Latent Gaussian process with composite likelihoods and numerical quadrature: Supplementary Materials

1 Derivation of the evidence lower bound (ELBO)

To obtain the ELBO, we can first write the log-likelihood as,

$$\log p(\boldsymbol{Y}) = \log \int p(\boldsymbol{X})p(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U})p(\boldsymbol{Y}|\boldsymbol{F})d\boldsymbol{X}d\boldsymbol{F}d\boldsymbol{U}.$$

Multiplying and dividing by $q(\boldsymbol{X},\boldsymbol{F},\boldsymbol{U})$, we can re-write the log-likelihood as,

$$\log p(\boldsymbol{Y}) = \log \int \frac{q(\boldsymbol{X},\boldsymbol{F},\boldsymbol{U})}{q(\boldsymbol{X},\boldsymbol{F},\boldsymbol{U})}p(\boldsymbol{X})p(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U})p(\boldsymbol{Y}|\boldsymbol{F})d\boldsymbol{X}d\boldsymbol{F}d\boldsymbol{U}. \tag{1}$$

The variational approximation to the posterior distribution, $q(\boldsymbol{X},\boldsymbol{F},\boldsymbol{U})$, can be factorised as follows:

$$q(\boldsymbol{X},\boldsymbol{F},\boldsymbol{U}) = q(\boldsymbol{X})q(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U}). \tag{2}$$

Substituting into Eq. (1), we get:

$$\log p(\boldsymbol{Y}) = \log \int q(\boldsymbol{X})q(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U})\frac{p(\boldsymbol{X})p(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U})p(\boldsymbol{Y}|\boldsymbol{F})}{q(\boldsymbol{X})q(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U})}d\boldsymbol{X}d\boldsymbol{F}d\boldsymbol{U}. \tag{3}$$

Jensen's inequality relates the value of a concave (or convex) function of an integral to the integral of the concave (or convex) function (Jensen et al., 1906). Assume φ is a concave function and X is a random variable. By the Jensen's inequality for a concave function, we can write:

$$\varphi(\mathbb{E}[X]) \geq \mathbb{E}[\varphi(X)]. \tag{4}$$

In our model, we have $\varphi = \log$. Substituting this in Eq. (4) and for a random variable X, we have:

$$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log(X)]. \tag{5}$$

We can now apply the Jensen's inequality from Eq. (5) to Eq. (3):

$$\log p(\boldsymbol{Y}) \geq \int q(\boldsymbol{X})p(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U})\log\frac{p(\boldsymbol{X})p(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U})p(\boldsymbol{Y}|\boldsymbol{F})}{q(\boldsymbol{X})p(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U})}d\boldsymbol{X}d\boldsymbol{F}d\boldsymbol{U}. \tag{6}$$

The Kullback-Leibler divergence (Kullback and Leibler, 1951) between $q(\boldsymbol{X})$ and $p(\boldsymbol{X})$ as well as between $q(\boldsymbol{U})$ and $p(\boldsymbol{U})$ can be written as

$$\begin{aligned}\mathrm{KL}(q(\boldsymbol{X})||p(\boldsymbol{X})) &= \int q(\boldsymbol{X})\log\frac{q(\boldsymbol{X})}{p(\boldsymbol{X})}d\boldsymbol{X},\\ \mathrm{KL}(q(\boldsymbol{U})||p(\boldsymbol{U})) &= \int q(\boldsymbol{U})\log\frac{q(\boldsymbol{U})}{p(\boldsymbol{U})}d\boldsymbol{X}.\end{aligned}$$

Substituting the KL divergences in Eq. (6) and unwrapping the remaining terms along the dimension d (i.e. the dimension of the data space) from their vectorised form, we get:

$$\begin{aligned}\log p(\boldsymbol{Y}) \geq &-\mathrm{KL}(q(\boldsymbol{X})||p(\boldsymbol{X})) - \mathrm{KL}(q(\boldsymbol{U})||p(\boldsymbol{U}))\\ &+\sum_{d=1}^{D}\int q(\boldsymbol{X})q(\boldsymbol{u}_d)p(\boldsymbol{f}_d|\boldsymbol{X},\boldsymbol{u}_d)\cdot\log p(\boldsymbol{y}_d|\boldsymbol{f}_d)d\boldsymbol{X}d\boldsymbol{f}_d d\boldsymbol{u}_d = \mathcal{L}.\end{aligned} \tag{7}$$

2 Likelihood model with categorical distribution

For the categorical distribution, we make of a formulation similar to Gal et al. (2015) which is a generalisation of the binomial distribution. In this case, the GPs produce the weights for each of the categories. We then make use of the *softmax* function to get probabilities for the categories in the range of $[0, 1]$. Assume all categorical variables to have the same cardinality, K. Hence, $P_d = K$. For the d^{th} variable of the n^{th} entry, we can write $\bar{f}_{n,d} = \{f_{n,d,1}, f_{n,d,2}, ..., f_{n,d,K}\}$. Following a similar notation to Gal et al. (2015), we can write $y_{n,d} \sim \text{softmax}(\bar{f}_{n,d})$, where

$$\text{softmax}(y_{n,d} = k; \bar{f}_{n,d}) = \text{categorical}\left(\frac{\exp(f_{n,d,k})}{\sum_{k'=1}^{K}\exp(f_{n,d,k'})}\right),$$

and categorical corresponds to the categorical distribution (or generalised Bernoulli distribution).

3 Origin-centred latent representations

From Eq. (12) in the main paper, we can write the lower bound $\mathcal{L}$ with our suggested modification as follows:

$$\mathcal{L} \approx -\eta\overbrace{\text{KL}(q(\boldsymbol{X})||p(\boldsymbol{X}))}^{\text{KL}_{\boldsymbol{X}}} - \overbrace{\text{KL}(q(\boldsymbol{U})||p(\boldsymbol{U}))}^{\text{KL}_{\boldsymbol{U}}} + \frac{1}{N_x}\sum_{i=1}^{N_x}\sum_{d=1}^{D}\mathbb{E}_{q(\boldsymbol{f}_d|\boldsymbol{X}_i)}[\log p(\boldsymbol{y}_d|\boldsymbol{f}_d)], \tag{8}$$

where η is the new hyper-parameter. If $\eta = 1$, we get the same equation as Eq. (12) in the main paper. Using $\eta > 1$ results in embeddings that are more centred about the origin but with qualitatively consistent results (i.e. qualitatively consistent with the hyper-parameter $\eta = 1$).

To obtain a centred embedding, the hyper-parameter η must be tuned. We have also observed that as the value of η begins to increase to larger values, the clustering structure begins to disappear as the KL term associated with X (i.e. $\text{KL}_{\boldsymbol{X}}$ in Eq. (8)) begins to dominate and the optimisation tries to move the latent points closer to zero while making them appear to be a sample from the standard normal distribution.

Hence, we can infer that incorporating a weight on $\text{KL}_{\boldsymbol{X}}$ (i.e. $\eta > 1$) results in latent embeddings that are generally centred around the origin with results that are qualitatively consistent with Eq. (12) in the main paper (i.e. $\eta = 1$).

4 Ablation study

We demonstrate that numerical quadrature has a comparable predictive performance to the standard sampling based inference, while having a significantly lower run time. We followed an approach similar to the 'Benchmark and performance comparisons' described in Section 3 of the main manuscript. We performed a 2-fold cross-validation with the predictive log-likelihood as the evaluation score to compare between the two techniques. This was done by sub-sampling 50% of the original data for training and then computing the predictive log-likelihoods on the remaining data. This process was repeated 30 times and the same folds were used for both analyses. In all the runs we used $Q = 2$. We can see from Fig. 11 that quadrature takes half the amount of time for each run. Fig. 12 visualises the trace plot of the ELBOs.

5 Dimensionality reduction techniques

Most dimensionality reduction techniques focus on preserving the distances between nearby objects than objects that are further apart (i.e. local distance preservation). However, it is also important to ensure that objects that are further apart in data space are also kept apart in the reduced space (i.e. dissimilarity preservation). Unfortunately, in most cases it is not possible to achieve both. In this work, we focus on feature projection which is the transformation of data from a high-dimensional space to lower dimensional manifold. There are many popular algorithms for dimensionality reduction such as PCA (Jolliffe, 2011), kernel PCA (Schölkopf et al., 1997)], locally linear embedding (Roweis and Saul, 2000), Isomap (Tenenbaum et al., 2000), GPLVM (Lawrence, 2003), etc. These algorithms are suitable in the homogeneous data setting and do not necessarily extend well to a heterogeneous data setting. We build upon the GPLVM to support a heterogeneous data setting.

PCA tries to identify an orthogonal linear projection of the data along the direction of maximum variance. Classical PCA has some shortcomings. It is not probabilistic as it has no likelihood model for the observed data. Moreover, computation of the covariance matrix and its associated eigendecomposition can be computationally intensive for large datasets with high dimensionality (Prasad and Bruce, 2008). Another issue with classical PCA is that it cannot handle missing data properly and is not robust to outliers (Kambhatla and Leen, 1997). Hence, it is not suitable in a generative model setting.

Probabilistic principal component analysis (PPCA), proposed by (Tipping and Bishop, 1999), is a generalisation of the classical PCA that tries to overcome its shortcomings. It incorporates a probabilistic model and obtains a linear projection by maximising the likelihood. Assume a D-dimensional dataset $\boldsymbol{Y}$ of N points, i.e., $\boldsymbol{Y} = \{\boldsymbol{y}_n\}_{n=1}^{N}$ such that $\boldsymbol{y}_n \in \mathbb{R}^D$ and is centred. We can denote each latent variable corresponding to each data point as $\boldsymbol{x}_n \in \mathbb{R}^Q$ such that $Q \leq D$. Also, let $\boldsymbol{W} \in \mathbb{R}^{D\times Q}$ denote the principal axes or weights. We can write the likelihood for an individual data point as

$$\begin{aligned} p(\boldsymbol{y}_n|\boldsymbol{W}, \sigma^2) &= \int p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{W}, \sigma^2) p(\boldsymbol{x}_n) d\boldsymbol{x}_n \\ p(\boldsymbol{x}_n) &= N(\boldsymbol{x}_n|0, \boldsymbol{I}) \\ p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{W}, \sigma^2) &= N(\boldsymbol{y}_n|\boldsymbol{W}\boldsymbol{x}_n, \sigma^2\boldsymbol{I}_D) \end{aligned}$$

where σ^2 is the noise variance and we assume an isotropic Gaussian noise model. To solve for $\boldsymbol{W}$ we assume that $\boldsymbol{y}_n$ is i.i.d. and maximise the likelihood for all data points

$$p(\boldsymbol{Y}|\boldsymbol{W}, \sigma^2) = \prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{W}\boldsymbol{x}_n, \sigma^2\boldsymbol{I}_D).$$

Marginalising out the latent variables, the distribution for each point can be written as

$$\boldsymbol{y}_n \sim N(0, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}_D).$$

The parameters are optimised to obtain the maximum likelihood. We can say that the classical PCA is a limiting case of PPCA when the covariance becomes infinitesimally small, i.e., $\sigma^2 \to 0$.

Lawrence (2003) proposed the Gaussian process latent variable model (GPLVM) as a generalisation of PPCA. It is an unsupervised learning algorithm that extends PPCA by making use of a less restrictive covariance function that allows for non-linear mappings. Building upon the derivation of PPCA and following Lawrence (2003), we can obtain the GPLVM formulation.

A prior distribution is defined for $\boldsymbol{W}$ as $p(\boldsymbol{W}) = \prod_{i=1}^{D} N(\boldsymbol{w}_i|0, \alpha^{-1}\boldsymbol{I})$. From the PPCA derivation, instead of integrating out the latent variables $\boldsymbol{X}$, the principal axes $\boldsymbol{W}$ is marginalised to give the marginal likelihood for $\boldsymbol{Y}$,

$$p(\boldsymbol{Y}|\boldsymbol{X}, \sigma) = \frac{1}{(2\pi)^{\frac{DN}{2}}|\boldsymbol{K}|^{\frac{D}{2}}} \exp\left(-\frac{1}{2}\operatorname{tr}(\boldsymbol{K}^{-1}\boldsymbol{Y}\boldsymbol{Y}^T)\right),$$

where tr corresponds to the trace of a matrix and $\boldsymbol{K} = \alpha\boldsymbol{X}\boldsymbol{X}^T + \sigma^2\boldsymbol{I}$. Hence, the marginal likelihood that is being optimised can be interpreted as the product of D independent Gaussian processes where $\boldsymbol{K}$ is given by the linear covariance function. Therefore to obtain the GPLVM formulation, a non-linear covariance function is introduced which corresponds to a non-linear mapping from latent space to data space. A common choice for the process prior is the radial basis function (RBF) kernel. The marginal likelihood is jointly marginalised with respect to the latent variables, $\boldsymbol{X}$ and other parameters. In this study, we make use of the automatic relevance determination radial basis function (ARD RBF) kernel that allows for separate length scales in each dimension of the latent space.

Hence, the optimisation problem can be written as,

$$\{\hat{\boldsymbol{X}}, \hat{\boldsymbol{\theta}}\} = \underset{\boldsymbol{X},\boldsymbol{\theta}}{\arg\max}\, p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ corresponds to the kernel parameters and $\hat{\boldsymbol{X}}$ as well as $\hat{\boldsymbol{\theta}}$ corresponds to the optimal values of the latent variables and kernel parameters, respectively.

6 Runtime performance

The mean wall clock times (on an Intel Xeon E5 processor) for each repetition of 1000 epochs in Fig. 4 of the main manuscript using the simulated data was as follows:

- Approach 1: 1.3 hours
- Approach 2 (our method): 1.35 hours
- Approach 3: 1.5 hours

Also, Suppl. Fig. 11 shows a comparison of the wall-clock time for 1000 epochs between numerical quadrature and sampling based inference in our model with the simulated data. Our results indicate that the numerical integration makes the inference about two-fold faster, and overall our model is approximately as fast as the Gaussian GPLVM. We had sub-sampled the MNIST to maintain a short computation time.

7 Neural network architecture

The mean and covariance of the variational distribution over $q(\boldsymbol{X})$ are parameterised by neural networks. In our experiment with clinical patient data, we utilised simple feedforward multilayered perceptrons (MLPs) as the recognition models. Concretely, we made use of two separate MLPs for the mean and covariance respectively. The hyperparameters for the networks are reported in Table 1.

	Hyperparameter	Value
Mean	Dimensionality of input	46
	Number of hidden layers	1
	Width of hidden layer	30
	Activation function of the hidden layer	TanH
	Dimensionality of output	$\mathcal{Q}$
	Activation function of output layer	Linear
	Weight initialisation	Xavier initialisation (Glorot and Bengio, 2010)
Covariance	Dimensionality of input	46
	Number of hidden layers	1
	Width of hidden layer	30
	Activation function of the hidden layer	TanH
	Dimensionality of output	$\mathcal{Q}$
	Activation function of output layer	Sigmoid
	Weight initialisation	Xavier initialisation (Glorot and Bengio, 2010)

Table 1: Hyperparameters used in the recognition models for the clinical patient dataset.

8 Supplementary figures

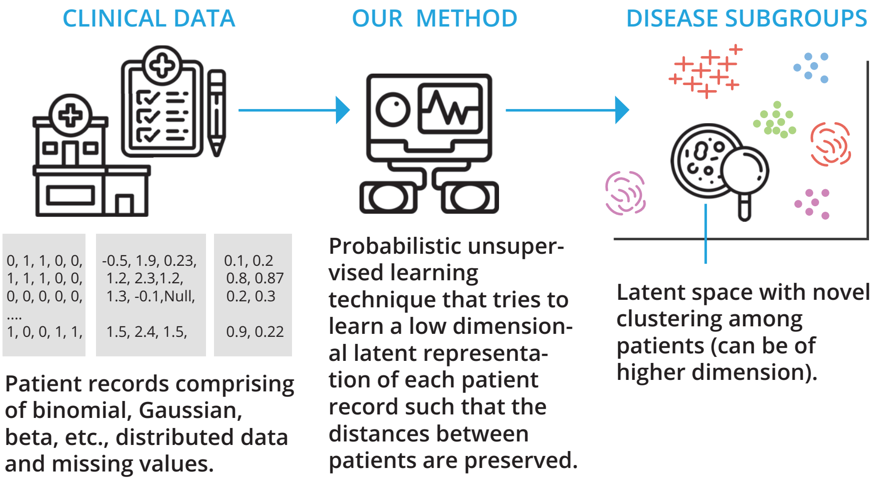
CLINICAL DATA
OUR METHOD
DISEASE SUBGROUPS
0, 1, 1, 0, 0,
1, 1, 1, 0, 0,
0, 0, 0, 0, 0,
....
1, 0, 0, 1, 1,
-0.5, 1.9, 0.23,
1.2, 2.3,1.2,
1.3, -0.1,Null,
1.5, 2.4, 1.5,
0.1, 0.2
0.8, 0.87
0.2, 0.3
0.9, 0.22
Patient records comprising of binomial, Gaussian, beta, etc., distributed data and missing values.
Probabilistic unsupervised learning technique that tries to learn a low dimensional latent representation of each patient record such that the distances between patients are preserved.
Latent space with novel clustering among patients (can be of higher dimension).

Figure 1: An overview of our unsupervised generative model for disease stratification.

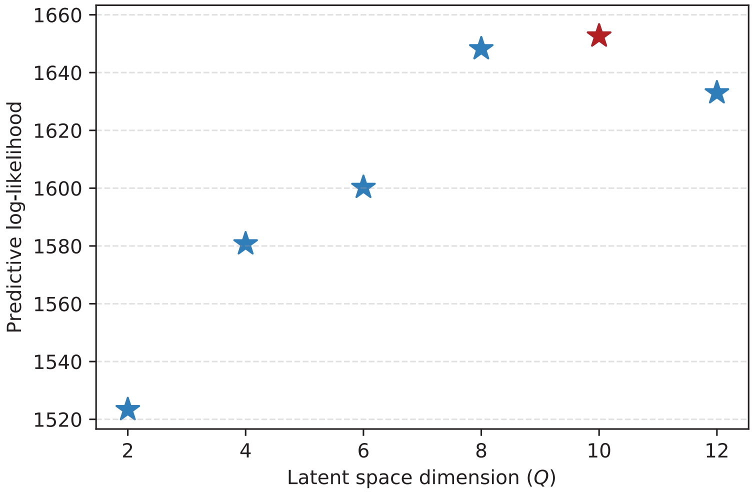

Figure 2: Assessment of optimal latent dimensionality using the predictive log-likelihood on the held-out test data for different latent space dimensions, Q in the clinical patient dataset experiment. The best predictive log-likelihood among three runs is shown.

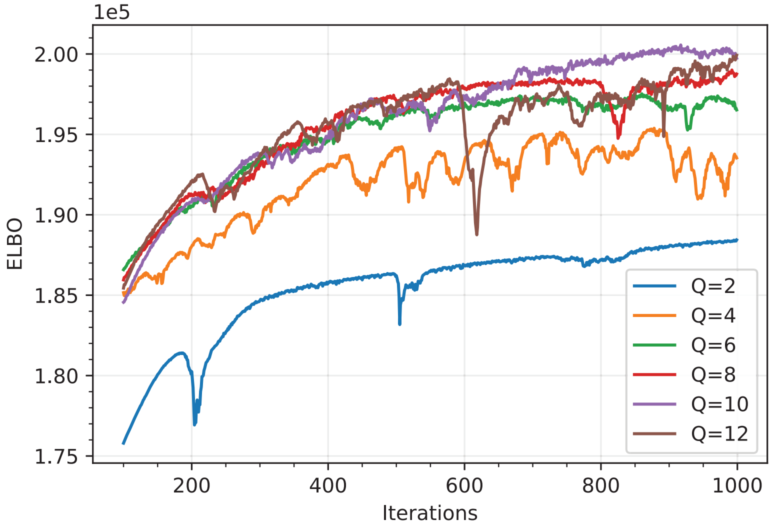

Figure 3: Trajectories of the evidence lower bounds (ELBO) from the best optimisation run for each latent dimension in Fig. 2.

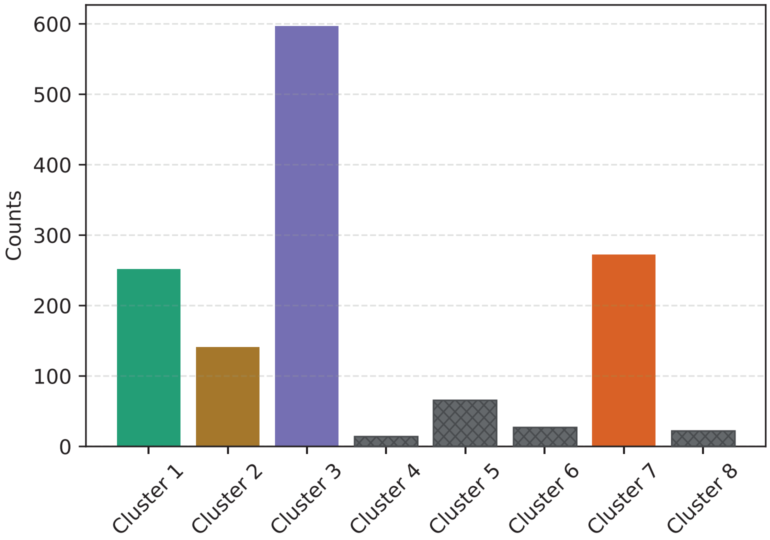

Figure 4: Visualisation of the number of patients assigned to each cluster. The optimal number of clusters and cluster membership was obtained using the described method. This figure is similar to Fig. 3(a) in the main manuscript, but includes the clusters deemed as outliers.

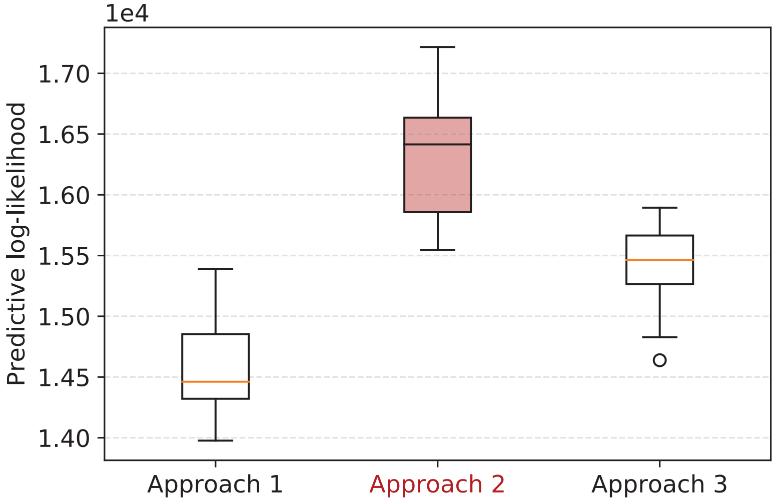

Figure 5: A box plot comparing three approaches of estimating the Gaussian distributed data similar to Fig. 4(b) in the main manuscript. **Approach 2** is our method. The predictive log-likelihood is computed on 30 sub-samples using 2-fold cross-validation (the partitions are the same across the analyses).

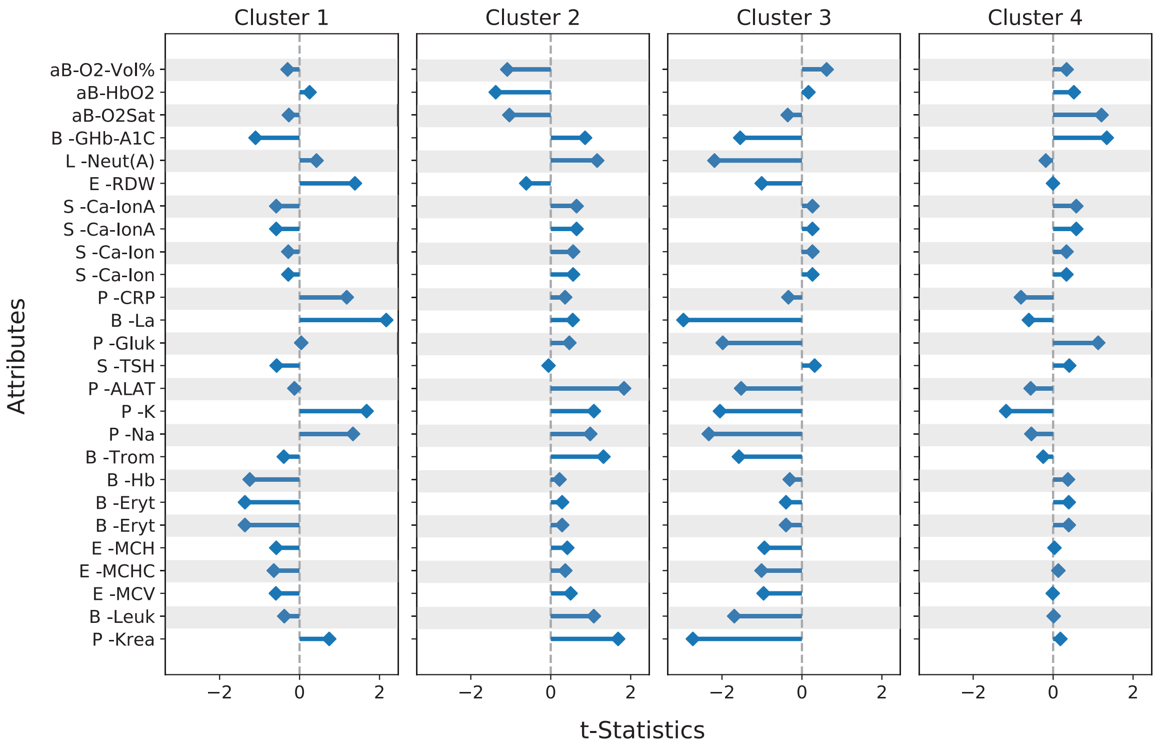

Figure 6: Evaluation of cluster characteristics using t-statistics in the clinical patient dataset.

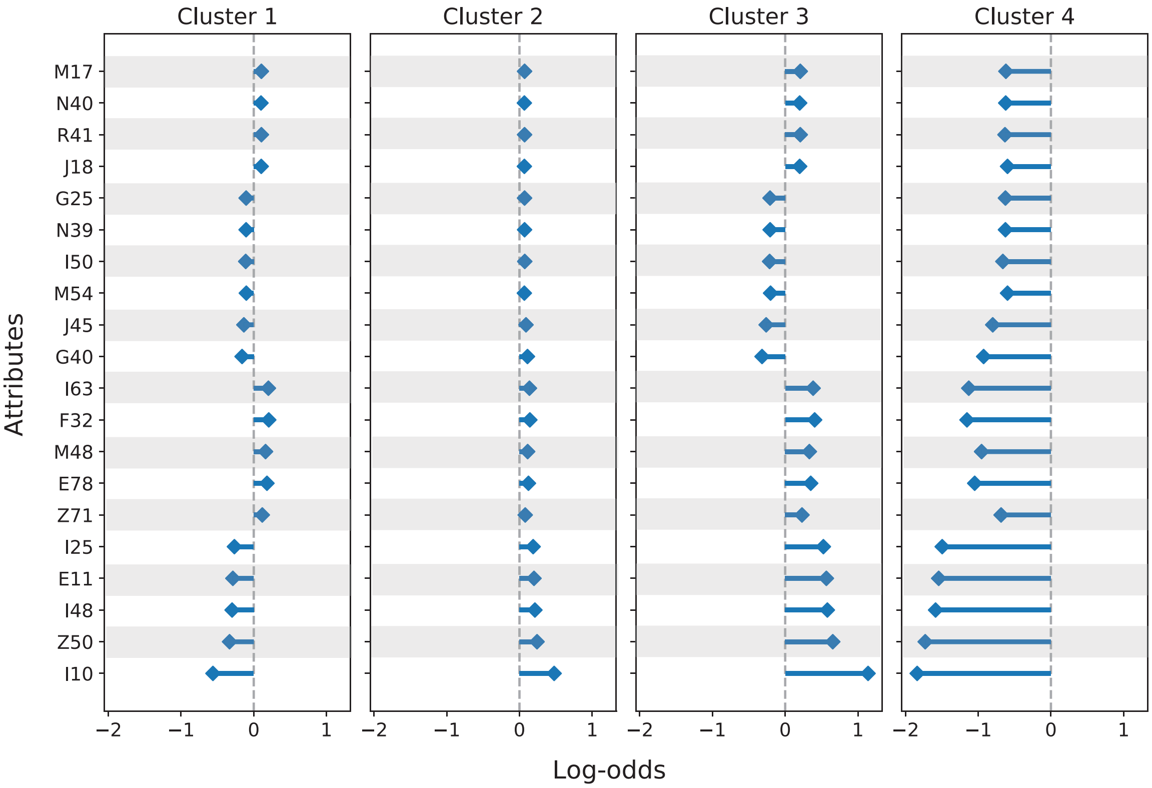

Figure 7: Evaluation of binomial cluster characteristics using log-odds ratio in the clinical patient dataset.

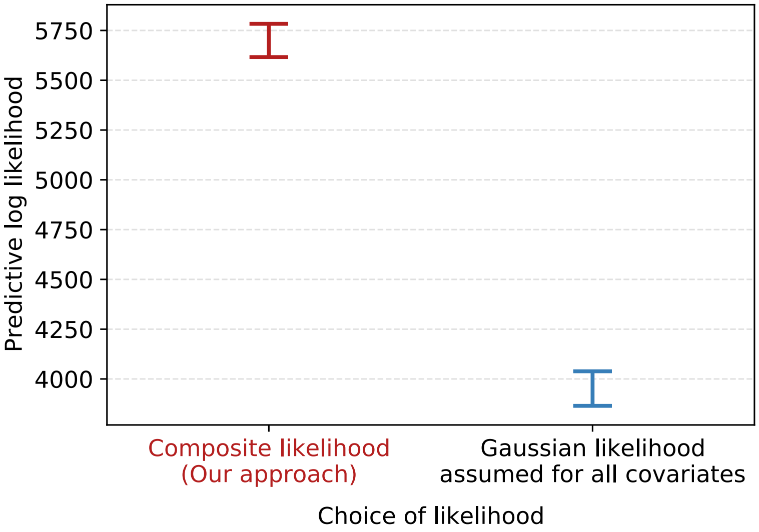

Figure 8: Comparison between our composite likelihood method and only Gaussian likelihood method for the simulated dataset.

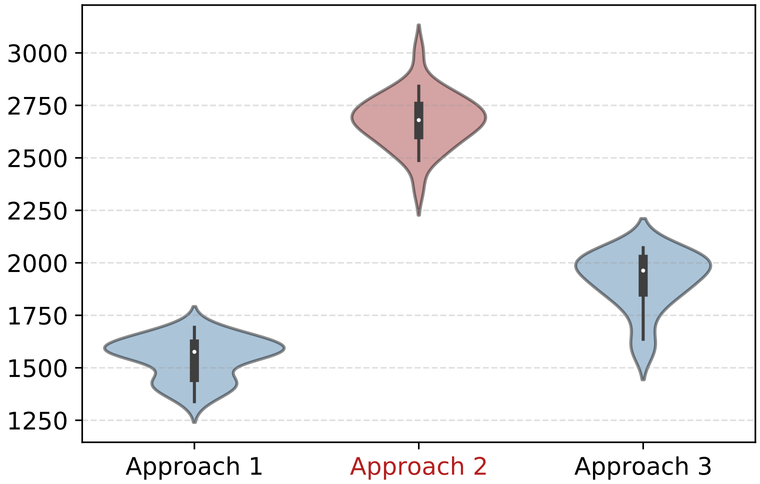

Figure 9: Violin plot comparing three approaches of estimating the Gaussian distributed covariates for the simulated dataset. **Approach 2** is our method.

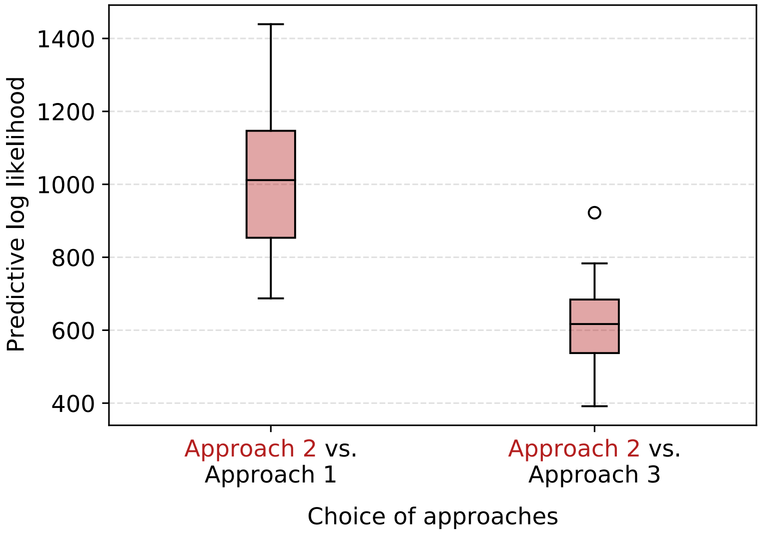

Figure 10: A pair-wise comparison of the difference in the predictive log-likelihood between our approach and the other approaches for the simulated dataset. The paired differences are computed on the same, matched sub-samples.

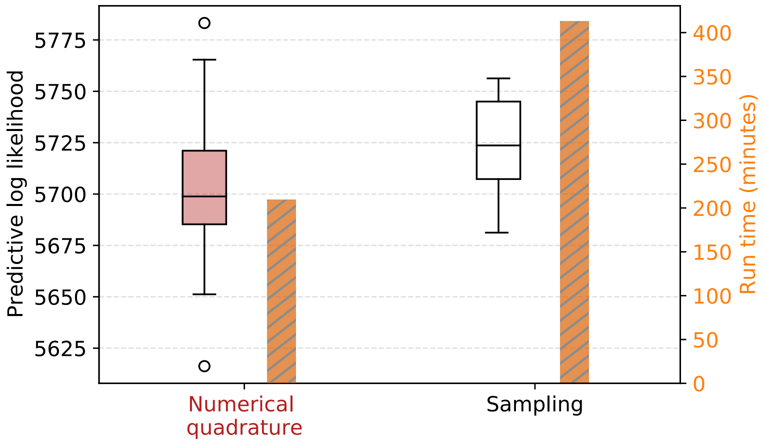

Figure 11: Comparison of the log predictive likelihood achieved by using either quadrature or sampling based inference for the simulated dataset. The wall-clock run times for 1000 epochs are shown on right.

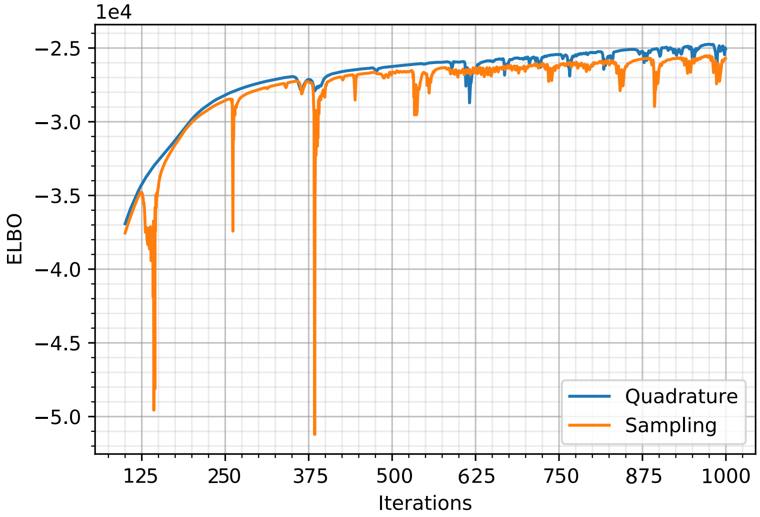

Figure 12: Trajectories of the evidence lower bounds (ELBO) for quadrature and sampling based inference with $Q = 2$ for the simulated dataset.

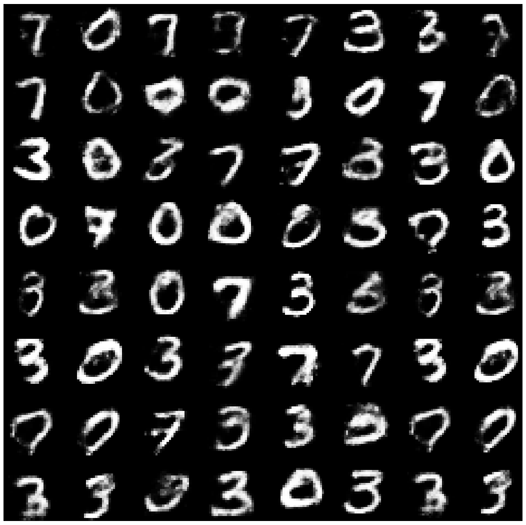

Figure 13: Reconstructions obtained in the simulated dataset experiment.

References

Y. Gal, Y. Chen, and Z. Ghahramani. Latent gaussian processes for distribution estimation of multivariate categorical data. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, 2015.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010*, JMLR Proceedings, 2010.

J. L. W. V. Jensen et al. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30:175–193, 1906.

I. Jolliffe. Principal component analysis. In *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer, 2011.

N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis. *Neural Computation*, 9(7), 1997.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems, NeurIPS 2003*, 2003.

S. Prasad and L. M. Bruce. Limitations of principal components analysis for hyperspectral target recognition. *IEEE Geoscience and Remote Sensing Letters*, 5(4), 2008.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500), 2000.

B. Schölkopf, A. J. Smola, and K. Müller. Kernel principal component analysis. In *7th International Conference on Artificial Neural Networks, ICANN*. Springer, 1997.

J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500), 2000.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 61(3):611–622, 1999.