Ornstein Auto-Encoders

Youngwon Choi and **Joong-Ho Won***
Department of Statistics, Seoul National University, Republic of Korea
muha@snu.ac.kr, wonj@stats.snu.ac.kr

Abstract

We propose the Ornstein auto-encoder (OAE), a representation learning model for correlated data. In many interesting applications, data have nested structures. Examples include the VGGFace and MNIST datasets. We view such data consist of i.i.d. copies of a stationary random process, and seek a latent space representation of the observed sequences. This viewpoint necessitates a distance measure between two random processes. We propose to use Orstein's d-bar distance, a process extension of Wasserstein's distance. We first show that the theorem by Bousquet et al. (2017) for Wasserstein auto-encoders extends to stationary random processes. This result, however, requires both encoder and decoder to map an entire sequence to another. We then show that, when exchangeability within a process, valid for VGGFace and MNIST, is assumed, these maps reduce to univariate ones, resulting in a much simpler, tractable optimization problem. Our experiments show that OAEs successfully separate individual sequences in the latent space, and can generate new variations of unknown, as well as known, identity. The latter has not been possible with other existing methods.

1 Introduction

Most machine learning algorithms implicitly or explicitly assume that samples in the training and test datasets are drawn independently and identically from an unknown data distribution. However, this i.i.d. assumption is violated in many real-world tasks with *nested* data structures, i.e., when data were collected from grouped observational units.

As a concrete example, consider the VGGFace2 dataset [Cao *et al.*, 2018], an expansion of the famous VGGFace dataset [Parkhi *et al.*, 2015]. VGGFace2 is a large-scale face dataset containing 3.31 million images of 9131 identities. For each person, it contains 362.6 images on average, with minimum of 30 and maximum of 843. These portraits are highly correlated within a single person, violating the i.i.d. assumption. A similar issue arises in classification. The images of the MNIST dataset show strong correlations within a digit.

When the categories are fixed like the MNIST data, a popular approach is to model the data distribution with a finite mixture model or class-conditional models. However, if the number of categories (classes) is too large or not even fixed, the use of these models may not be desirable. For example, in the VGGFace2 data the number of classes is 9,131. Since the identities are randomly sampled, any model trained with this dataset must deal with the increasing number of classes for generalizability. Even with fixed categories, class imbalance is a big problem in learning with these models.

Random effects models [Diggle *et al.*, 2002; Fitzmaurice *et al.*, 2012] provide a flexible framework for handling both regimes. Applying those models is a standard approach in statistics when there are correlations among observational units within a group or subject. As an example, consider the random intercept model (with no slope):

$$y_j^i = \mu_0 + b^i + \varepsilon_j^i, \quad \varepsilon_j^i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2),$$
$$b^i \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_0^2), \quad b^i \perp\!\!\!\perp \varepsilon_j^i,$$

where y_j^i represent the jth observation within subject i. Note due to the presence of the random intercept b^i, the sequence of observations $\{y_j^i\}_{j=1}^{n_i}$ in subject i are correlated with correlation coefficient $\tau_0^2/(\sigma_0^2 + \tau_0^2)$. This is the simplest example of linear mixed effects models. In machine learning, Dundar et al. [2007] show that classifiers with a linear mixed effects model allow us to explicitly model the dependence in non-i.i.d. data. Differing number of samples between groups is naturally handled.

The reader may have noticed that the random intercept model defines an infinite exchangeable random process. Let $Y = (\dots, Y_{-1}, Y_0, Y_1, \dots)$ be a (doubly) infinite sequence with coordinates Y_j's are conditionally independent given B. If $B \sim \mathcal{N}(0, \tau_0^2)$ and $Y_j | \{B = b\} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0 + b, \sigma_0^2)$, then $\{y_j^i\}_{j=-\infty}^{\infty}$ is a realization of the ith i.i.d. copy of the random process Y. On the other hand, both VGGFace2 and MNIST data consist of exchangeable sequences nested within subjects or classes: the order of portraits of any given person does not affect any conceivable learning task.

The goal of this paper is to bring the nested data structure that arises from various applications down to generative la-

*Contact Author

tent variable modeling. If the latent variables share the nested structure of the observed variables, then the generative power of the latent space representation is likely to increase. As discussed above, this nesting often translates to i.i.d. observations of a correlated random process. Our main contributions are as follows:

- We introduce Ornstein's d-bar distance to the community, which is an optimal transport distance between random processes.
- We show that the theorem by [Bousquet *et al.*, 2017] extends to stationary processes and propose the Ornstein auto-encoder (OAE), which can be thought of as a stationary random process version of the Wasserstain auto-encoder (WAE) [Tolstikhin *et al.*, 2018].
- When exchangeability is assumed, we show that the optimization problem for OAE greatly reduces to almost the same as that of WAE, enabling a simple algorithm.
- We empirically show that the generative power of OAE surpasses state-of-the-arts. Importantly, OAE is robust to data imbalance and can generate new variations of unknown, out-of-training-set subjects, which has been impossible with other methods.
- We demonstrate that OAE can provide disentangled representations, i.e., latent variables are well-clustered by subjects. This capability has a potential applications in classification and recognition.

In the next section, we provide a necessary background. Section 3 introduces Ornstein's d-bar distance and develop OAE for stationary and exchangeable processes, respectively. In Section 4 we demonstrate the power of OAE using VGGFace2 and MNIST data. We conclude the paper in Section 5. Proofs, additional examples, and details of implementation are given in the Online Supplement available at https://tinyurl.com/y5x6ufuj.

2 Preliminaries

2.1 Notation

The space of observable variables is denoted by $\mathcal{X}$. The space of latent variables is denoted by $\mathcal{Z}$. We assume that both $\mathcal{X}$ and $\mathcal{Z}$ are standard measurable spaces so that conditional probability distributions are well-defined. Unless necessary, the associated event spaces $\mathcal{B}_X$ and $\mathcal{B}_Z$ are suppressed. We also assume that $\mathcal{X}$ is a complete, separable metric space equipped with metric d. Its Cartesian product space is denoted by $\mathcal{X}^n$ for $n = 1, 2, \ldots$; $n = \infty$ is allowed. Both random variables and random processes are represented by capital letters (e.g., X), and their realizations by lower case letters (e.g., x). A random process is always two-sided. An i.i.d. copy of X for subject i is denoted by superscript X^i. If X^i is a random process, its coordinate random variable is written using a subscript, e.g., X^i_j. The probability distribution of random variable or process X is denoted by P_X. The joint distribution of X and Y is denoted by P_{XY}; conditional distributions are written as $P_{X|Y}$. A finite-length random vector induced by random process X is written as $X_{1:n}$ etc.

2.2 Generative Latent Variable Models

Generative latent variable models (LVMs) are a family of parametric models trained to transform samples drawn from an unknown distribution P_X on $\mathcal{X}$ to latent variables in a lower dimensional space $\mathcal{Z}$. In many real-world data, especially images, we cannot estimate the density of P_X, which may not exist because the distribution is supported by low dimensional manifolds. To overcome this problem, LVMs define a latent random variable $Z \in \mathcal{Z}$ with a prior distribution P_Z such as the standard Gaussian, and learn a "decoder" $Q_{\hat{X}|Z}$, or a conditional distribution of the reconstructed input $\hat{X} \in \mathcal{X}$ given Z. The marginal distribution of the reconstruction $\hat{X}$ is given by $P_{\hat{X}} = \int Q_{\hat{X}|Z} dP_Z$, and we learn the decoder $Q_{\hat{X}|Z}$ by solving the following optimization problem

$$\inf_{Q_{\hat{X}|Z}} \mathcal{D}(P_X, P_{\hat{X}}) \tag{1}$$

for some "distance" measure $\mathcal{D}$ between the data and reconstruction distributions, with a possible addition of a regularization term. Different choices of $\mathcal{D}$ and regularizer yield different model. For example, Wasserstein auto-encoders (WAE) [Tolstikhin *et al.*, 2018] utilizes the p-Wasserstein distance between X and $\hat{X}$ on the metric space $(\mathcal{X}, d)$ [Villani, 2008]

$$\bar{d}_p(P_X, P_{\hat{X}}) \triangleq \left(\inf_{\pi \in \mathcal{P}(P_X, P_{\hat{X}})} \mathrm{E}_\pi d^p(X, \hat{X}) \right)^{\min(1, 1/p)} \tag{2}$$

but the pth power of $\bar{d}_p$:

$$\mathcal{D}_{\mathrm{WAE}}(P_X, P_{\hat{X}}) \triangleq \inf_{\pi \in \mathcal{P}(P_X, P_{\hat{X}})} \mathrm{E}_\pi d^p(X, \hat{X}). \tag{3}$$

when $p \geq 1$; $\mathcal{P}(P_X, P_{\hat{X}})$ is the set of joint distributions on $(X, \hat{X})$ whose marginals on X and $\hat{X}$ are P_X and $P_{\hat{X}}$, respectively. Tolstikhin et al. [2018] use Theorem 1 of Bousquet et al. [2017] to reparametrize (3) in terms of probabilistic encoder $Q_{Z|X}$. Write $Q_Z = \int Q_{Z|X} dP_X$. Then we have

$$\mathcal{D}_{\mathrm{WAE}}(P_X, P_{\hat{X}}) = \inf_{Q_{Z|X} : Q_Z = P_Z} \mathrm{E}_{P_X} \mathrm{E}_{Q_{Z|X}} d^p(X, g(Z)), \tag{4}$$

when the decoder is deterministic, i.e., $Q_{\hat{X}|Z}(\cdot|z)$ is a Dirac measure on $g(z)$ for all $z \in \mathcal{Z}$. In practice, the resulting constrained optimization problem is relaxed to an unconstrained one:

$$\inf_g \inf_{Q_{Z|X}} \mathrm{E}_{P_X} \mathrm{E}_{Q_{Z|X}} d^p(X, g(Z)) + \lambda \mathcal{D}_Z(P_Z, Q_Z) \tag{5}$$

for some divergence measure $\mathcal{D}_Z$ and $\lambda > 0$.

Relaxation (5) of WAE is equivalent to the adversarial auto-encoders (AAE) [Makhzani *et al.*, 2016] if $p = 2$, $\mathcal{X}$ is Euclidean, $d(x, y) = \|x - y\|$ is the standard Euclidean norm, and $\mathcal{D}_Z$ is

$$\mathcal{D}_{\mathrm{GAN}}(P_X, P_{\hat{X}}) \triangleq \sup_{f \in \mathcal{F}} \mathrm{E}_{P_X} \log f(X) + \mathrm{E}_{P_Z} \log(1 - f(g(Z)))$$

where $f : \mathcal{X} \to (0, 1)$ is the "discriminator" [Bousquet *et al.*, 2017]. In addition, the conditional AAE (cAAE) minimizes a class-conditional version of AAE.

3 The Ornstein auto-encoder (OAE)

3.1 From Ornstein's d-bar Distance to OAE

In order to extend WAE to random processes, we need a distance metric between two random sequences $X = (\dots, X_{-1}, X_0, X_1, \dots)$ and $Y = (\dots, Y_{-1}, Y_0, Y_1, \dots)$, both defined in $\mathcal{X}^\infty$. Let

$$\rho_n(x^n, y^n) \triangleq \sum_{j=1}^{n} d^p(x_j, y_j),$$

where d is a metric on $\mathcal{X}$ and $p \geq 0$; d^0 denotes the 0-1 loss. A possible distance measure between the latter two is

$$\bar{\rho}_n(P_{X_{1:n}}, P_{Y_{1:n}}) \triangleq \inf_{\pi \in \mathcal{P}(P_{X_{1:n}}, P_{Y_{1:n}})} \mathrm{E}_\pi \rho_n(X_{1:n}, Y_{1:n}).$$

Then a process distance between P_X and P_Y is defined by

$$\bar{\rho}(P_X, P_Y) \triangleq \sup_n \frac{1}{n} \bar{\rho}_n(P_{X_{1:n}}, P_{Y_{1:n}}).$$

It is known that $\bar{d}_p(P_X, P_Y) \triangleq \bar{\rho}^{\min(1,1/p)}(P_X, P_Y)$ is a metric on the space of all possible stationary processes in $\mathcal{X}$, so $\bar{d}_p$ is a true distance. The $\bar{d}_p$ or d-bar distance for random processes was introduced by Ornstein [1973] for the special case of $p = 0$ and discrete $\mathcal{X}$, and was extended to $p \geq 0$ with more general $\mathcal{X}$ by Gray, Neuhoff, and Shields [1975]. Furthermore, if P_X and P_Y are stationary, then we have

$$\bar{\rho}(P_X, P_Y) = \inf_{\pi \in \mathcal{P}_s(P_X, P_Y)} \mathrm{E}_\pi d^p(X_0, Y_0), \tag{6}$$

where $\mathcal{P}_s(P_X, P_Y)$ is the set of all jointly stationary distributions on $(X, Y) \in \mathcal{X}^\infty \times \mathcal{X}^\infty$ having P_X and P_Y as marginals [Gray *et al.*, 1975].

From the resemblance of equation (6) to the finite-dimensional Wasserstein metric (2), a reparametrization similar to (4) can be made:

Theorem 1. *Assume process distributions P_X on $\mathcal{X}^\infty$ and P_Z on $\mathcal{Z}^\infty$ are both stationary. Also assume that $Q_{\hat{X}|Z}(\cdot|z)$ is the Dirac measure on $g(z)$ for all z, i.e., $\hat{X} = g(Z)$ with probability 1 for $g : \mathcal{Z}^\infty \to \mathcal{X}^\infty$ that maps a stationary sequence to a stationary sequence. Then,*

$$\bar{\rho}(P_X, P_{\hat{X}}) = \inf_{Q_{Z|X} \in \mathcal{Q}_{Z|X}} \mathrm{E}_{P_X} \mathrm{E}_{Q_{Z|X}} d^p(X_0, g(Z)_0),$$

where $P_{\hat{X}} = \int Q_{\hat{X}|Z} dP_Z$ and $\mathcal{Q}_{Z|X}$ is the set of encoders $Q_{Z|X}$ such that $Q_{Z|X} P_X$ is jointly stationary in (X, Z) and $\int Q_{Z|X} dP_X = P_Z$.

By defining

$$\mathcal{D}_{\mathrm{OAE}}(P_X, P_{\hat{X}}) = \inf_{Q_{Z|X} \in \mathcal{Q}_{Z|X}} \mathrm{E}_{P_X} \mathrm{E}_{Q_{Z|X}} d^p(X_0, g(Z)_0)$$

and minimizing it over g, we obtain the OAE model. Similar to relaxation (5), we may solve an unconstrained problem

$$\inf_g \inf_{Q_{Z|X}} \mathrm{E}_{P_X} \mathrm{E}_{Q_{Z|X}} d^p(X_0, g(Z)_0) + \lambda \mathcal{D}_Z(P_Z, Q_Z), \tag{7}$$

where $Q_Z = \int Q_{Z|X} dP_X$. The additional constraint of $Q_{Z|X} P_X$ being stationary can be satisfied by restricting $Q_{Z|X}$ to be stationary (the latter implies the former).

Despite the apparent similarity to WAE (5), problem (7) has two practical issues. First, the decoder g for OAE needs to map an *infinite* sequence to another infinite sequence. Learning such a map with infinite memory may face computational challenges. Second, since P_Z and Q_Z are both *process* distributions, computing the divergence $\mathcal{D}_Z$ may also run into trouble.

3.2 OAE for Exchangeable Data

If we can assume that the *pair process* $\{(X_j, Y_j)\}$ is exchangeable, then computation of $\tilde{\rho}(P_X, P_Y)$ amounts to that of WAE (4), a great simplification:

Theorem 2. *Assume pair process $\{(X_j, Y_j)\}$ in $\mathcal{X}^\infty \times \mathcal{X}^\infty$ is exchangeable. Let P_X and P_Y denote its marginal distributions on $\{X_j\}$ and $\{Y_j\}$, respectively. Then*

$$\bar{\rho}(P_X, P_{\hat{X}}) = \bar{\rho}_1(P_{X_0}, P_{Y_0}) = \inf_{\pi \in \mathcal{P}(P_{X_0}, P_{Y_0})} \mathrm{E}_\pi d^p(X_0, Y_0).$$

Exchangeability of the pair process is valid for our applications, because the jth observation X_j^i of subject i and its reconstruction $\hat{X}_j^i$ must be exchangeable with the kth observation-reconstruction pair $(X_k^i, \hat{X}_k^i)$ of the same subject.

Theorem 2 ensures an alternative parametrization (cf. (4))

$$\bar{\rho}(P_X, P_Y) = \inf_{Q_{Z_0|X_0} : Q_{Z_0} = P_{Z_0}} \mathrm{E}_{P_{X_0}} \mathrm{E}_{Q_{Z_0|X_0}} d^p(X_0, g(Z_0)),$$

and the optimization problem of the form (5). Here the decoder g only takes a *single coordinate* of the latent process Z as input and outputs a *single coordinate* of Y.

We explicitly model exchangeability in the latent space by introducing a random variable B and conditioning Z_j on B: $P_{Z_{1:n}} = \int \prod_{j=1}^n P_{Z_0|B} dP_B$ for all n. The probabilistic encoder is a pair $(Q_{Z_0|B,X_0}, Q_{B|X_0})$. Constraining $\int Q_{Z_0|B,X_0} dP_{X_0} = P_{Z_0|B}$ and $\int Q_{B|X_0} dP_X = P_B$ ensures $Q_Z = P_Z$. A relaxation like (5) yields

$$\inf_g \inf_{Q_{Z_0|B,X_0}} \inf_{Q_{b|X_0}} \big[\mathrm{E}_{P_{X_0}} \mathrm{E}_{Q_{Z_0|B,X_0}} \mathrm{E}_{Q_{B|X_0}} d^p(X_0, g(Z_0)) + \lambda_1 \mathcal{D}_{Z_0|B}(P_{Z_0|B}, Q_{Z_0|B}) + \lambda_2 \mathcal{D}_B(P_B, Q_B) \big], \tag{8}$$

where $Q_{Z_0|B} = \int Q_{Z_0|B,X_0} dP_{X_0} = P_{Z_0|B}$ and $Q_B = \int Q_{B|X_0} dP_X = P_B$, for appropriate choices of divergence measures $\mathcal{D}_{Z_0|B}$ and $\mathcal{D}_B$.

If we use $\mathcal{D}_{Z_0|B} = \mathcal{D}_{\mathrm{GAN}}$ and $\mathcal{D}_B = \mathcal{D}_{\mathrm{MMD},\kappa}$ where

$$\mathcal{D}_{\mathrm{MMD},\kappa}(P_B, Q_B) = \| \mathrm{E}_{P_B} \kappa(\cdot, B) - \mathrm{E}_{Q_B} \kappa(\cdot, B) \|_{\mathcal{H}}^2$$

is the maximum mean discrepancy (MMD) [Gretton *et al.*, 2012] for a positive definite kernel $\kappa : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ that induces a reproducing kernel Hilbert space $\mathcal{H}$ equipped with the norm $\|\cdot\|_{\mathcal{H}}$, then we obtain a training algorithm described in Algorithm 1, based on the sample estimates of the terms in (8).

Lines 6 and 7 of Algorithm 1 need some explanation. Since the encoder $Q_{B|X_0}$ takes only a single coordinate as its input, it yields $\tilde{b}_j^i \sim Q_{B|X_0}(\cdot|x_j^i)$ for each $j = 1, \dots, m_i$. In order to obtain a single sample, we aggregate $\tilde{b}_j^i$'s so that $\tilde{b}^i = \frac{1}{m_i} \sum_{j=1}^{m_i} \tilde{b}_j^i$, as used in Line 6. In Line 7, we sample $\tilde{z}_j^i$ independently from $Q_{Z|B,X_0}(\cdot|\tilde{b}^i, x_j^i)$ given this $\tilde{b}^i$ and the data x_j^i, for $j = 1, \dots, m_i$.

Algorithm 1 Ornstein Auto-Encoder for Exchangeable Data

Input: Exchangeable sequences $(x_1^i, ..., x_{n_i}^i)$ for $i = 1, ..., L$
Output: Encoder pair $(Q_{Z|B,X_0}, Q_{B|X_0})$ and decoder g
Require: Latent variable distributions $P_B, P_{Z_0|B}$, regularization coefficients λ_1, λ_2, positive definite kernel κ

1: Initialize: parameters of $(Q_{Z|B,X_0}, Q_{B|X_0})$, g, and discriminator f
2: **while** $Q_{Z|B,X_0}, Q_{B|X_0}, f, g$ not converged **do**
3: Sample subjects $i = 1, \ldots, n$ and sequence $(x_1^i, \ldots, x_{m_i}^i)$ for each subject i from the training set
4: Sample b^i from P_B for $i = 1, \ldots, n$
5: Sample $(z_1^i, \ldots, z_{m_i}^i)$ from $P_{Z_0|B}$ given b^i for $i = 1, \ldots, n$
6: Sample $\tilde{b}^i$ from $Q_{B|X_0}$ given $(x_1^i, \cdots, x_{m_i}^i)$ for $i = 1, \ldots, n$.
7: Sample $(\tilde{z}_1^i, \cdots, \tilde{z}_{m_i}^i)$ from $Q_{Z|B,X_0}$ given $\tilde{b}^i$ and $(x_1^i, \cdots, x_{m_i}^i)$ for $i = 1, ..., n$.
8: Update $Q_{Z|B,X}$, $Q_{B|X}$, and g by descending:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} d^p(x_j^i, g(\tilde{z}_j^i)) - \frac{\lambda_1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\log f(\tilde{z}_j^i) +$$

$$\frac{\lambda_2}{n(n-1)}\Big(\sum_{i\neq l}\kappa(b^i, b^l) + \sum_{i\neq l}\kappa(\tilde{b}^i, \tilde{b}^l)\Big) - \frac{2\lambda_2}{n^2}\sum_{i,l}\kappa(b^i, \tilde{b}^l)$$

9: Update f by ascending:

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\log f(z_j^i) + \log(1 - f(\tilde{z}_j^i))$$

10: **end while**

3.3 Generating Variations of Unknown Subjects

A trained OAE can be used to generate a sequence of variations for a new, unknown subject *out of* the training examples. Suppose one or few input(s) $(x_1^{\text{new}}, \ldots, x_{m_{\text{new}}}^{\text{new}})$ from a new subject are given to OAE. Then we sample b_j^{new} from $Q_{B|X_0}(\cdot|x_j^{\text{new}})$ for $j = 1, 2, \ldots$, get $b^{\text{new}} = \frac{1}{m_{\text{new}}}\sum_{j=1}^{m_{\text{new}}} b_j^{\text{new}}$, and sample z_j^{new} from $P_{Z_0|B}(\cdot|b^{\text{new}})$. Then new variations $(\hat{x}_1^{\text{new}}, \hat{x}_2^{\text{new}}, \ldots)$ are obtained by passing $(z_1^{\text{new}}, z_2^{\text{new}}, \ldots)$ through the trained decoder g. Fine control on the variations is possible if further assumptions on the encoder are made; see §4. Generating new variations of a *known*, in-training subject can also be conducted in the same fashion.

Note that generating images from an unknown subject is impossible for the existing conditional LVMs, e.g., cAAE, because they require a fixed number of conditional distributions. When data imbalance is present, OAE has an advantage over conditional LVMs because the latter have to train all the conditional encoders, which is hard for minority groups with small sample sizes. OAE handles this problem by sharing a variance component.

4 Experiments

4.1 Implementation

In all the experiments in the following, we assumed $\mathcal{X}$ and $\mathcal{Z}$ are Euclidean spaces with dimensions d_x and d_z, respectively; accompanied Euclidean metric $d(x, x') = \|x - x'\|_2$ on $\mathcal{X}$ and $p = 2$ were used. We set the prior distribution P_Z of the latent variable Z as a random intercept model:

$$Z_j^i|\{B^i = b^i\} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0\mathbf{1} + b^i, \sigma_0^2\mathbf{I}), \quad B^i \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_0^2\mathbf{I}).$$

The encoder pair $(Q_{Z|B,X_0}, Q_{B|X_0})$ was designed to be another random intercept model:

$$\begin{aligned} Z_j^i|\{B^i = \tilde{b}^i, X_j^i = x_j^i\} &\overset{\text{iid}}{\sim} \mathcal{N}(\mu(x_j^i) + \tilde{b}^i, \sigma^2(x_j^i)\mathbf{I}) \\ B^i|\{X_j^i = x_j^i\} &\overset{\text{iid}}{\sim} \mathcal{N}(\nu(x_j^i), \tau^2\mathbf{I}), \end{aligned} \tag{9}$$

where the mean functions $\mu : \mathcal{X} \to \mathcal{Z}$, $\nu : \mathcal{X} \to \mathcal{X}$, and the variance function $\sigma^2 : \mathcal{X} \to \mathbb{R}_{++}$ were parameterized by deep neural networks. The hyperparameter τ was kept small. Although Gaussian encoders are suboptimal to our optimization problem (8) due to the restricted search space, Rubenstein et al. [2018] has shown empirically that such a restriction produces better outcomes when the appropriate number of dimensions for the latent space is not known. The decoder g was also parameterized by deep neural networks.

Interpreting each subject as a class, we compared OAE with cAAE with conditional Gaussian latent variables:

$$Z_j^i|\{Y^i = k\} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{0k}\mathbf{1}, \sigma_{0k}^2\mathbf{I}),$$

where C is the number of subjects, and μ_{0k}, σ_{0k}^2 are prespecified for $k = 1, \ldots, C$. Similar to OAE, we used a Gaussian encoder for $Q_{Z|Y,X}$:

$$Z^i|\{X^i = x^i, Y^i = k\} \sim \mathcal{N}\big(\mu_k(x^i), \sigma_k^2(x^i)\big),$$

where $\mu_k : \mathcal{X} \to \mathcal{X}$, $\sigma_k^2 : \mathcal{X} \to \mathbb{R}_{++}$ are parameterized by deep neural networks for $k = 1, \ldots, C$.

For optimization, we used the Adam [Kingma and Ba, 2014] optimizer with $\beta_1 = 0.5$ for updating the first moment estimate and $\beta_2 = 0.999$ for updating the second moment estimate. When generating new variations of a given subject from the test dataset, we used one image per subject. For all convolutional layers, we used the batch normalization [Ioffe and Szegedy, 2015], padding, and truncated normal initialization.

4.2 A Toy Model

To see if OAE can learn a known low dimensional distribution embeded in a higher dimension, we generated training samples $Z_j^i = b^i + \varepsilon_j^i$ from the two-dimensional latent space for $i = 1, 2, ..., 100$, $j = 1, 2, ..., 5000$ with

$$\varepsilon_j^i \sim \mathcal{N}\Big(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.009 & 0\\0 & 0.007\end{bmatrix}\Big), b^i \sim \mathcal{N}\Big(\begin{bmatrix}0.2\\-0.4\end{bmatrix}, \begin{bmatrix}1.018 & 0.12\\0.12 & 0.745\end{bmatrix}\Big),$$

and embedded them into four-dimensional Euclidean space by $X_j^i = AZ_j^i$ with

$$A = \begin{bmatrix}0.027 & 0.171 & 0.084 & 0.290\\0.252 & 0.388 & 0.248 & 0.371\end{bmatrix}^T.$$

For learning the representation, we misspecified the two-dimensional latent variable for $i = 1, 2, ..., n$, $j = 1, 2, ..., n_i$ as

$$Z_j^i \sim b^i + \varepsilon_j^i, \quad \varepsilon_j^i \sim \mathcal{N}\big(\mathbf{0}, 0.01\mathbf{I}\big), \quad b^i \sim \mathcal{N}\big(\mathbf{0}, \mathbf{I}\big).$$

and trained the OAE with a simple architectures with 4.8k parameters. We used the linear decoder to restrict the generated sample distribution to be normal. The model was trained for 50 epochs with mini-batch size 3000, $\lambda_1 = 10$, $\lambda_2 = 10$ and the learning rates of 0.01 for the encoder-decoder and 0.005 for the discriminator. After training, we generated samples through the decoder with $n = 100$ and $n_i = 500$, and measured the error between the samples and the true moments. The root-mean squared error (RMSE) of the mean $\mathrm{E}_{b_i}[\mathrm{E}[\hat{X}_{ij}|b_i]]$ was 0.0233, and the RMSE of the covariance matrix $E_{b_i}[\mathrm{cov}[\hat{X}_{ij}|b_i]]$ was 0.0001. This result shows that the OAE works well on this toy but informative model.

4.3 VGGFace2 Dataset

Recall that in the VGGFace2 dataset the portraits of each individual are highly correlated and exchangeable. It is also highly imbalanced, with number of portraits per person varying from 30 to 843. The goal of this experiment is to examine the capability of OAE in generating new variation of portraits of both known and unknown subjects in the presence of many subjects (classes) and data imbalance. As emphasized in the previous section, generating images from an unknown subject is impossible with existing (conditional) LVMs, e.g., cAAE. For known subjects, we compared the sample quality of OAE with that of cAAE. For unknown subject, cAAE cannot generate samples, and we compare the quality of the generated samples with WAE, which ignores the subject information.

Algorithm Parameters. We chose $d_z = 128$ as the latent space dimension, and used hyperparameters $\mu_0 = 0$, $\sigma_0^2 = 1$, $\tau_0^2 = 100$. The encoder-decoder architecture had 13.6M parameters and the discriminator had 12.8M parameters. We set $\lambda_1 = 10$, $\lambda_2 = 10$ for OAE, and $\lambda = 10$ for WAE and cAAE. All models were trained for 100 epochs with a constant learning rate of 0.0005 for the encoder and decoder, and 0.001 for the discriminator. We used mini-batches of size 200.

Training. As a pre-processing, we cropped the faces and rescaled them to a common size of 64 by 64. We constructed a training set of 146,519 images from 500 randomly chosen subjects. Since the number of subjects far exceeded the mini-batch size and the dataset is highly imbalanced, we used importance sampling to limit both the number of subjects and the maximum number of variations per mini-batch in early training epochs. For data augmentation, we either added white Gaussian noise to or vertically flipped randomly chosen images in a mini-batch.

Evaluation Measures. The quality of reconstruction of a given image was measured by the mean squared error (MSE). The quality of generated samples was quantified by the sharpness using the Laplace filter [Rubenstein *et al.*, 2018], and the Frechet inception distance (FID) between image distributions [Heusel *et al.*, 2017]. Both are commonly used in the LVM literature. For FID, we picked 100 images from the generated samples and the test dataset.

Generating New Portraits of Known Subjects. We constructed a test dataset (Testset 1) with 11,250 images of 49 subjects from training dataset. We generated 100 new variations for each subject using OAE and cAAE. Table 1 suggests that OAE could generate quality variations for known identities better than cAAE.

	Known subjects (Testset 1)			Unknown subjects (Testset 2)		
	MSE	FID	Sharpness	MSE	FID	Sharpness
OAE	28.551	151.994	1×10^{-4}	34.492	156.935	1×10^{-4}
cAAE	46.020	152.077	1×10^{-4}	-	-	-
WAE	-	-	-	33.469	163.612	1×10^{-4}
Testset	-	-	4×10^{-3}	-	-	3×10^{-3}

Table 1: VGGFace2 evaluation. MSE (lower is better), FID (lower is better), sharpness (similar to testset is better).

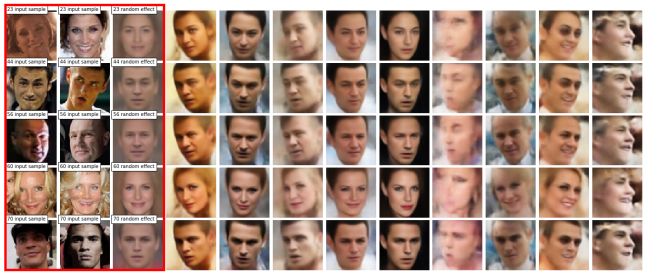

Figure 1: Generated new variations from unknown subjects of VGGFace2. Each row corresponds to a subject. Columns 1 and 2 show randomly chosen test images from the person. Column 3 shows generated images from the estimated random intercept. Columns 4 and on represent the generated images using common variations.

Generating New Portraits of Unknown Subjects. We constructed another test dataset (Testset 2) with 11,250 images of 49 subjects from randomly chosen 500 subjects not used for training. We generated 100 new variations for each subject from OAE, and 4,900 images from WAE, in which subject identity cannot be used. Table 1 shows that OAE can generate new variations for given but unknown identities with sample quality comparable to WAE, which can only generate random identities. Figure 1 presents some generated variations of unknown subjects.

Vector Arithmetic. The random intercept modeling of the encoder (9) allows an additional advantage of performing vector arithmetic on the portraits. Suppose an image x_0^i of person i is given and we want to generate a variation similar to the lth image of another person k. If $\tilde{b}^i$ is the intercept of person i in the latent space obtained by applying encoder $Q_{B|X_0}$ to x_0^i, and $(z_l^k, \tilde{b}^k)$, $(z_0^i, \tilde{b}^i)$ and $(z_l^k, \tilde{b}^k)$ are the encoding of x_0^i and x_l^k, then $z_l^k - \tilde{b}^k + \tilde{b}^i$ exchange the mean of z^k to the mean of z^i. Hence decoding

$$\hat{x}_l^i = g(z_l^k - \tilde{b}^k + \tilde{b}^i)$$

amounts to switching the identity of x_l^k to that of person i. Figure 2 demonstrates some results when both target and base persons are chosen from unknown subjects. This generalizability is unique to OAE, and suggests that OAE can be a useful data augmentation tool for many applications such as face recognition in the presence of high imbalance.

Subject-level Disentanglement in Representation. Another benefit of our random process modeling is that subjects

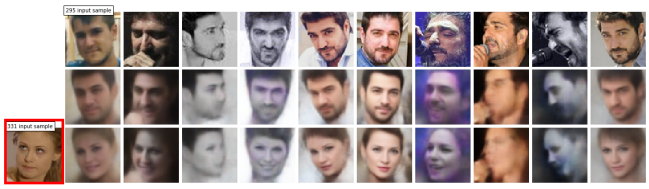

Figure 2: Vector arithmetic results for VGGFace2. Row 1, images of the base person. Row 2, reconstruction of Row 1. Row 3, input (highlighted) and generated images using vector arithmetic (rest).

(a) OAE (known) SSW/SSB = 0.362 (b) OAE (unknown) SSW/SSB = 0.551 (c) WAE SSW/SSB = 40.234

Figure 3: t-SNE map of the encoded images from VGGFace2. (a) known subjects. (b) unknown subjects. (c) WAE. Each color represents a single preson.

can be well-separated in the latent space. Figure 3 shows t-SNE maps [Maaten and Hinton, 2008] of the latent space representation of randomly selected 225 images of 10 subjects, known or unknown. For known subjects, clustering by subject is clear. Unknown subjects are also separated well, judged by visual inspection and by the ratio of within-group sum of squares (SSW) to between-group sum of squares (SSB); SSW/SSB is less than 1 for both cases. By design, WAE could not separate subjects at all.

4.4 MNIST Dataset

The goal of this experiment is to see how well OAE performs when the number of subjects is given and fixed. With balanced data, conditional methods such as cAAE are expected to perform well. In the presence of class imbalance, however, random process-based OAE has an advantage due to its generalizability.

Algorithm Parameters. We chose $d_z = 8$ as the latent space dimension, and used hyperparameters $\mu_0 = 0$, $\sigma_0^2 = 1$, $\tau_0^2 = 100$. The encoder-decoder architecture had 6.1M parameters and the discriminator had 265k parameters. We set $\lambda_1 = 10$, $\lambda_2 = 10$, and $\lambda = 10$. All models were trained for 100 epochs with mini-batch size 100, with learning rates of 0.01 for the encoder-decoder and 0.005 for the discriminator which were manually halved at the 30th and 50th epochs. The network architectures for cAAE and WAE were mostly the same as OAE except for the random intercept part.

Evaluation Measures. Similar to the VGGFace2 experiment, we evaluated the MSE of the reconstruction of given images and measured the sharpness of generated images. To compare the class-conditional generation quality, we generated class-conditional samples from OAE and cAAE, than calculated the classification accuracy of the generated digits, measured by a pre-trained deep MNIST digit classifier with 99.2% accuracy. Additionally, we compared the diversity of the generated samples per class by evaluating the structural similarity (SSIM), which is a perceptual similarity metric range between 0 and 1 [Wang *et al.*, 2004; Odena *et al.*, 2017]. We evaluated the mean SSIM score of 50 randomly chosen image pairs conditioned on each digit, and took the average of the digit-wise mean SSIM scores.

	Balanced training data			Imbalanced training data		
	MSE	Accuracy	SSIM	MSE	Accuracy	SSIM
OAE	0.793	0.992	0.318	0.977	0.972	0.320
cAAE	0.572	0.877	0.224	0.661	0.839	0.190
WAE	0.646	-	-	0.759	-	-
Testset	-	0.999	0.235	-	0.999	0.235

Table 2: MNIST evaluation. MSE (lower is better), accuracy (larger is better), sharpness (similar to testset is better), SSIM (similar to testset is better).

Balanced Training Data. We used a balanced training data with 10 classes of 56,000 images and a balanced test data with 10 classes of 1,000 images. We also generated 10 classes of 1,000 images from cAAE and OAE, and 10,000 images from WAE ignoring classes. The accuracy shown in Table 2 suggests that OAE mostly generated correct digits whereas cAAE sometimes failed. The slightly higher reconstruction error did not harm the classifier. The diversity of generated samples were similar.

Imbalanced Training Data. In order to create an imbalanced dataset, we dropped 90% of images in randomly chosen three classes (digits of 0, 3, and 4) from the balanced training set. The resulting set had 10 classes, 40,933 images. Table 2 reveals that the accuracy gap between OAE and cAAE for the generated samples widened in the imbalanced setting.

Additional Examples. Online Supplementary Material https://tinyurl.com/y3ghw3yp contains additional visualizations for generating new variations of digits and disentanglement in the representation space.

5 Conclusion

In this work we paid attention to the nested data structure of common machine learning datasets, which led us to view the data as a collection of i.i.d. observations of exchangeable random processes. We then introduced the optimal transport distance between stationary random processes. Using this, we proposed the Ornstein auto-encoder, which, under exchangeability inherently residing in the data, reduces to a tractable optimization problem. Our random process approach allowed us to generate correlated samples for the unknown subjects never used in training, which has been impossible for previous works on generative latent variable models.

In the future, we plan to expand this work to non-exchangeable stationary random processes. Another helpful direction would be latent variable modeling of multilevel data, which often arise in biomedical applications.

Acknowledgments

This work is a part of SNU-Samsung Smart Campus research program, supported by Samsung Electronics Co., Ltd.

References

[Bousquet *et al.*, 2017] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the VEGAN cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

[Cao *et al.*, 2018] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74, 2018.

[Diggle *et al.*, 2002] Peter Diggle, Peter J Diggle, Patrick Heagerty, Patrick J Heagerty, Kung-Yee Liang, Scott Zeger, et al. *Analysis of longitudinal data*. Oxford University Press, 2002.

[Dundar *et al.*, 2007] Murat Dundar, Balaji Krishnapuram, Jinbo Bi, and R. Bharat Rao. Learning classifiers when the training data is not IID. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 756–761. Morgan Kaufmann Publishers Inc., 2007.

[Fitzmaurice *et al.*, 2012] Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied longitudinal analysis*, volume 998. John Wiley & Sons, 2012.

[Gray *et al.*, 1975] Robert M Gray, David L Neuhoff, and Paul C Shields. A generalization of Ornstein's $\bar{d}$ distance with applications to information theory. *The Annals of Probability*, pages 315–328, 1975.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456. PMLR, 2015.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Makhzani *et al.*, 2016] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. In *International Conference on Learning Representations*, 2016.

[Odena *et al.*, 2017] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 2642–2651. JMLR.org, 2017.

[Ornstein, 1973] Donald S Ornstein. An application of ergodic theory to probability theory. *The Annals of Probability*, 1(1):43–58, 1973.

[Parkhi *et al.*, 2015] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[Rubenstein *et al.*, 2018] Paul K. Rubenstein, Bernhard Schölkopf, and Ilya Tolstikhin. Wasserstein auto-encoders: Latent dimensionality and random encoders. In *Workshop at the 6th International Conference on Learning Representations (ICLR)*, May 2018.

[Tolstikhin *et al.*, 2018] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

[Villani, 2008] Cédric Villani. *Optimal Transport: Old and new*, volume 338. Springer Science & Business Media, 2008.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.