Variational Autoencoders on Hilbert spaces

Generating functional data

Lorenz Wolf
Supervisor Dr Andrew Duncan

Department of Mathematics
Imperial College London

16 September 2021

Imperial College London

Motivation

- Abundance of functional data
- Successes of Deep Learning
- Applications of Generative Modelling

Outline

- Functional Data
- Background and Theory
- Variational Autoencoders on Hilbert spaces
- Simulations
- Application
- Conclusion

Functional Data

- Observations are functions in $L^2([0,1])$
- Infinite degrees of freedom
- $\{x_i\}_{i=1}^N$, where $x_i = (X_i(t_{ij}))_{j=1}^M$ for $t_{i1}, \ldots, t_{iM} \in \mathcal{I}$

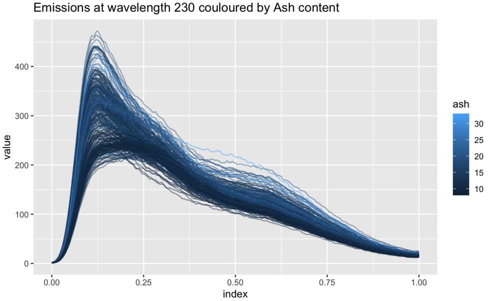

Figure: Sugar spectra (Source[1])

[1] http://jeffgoldsmith.com/IWAFDA/shortcourse$_s$*ofr.html*

Why take a functional approach?

- Allows evaluation at any point in time
- Continuity, smoothness, and derivatives
- Multivariate methods may not be robust $M \gg N$

Grid Refinement Invariance Principle (GRIP)

Theorem (GRIP)

Methods for functional data should be robust under changes of the dimension of the representation as long as the dimension is large enough to give an accurate representation.

How to devise methods appropriately?

$\rightarrow$ Method for functional data and project into finite dimensions

$\rightarrow$ Method for multivariate data and check limit as $M \rightarrow \infty$

Basis Expansions

Proposition

Any $f \in L^2(\mathcal{D})$ with an orthonormal basis $\{e_n\}_{n=1}^{\infty}$ can be written as

$$f(t) = \sum_{n=1}^{\infty} f_n e_n(t),$$

where $f_n = \int_{\mathcal{D}} f(t) e_n(t) dt$.

- In practice truncated to a suitable number of basis functions B

$$f(t) \approx \sum_{n=1}^{B} f_n e_n(t)$$

Different bases - Fourier basis

$e_1(t) = 1,$
$e_2(t) = \sqrt{2} sin(2\pi t),$
$e_3(t) = \sqrt{2} cos(2\pi t)$, etc.

- Fast computation
- Assumes periodicity

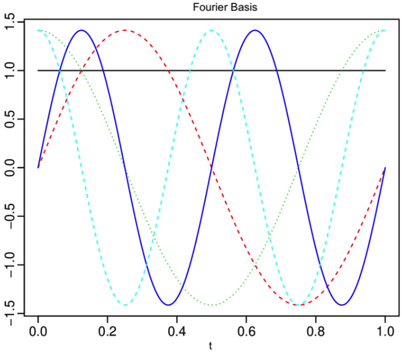

Figure: First 5 Fourier basis functions on $[0, 1]$.

Different bases - Kernel induced basis

Kernel induced basis

- Let k be a non-negative definite kernel
- Kernel induced basis (Mercer's theorem)
- Eigenvectors of Gram matrix
- Matérn kernel yields Sobolev space [Bac20]

Can we choose a basis **optimally** for a specific problem / data set?

Functional Principal Components (FPCA)

- Let X be a zero mean, square integrable random variable in $L^2([0,1])$.
- Can we find basis functions $u_1, \ldots, u_B$ minimising the loss

$$S(u_1, \ldots, u_B) = \mathbb{E}\left\|X - \sum_{i=1}^{B} \langle X, u_i \rangle u_i\right\|^2?$$

- In fact taking $u_1, \ldots, u_B$ as in the Karhunen Loève expansion is solution, i.e. first B eigenfunctions of

$$(T_\gamma f)(t) = \int_0^1 \gamma(t,s) f(s) ds$$

where $\gamma(t,s)$ is the covariance function [Dun21].

Functional Principal Components (FPCA)

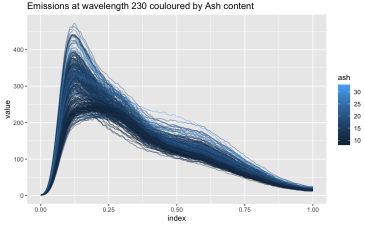

Figure: Sugar emission spectra at wavelength 230.

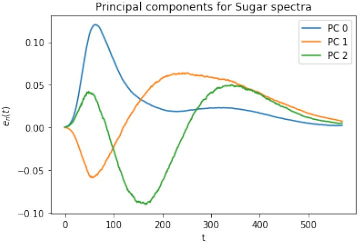

Figure: First 3 normalised functional principal components.

Generative Modelling

- Given a training set X learn the distribution $p(X)$
- Unsupervised learning $\rightarrow$ leverage unlabelled data
- Data augmentation, data privacy, density estimation, out of distribution detection

Variational Autoencoder (VAE)

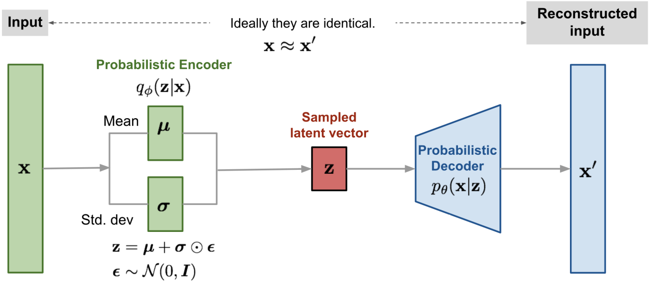

Figure: Gaussian VAE model architecture (Source[2])

Maximise the ELBO (see [KW13]):

$$\mathcal{L}(\theta, \phi; x) = -D_{KL}\left(q_\phi(z \mid x) \,||\, p_\theta(z)\right) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right]$$

[2] https://blog.bayeslabs.co/2019/06/04/All-you-need-to-know-about-Vae.html

Variational Autoencoder on Hilbert spaces

Given function evaluation x_i^k at point s_i^k ([MFB20]):

$$\hat{x}_{e,i}^k = \beta_i^\top \Phi\left(s_i^k\right) \tag{1}$$

$$[z_\mu, z_{sd}]^\top = Encoder\left(\phi, \beta_i\right) \tag{2}$$

$$\mathcal{Z} \sim \mathcal{N}\left(z_\mu, z_{sd}^2 \mathbb{I}\right) \tag{3}$$

$$\hat{\beta}_i = Decoder\left(\theta, \mathcal{Z}\right) \tag{4}$$

$$\hat{x}_{d,i}^k = \hat{\beta}_i^\top \Phi\left(s_i^k\right). \tag{5}$$

Where $\Phi(s_i^k) = (\varphi_1(s_i^k) \ \ldots \ \varphi_B(s_i^k))^T$ with $\{\varphi_j\}_{j=1}^B$ a set of B basis functions.

Data sets

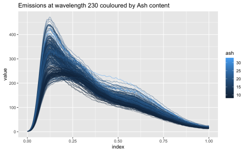

Figure: Sugar emission spectra.

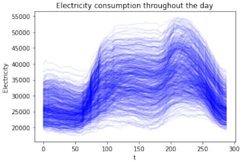

Figure: Electricity consumption.

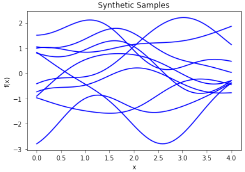

Figure: Simulated data.

Data set	N	M
Simulated	4000	100
Sugar spectra	268	571
Gridwatch	532	288

Table: Summary of data sets

Imperial College London

Assessing the model performance

- Synthetic sample diversity - Visual comparison
- Discriminative comparison - Classifier and maxmimum mean discrepancy (MMD)
- Usefulness of synthetic samples - Application

Which basis should we choose?

Basis	Data	$\widehat{\text{MMD}}^2_{CEXP}$	Loss	Accuracy
Fourier	Simulated GP	0.0001	0.688	0.534
	Sugar Spectra	0.036	0.577	0.701
	Gridwatch	3.49×10^{-22}	0.697	0.476
FPCA	Simulated GP	0.011	0.655	0.604
	Sugar Spectra	0.002	0.697	0.515
	Gridwatch	6.67×10^{-15}	0.693	0.494
Matérn	Simulated GP	0.001	0.692	0.519
	Sugar Spectra	0.045	0.640	0.746
	Gridwatch	3.34×10^{-10}	0.694	0.5

Table: Results of the discriminative comparison.

Which basis should we choose?

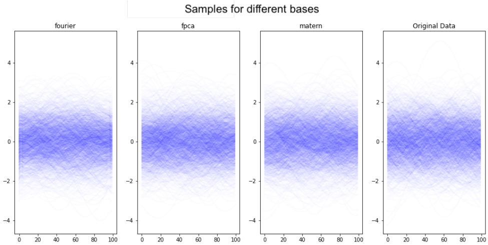

Figure: Synthetic samples for simulated data.

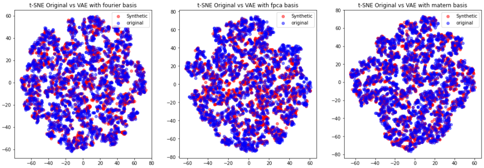

Figure: t-SNE projection for simulated data.

Imperial College London

Which basis should we choose?

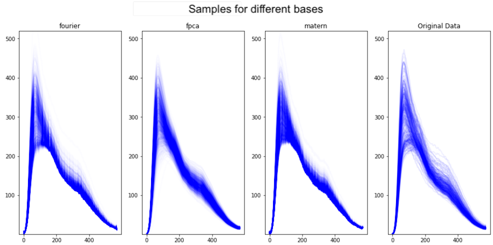

Figure: Synthetic samples for sugar spectra data.

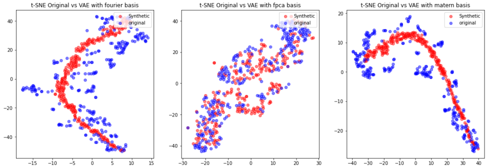

Figure: t-SNE projection for sugar spectra.

A conditional model to battle mode collapse

- condition on other knowledge c
- $q_\phi(z|x, c)$ and $p_\theta(x|z, c)$ as the variational and inference models of the conditional VAE [SLY15].
- The ELBO becomes:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; x, c) = \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z, c)] - D_{KL}\left(q_\phi(z|x, c) \| p_\theta(z \mid c)\right).$$

- Sugar - spectra modes correspond wavelengths
- Gridwatch - modes correspond to day of the week

A conditional model to battle mode collapse

Data set	Basis	Model type	$\widehat{\text{MMD}}^2_{\text{CEXP}}$	Loss	Accuracy
Sugar	FPCA	Conditional	0.004	0.369	0.792
		Standard	0.001	0.416	0.799
	Matérn	Conditional	0.019	0.163	0.954
		Standard	0.003	0.133	0.96
	Fourier	Conditional	0.012	0.055	0.994
		Standard	0.007	0.024	0.999
Gridwatch	FPCA	Conditional	5.19×10^{-163}	0.6932	0.5
		Standard	3.8×10^{-93}	0.6932	0.5
	Matérn	Conditional	8.16×10^{-198}	0.6932	0.5
		Standard	7.74×10^{-73}	0.6932	0.5
	Fourier	Conditional	1.64×10^{-265}	0.6932	0.5
		Standard	7.34×10^{-70}	0.6932	0.5

Table: Simulation results for the discriminative comparison of conditional and standard models with different basis.

A conditional model to battle mode collapse

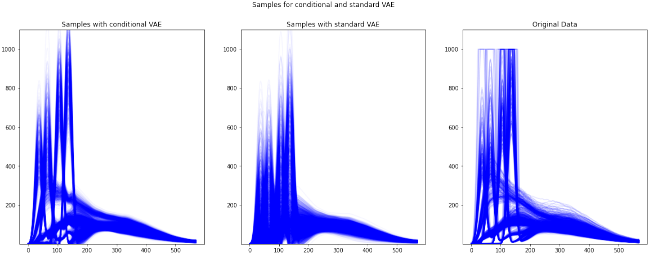

Figure: Synthetic samples with FPCA based basis for sugar data.

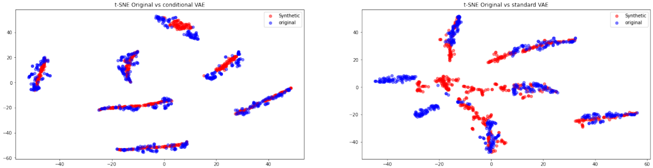

Figure: t-SNE projection for sugar data.

Imperial College London

A conditional model to battle mode collapse

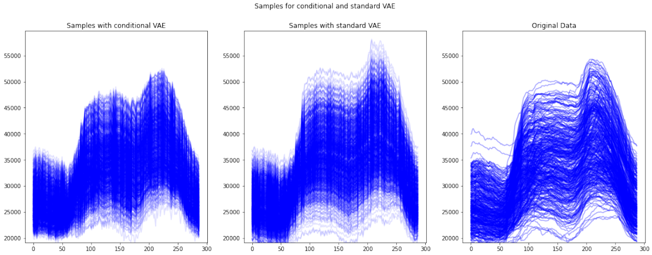

Figure: Synthetic samples with Fourier basis for gridwatch data.

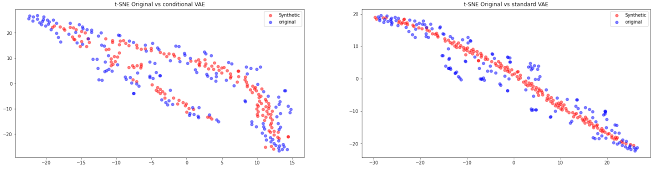

Figure: t-SNE projection for sugar data.

Application

Ash content	> 18	≤ 18
Number of observations	31	237

Table: Classification data set

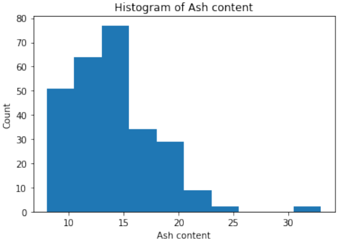

Figure: Histogram of ash content in sugar samples.

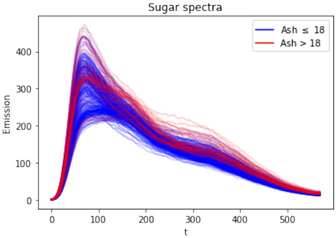

Figure: Sugar spectra coloured corresponding to high or low ash content.

Application

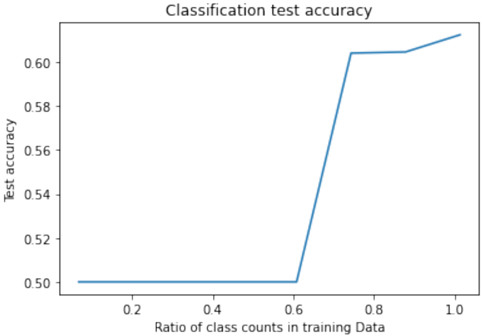

Figure: Averaged accuracy of 4 independent runs on test set plotted against the ratio of number of samples containing a high ash content and number of samples containing a low ash content.

Conclusions

- Basis choice is dependent on data
- FPCA based basis yields good sample diversity
- Prevent mode collapse by conditioning on classes
- Significant improvements by augmenting data set with synthetic samples

Future Work:

- Fully functional VAE
- Other bases e.g. B-splines
- Method for sparse functional data

Francis Bach.
Learning theory from first principles.
Technical report, 2020.

Andrew Duncan.
Functional Data Analysis Notes.
Technical report, 2021.

Diederik P Kingma and Max Welling.
Auto-Encoding Variational Bayes.
12 2013.

Swapnil Mishra, Seth Flaxman, and Samir Bhatt.
piVAE: Encoding stochastic process priors with variational autoencoders, 2 2020.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan.
Learning Structured Output Representation using Deep Conditional Generative Models.

Imperial College London

In C Cortes, N Lawrence, D Lee, M Sugiyama, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

George Wynne and Andrew B. Duncan.
A Kernel Two-Sample Test for Functional Data.
8 2020.

Question - MMD

Given a kernel k and the associated RKHS $\mathcal{H}_k(\mathcal{X})$ we define $\mathcal{P}$ the set of Borel probability measures on $\mathcal{X}$. Furthermore, assuming k is measurable define $\mathcal{P}_k \subset \mathcal{P}$ as the set of all $P \in \mathcal{P}_k$ such that $\int k(x,x)^{\frac{1}{2}} dP(x) < \infty$. For $P, Q \in \mathcal{P}_k$ we define the Maximum Mean Discrepancy denoted $\mathrm{MMD}_k(P,Q)$ as follows

$$\mathrm{MMD}_k(P,Q) = \sup_{\|f\|_{\mathcal{H}_k(\mathcal{X})} \leq 1} \left| \int f dP - \int f dQ \right|.$$

CEXP:

$$k_{\mathrm{c-exp}(F,l)}(s,t) = e^{-\frac{1}{2l^2}(s-t)^2} k_{\cos(F)}(s,t)$$

where $F \in \mathbb{N}$ and

$$k_{\cos(F)}(s,t) = \sum_{n=0}^{F-1} \cos(2\pi n(s-t)) \text{ on } [0,1]^2.$$

Maximum mean discrepancy (MMD)

- Enables comparison of two samples on $L^2(\mathcal{D})$ with $\mathcal{D} \subset \mathbb{R}^d$
- Can be estimated unbiasedly by the Monte Carlo estimator

$$\widehat{\mathrm{MMD}}_k\left(X_n, Y_n\right)^2 := \frac{1}{n(n-1)} \sum_{i \neq j}^{n} h\left(z_i, z_j\right),$$

where $h\left(z_i, z_j\right) = k\left(x_i, x_j\right) + k\left(y_i, y_j\right) - k\left(x_i, y_j\right) - k\left(x_j, y_i\right)$ and k is a kernel [WD20].

Question - Matérn kernel basis

The Matérn is kernel defined by

$$k_M(\mathrm{x}, \mathrm{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} \, \|\mathrm{x} - \mathrm{x}'\|_2}{\sigma} \right) K_\nu \left(\frac{\sqrt{2\nu} \, \|\mathrm{x} - \mathrm{x}'\|_2}{\sigma} \right), \ \sigma, \nu > 0,$$

where K_ν is the modified Bessel function of second kind of order ν and Γ is the Gamma function.

- Translation invariant $\rightarrow$ Bochner's theorem
- Evaluate the kernel on a dense grid $t_1, \ldots, t_n \in \mathcal{I}$
- Obtain the Gram matrix defined by $G_{ij} = k(t_i, t_j)$ for $i, j = 1, \ldots, n$
- Compute the eigenvectors of the Gram matrix
- Interpolate eigenvectors to obtain eigenfunctions

Question - Karhunen Loève expansion

Theorem (Karhunen Loève)

Let $\{X_t,\ t \in [0,1]$ be a zero mean process on $L^2([0,1])$ with continuous covariance function $\gamma(s,t)$. Then

$$X_t = \sum_{n=1}^{\infty} \xi_n e_n(t), \quad t \in [0,1],$$

where $\xi_n = \int_0^1 X_t e_n(t) dt$ and $\{\lambda_n, e_n(t)\}_{n=1}^{\infty}$ are the eigenvalues and eigenfunctions of T_γ. Furthermore, we have that $\mathbb{E}\xi_n = 0$ and $\mathbb{E}(\xi_n \xi_m) = \lambda_n \delta_{n,m}$.

Here the integral operator T_γ associated with γ is defined by

$$(T_\gamma f)(t) = \int_0^1 \gamma(t,s) f(s) ds.$$

Question - Karhunen Loève expansion

- Series converges in L^2 to $X(t)$, uniformly in t
- Coefficients are random variables and contain information about the variability around the eigenfunctions
- Represent realisations of the stochastic process as realisations of random coefficients
- Uncorrelated coefficients are independent for a Gaussian processes:

$$X_t = \sum_{n=1}^{\infty} \sqrt{\lambda_n} \xi_n e_n(t),$$

where $\{\xi_n\}_{n=1}^{\infty}$ are independent $\mathcal{N}(0, 1)$

Question - t-SNE

- t-distributed stochastic neighbour embedding
- Construct probability distribution over pairs of observations, similar pairs yield high probability
- Construct similar distribution over lower dimensional representation
- minimise Kullback-Leibler Divergence between distirbutions