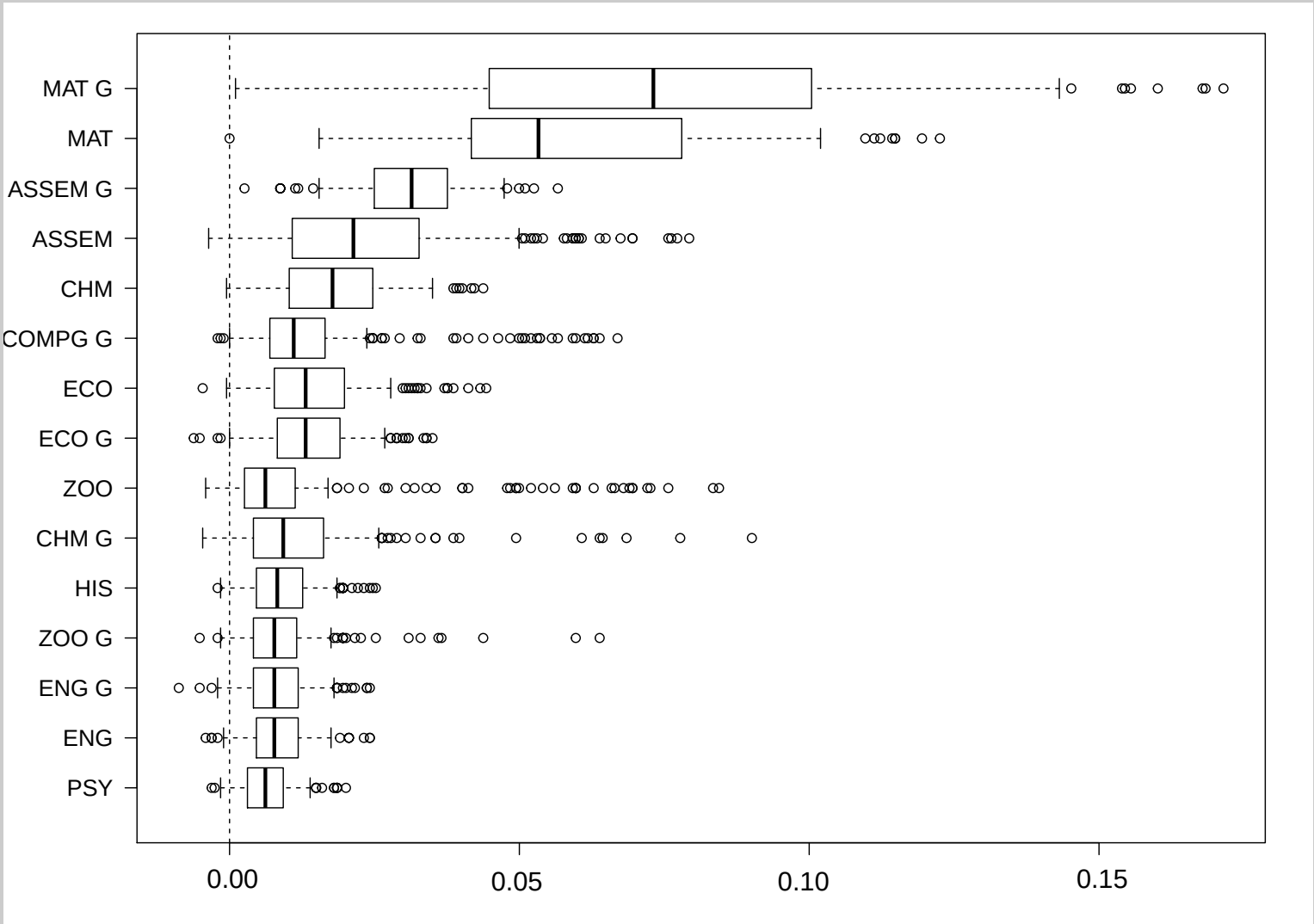
Figure: Example with ID70.

Results : controlled image generation

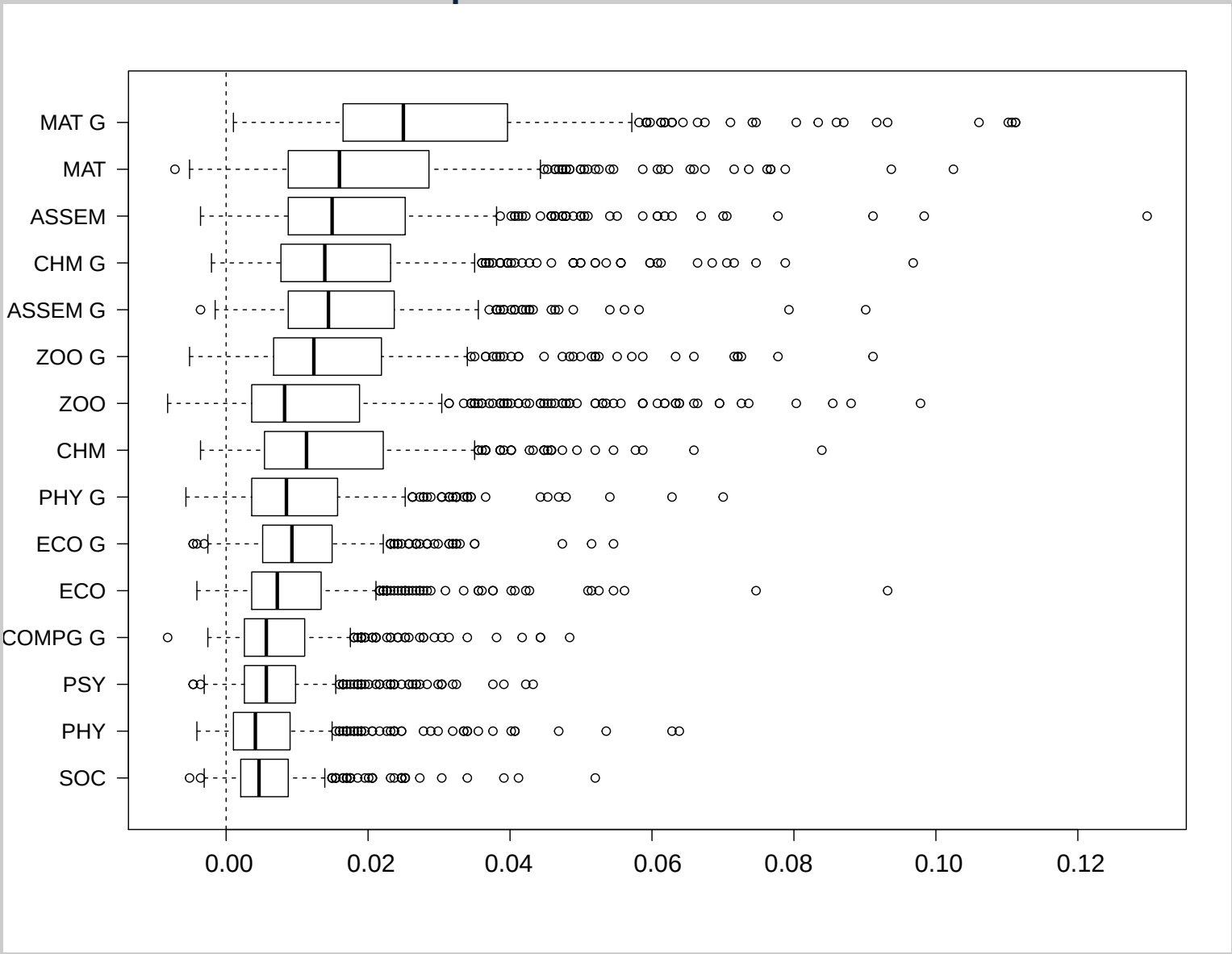


Thank you!

Results : variable importance



Results : variable importance



Theoretical intuition

- Assuming $\mathbf{X} \times \mathbf{Y} \sim \mathcal{D}$ (i.i.d)
- The true loss of a classifier h is: $L_{\mathcal{D}}(h) = \mathbf{P}_{\mathcal{D}}[h(x) \neq y]$
- Which we estimate with the empirical loss on data set S: $L_S(h) = \frac{|\{i \in [n] : h(x_i) \neq y_i\}|}{n}$
- The true loss can be decomposed (bias-complexity trade-off)

$$\begin{aligned} L_D(h_S) &= \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + (L_D(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)). \\ &= e_{\text{app}}(\mathcal{H}) + e_{\text{est}}(h_S). \end{aligned} \tag{3}$$

where $\mathcal{H}$ is the set of all classifiers considered (hypothesis class).

Theoretical intuition

- When using BESTree, we select a subset of decision trees $\mathcal{H}_{BESTree} \in \mathcal{H}_{Trees}$.
- Which reduces the estimation error.
- **Bold assumption**: $\underset{h \in \mathcal{H}_T}{\operatorname{argmin}} L_D(h) \in \mathcal{H}_B$, the best tree classifier belong in subset defined.
- We reduce the estimation error and we hope to keep the approximation error the same.

BESTree: Theoretical intuition

$$\begin{aligned} L_D(h_S(\mathcal{H}_T)) &= \min_{h \in \mathcal{H}_T} L_{\mathcal{D}}(h) + e_{\text{est}}(h_S(\mathcal{H}_T)). \\ &= \min_{h \in \mathcal{H}_B} L_{\mathcal{D}}(h) + e_{\text{est}}(h_S(\mathcal{H}_T)). \\ &\geq \min_{h \in \mathcal{H}_B} L_{\mathcal{D}}(h) + e_{\text{est}}(h_S(\mathcal{H}_B)). \\ &= L_D(h_S(\mathcal{H}_B)), \end{aligned} \tag{4}$$

(More details in the article)

Quick introduction to VAEs

- A VAE is a latent variable model (such as GMMs).
- The latent variable increase the expressiveness and complexity of the observed-data distribution.
- Assume there is an hidden (unobserved) phenomenon affecting the variables we observe.
- $\mathbf{x}$ is the observed variable of size m and $\mathbf{z}$ the latent variable of size d. However, $\mathbf{z}$ is continuous.
- Assuming $p_\theta(\mathbf{z}) = N(0, I)$ and $p_\theta(\mathbf{x}|z) = N(\mu_x, I\sigma_x^2)$.

Quick introduction to VAEs

- We suppose the set of paramaters $\theta = [\mu_x, \sigma_x]$ is a continuous function of z.
- $\theta(z) = [\mu_x(z), \sigma_x(z)]$ is a continuous function.
  $\theta(z) : \mathbb{R}^d \to \mathbb{R} \times \mathbb{R}^+$.
- θ is a neural network function that requires training (parameter estimation).
  1. Flexible and expressive: universal function estimator in specific cases.
  2. Quite easy to train with back-propagation (lots of packages and online support)
  3. However turn the posterior $p_\theta(\mathbf{z}|x)$ intractable analytically.

Quick introduction to VAEs

- Solution is borrowed from variational Bayes: we approximate $p_\theta(\mathbf{z}|x)$ with a variational distribution $q_\varphi(\mathbf{z}|x)$.
- We chose $q_\varphi(\mathbf{z}|x) = N(\mu_z, \sigma_z^2 I)$.
- The parameters $\varphi(x) = [\mu_z(x), \sigma_z(x)]$ are also obtained as a neural network function taking x as input.
- We cannot maximize directly the log-likelihood nor use EM.
- The proposed solution is to maximize a lower bound of $\log p(x)$, the ELBO (*Evidence Lower BOund*).

ELBO

$$
\begin{aligned}
\log p(x) &= \mathbf{E}_{q(z|x)}[\log p(x)] \\
&= \mathbf{E}_{q_\varphi(z|x)}\left[\log\left(\frac{p(x,z)}{p(z|x)}\right)\right] \\
&= \mathbf{E}_{q(z|x)}\left[\log\left(\frac{p(x,z)q(z|x)}{q(z|x)p(z|x)}\right)\right] \\
&= \mathbf{E}_{q(z|x)}\left[\log\left(\frac{p(x,z)}{q(z|x)}\right)\right] - \mathbf{E}_{q(z|x)}\left[\log\left(\frac{p(z|x)}{q(z|x)}\right)\right] \\
&= \mathcal{L}(q_\varphi, p_\theta) + KL(q_\varphi||p_\theta).
\end{aligned}
\tag{5}
$$

ELBO

$$\mathcal{L}(q_\varphi, p_\theta) = \mathbf{E}_{q_\varphi(z|x)} \left[\log p_\theta(z) + \log p_\theta(x|z) - \log q_\varphi(z|x)\right] \quad (6)$$

- The difference between $\log p(x)$ and $\mathcal{L}(q_\varphi, p_\theta)$ is $KL(q_\varphi || p_\theta)$
- This expectation is in term estimated with a Monte Carlo sample.

VAE : Training algorithm

Algorithm : Train a VAE

1) Process observation x through the NNs φ
and obtain $\mu_z(x)$ and $\sigma_z(x)$
2) Sample z from $q_{\varphi(x)}(z|x)$ (Monte Carlo Sample)
3) Process the Monte Carlo sample z through the NNs θ
and obtain $\mu_x(z)$ and $\sigma_x(z)$
4) Compute $\ln p_{\theta(x)}(z) + \ln p_{\theta(z)}(x|z) - \ln q_{\varphi}(z|x)$
the ELBO Monte Carlo estimate.
5) Maximize the ELBO with respect to the parameters
of φ and σ using back propagation
Repeat 1-5 until convergence.

SAVAE: ELBO

$$\begin{aligned}\text{ELBO} &= \mathbf{E}_{q_\phi}\left[\ln \frac{p_\theta(\mathbf{x}, t, y, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x}, y)}\right] = \mathbf{E}_{q_\phi}\left[\ln p_\theta(\mathbf{x}, t, y, \mathbf{z}) - \ln q_\phi(\mathbf{z}|\mathbf{x}, y)\right] \\ &= \mathbf{E}_{q_\phi}[\ln p_\theta(\mathbf{z}) + \ln p_\theta(\mathbf{x}|\mathbf{z}) + \ln p_\theta(t|\mathbf{x}) + \ln p_\theta(y|t, \mathbf{z}) \\ &- \ln q_\phi(\mathbf{z}|\mathbf{x}, y)].\end{aligned} \tag{7}$$

where

$$\log p_\theta(y|t, \mathbf{z}) = \delta \log f_\theta(y|t, \mathbf{z}) + (1 - \delta) \log S_\theta(y|t, \mathbf{z}), \tag{8}$$

with $\delta = 1$ if y is observed and 0 fi y is right-censored.

Distributions

$$p_\theta(\mathbf{x}|\mathbf{z}) = \prod_{j=1}^{D_x} p_\theta(x_j|\mathbf{z}) \tag{9}$$

$$p(\mathbf{t}_i|x) = \text{Ber}(\hat{\pi}_i) \text{ for } i \in \{1, 2\}. \tag{10}$$

$$p(\mathbf{y}|t, z) = \text{Weibull}(\lambda, K) \tag{11}$$

$$q(\mathbf{z}|\mathbf{x}, y) = \mathcal{N}(\mathbf{z}|\mu, \sigma^2 I) \tag{12}$$

$$\theta = f_2(\mathbf{W}_2 f_1(\mathbf{W}_1 \mathbf{z})) \tag{13}$$

$$[\pi_1, \pi_2] = f_4(\mathbf{W}_4 f_3(\mathbf{W}_3 \mathbf{x})) \tag{14}$$

$$[\lambda, K] = f_6(\mathbf{W}_6 f_5(\mathbf{W}_5 [\mathbf{z}, t])) \tag{15}$$

$$[\mu, \sigma] = f_8(\mathbf{W}_8 f_7(\mathbf{W}_7 [\mathbf{x}, y])). \tag{16}$$

Prediction

Finally we can estimate the complete survival distribution (λ,K) $p(y|t,\mathbf{x})$ using importance sampling:

$$p(y|t,\mathbf{x}) \approx \sum_{l=1}^{L} w_l p_\theta(y|t,\mathbf{z}_l) \tag{17}$$

where :

$$w_l = \frac{p_\theta(\mathbf{x}|\mathbf{z}_l)}{\sum_{k=1}^{L} p_\theta(\mathbf{x}|\mathbf{z}_k)} \tag{18}$$

Sophisticated classifiers

Simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification paradigm.

- David J. Hand

Motivation

Using a simple VAE we can see digits from similar *style* are close to one another.

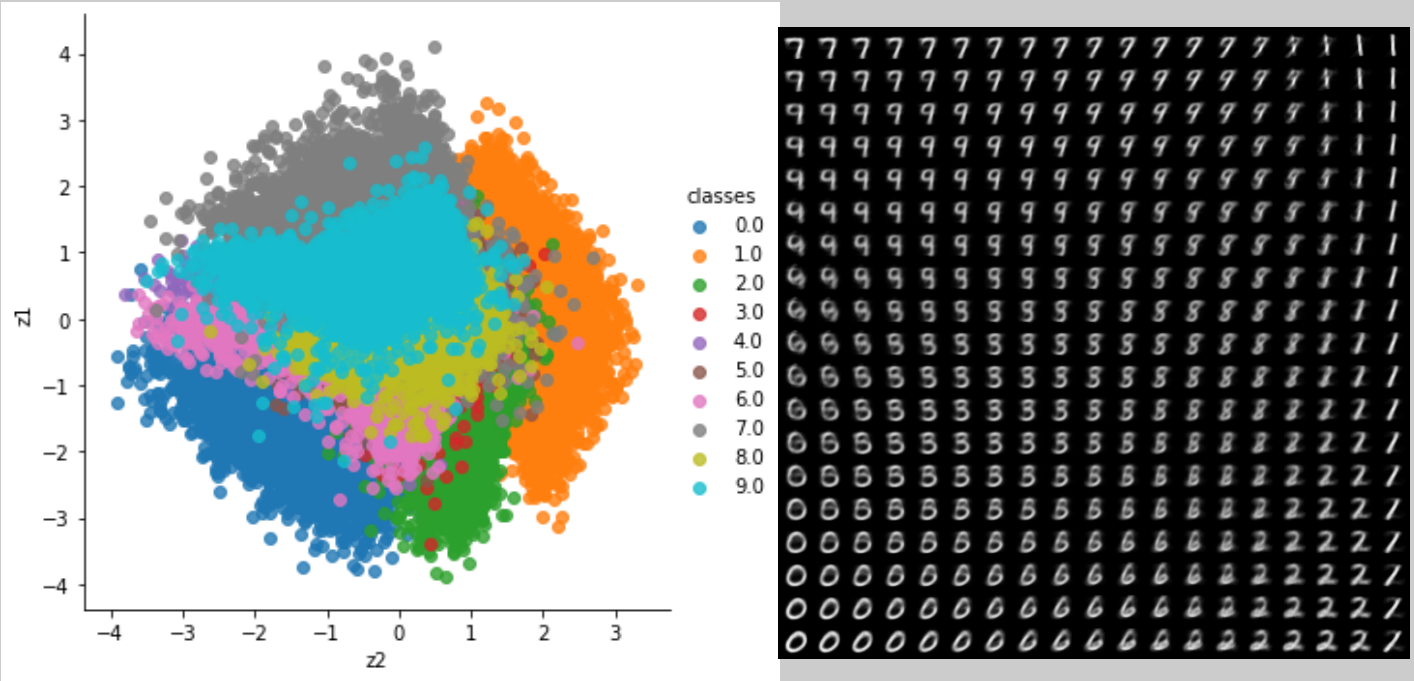


Figure: A Latent representation of the *MNIST data set*.

Data collection

- Our target was to gather handwritten digits from 200-300 students at UofT.
- We were on the process of booking room for late March.
- The COVID strikes back.
- We had to settle for mail packages and ended up with 97 writers.

Data collection

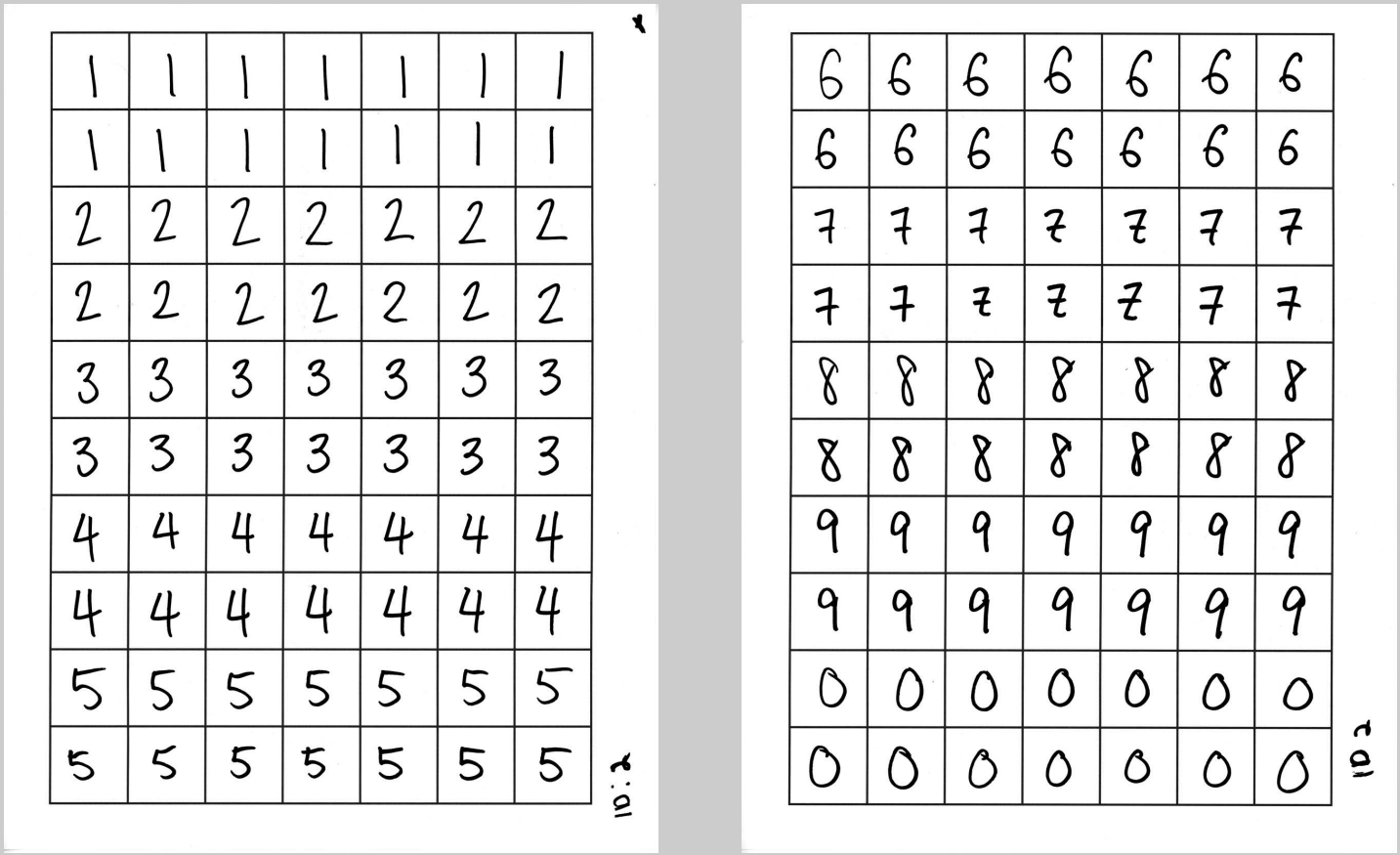


Figure: Data sheets example.

Results

1. Wide range of signal: High signal for the digit, medium for the ID, the first language, the hand and education level and no signal for the gender and writing medium.
2. For a fixed run time, using more complex technique on low-resolution data set leads to better accuracy than simple techniques on high-resolution data set.
3. Using MNIST as unlabelled data set significantly increases the accuracy on the classification tasks.

First look at the variable predictability

	CNN		Committee CNN	
	Mean	S.D.	Mean	S.D.
Digit	0.9399	0.0143	**0.9762**	0.0013
ID	0.3473	0.0136	**0.6195**	0.0063
Gender	0.5367	0.0183	0.5483	0.0372
Language	0.6792	0.0322	**0.7621**	0.0626
Hand	0.7940	0.0285	**0.8304**	0.0499
Education	0.4117	0.0222	**0.4726**	0.0343
Writing	0.4585	0.0225	0.4782	0.0372

Semi-supervised learning with VAEs

- Can we integrate unlabelled data points (S_u) to an existing labelled data set (S_l) in order to improve the performance.
- Our data base is different from the *MNIST data set*, but sufficiently similar to attempt such experiment.
- We used the M2 model (Kingma 2014).
- The unlabelled data points serve as regularization.

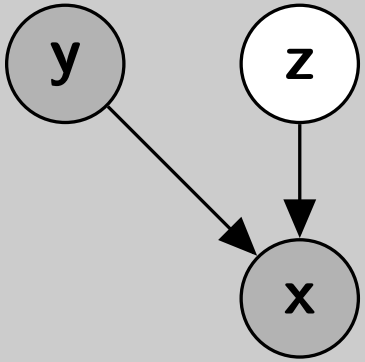
Semi-supervised classification: results

	CNN		M2	
	Mean	S.D.	Mean	S.D.
Digit	0.9399	0.0143	**0.9542**	0.0060
ID	0.3473	0.0136	**0.4174**	0.0099

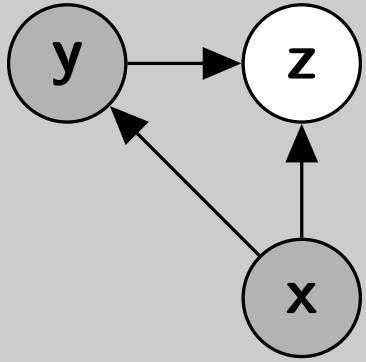
Higher resolution

	CNN(28x28)		CNN(100x100)	
	Mean	S.D.	Mean	S.D.
Digit	0.9399	0.0143	**0.9683**	0.0044
ID	0.3473	0.0136	**0.3675**	0.0224
Gender	0.5367	0.0183	0.5354	0.0410
Language	0.6792	0.0322	**0.7284**	0.0441
Hand	0.7940	0.0285	**0.8129**	0.0355
Education	0.4117	0.0222	**0.4466**	0.0368
Writing	0.4585	0.0225	0.4612	0.0234

VAE: M2 model



(a) Generative network. Assumes $p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{y}) = p_\theta(\mathbf{z})p_\theta(\mathbf{y})p_\theta(\mathbf{x}|z, y)$.



(b) Inference network. Given x and y we can estimate get z with $q_\varphi(\mathbf{z}|x, y)$. If y is missing, we predict it with $q_\varphi(\mathbf{y}|x)$.

Figure: Graphical representation of the M2 model.

VAE: M2 model

$$\begin{aligned}\log p_\theta(\mathbf{x}, \mathbf{y}) &\geq \mathbf{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log p_\theta(\mathbf{z}) + p_\theta(\mathbf{y}) + p_\theta(\mathbf{x}|z, y) - \log q_\varphi(\mathbf{z}|x, y)\right] \\ &= \mathcal{L}(x, y)\end{aligned} \tag{19}$$

$$\begin{aligned}\log p_\theta(\mathbf{x}) &\geq \mathbf{E}_{q(\mathbf{z},\mathbf{y}|x)}\left[\log p_\theta(\mathbf{z}) + p_\theta(\mathbf{y}) + p_\theta(\mathbf{x}|z, y) - \log q_\varphi(\mathbf{z}, \mathbf{y}|x)\right] \\ &= \sum_y \left[q_\varphi(\mathbf{y}|x)(\mathcal{L}(x, y))\right] + \mathcal{H}(q_\varphi(\mathbf{y}|x)) \\ &= \mathcal{U}(x)\end{aligned} \tag{20}$$

$$\mathcal{J} = \sum_{S_l} \mathcal{L}(x, y) + \sum_{S_u} \mathcal{U}(x) \tag{21}$$

VAE: M2 model

However, this way we train the classifier $q_\varphi(\mathbf{y}|x)$ strictly on unlabelled data. The proposed solution (Kingma 2014) is to modify the objective function:

$$\mathcal{J}^\alpha = \mathcal{J} + \alpha \mathbf{E}_{S_l} \left[-\log q_\varphi(\mathbf{y}|x) \right]. \tag{22}$$

This can be framed as a penalized optimization problem:

$$\mathcal{J}^\alpha = \alpha \mathcal{J} + \mathbf{E}_{S_l} \left[-\log q_\varphi(\mathbf{y}|x) \right] \tag{23}$$

Projet

$$\mathcal{J}^{\alpha} = \alpha\mathcal{J} + \mathbf{E}_{S_l}\left[-\log q_{\varphi}(\mathbf{y}|x)\right] \tag{24}$$

When $\alpha = 0$ we go back to the fully supervised classifier.

We assume that the VAE *component* serves as regularization (penalization).

Image generation

- Since the distribution p_θ is fully defined we can sample from it: allows us to generate new images x.
- Ancestral sampling : $z \sim p_\theta(z)$ then $x \sim p_\theta(x|z)$.
- This simple process will generate images of random digits with a random style.

Controllable generation : project

- Modèles conçus pour l'analyse semi-supervisée: une grande précision est atteinte avec peu d'observations annotées.
- *Hypothèse:* Si un signal existe entre une variable et son image, nous pouvons utiliser cette variable lors de la génération.
- *Notre idée*: Nous voulons utiliser ce principe pour décider des caractéristiques de l'image que l'on contrôle.
- *Exemple*: Télécharger des images de ciel sur internet et assigner une variable binaire pour caractériser *dégagé* ou *nuageux*.
- S'il existe un signal, nous pourrons générer une image d'un ciel et contrôler si celui-ci est dégagé ou nuageux.

Controllable generation : project

- Nous voulons *mathématiser* certains concepts . Comment évaluer notre *contrôle* ?
- Par exemple, *la force du contrôle* peut être évaluée à l'aide de l'information mutuelle :

$$\begin{aligned} I(Y,X) &= \int_y \int_x p(y,x) \log \frac{p(y,x)}{p(y)p(x)} dxdy \\ &= H(X) - H(X|Y) \end{aligned} \tag{25}$$

Génération contrôlée : futur projet

- Nous sommes aussi intéressés par l'interpolation et l'extrapolation
- et par l'interprétation et la séparabilité des variables de contrôle.
- Nous voulons fournir une définition mathématique de ces composantes en parallèle d'une définition intuitive.

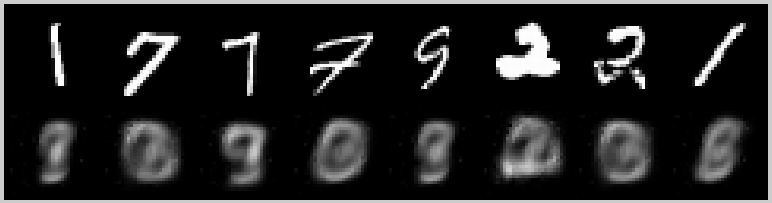
Projet #5 : Problèmes et piste de solutions

Les problèmes que révèlent nos expériences

Les problèmes que révèlent nos expériences

- Le problème est que les auteurs ont réussi a implémenter un VAE qui fonctionne mais que celui-ci n'est pas vraiment un VAE.
- La communauté d'apprentissage automatique est centré sur les implémentations, eux n'y voit pas de problèmes.
- Je suis un statisticien et j'y vois un problème.

Démonstration empirique: reconstruction



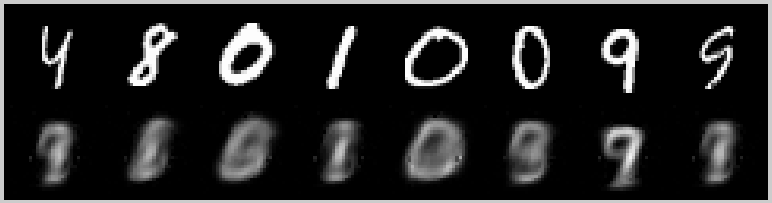
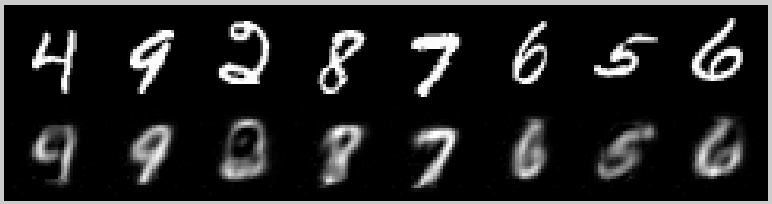


Figure: Images x et ses reconstructions $\mu_x(q_\varphi(x))$ produient par un VAE ou z est de dimension $d = 2$



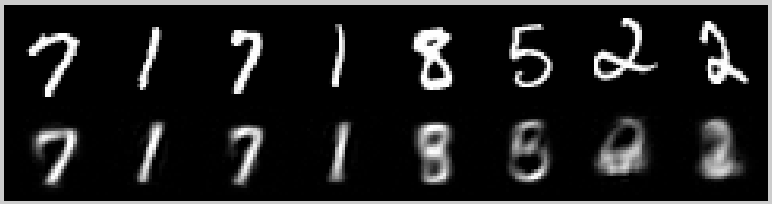


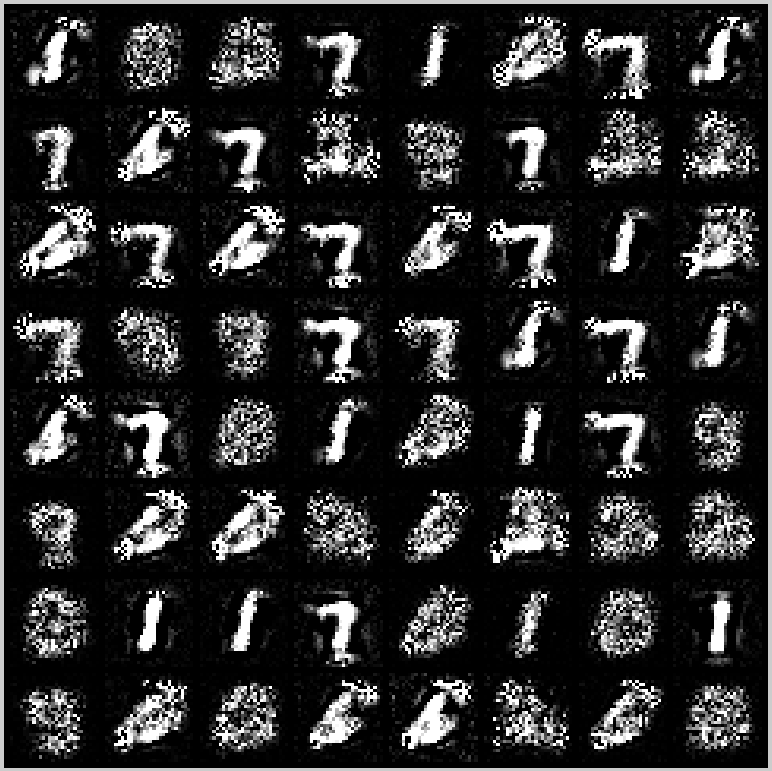
Figure: Images x et ses reconstructions $\mu_x(q_\varphi(x))$ produient par un VAE ou z est de dimension $d = 20$

Comparaison empirique: reconstruction

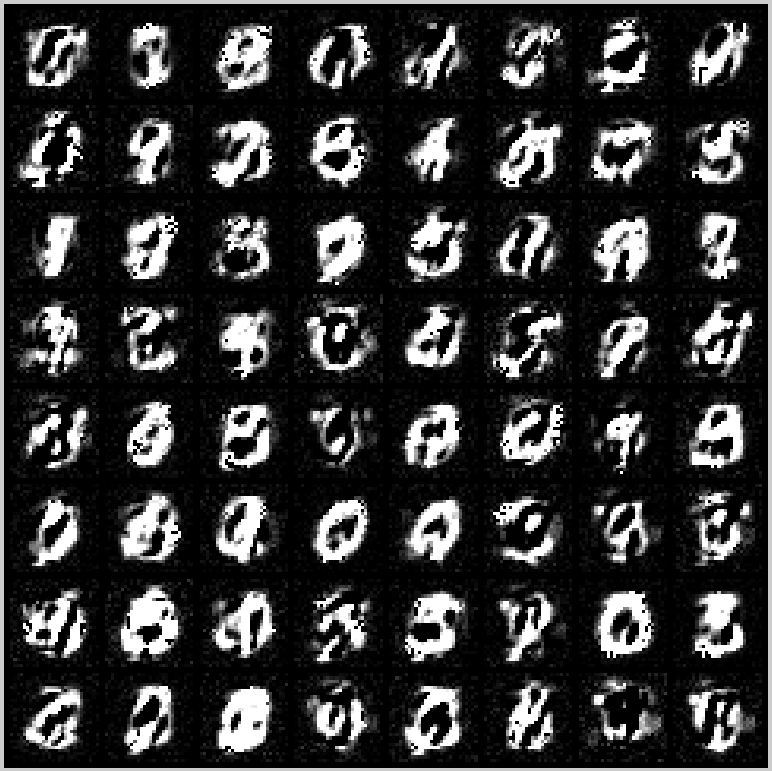
Figure: Reconstruction de l'image à gauche par PCA avec $d = 2, 10, 50, 250$.

L'analyse en composante principale fait mieux.

Démonstration empirique: génération



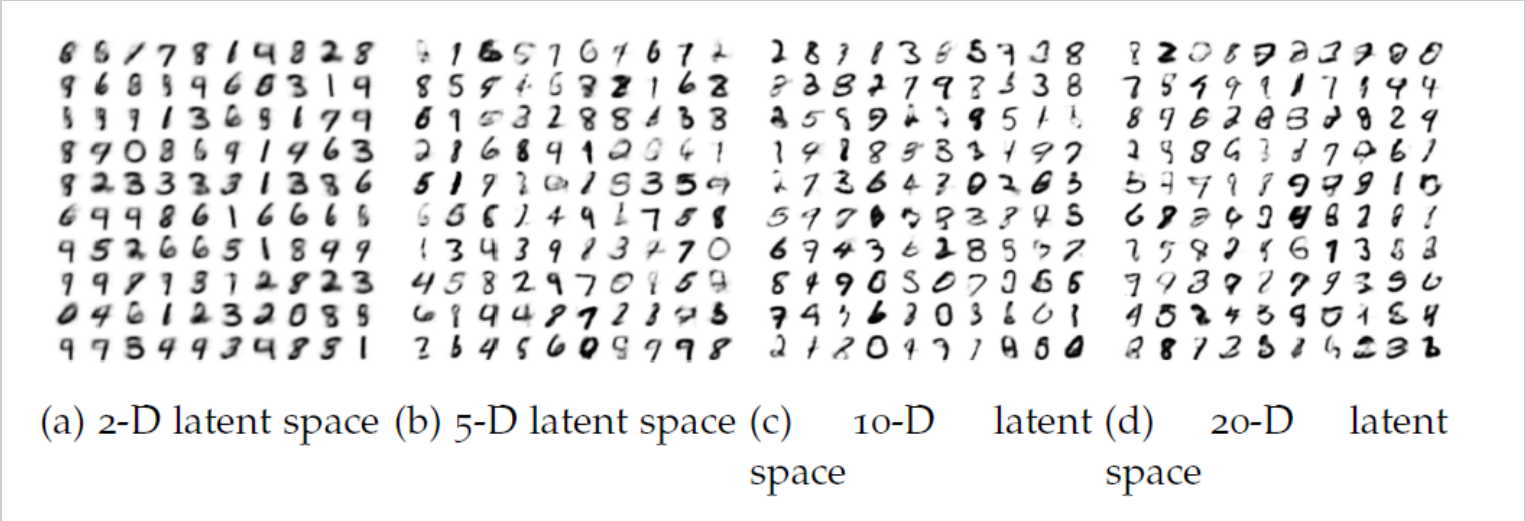
(a) $d = 2$



(b) $d = 20$

Figure: Sample obtained from the ancestral sampling described in the previous section.

Comparaison empirique: génération



(a) 2-D latent space (b) 5-D latent space (c) 10-D latent space (d) 20-D latent space

Figure: Snapshot of the result section of Kingma's thesis.

Les auteurs obtiennent un résultat très différent: que faisons-nous de mal ?

Que se passe-t-il ?

- Nous avons implémenter le modèle présenter à la section précédente, celui introduit dans l'article (et autres référence).
- Mes collègues (Vector Institue/Compute scientist) me suggère de petits *coding tricks*.
- Les implémentations en ligne, fournis par les auteurs et collaborateur, utilisent ces *coding tricks*.
- Le problème c'est que ces *coding tricks* modifient drastiquement le modèle mais personne ne l'admet.

Modification à l'implémentation

Nous allons discuter trois modifications majeurs faites lors de l'implémentation :

- β-VAE: Ajuster la régularisation du modèle
- Modifier la distribution observationelle.
- Modifer la procédure d'échantiollonage.

β-VAE

Rapel :

$$\begin{aligned}\mathcal{L}(\varphi,\theta) &= \mathbf{E}_{q_\varphi(z|x)}\left[\ln p_\theta(z) + \ln p_\theta(x|z) - \ln q_\varphi(z|x)\right] \\ &= \underbrace{\mathbf{E}_{q_\varphi(z|x)}\left[\ln p_\theta(x|z)\right]}_{\text{Reconstruction error}} - \underbrace{KL\left(q_\varphi(z|x)|p_\theta(z)\right)}_{\text{Regularization term}}\end{aligned} \tag{26}$$

β-VAE

L'objectif est de se donner le pouvoir de controller la ratio entre ces deux composantes :

$$\mathbf{E}_{q_\varphi(z|x)}\left[\ln p_\theta(x|z)\right] - \beta KL\left(q_\varphi(z|x)|p_\theta(z)\right) \tag{27}$$

β-VAE

Pour améliorer la reconstruction, il est proposé de sélectionner un $\beta < 1$

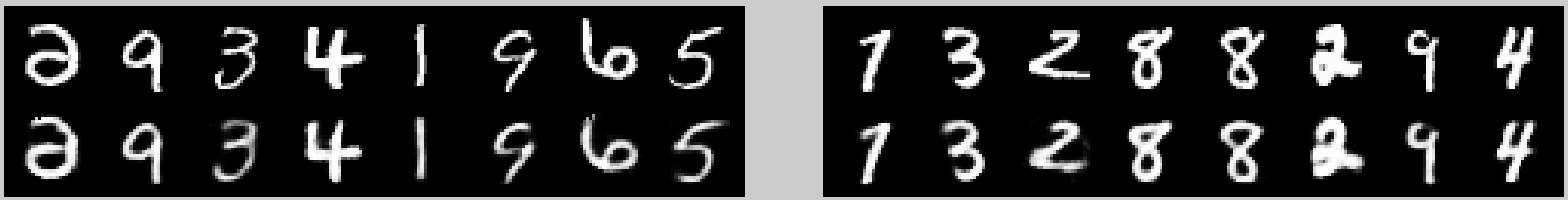


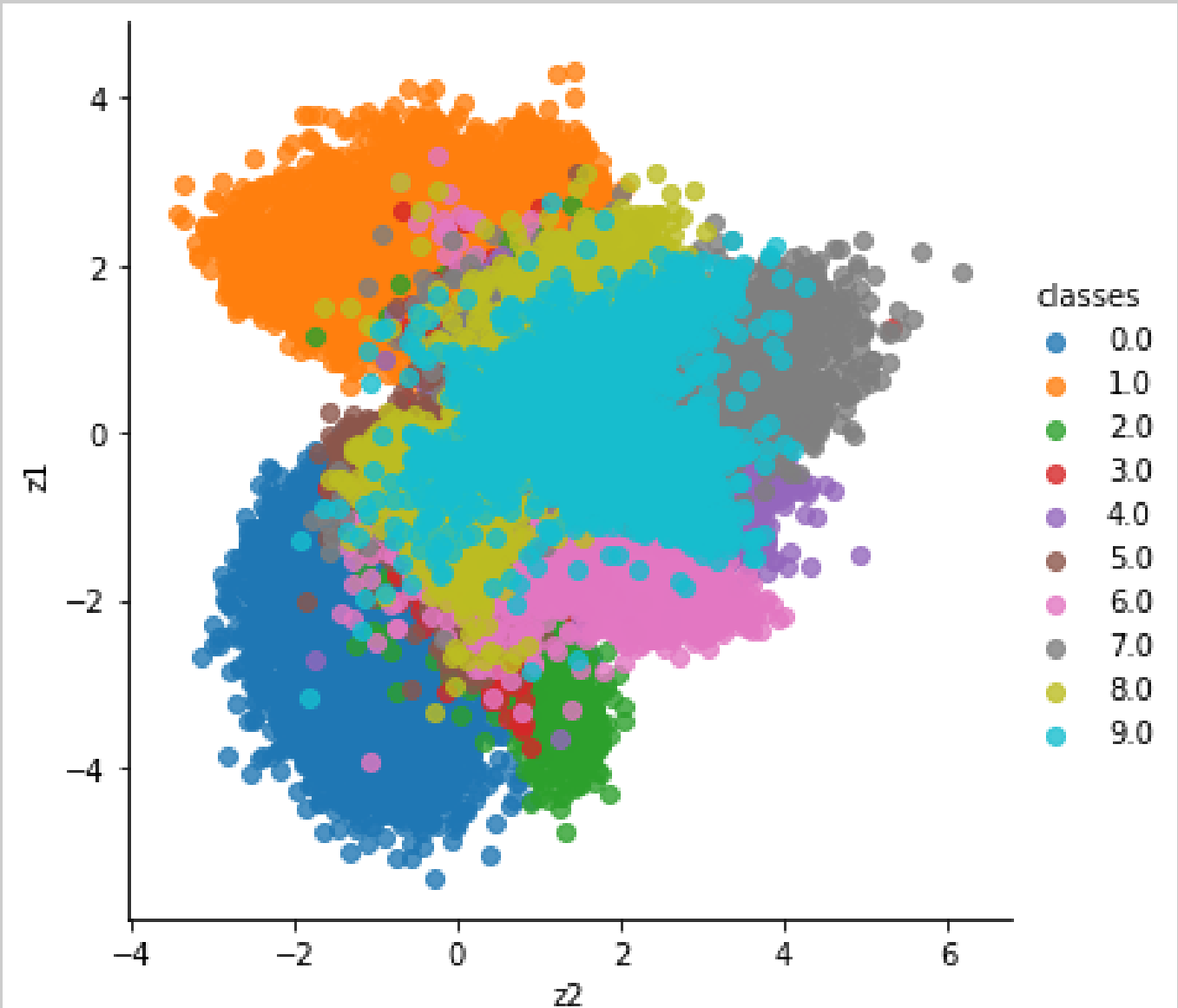
Figure: Images x on the top row and its reconstruction $\mu_x(q_\varphi(x))$ on the bottom row produced with a β-VAE with latent space of dimension $d = 20$

β-VAE

Nous démontrons que cette solution est probablématique pour trois raisons :

- La fonction objectif n'est plus une bonne inférieur de $\log p(x)$.
- β est un nouveau hyper-paramètres (encore!) difficile a fixer.
- Comme nous tirons un échantillon Monte Carlo de $q(z|x)$ pour entraîner $p(x|z)$, le manque de régularisation nuit aux capacités génératrice du VAE.

β-VAE: Visualisation de $q(z)$



β-VAE: cas limite

Discutons le cas limite où $\beta = 0$.

- Produit les meilleurs reconstructions.
- Élimine complètement la composante probabiliste sur z.
- Pour améliorer un VAE, on se rapproche d'un AE.

Distribution observationelle

Rapel :

$$\mathcal{L}(\varphi, \theta) = \underbrace{\mathbf{E}_{q_\varphi(z|x)}\left[\ln p_\theta(x|z)\right]}_{\text{Reconstruction error}} - \underbrace{KL\left(q_\varphi(z|x)|p_\theta(z)\right)}_{\text{Regularization term}} \tag{28}$$

- Les implémentation en ligne (tutorielle officielle de PyTorch & TensorFlow) remplace simplement $\ln p_\theta(x|z)$ par l'erreur quadratique $(x - \mu_x(z))^2$.
- Ceci est equivalent à fixer $\sigma_x = 1$:

Distribution observationelle

$$\begin{aligned} \ln p_\theta(x|z) &= \ln\left(\frac{1}{\sqrt{2\pi\sigma(z)^2}}\exp\left(\frac{-(x-\mu(z))^2}{2\sigma(z)^2}\right)\right) \\ &= -\frac{1}{2}\ln\left(2\pi\sigma(z)^2\right) - \frac{(x-\mu(z))^2}{2\sigma(z)^2} \end{aligned} \tag{29}$$

Si $\sigma_x(z) = 1$

$$-\frac{1}{2}\ln(2\pi) - \frac{(x-\mu(z))^2}{2} \propto -(x-\mu(z))^2$$

Échantillonage ancestral... ou pas

1. Échantilloner z de $N(0, I)$
2. Passer z dans le NN θ pour obtenir $\mu_x(z)$ et $\sigma_x(z)$
3. Échantilloner x de $N(\mu_x(z), \sigma_x(z))$
4. Retourner x

1. Échantilloner z de $N(0, I)$
2. Passer z dans le NN θ pour obtenir $\mu_x(z)$ et $\sigma_x(z)$
3. Retourner $\mu_x(z)$.

Échantillonage ancestral... ou pas

(a) $d = 2$

(b) $d = 20$

Figure: Sample obtained from $\mu_x(z)$ where $z \sim N(0, I)$.

VAE: notre perspective

Si nous utilisons l'erreur quadratique comme erreur de reconstruction ET que nous retournons $\mu_x(z)$ lors de la génération. x devient une fonction déterministique de z; x n'est plus aléartoire. $\tilde{x} = \mu(z)$ et cette reconstruction $\tilde{x}$ optimal en minimisant $\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x_i} - \tilde{\mathbf{x}}_\mathbf{i}||^2$.

Nous avons complétemment éliminé la composante probabliste de x et nous rapprochons d'un AE traditionnel.

VAE: notre perspective

En combinant ces résultats avec ceux obtenus précédemment lors de notre discussion sur les β-VAE; fixer $\beta = 0$ nous éliminons la composante probabililiste sur le code z.

Pour permettre au VAE d'avoir la meilleur reconstruction nous l'avons transformer en AE.

Il doit y avoir une autre solution.

VAE: piste de solution

- Nous croyons qu'il s'agit d'un problème d'intentifiabilité de variance.
- Dans un modèle à variables latentes, la variance observationnelle est divisé en deux partie: z et $x|z$.
- Ces problèmes ne sont pas présent dans PCA ou pPCA où une distribution de la variance est suggéré par le modèle (on maximise la variance de l'espace latent).

VAE: piste de solution

1. Il faut trouver une manière d'optimiser naturellement comment diviser la variabilité.
2. Permettre à la variabilité observationelle de s'exprimer différement: utiliser une matrice de coviariance plutot qu'une simple diagonal principale.

VAE: piste de solution

1. S'inspiré de pPCA.
2. S'inspiré de la recherche en statistique spatiale.