Learning high-dimensional mixed models via amortized variational inference

Priscilla Ong[1] **Manuel Haußmann**[1 2] **Harri Lähdesmäki**[1]

Abstract

Modelling longitudinal data is an important yet challenging task. These datasets can be high-dimensional, consist of non-linear effects, and contain time-varying covariates. In this work, we leverage linear mixed models (LMMs) and amortized variational inference to provide conditional priors for VAEs, and propose LMM-VAE, a model that is scalable, interpretable, and shares theoretical connections to the GP-based VAEs. We empirically demonstrate that LMM-VAE performs competitively compared to existing approaches.

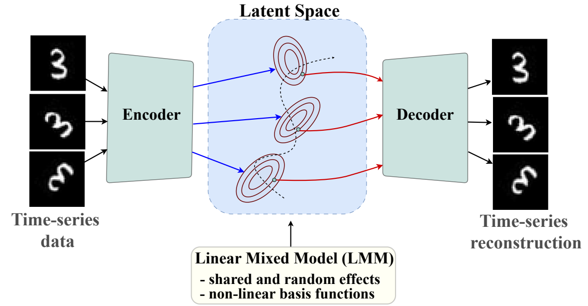
Latent Space
Encoder
Decoder
Time-series data
Time-series reconstruction
Linear Mixed Model (LMM)
- shared and random effects
- non-linear basis functions

Figure 1: Overview of LMM-VAE. The latent space is modulated by auxiliary covariates parametrized by a linear mixed model.

1. Introduction

Longitudinal datasets, which contain repeated measurements of individuals typically over time, have applications in numerous fields, such as the social sciences and the biomedical field. Longitudinal study designs are particularly useful for revealing associations between explanatory covariates and a response variable, such as the relationship between risk factors and disease progression (Caruana et al., 2015), but require appropriate statistical tools that can account for correlations both within subjects as well as across subjects. Analysis of univariate or low-dimensional longitudinal data is currently dominated by various linear mixed models (LMMs) and other additive models. However, existing methods scale poorly to currently popular longitudinal datasets, such as electronic health records, as the datasets are often high-dimensional (Zipunnikov et al., 2014), contain non-linear effects and time-varying covariates, and may contain missing values (Ramchandran et al., 2021).

Variational autoencoders (VAEs) are a class of models commonly used for representation learning and generative modelling (Rezende et al., 2014; Kingma & Welling, 2022). Nevertheless, they cannot be directly applied to longitudinal data as they assume that observations are independent and identically distributed, thereby failing to capture correlations between samples. Another challenge concerns fully exploiting the rich auxiliary covariates available for modelling. While conditional VAEs (CVAEs) can easily incorporate any number of covariates, limited work has tackled the problem of finding an appropriate time-series or longitudinal model that can effectively scale to large number of covariates in the VAE prior. Considering the potential upside in model performance, addressing the prior model's scalability, specifically to include more covariates, is a fruitful avenue of research. Different from previous work, we propose modelling the prior using LMMs. While simple in idea, this model class is scalable, vastly simplifies the training procedure within standard deep learning frameworks, and enjoys several advantages as summarized below.

Contributions. We propose the *Linear Mixed Model VAE (LMM-VAE)*, a natural extension of the commonly used univariate LMM, which is capable of handling high-dimensional data and large dataset sizes, modelling an arbitrary number of covariates in the prior, can be adapted to problems of various complexities via basis functions, and enjoys the advantages of being *interpretable* by way of parametrization. We demonstrate that LMM-VAE is competitive against commonly used GP-based VAE methods for integrating auxiliary covariates into the prior. From a practical standpoint, LMMs show promise as an alternative VAE prior for longitudinal modelling, especially in the presence of high dimensional covariates.

[1] Department of Computer Science, Aalto University, Espoo, Finland [2] Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. Correspondence to: Priscilla Ong <priscilla.ong@aalto.fi>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).

2. Background

Linear Mixed Models. Consider a pair $(\boldsymbol{x}, \boldsymbol{z})$, where $\boldsymbol{x} = (x_1, \dots, x_Q)^T \in \mathcal{X} = \prod_{i=1}^{Q} \mathcal{X}_i$ is Q-dimensional covariate vector and $\boldsymbol{z} \in \mathcal{Z} = \mathbb{R}^L$ is a L-dimensional response variable. The standard *linear model (LM)* for $(\boldsymbol{x}, \boldsymbol{z})$ is

$$\boldsymbol{z} = \boldsymbol{a}_1 x_1 + \cdots + \boldsymbol{a}_Q x_Q + \boldsymbol{\epsilon} = A\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{a}_i \in \mathbb{R}^L$, $A = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_Q) \in \mathbb{R}^{L \times Q}$, and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma_z^2 I)$ assuming equal variance across L dimensions. For N pairs of covariates and response variables, $\{(\boldsymbol{x}_n, \boldsymbol{z}_n)\}_{n=1}^{N}$, the linear model is given as

$$Z = AX + E,$$

where $Z = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_N)$, $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ and $E = (\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_N)$ such that

$$p(Z|X) \triangleq \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{z}_n | A\boldsymbol{x}_n, \sigma_z^2 I). \tag{2}$$

Without loss of generality (w.l.o.g.), we assume that each covariate is either continuous or binary since categorical covariates can be one-hot encoded into binary values. To model non-linear effects in $\mathcal{Z}$, we follow common practice and extend the standard linear model with non-linear basis functions for continuous covariates, $x_j = \phi(x)$, such as $x_j = x_i^2$, $x_k = \sin(x_i)$, $x_l = \cos(x_i)$, etc.

Longitudinal datasets consist of repeated measurements of instances (e.g. patients) over time and are commonly modeled using *linear mixed models (LMM)* (Laird & Ware, 1982). While standard LMs are designed to model effects that are shared across all instances (shared effects), LMMs can simultaneously model so-called random effects, i.e., effects that are specific to subsets of instances. Again, w.l.o.g., we assume that the covariates are ordered as

$$\boldsymbol{x} = (\underbrace{x_1, \dots, x_S}_{\boldsymbol{x}_S^T}, \underbrace{x_{S+1}, \dots, x_{S+R}}_{\boldsymbol{x}_R^T})^T,$$

where $Q = S + R$ and we use the first S covariates to model shared effects and the remaining R to model random effects. For example, $\boldsymbol{x}_S$ may include the age of an individual that we would like to model as a shared effect. Similarly, $\boldsymbol{x}_R$ may include binary covariates that correspond to the identity of all instances (patients) and we would include instance-specific random offset terms into the model. Covariates $\boldsymbol{x}_R$ can also include other types of variables, such as interaction terms that can be used to specify random effects for arbitrary subgroups of individuals. We write the LMM as

$$\begin{aligned} \boldsymbol{z} &= \underbrace{(\boldsymbol{a}_1, \dots, \boldsymbol{a}_S)}_{A_S} \boldsymbol{x}_S + \underbrace{(\boldsymbol{a}_{S+1}, \dots, \boldsymbol{a}_{S+R})}_{A_R} \boldsymbol{x}_R + \boldsymbol{\epsilon} \\ &= \underbrace{A_S \boldsymbol{x}_S}_{\text{shared effects}} + \underbrace{A_R \boldsymbol{x}_R}_{\text{random effects}} + \boldsymbol{\epsilon} = \underbrace{(A_S, A_R)}_{A} \boldsymbol{x} + \boldsymbol{\epsilon}. \end{aligned}$$

Variational Autoencoders. We assume D-dimensional observations, $\boldsymbol{y} \in \mathcal{Y} = \mathbb{R}^D$, and L-dimensional latent variables, $\boldsymbol{z} \in \mathcal{Z} = \mathbb{R}^L$, where $L \ll D$. Given a dataset $Y = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_N)$ containing N observations, we assume that Y is generated by latent variables $Z = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_N)$. and write the joint generative model for a single observation as $p_\omega(\boldsymbol{y}, \boldsymbol{z}) = p_\theta(\boldsymbol{y}|\boldsymbol{z}) p_\varphi(\boldsymbol{z})$, where $\omega = \{\theta, \varphi\}$. For vanilla latent variable models, $\boldsymbol{z}$ typically assumes a standard normal prior, i.e., $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I)$, where I is the L-by-L identity matrix.

VAEs (Rezende & Mohamed, 2016; Kingma & Welling, 2022) rely on amortized variational inference, by specifying a parameterized approximation $q_\phi(\boldsymbol{z}|\mathbf{y})$ to the intractable true posterior $p(\boldsymbol{z}|\mathbf{y})$, commonly known as an encoder. Its parameters ϕ are optimized jointly with ω by maximizing the *evidence lower bound (ELBO)* as given by

$$\log p_\omega(Y) \geq \mathbb{E}_{q_\phi}\left[\log p_\theta(Y|Z)\right] - \mathrm{KL}\left(q_\phi(Z|Y) \parallel p_\varphi(Z)\right),$$

where KL denotes the Kullback-Leibler divergence.

CVAEs (Sohn et al., 2015) extend the standard VAEs by conditioning the generative model with auxiliary covariates, $\boldsymbol{x} \in \mathcal{X}$. The joint distribution of CVAEs can be written, e.g, as $p_\omega(\mathbf{y}, \mathbf{z}|\boldsymbol{x}) = p_\theta(\mathbf{y}|\mathbf{z}, \boldsymbol{x}) p_\varphi(\mathbf{z}|\boldsymbol{x})$. The ELBO objective for a CVAE is obtained, as the standard VAE above, by conditioning the probabilities of the generative model as well as the encoder network with covariates $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$.

3. Linear Mixed Model VAEs

We incorporate Linear Mixed Models (LMMs) into the prior of the VAE, and propose the LMM-VAE. This model can accommodate an arbitrary number of auxiliary covariates in the prior, model both shared and random effects, and allow efficient model learning via the global parametrization. Assuming a prior on the matrix A and an optionally probabilistic decoder, we formulate the generative model for a sample $\mathbf{y}$ with covariates $\boldsymbol{x}$ as

$$\theta \sim p(\theta) \tag{3}$$

$$A \sim p(A) \tag{4}$$

$$\boldsymbol{z}|A, \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{z}|A\boldsymbol{x}, \sigma_z^2 I) \tag{5}$$

$$\mathbf{y}|\boldsymbol{z}, \theta \sim p(\mathbf{y}|\boldsymbol{z}, \theta), \tag{6}$$

where θ parameterizes a decoder $f_\theta : \mathcal{Z} \rightarrow \mathcal{Y}$ representing the Gaussian (or other) likelihood model. See Figure 2 for the corresponding plate. We can define different priors for shared and random effects, $p(A) = p(A_S)p(A_R)$.

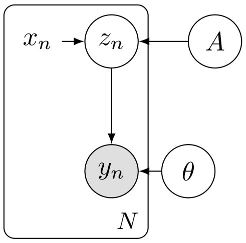

Figure 2: *Priors.* A plate diagram of the LMM-VAE with probabilistic priors on all parameters. Shaded and blank circles refer to observed and latent variables, respectively.

We assume a factorizing variational posterior $q_\phi(A, \theta, Z|Y) = \prod_n q_\phi(\boldsymbol{z}_n|\mathbf{y}_n)q(A)q(\theta)$, where all approximate posteriors are mean-field Gaussian, and q_ϕ denotes an encoder network that maps an individual observation $\mathbf{y}_n$ into parameters of the Gaussians, mean $\boldsymbol{\mu}_n$ and diagonal covariance $\boldsymbol{\sigma}_n^2$. The ELBO is given as

$$\begin{aligned}\log p(Y|X) \geq \sum_n \Big(& \mathbb{E}_{q(\theta)q_\phi(\boldsymbol{z}_n|\mathbf{y}_n)} \left[\log p(\mathbf{y}_n|\boldsymbol{z}_n, \theta)\right] \\ & - \mathbb{E}_{q(A)} \left[\mathrm{KL}\left(q_\phi(\boldsymbol{z}_n|\mathbf{y}_n) \parallel p(\boldsymbol{z}_n|A, \boldsymbol{x}_n)\right)\right] \Big) \\ & - \mathrm{KL}\left(q(A) \parallel p(A)\right) - \mathrm{KL}\left(q(\theta) \parallel p(\theta)\right).\end{aligned}$$

It is straightforward to optimize this ELBO with mini-batch-based stochastic gradient descent (SGD) because the parametrization of the LMM-VAE model is global. The first expectation can be approximated via Monte Carlo sampling, while the remaining terms are analytically tractable assuming Gaussian priors and posteriors.

4. GP-prior VAEs as LMM-VAEs

LMs and LMMs can be made arbitrarily complex by incorporating additional basis functions, while at the same time preserving linearity with respect to the parameters. Using a basis function extension, we draw a useful link between LMM-VAEs and GP-prior VAEs, which have received considerable attention in the existing literature. This connection builds upon a well-known result that GPs correspond to Bayesian linear regression with infinitely many basis functions (see, e.g., Rasmussen & Williams, 2006). Here, we use the spectral domain representation, but similar constructions could also be derived for other basis, such as eigenfunctions from the Mercer's theorem (Rasmussen & Williams, 2006) or Laplace eigenfunctions (Solin & Särkkä, 2019).

Based on the spectral domain representation, a univariate stationary covariance function $k(r)$, where $r = x - x'$, can be approximated using a finite basis function expansion (Rasmussen & Williams, 2006)

$$k(r) = \frac{1}{2\pi} \int_{-\infty}^{\infty} s(\omega) e^{i\omega r} d\omega \approx \frac{\sigma^2}{M} \sum_{m=1}^{M} \cos(\omega_m r) = \tilde{k}(r),$$

where $s(\omega)$ is the kernel's spectral density at frequency ω, i is the imaginary unit, σ^2 is the kernel variance, ω_m are either regular or random Fourier frequencies (with a distribution proportional to the spectral density $s(\omega)$), and M is the number of terms in the approximation. To construct the basis functions, we can exploit the trigonometric identity $\cos(u - v) = \cos(u)\cos(v) + \sin(u)\sin(v)$ (Tompkins & Ramos, 2018), and represent $\tilde{k}(r)$ using the feature map

$$\begin{aligned}\boldsymbol{\phi}(x) = (\, & \cos(\omega_1 x), \dots, \cos(\omega_M x), \\ & \sin(\omega_1 x), \dots, \sin(\omega_M x))^T.\end{aligned}$$

Following Hensman et al. (2018), the approximate GP is then given as

$$f(x) \sim \mathcal{GP}\left(0, \frac{\sigma^2}{M} \boldsymbol{\phi}(x)^T \boldsymbol{\phi}(x')\right) = \mathcal{GP}\left(0, \tilde{k}(x - x')\right)$$

with an equivalent parametric expression

$$f(x) = \phi(\mathbf{x})^T \mathbf{a} = \mathbf{a}^T \phi(\mathbf{x}) = \bar{\mathbf{a}} \cdot \phi(\mathbf{x}),$$

where $\bar{\mathbf{a}} = \mathbf{a}^T$, $p(\mathbf{a}) = \mathcal{N}(\mathbf{0}, \boldsymbol{S})$ with $\boldsymbol{S} = \mathrm{diag}(s(\omega_1), \dots, s(\omega_M), s(\omega_1), \dots, s(\omega_M))$ for regular Fourier features, and $p(\mathbf{a}) = \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{M} I\right)$ for random Fourier features.

For notational brevity, consider a GP-prior VAE with a stationary covariance that is conditioned with a single continuous covariate x. Existing GP-prior VAE models assume that the prior for the latent space factorizes across the dimensions, i.e., $p(\mathbf{z}|x) = \prod_{l=1}^{L} \mathcal{GP}(0, k_l(x, x'))$. Assuming the reduced-rank approximation for a GP-prior with M Fourier features, the GP-prior VAE can be directly implemented by LMM-VAE by setting $\mathbf{z} = A\phi(\mathbf{x})$, where

$$A = \left(\bar{\mathbf{a}}_1^T, \dots, \bar{\mathbf{a}}_L^T\right)^T,$$

$p(A) = \prod_{l=1}^{L} p(\bar{\mathbf{a}}_l)$ as defined above, and letting $\sigma_z^2 \to 0$.

It is straightforward to establish a similar connection between the longitudinal GP-prior VAE (LVAE, Ramchandran et al. (2021)) and LMM-VAE. Assuming the same Fourier features $\boldsymbol{\phi}(\boldsymbol{x})$ for all additive kernels, we can construct a linear basis function approximation $\boldsymbol{z}^{(j)} = A^{(j)}\phi(\boldsymbol{x})$ for each of the kernels j and define the final model as

$$\boldsymbol{z} = \sum_{j=1}^{R} \boldsymbol{z}^{(j)} = \left(A^{(1)}, \dots, A^{(R)}\right) \phi(\boldsymbol{x}), \tag{7}$$

where R is the number of kernels in a L-VAE model and $p(A) = \prod_{j=1}^{R} p(A^{(j)})$.

While theoretically connected to the GP-based VAEs, LMM-VAE greatly simplifies the overall training procedure by sidestepping the $\mathcal{O}(N^3)$ computations a GP requires. This is done via the spectral representation and global parametrization that allows for straightforward SGD optimization.

5. Experiments

We compare LMM-VAE to the GP-prior VAEs, which are most closely aligned with our work in modelling the prior via regression methods. We prioritize this approach as regression techniques remain important and extensively used in longitudinal studies (see, e.g., Sauty & Durrleman, 2022).

Table 1: Imputation MSEs for the GP-based models and LMM-VAE on the Health MNIST Dataset with latent dimension 32.

Model	Imputation MSE ↓
GPP-VAE (Casale et al., 2018)	0.021 ±0.0012
SVGP-VAE (Jazbec et al., 2021)	0.015 ±0.0012
LVAE (Ramchandran et al., 2021)	0.018 ±0.0006
LMM-VAE (Ours)	0.002 ±0.0000
LMM-VAE (Ours)	0.002 ±0.0000
LMM-VAE (Ours)	0.002 ±0.0000

5.1. Health MNIST

Following prior work (Ramchandran et al., 2021), we generate a longitudinal dataset with missing pixels by augmenting the MNIST dataset. This modified dataset replicates various characteristics present in real medical data, where each snapshot of the time series corresponds to the health state of a patient. We select the digits ‘3’ and ‘6’ to represent two biological sexes. In total, there are $Q = 6$ covariates describing the dataset, which are age, id, diseasePresence, diseaseAge, sex, and location. Further details of the dataset and train-test splits are discussed in Appendix B.1.

GPP-VAE and SVGP-VAE are parametrized by two kernels describing the object and view. As such, we train these two models using the id and age covariates per respective kernel. Meanwhile, we parametrize LVAE using the optimal set-up as reported in their paper, i.e. id, age, sex×age, and diseasePresence×diseaseAge (Ramchandran et al., 2021). To demonstrate different ways of parametrizing LMM-VAE, we train three model variants, and summarize the included covariates per configuration in Figure 3. Note that the color-coding in Table 1 follows this mapping.

Missing Value Imputation. We report missing data imputation performance based on the *training* set in Table 1. Note that LMM-VAE’s imputation MSE remains robust across the different parametrizations of the linear model. In addition, LMM-VAE’s imputation MSE is significantly smaller than those of the the GP-based baselines. The results obtained with a smaller latent dimension for LMM-VAE can be found in Appendix B.4.

Future Prediction. By design, the data generation mechanism is a complex function of several key covariates. As shown in Figure 3, given LMM-VAE’s ability to model all of these covariates, it is unsurprising to learn that it achieves a lower conditional test MSE compared to GPP-VAE and SVGP-VAE. In addition, the sensitivity of LMM-VAE’s results to different prior parametrizations demonstrates the importance of including relevant covariates that could reveal information about the data generation process.

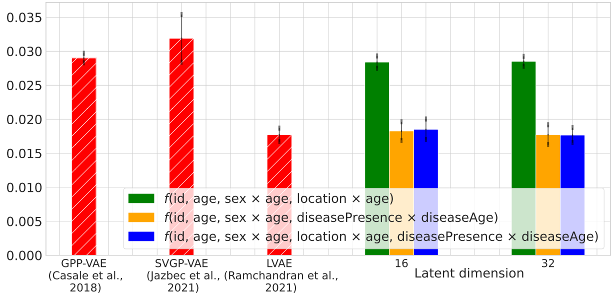

Figure 3: Predictive test MSEs on the Health MNIST dataset. The latent dimension used for the baselines is 32.

In addition, Figure 3 shows that LMM-VAE reports competitive MSE as compared to LVAE when the covariates used for training are identical.[1] Interestingly, the flexibility provided by a GP-prior may not translate into significant gain in predictive performance.

6. Conclusion

We present LMM-VAE, a novel method for modelling high dimensional longitudinal data that scales to large datasets with numerous covariates, and inherits interpretability from the additive LMM prior, which could appeal to practitioners. Theoretical analysis demonstrates connections to GP-based VAEs. This connection provides a foundation to adapt LMM-VAE to different modeling tasks by adaptively incorporating basis functions, as well as establishes LMM-VAE as a reduced-rank approximation method for GP prior VAEs.

Acknowledgements

We are grateful to the Aalto Science-IT project for the generous computational resources. This work has been supported by the Research Council of Finland (decision number: 359135).

References

Ashman, M., So, J., Tebbutt, W., Fortuin, V., Pearce, M., and Turner, R. E. Sparse gaussian process variational autoencoders, 2020.

Bauer, M., van der Wilk, M., and Rasmussen, C. E. Understanding probabilistic sparse gaussian process approximations, 2017.

Caruana, E. J., Roman, M., Hernández-Sánchez, J., and Solli, P. Longitudinal studies. *Journal of Thoracic Disease*, 7(11), 2015. ISSN 2077-6624.

[1] LVAE contains an additional interaction term between id and age, due to its modelling assumptions (Ramchandran et al., 2021).

Casale, F. P., Dalca, A. V., Saglietti, L., Listgarten, J., and Fusi, N. Gaussian process prior variational autoencoders. *CoRR*, abs/1810.11738, 2018.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. *CoRR*, abs/1309.6835, 2013.

Hensman, J., Durrande, N., and Solin, A. Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

Jazbec, M., Ashman, M., Fortuin, V., Pearce, M., Mandt, S., and Rätsch, G. Scalable gaussian process variational autoencoders, 2021.

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. Composing graphical models with neural networks for structured representations and fast inference. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Kingma, D. P. and Welling, M. An introduction to variational autoencoders. *CoRR*, abs/1906.02691, 2019.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2022.

Laird, N. M. and Ware, J. H. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982. ISSN 0006341X, 15410420.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Ramchandran, S., Tikhonov, G., Kujanpää, K., Koskinen, M., and Lähdesmäki, H. Longitudinal variational autoencoder. In *International Conference on Artificial Intelligence and Statistics*, pp. 3898–3906. PMLR, 2021.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows, 2016.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

Sauty, B. and Durrleman, S. Progression models for imaging data with longitudinal variational auto encoders. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pp. 3–13, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-16430-9. doi: 10.1007/978-3-031-16431-6_1.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Solin, A. and Särkkä, S. Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing*, 30(2):419–446, August 2019. ISSN 1573-1375. doi: 10.1007/s11222-019-09886-w.

Titsias, M. Variational learning of inducing variables in sparse gaussian processes. In van Dyk, D. and Welling, M. (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

Tomczak, J. M. and Welling, M. VAE with a vampprior. *CoRR*, abs/1705.07120, 2017.

Tompkins, A. and Ramos, F. Fourier feature approximations for periodic kernels in time-series modelling. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Zhao, S., Song, J., and Ermon, S. Infovae: Information maximizing variational autoencoders, 2018.

Zhu, H., Balsells-Rodas, C., and Li, Y. Markovian gaussian process variational autoencoders. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Zipunnikov, V., Greven, S., Shou, H., Caffo, B. S., Reich, D. S., and Crainiceanu, C. M. Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis. *The Annals of Applied Statistics*, 8(4):2175 – 2202, 2014. doi: 10.1214/14-AOAS748.

Appendix

A. Related Work

The literature on extending VAEs' modelling capacity is vast. This spans from extending the flexibility of variational posterior distributions, e.g., via normalizing flows (Rezende & Mohamed, 2016), to enabling a more informative prior over the standard normal used in plain VAEs (Sohn et al., 2015; Tomczak & Welling, 2017; Kingma & Welling, 2019). Additionally, alternative research efforts focus on endowing the model with desirable characteristics, such as a disentangled latent space (Higgins et al., 2016; Zhao et al., 2018). Our work enhances the latent space representation by incorporating auxiliary covariates into the prior, such as time-varying side-profile information.

A.1. Structured Variational Autoencoder

Our work builds upon the ideas presented in Johnson et al. (2016), wherein neural networks and probabilistic graphical models are combined to provide structured latent representations. LMM-VAE can be seen as a specific case of the Structured VAE (SVAE), but with enhancements designed to induce structure using arbitrarily many covariates characteristic of longitudinal data. In real-world longitudinal applications, the incorporation of such structure remains relevant. For instance, Sauty & Durrleman (2022) characterize the progression of Alzheimer's disease via a mixed-effect longitudinal model in the VAE setting. However, this model's parametric structure is confined to disease progression and necessitates an internal procedure of Markov chain Monte Carlo sampling. LMM-VAE, in contrast, emerges as a more generalized rendition of this approach.

A.2. Gaussian Process Variational Autoencoder

The family of GP-based VAEs is most relevant as related work in modelling the prior via regression methods. Similar to LMM-VAE, the GP-VAEs fall under the broad umbrella of SVAEs (Johnson et al., 2016). Already, within the GP-VAE framework, there have been multiple developments (Casale et al., 2018; Ramchandran et al., 2021; Jazbec et al., 2021; Ashman et al., 2020; Zhu et al., 2023), where the GP's expressiveness and smoothness are leveraged to enable more flexible, yet robust VAE priors. Here, we focus on the GP-VAEs that are most compatible with modelling longitudinal data, i.e. repeated measurements with auxiliary covariates, including id-specific information.

While GP-based priors such as those in Casale et al. (2018); Jazbec et al. (2021); Ramchandran et al. (2021) can model longitudinal data, using a GP model component comes with a host of difficulties. *Firstly*, learning scales cubically with respect to the number of samples (Rasmussen & Williams, 2006), thereby constituting a bottleneck to the model's scalability. To this end, attempts have been made to reduce the training complexity via approximations of the GP-priors, such as by Taylor approximation (Casale et al., 2018), or through inducing points (Ramchandran et al., 2021; Jazbec et al., 2021). For the latter, optimizing the inducing point locations is not straightforward (Bauer et al., 2017) and may complicate training due to the coupled learning of both the latent variables and the inducing points (Titsias, 2009; Hensman et al., 2013). This complication may be exacerbated with categorical inputs, which are typically modelled in longitudinal set-ups. In addition, unless appropriately designed and implemented, these approximations may result in diminished expressiveness of the GP-priors.

Secondly, modelling all the available covariates via a GP is non-trivial. Jazbec et al. (2021)'s approach may be ill-suited for this task as it assumes low-dimensional covariates. Casale et al. (2018) posit that the auxiliary information can be represented by an object and a view kernel, where the dataset comprises of objects in different views. However, constructing these kernels from high dimensional side-profile information consisting of continuous and categorical covariates remains unclear. Meanwhile, Ramchandran et al. (2021) propose using additive kernels to include all available covariates, where each component or pair of components implements one of the additive kernels. Nevertheless, any potential performance gain may be limited by the challenges associated with training the GPs as a VAE prior.

B. Health MNIST

B.1. Dataset Description

We rely on the dataset construction script from Ramchandran et al. (2021).

Table 2: Comparison of related methods.

MODEL	PRIOR	COVARIATES	MINIBATCHING	REFERENCES
CVAE	(I.I.D)^{2}GAUSSIAN	ARBITRARY	✓	SOHN ET AL. (2015)
GPP-VAE	GP	LIMITED	PSEUDO	CASALE ET AL. (2018)
SVGP-VAE	GP	LIMITED	✓	JAZBEC ET AL. (2021)
LVAE	GP	ARBITRARY	✓	RAMCHANDRAN ET AL. (2021)
LONGITUDINAL VAE	LM	LIMITED	✓	SAUTY & DURRLEMAN (2022)
LMM-VAE	LMM (OR GP)	ARBITRARY	✓	OUR WORK

To generate a shared age-related effect, we gradually shift all digit instances towards the right corner over time. Half of the instances of '3' and '6' are assumed to be healthy (diseasePresence $= 0$), while the other half are inflicted with a disease (diseasePresence $= 1$). The diseased instances are rotated across 20 timepoints, with the rotation degree determined by the time to disease diagnosis (diseaseAge).

We further include a binary noise covariate location, which is randomly assigned to each unique instance, and apply a random rotational jitter to each data point to simulate noisy observations. Additionally, we mask out 25% of each image's pixels (to assess imputation capabilities).

In total, there are 1300 unique instances present in the dataset, where 650 correspond to the biological sex Male, and the remaining 650 correspond to Female. Each of these unique instances have a sequence length of 20. We withhold the last 15 timepoints of 100 subjects to construct the test set. The first five timepoints of these aforementioned subjects are included in the training set. The remaining dataset is then randomly split to construct the train and validation sets, in an approximate ratio of $85 : 15$.

B.2. Experimental details

We took the implementation of the baseline SVGP-VAE (Jazbec et al., 2021) from `https://github.com/ratschlab/SVGP-VAE`, GPP-VAE (Casale et al., 2018) from `https://github.com/fpcasale/GPPVAE`, and LVAE (Ramchandran et al., 2021) from `https://github.com/SidRama/Longitudinal-VAE`. To perform experiments, we use the default hyperparameter setting specified in the respective repositories. Across all baselines, the architecture used for the experiment can be found in Appendix C.

For experiments regarding LMM-VAE, we use Adam optimizer (Kingma & Ba, 2017) with a learning rate of 0.001. We also use a step learning rate scheduler with a step size of 500, and a learning rate decay factor of 0.99. We monitor the loss on the validation set and employ a strategy similar in spirit to early stopping, where we save the weights of the model with the optimal validation loss. LMM-VAE was allowed to run for a maximum of 2500 epochs. We define $\sigma_z = 1$.

For all experiments, we report the mean and standard deviation obtained across five runs.

B.3. Additional illustrations of LMM-VAE's predictions

We visualize two sets of trajectories corresponding to 2 individuals in Figure 4.

B.4. Supplementary tables for Health MNIST

The tables containing the experimental results for Health MNIST are described in Table 3 and Table 4.

C. Model Architectures

Table 5 contains the neural network architecture used in the Health MNIST experiments, which follows Ramchandran et al. (2021).

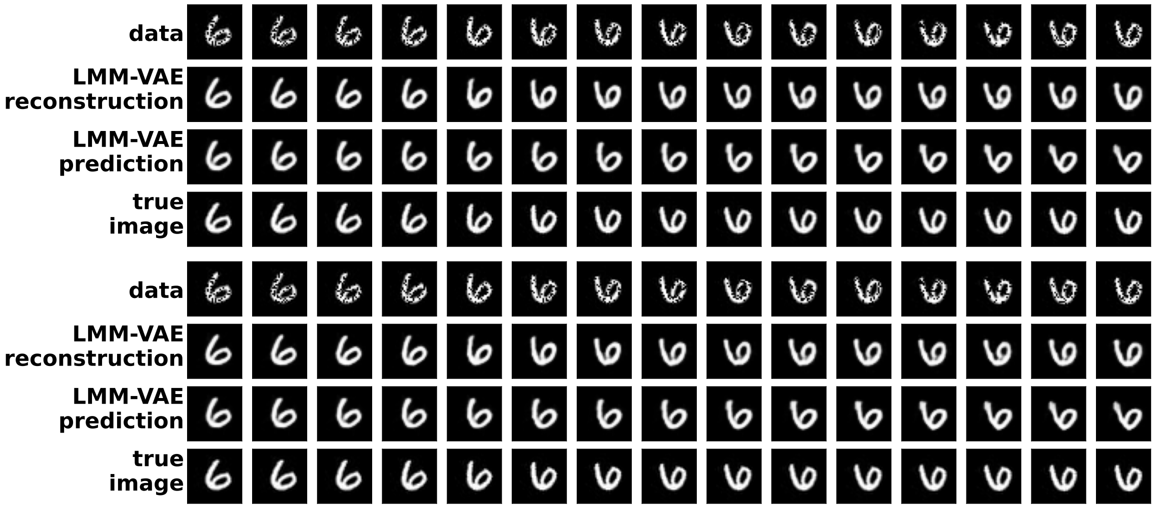

Figure 4: We illustrate 2 sets of images here, each corresponding to a different biological sex. Per image set, we visualize the reconstructions and predictions obtained on the test set by LMM-VAE with 16 latent dimensions in the second and third rows. The noisy data along with the original uncorrupted images are also depicted in the first and last rows, respectively.

Table 3: Imputation MSEs for LMM-VAE on the Health MNIST Dataset.

MODEL	LATENT DIMENSION	IMPUTATION MSE ↓
LMM-VAE	16	0.002 ±0.0000
LMM-VAE		0.002 ±0.0000
LMM-VAE		0.002 ±0.0000
LMM-VAE	32	0.002 ±0.0000
LMM-VAE		0.002 ±0.0000
LMM-VAE		0.002 ±0.0000

Table 4: Predictive Test MSEs for LMM-VAE on the Health MNIST Dataset.

MODEL	LATENT DIMENSION	PREDICTIVE MSE ↓
LMM-VAE	16	0.0284 ±0.0007
LMM-VAE		0.0183 ±0.0010
LMM-VAE		0.0185 ±0.0014
LMM-VAE	32	0.0285 ±0.0009
LMM-VAE		0.0177 ±0.0013
LMM-VAE		0.0177 ±0.0008

Table 5: Neural Network Architecture used for the model in the Health MNIST experiments.

	Hyperparameter	Value
Inference Network	Dimensionality of input	36×36
	Number of convolution layers	2
	Kernel size	3×3
	Stride	2
	Pooling	Max Pooling
	Pooling kernel size	2×2
	Pooling stride	2
	Number of feedforward layers	2
	Width of feedforward layers	300, 30
	Dimensionality of latent space	L
	Activation function of layers	ReLU
Generative Network	Dimensionality of input	L
	Number of transposed convolution layers	2
	Kernel size	4×4
	Stride	2
	Number of feedforward layers	2
	Width of feedforward layer	30, 300
	Activation function of layers	ReLU, Sigmoid