Amortized Structured Stochastic Variational Inference for Gaussian Process Latent Variable Models

Maksym Tretiakov[1,2], **Sarah Lucie Filippi**[3], **Vincent Fortuin**[2,4,5], **Ruth Misener**[3]
Ruby Sedgwick[6], **James Odgers**[2,4,5]

[1]LMU Munich, [2]Munich Center for Machine Learning (MCML), [3]Imperial College London
[4]University of Technology Nuremberg (UTN), [5]Helmholtz AI, [6]Xyme

Abstract

Many machine learning methods aim to approximate the lower-dimensional manifold on which the data lives. A desirable feature of such methods is that they should capture the epistemic uncertainty of this learned manifold. One model that achieves this is the Gaussian Process Latent Variable Model, in which a Gaussian Process (GP) mapping from the latent space provides an estimate of the uncertainty of the manifold. However, the effectiveness of this uncertainty estimation is limited by the mean-field variational approximation between the GP inducing points and the latent variables. In this work, we apply Amortized Structured Stochastic Variational Inference to allow the variational posterior for the latent space to be conditionally dependent on the value of the inducing points. We demonstrate that this more flexible variational posterior improves several metrics relating to the reconstruction of points on the data manifold.

1 Introduction

In many real-world settings, observed high-dimensional data has a complex structure that can be naturally explained by a smaller number of unobserved factors. This general formulation is relevant for a broad class of models, starting from classical models, such as Probabilistic Principal Component Analysis [PPCA; Tipping and Bishop, 1999], to modern deep generative models, such as Variational Auto-Encoders [VAEs; Kingma and Welling, 2014]. These approaches can uncover interpretable low-dimensional structure and enable downstream inference tasks.

Unfortunately, in many practical tasks where learning a data manifold is desirable, the data is likely to be both noisy and limited in number. In such cases, it is important to learn not only point estimates for the data manifold, but also to estimate the epistemic uncertainty of the manifold. One model that can achieve this is the Gaussian Process Latent Variable Model [GPLVM; Lawrence, 2005, Titsias and Lawrence, 2010, Damianou et al., 2016]. Thanks to its non-parametric Bayesian approach, this method can both learn arbitrarily complex data manifolds and remain robust to overfitting. These properties have allowed the GPLVM and its variants to be applied to a diverse set of applications, including biology [Reid and Wernisch, 2016, Ahmed et al., 2019, Sedgwick et al., 2025], chemical engineering [Odgers et al., 2025], causal learning [Dhir et al., 2024, 2025], and robotics [Jaquier et al., 2024, Rozo et al., 2025].

However, all approaches to the GPLVM share a shortcoming: a Mean Field (MF) approximation in the variational family. Specifically, current methods approximate the posterior of the positions of the data in the latent space and the mapping from the latent space to the observations to be independent of one another [Lalchand et al., 2022]. This approach results in the model underestimating the epistemic uncertainty in the GP mapping. In this work, we apply Structured Stochastic Variational Inference [SSVI; Hoffman and Blei, 2015] to introduce dependence of the latent variable on the Sparse GP inducing points, thereby improving the approximation to the true posterior.

The rest of the paper is laid out as follows: first, in Section 2, we provide an overview of the background required for this paper. Following this, in Section 3 we introduce our method and derive the Evidence Lower Bound (ELBO) and discuss the relationship of our approach to existing methods in Section 4. In Section 5, we demonstrate that our method outperforms the existing MF methods and argue that it provides a better estimate of the epistemic uncertainty. Finally, in Sections 6 and 7, we discuss some of the current limitations of our approach and suggest further work which could take advantage of our variational family.

2 Background

The problem of learning the latent structure from high-dimensional data has been studied for a significant amount of time [Tipping and Bishop, 1999]. We follow a standard setup by assuming that we have high-dimensional observations, $y_n \in \mathbb{R}^P$, which we seek to describe by a lower-dimensional latent variable, $x_n \in \mathbb{R}^Q$, where $P \gg Q$. We assume that there exists a mapping $f : \mathbb{R}^Q \to \mathbb{R}^P$, such that the relationship between x_n and y_n is

$$y_n = f(x_n) + \epsilon_n, \tag{1}$$

with $\epsilon_n \in \mathbb{R}^P$ representing random observation noise. For the duration of this paper, we assume it follows the distribution $\epsilon_n \sim N(0, \sigma^2 \mathbb{I}_P)$, and place a prior on $x_n \sim N(0, \mathbb{I}_Q)$. For notational convenience, we will also use $f_n = f(x_n) \in \mathbb{R}^P$.

We assume that we have a large number of high-dimensional data points, $Y \in \mathbb{R}^{N \times P}$ which are associated with values for the latent variables, $X \in \mathbb{R}^{N \times Q}$, and function values, $F \in \mathbb{R}^{N \times P}$. The matrices are constructed such that x_n, f_n, and y_n are the nth row of X, F, and Y respectively. We assume that we have observed Y, and that the goal of our probabilistic model is to infer probability distributions for the possible values of X and F.

We take a Bayesian approach in this paper, ideally aiming to learn a posterior distribution $p(F, X|Y)$. This requires us to specify a prior for both the function and the latent variables. While the prior over the latent space can be used to encode useful information [Damianou et al., 2016], we will use an independent spherical Normal prior for each x_n, so

$$p(x_n) = N(0, \mathbb{I}_Q). \tag{2}$$

Instead of attempting to learn a parametric representation of $f(\cdot)$, we model it using Gaussian Processes (GPs). GPs are a non-parametric Bayesian technique in which each finite set of function values is assumed to follow a Multivariate Normal Distribution [Rasmussen and Williams, 2006]. We apply an independent zero-mean GP prior over each of the output dimensions of our high-dimensional observation, *i.e.*, we take $f_p \in \mathbb{R}^N$ to be the pth column of F and use the prior

$$p(f_p|X) = N(0, K_{XX}). \tag{3}$$

The covariance matrix, K_{XX} is given by a kernel function, $[K_{XX}]_{ij} = k(x_i, x_j)$, which controls properties of the function and can be fit to data using the marginal likelihood [Rasmussen and Williams, 2006], which, following Titsias and Lawrence [2010], we assume to be an Automatic Relevance Determination (ARD) kernel for this paper.

With these priors, we can factorize the model as

$$p(Y, F, X) = \prod_{n=1}^{N} p(y_n|f_n)p(x_n) \prod_{p=1}^{P} p(f_p|X). \tag{4}$$

Ideally, a GPLVM is the posterior distribution for both the latent variables and the functions, *i.e.* $p(X, F|Y)$. Unfortunately, this is not possible as it would require $p(Y) = \int p(Y|F)p(F|X)p(X)dXdF$, which requires an intractable integral over the inverse of the kernel matrix K_{XX}^{-1}. Furthermore, as evaluating the likelihood of any new combination of latent variables would require computing the same inverse matrix K_{XX}^{-1}, sampling-based MCMC approaches would be prohibitively computationally expensive. When GPLVMs were first introduced by Lawrence [2005], the uncertainty of the position of the inputs x_i was ignored, resulting in a MAP estimate for the latent space.

In order to approximate the uncertainty in the latent space, we need to use a Variational Inference (VI) approach. VI is a method in machine learning and statistics that approximates the posterior by minimizing the KL divergence between an approximate posterior, $q(\cdot)$, and the true posterior [Blei et al., 2017]. This is achieved by maximizing an objective known as the Evidence Lower Bound (ELBO). To apply Variational Inference to GPs, it is necessary to augment the GP with a set of fake data points, known as *inducing points* [Titsias, 2009, Hensman et al., 2013]. These inducing points are drawn from the stochastic process but are not associated with any observation. Mathematically, we augment each of the P independent GP mappings from the latent space to one dimension of the output space, $f_p : \mathbb{R}^Q \to \mathbb{R}$, with M additional points. Each output dimension's M inducing points have inputs in the latent space $Z_p \in \mathbb{R}^{M\times Q}$ and random output variables $u_p \in \mathbb{R}^M$. For practical reasons of saving compute, each of the Z_p inputs is assumed to have a single value, so $Z = Z_1 = \cdots = Z_P$. For notational convenience, we also introduce $U \in \mathbb{R}^{P\times M}$, which is the matrix formed by stacking the set of the output vectors, u_p.

As u_p is drawn from the same GP prior as f_p, the augmented prior for each output dimension is

$$(u_{p1}, \ldots, u_{pM}, f_{p1}, \ldots, f_{pN})^\top \sim \mathcal{N}\big(0, K_{(XZ)(XZ)}\big)\,.$$

where

$$K_{(XZ)(XZ)} = \begin{pmatrix} K_{ZZ} & K_{ZX} \\ K_{XZ} & K_{XX} \end{pmatrix}. \tag{5}$$

Furthermore, it is possible to write down each of the P conditional distributions for function values for each observed dimension, $f_p \in \mathbb{R}^N$, given a particular value of the inducing points. This distribution is identical to the distribution of predictions in the standard GP, so

$$p(f_p|u_p, X) = N(\mu_p^f, \Sigma_p^f) \tag{6}$$

where the values for the mean and covariance matrix are given by

$$\mu_p^f = K_{XZ}K_{ZZ}^{-1}u_p$$
$$\Sigma_p^f = K_{XX} - K_{XZ}K_{ZZ}^{-1}K_{ZX}.$$

This dependence has been exploited in various ways, notably by Hensman et al. [2013] to apply Stochastic Variational Inference (SVI) to GPs.

In the context of GPLVMs, it allows us to write a tractable ELBO which includes an approximation of the uncertainty in the latent space [Titsias and Lawrence, 2010]. In order to derive an ELBO, it is necessary to specify a variational family to approximate the posterior. To the best of our knowledge, all prior work has made a simplifying assumption at this point: they have made a Mean Field (MF) restriction between the latent variable and the GP posterior, so the variational posterior factorizes as

$$q(F, X, U) = \prod_{p=1}^{P} p(f_p \mid X, u_p)q(u_p) \prod_{n=1}^{N} q(x_n). \tag{7}$$

By noting that the contribution to the ELBO from the KL divergence between the variational posterior and prior of $p(f_p|u_p, X)$ is zero, as the two terms are equal [Titsias and Lawrence, 2010, Damianou et al., 2016], the ELBO for MF-GPLVM can be written as

$$\begin{aligned}\mathcal{L}_{MF} = \sum_{n=1}^{N} \Big[&\mathbb{E}_{p(f_n|x_n,U)q(x_n)q(U)}[\log\left(p(y_n \mid f_n, x_n, U)\right)] \\ &- D_{KL}(q(x_n)||p(x_n))\Big] - D_{KL}(q(U)||p(U)).\end{aligned} \tag{8}$$

Typically, a normal distribution is used for all of the variational posterior distributions, so $q(u_p) = N(\mu_p, \Sigma_p)$ and $q(x_n) = N(\mu_n, \Sigma_n)$. The ELBO can either be optimized directly by maximizing this objective over the constants $\{\mu_p, \Sigma_p\}_{p=1}^P$ and $\{\mu_n, \Sigma_n\}_{n=1}^N$, or by using an amortized strategy in which the parameters of the latent variable variational posterior are defined as the output of a function of the observations, *i.e.* $(\mu_n, \Sigma_n) = g(y_n)$ with $g : \mathbb{R}^P \to \mathbb{R}^Q \times \mathbb{R}^{Q\times Q}$, and the function $g(\cdot)$ being parametrized by a neural network which is learned during training [Lalchand et al., 2022].

The key issue with this variational posterior is that every likely value for every latent variable, $q(x_n)$, is combined with every likely value of the inducing points, $q(U)$, when evaluating the expected log likelihood of y_n. This enforces a trade-off between accurately representing the uncertainty of the GP mapping and data points in the latent space. As VI is generally mode-seeking [Minka, 2005], it tends to underestimate the epistemic uncertainty of the data manifold.

3 Structured Stochastic Variational Inference for GPLVMs

Although the MF-GPLVM provides a tractable ELBO, which has been shown to achieve a good approximation of the latent space, the MF assumption is a strong and incorrect one. In essence we are asserting that the latent variable posterior, $q(X)$, should not depend on the mapping from the latent space to the observed space. This seems intuitively incorrect, and the more precise mathematical issue can be seen in the expression for $p(f_p|u_p, X)$ in (6). In that equation, the mean of the distribution of f_p depends on both the precise value of the inducing points, u_p, and the kernel matrix, K_{XZ}. After conditioning on the observed data through the likelihood, $p(y_n|f_n)$, the posterior distribution for U and X becomes conditionally dependent.

The main contribution of this paper is to allow the variational posterior to capture this dependence, by relaxing the MF assumption explained in Equation 7. We do this by applying Structured Stochastic Variational Inference [Hoffman and Blei, 2015], a general framework that relaxes the independence assumptions of MF inference. The insight of SSVI is to take the conditional independence structure of the full posterior and mirror it with a conditional independence structure of the variational posterior. Concretely, the variational posterior is assumed to have the factorization

$$q(F, X, U) = \prod_{p=1}^{P} p(f_p|u_p, X)q(u_p) \prod_{n=1}^{N} q(x_n \mid U). \tag{9}$$

The key difference between this and Equation (7) is that now the variational posterior of the latent variables, x_n, depends on the global variables, U. This can be done while maintaining a parametric form of the posterior by keeping the same parametrization of the inducing points as the MF case where $q(u_p) = N(\mu_p, \Sigma_p)$, with μ_p, Σ_p being constant values, but now making the parameters of $q(x_n|U) = N(\mu_n, \Sigma_n)$ functions of the inducing points, *i.e.* $(\mu_n, \Sigma_n) = h(y_n, U)$ with $h : \mathbb{R}^P \times \mathbb{R}^{M\times P} \to \mathbb{R}^Q \times \mathbb{R}^{Q\times Q}$. The goal of the VI procedure is now to learn the constant values μ_p, Σ_p, as well as the mapping $h(\cdot, \cdot)$ which most closely matches the true posterior.

By once again noting that the KL contribution from the GP values, F, is zero due to the shared conditional distribution of F given U and X, the ELBO for the SSVI objective can be written as

$$\mathcal{L}_S = \sum_{n=1}^{N} \mathbb{E}_{q(U)} \left[\hat{\mathcal{L}}_n(U) \right] - D_{\mathrm{KL}}(q(U) \,\|\, p(U)) \tag{10}$$

where

$$\hat{\mathcal{L}}_n(U) = \mathbb{E}_{p(f_n|x_n,U)\, q(x_n|U)}[\log p(y_n \mid f_n, x_n, U)] - D_{\mathrm{KL}}(q(x_n \mid U) \,\|\, p(x_n)) \,. \tag{11}$$

The term within the expectation of the inducing points, $\hat{\mathcal{L}}_n(U)$[1], is called the Local ELBO. As in the Mean Field case, the SSVI ELBO decomposes into a term depending on the global variables $q(U)$, and a sum over terms which depend on the local variables y_n. This decomposition allows the objective to be stochastically optimized through mini-batching over the observations, allowing our method to scale to large datasets. The conditional variational distribution for the latent space, $q(x_n \mid U)$, may depend on U in an arbitrarily complex way, making the expectation $\mathbb{E}_{q(U)}$ intractable. We therefore approximate it using K Monte Carlo samples $U_k \sim q(U)$, yielding

$$\mathcal{L}_S \approx \frac{1}{K} \sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\mathcal{L}}_n(U_k) - D_{\mathrm{KL}}\big(q(U) \,\|\, p(U)\big). \tag{12}$$

This expression can be evaluated either using Monte Carlo techniques to evaluate the expectations, or, for special cases such as the ones we present in our examples, the expressions are analytically tractable. We present further details of our ELBO in Appendix B.

The structured variational family is more flexible than the mean-field alternative, as it includes the mean-field approximation as a special case. This gives the following guarantee on the optimal ELBO.

[1] In order to fully define $\hat{\mathcal{L}}_n$ values for y_n, μ_n, Σ_n, and parameters of the GP are all required in addition to U. The inclusion of U in this notation is to emphasize the dependence of $\hat{\mathcal{L}}_n(\cdot)$ on a particular value of the random variable, which is the key difference between the local ELBO and the standard ELBO.

Table 1: Performance comparison on the Oilflow, qPCR, MNIST, and FMNIST datasets under the standard configuration. Values are reported as mean ± standard deviation over 5 runs. The final negative ELBO is not reported for IWVI-GPLVM because it optimizes a different importance-weighted objective that is not directly comparable to the standard ELBO.

Dataset	GPLVM type	Train RMSE	Test RMSE	Train NLPD	Test NLPD	Final neg. ELBO
Oilflow	MF	0.0528 ± 0.0040	0.1581 ± 0.0088	-15.9416 ± 0.5589	28.5091 ± 4.6286	-5.2531 ± 0.2447
	A-MF	0.0460 ± 0.0045	0.0496 ± 0.0053	-14.8375 ± 0.5922	-14.3680 ± 0.9450	-4.8859 ± 0.2265
	IWVI-GPLVM	$\mathbf{0.0343 \pm 0.0025}$	$\mathbf{0.0375 \pm 0.0023}$	-17.8099 ± 0.7425	-17.4136 ± 0.4695	–
	A-S *(Ours)*	0.0418 ± 0.0032	0.0479 ± 0.0028	$\mathbf{-19.8808 \pm 0.7041}$	$\mathbf{-18.8218 \pm 0.6210}$	$\mathbf{-6.4527 \pm 0.0449}$
qPCR	MF	0.4993 ± 0.0141	0.6018 ± 0.0405	36.1396 ± 1.0722	50.4384 ± 5.1887	46.9602 ± 0.2065
	A-MF	0.4795 ± 0.0109	0.5190 ± 0.0061	35.3542 ± 0.6533	38.7142 ± 0.8894	47.4945 ± 0.2714
	IWVI-GPLVM	0.4354 ± 0.0053	0.4731 ± 0.0175	32.4221 ± 0.4552	35.5986 ± 1.3582	–
	A-S *(Ours)*	$\mathbf{0.4282 \pm 0.0047}$	$\mathbf{0.4701 \pm 0.0132}$	$\mathbf{29.8229 \pm 0.4723}$	$\mathbf{33.5050 \pm 1.2980}$	$\mathbf{44.9380 \pm 0.3163}$
MNIST	A-MF	0.1550 ± 0.0006	0.1568 ± 0.0003	-342.5288 ± 2.8080	-333.9984 ± 1.5725	-294.0345 ± 2.8327
	IWVI-GPLVM	0.1540 ± 0.0010	0.1581 ± 0.0011	-348.0312 ± 5.5166	-328.9493 ± 5.6709	–
	A-S *(Ours)*	$\mathbf{0.1515 \pm 0.0001}$	$\mathbf{0.1537 \pm 0.0007}$	$\mathbf{-359.8706 \pm 2.1250}$	$\mathbf{-348.4373 \pm 3.0311}$	$\mathbf{-306.3421 \pm 1.0179}$
FMNIST	A-MF	0.1339 ± 0.0004	0.1342 ± 0.0006	-456.6282 ± 3.1032	-454.4614 ± 4.2266	-409.8799 ± 2.5766
	IWVI-GPLVM	0.1340 ± 0.0009	0.1380 ± 0.0011	-458.0779 ± 5.4105	-436.8031 ± 6.0015	–
	A-S *(Ours)*	$\mathbf{0.1294 \pm 0.0003}$	$\mathbf{0.1299 \pm 0.0004}$	$\mathbf{-484.1356 \pm 1.8203}$	$\mathbf{-481.3409 \pm 2.7219}$	$\mathbf{-425.9231 \pm 1.0856}$

Remark 3.1. Let $\mathcal{Q}_{\mathrm{MF}}$ denote the set of variational posteriors of the form of Equation (7), and $\mathcal{Q}_{\mathrm{S}}$ those of the form of Equation (9), where both use the same parametrisation $q(u_p) = \mathcal{N}(\mu_p, \Sigma_p)$ and the same class of encoder networks. Then $\mathcal{Q}_{\mathrm{MF}} \subseteq \mathcal{Q}_{\mathrm{S}}$, and for any fixed setting of the GP hyperparameters,

$$\sup_{q \in \mathcal{Q}_{\mathrm{S}}} \mathcal{L}_{\mathrm{S}}(q) \geq \sup_{q \in \mathcal{Q}_{\mathrm{MF}}} \mathcal{L}_{\mathrm{MF}}(q). \tag{13}$$

This guarantees that the optimal ELBO of the A-S-GPLVM cannot be lower than that of the A-MF-GPLVM. However, this guarantee applies to the global optimum and does not necessarily hold for the solutions found by numerical optimization. A formal proof is provided in Appendix F.

3.1 Local and Global Variables for Stochastic Variational Inference

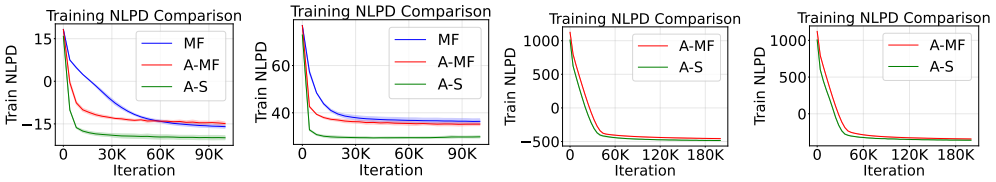
One important choice of our variational family which we have not discussed so far is the choice to specify our variational posterior as $q(X|U)q(U)$, rather than $q(U|X)q(X)$. The structure of the exact posterior can be written equally well in both forms: $p(X, U, F|Y) = p(F|U, X)p(U|X)p(X) = p(F|U, X)p(X|U)p(U)$. Instead, the reason to prefer $q(X|U)q(U)$ over the alternative is computational.

We are interested in developing a method which can scale to a larger number of observations than can reasonably be expected to all be fit into memory at once. To do this, we need to use Stochastic Variational Inference [SVI; Hoffman et al., 2013]. SVI relies on models whose variables can be split into two sets: local variables, which are associated with individual data points, and global variables, which are associated with many or all data points. The key insight of SVI is to realize that it is possible to get an unbiased estimate of the gradient for the parameters of the global variables by using only a subset of the data. Structured Stochastic Variational Inference [Hoffman and Blei, 2015] relies on this same insight, but with additional dependencies in the variational posterior.

In GPs, U are the global variables shared across the full dataset [Hensman et al., 2013], and in GPLVMs the local variables are latent variables, x_n. This means that the ELBO decomposes into a sum of per-data point contributions, which in turn allows us to estimate $\mathcal{L}_S$ using only an index set $\mathcal{I}$, where $|\mathcal{I}| \ll N$, as

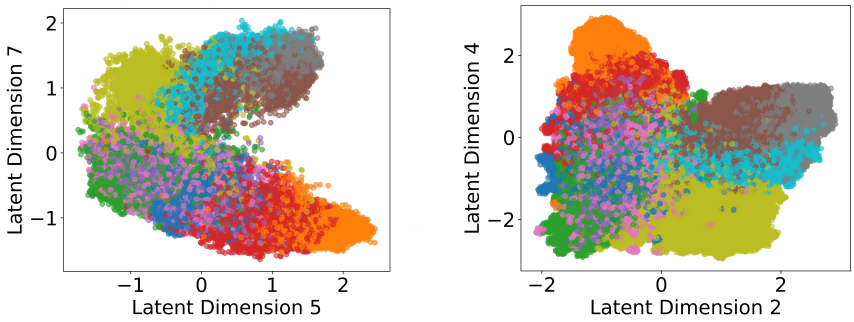
$$\tilde{\mathcal{L}}_S = \frac{N}{K|\mathcal{I}|} \sum_{k=1}^{K} \sum_{n \in \mathcal{I}} \hat{\mathcal{L}}_n(U_k) - D_{\mathrm{KL}}\big(q(U) \,\|\, p(U)\big). \tag{14}$$

This mini-batched version of the objective is the one we use in our experiments. Importantly, our structured variational family preserves the same mini-batching capability as the MF-GPLVM, while providing a more flexible variational posterior.



(a) Oilflow (b) qPCR (c) FMNIST (d) MNIST

Figure 1: The training NLPD progression during training across all datasets. The A-S-GPLVM and A-MF-GPLVM converge at similar rates, but the A-S-GPLVM converges to a better optimum.



(a) A-S-GPLVM. (b) A-MF-GPLVM.

Figure 2: The distribution of training points in latent space for FMNIST for the A-S-GPLVM and A-MF-GPLVM. The A-S-GPLVM latent space has learned substantially more structure than the A-MF-GPLVM. In fact, the A-S-GPLVM appears to have collapsed the latent predictions into a semantically meaningful curve within the latent space, capturing the relative similarity of the classes with sandals, sneakers, and ankle boots at one end and trousers and dresses at the other. The A-MF-GPLVM seems to try to capture a similar shape, but achieves a much less clear separation between the groupings.

4 Related work

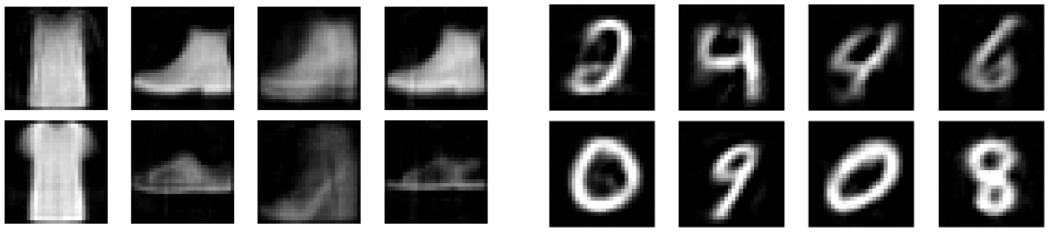
A closely related work is the GP state-space model work of Ialongo et al. [2019], who remove the mean-field factorization between latent state trajectories and the GP transition posterior in recurrent GP models. Their coupled approximation improves calibration while preserving similar time and space complexity as the mean-field alternative. This work approaches the same conceptual challenge: that decoupling latent variables from the GP posterior can be too restrictive, but Ialongo et al. [2019] study sequential GP state-space models, whereas we consider the latent variable setting.

A complementary approach to improving variational inference in GP models is to tighten the variational objective through importance weighting. Salimbeni et al. [2019] develop this approach for deep Gaussian processes, providing a mechanism to trade additional computation for a tighter lower bound. Adapted to the sparse GPLVM setting, importance-weighted variational inference offers an alternative to our explicit coupling of the latent-variable and inducing-variable posteriors. An amortized IWVI-GPLVM therefore provides a relevant baseline for comparing these two approaches to improving variational inference while retaining the same underlying sparse GPLVM model.

Besides the Bayesian GPLVM [Titsias and Lawrence, 2010] and Generalised GPLVM [Lalchand et al., 2022], discussed in the background section, there have been several other lines of work within the GPLVM field. One such line of work is the Random Feature Latent Variable Model (RFLVM), which replaces the sparse variational GP approximation with Random Fourier Feature (RFF) approximation of the mapping from latent variables to observations [Gundersen et al., 2021]. In their approach Gundersen et al. [2021] used a point MAP estimate for the latent variables X, which was conditioned

Table 2: Predictive uncertainty evaluation on the Oilflow, qPCR, MNIST, and FMNIST datasets. Values are reported as mean ± standard deviation over 5 runs. The best result for each dataset and metric is shown in bold.

Dataset	GPLVM type	Train CRPS ↓	Test CRPS ↓	Train coverage	Test coverage
Oilflow	MF	0.0423 ± 0.0031	0.0832 ± 0.0114	0.9867 ± 0.0026	$\mathbf{0.9385 \pm 0.0191}$
	A-MF	0.0400 ± 0.0020	0.0416 ± 0.0030	0.9912 ± 0.0045	0.9862 ± 0.0074
	IWVI-GPLVM	$\mathbf{0.0238 \pm 0.0020}$	$\mathbf{0.0251 \pm 0.0015}$	0.9979 ± 0.0006	0.9954 ± 0.0026
	A-S *(Ours)*	0.0247 ± 0.0021	0.0274 ± 0.0021	$\mathbf{0.9863 \pm 0.0019}$	0.9705 ± 0.0097
qPCR	MF	0.3454 ± 0.0061	0.4077 ± 0.0165	$\mathbf{0.9587 \pm 0.0024}$	0.9383 ± 0.0116
	A-MF	0.3364 ± 0.0031	0.3573 ± 0.0089	0.9650 ± 0.0020	$\mathbf{0.9477 \pm 0.0076}$
	IWVI-GPLVM	0.2459 ± 0.0030	0.2613 ± 0.0049	0.9763 ± 0.0005	0.9689 ± 0.0047
	A-S *(Ours)*	$\mathbf{0.2381 \pm 0.0015}$	$\mathbf{0.2558 \pm 0.0079}$	0.9686 ± 0.0007	0.9589 ± 0.0056
MNIST	A-MF	0.1007 ± 0.0001	0.1016 ± 0.0001	0.9269 ± 0.0001	0.9247 ± 0.0004
	IWVI-GPLVM	0.1008 ± 0.0009	0.1024 ± 0.0010	$\mathbf{0.9314 \pm 0.0008}$	$\mathbf{0.9272 \pm 0.0008}$
	A-S *(Ours)*	$\mathbf{0.0978 \pm 0.0001}$	$\mathbf{0.0989 \pm 0.0002}$	0.9270 ± 0.0002	0.9248 ± 0.0004
FMNIST	A-MF	0.0889 ± 0.0001	0.0890 ± 0.0002	0.9330 ± 0.0001	0.9324 ± 0.0004
	IWVI-GPLVM	0.0898 ± 0.0008	0.0916 ± 0.0005	$\mathbf{0.9371 \pm 0.0008}$	0.9321 ± 0.0027
	A-S *(Ours)*	$\mathbf{0.0859 \pm 0.0003}$	$\mathbf{0.0861 \pm 0.0001}$	0.9337 ± 0.0002	$\mathbf{0.9333 \pm 0.0007}$



(a) FMNIST (b) MNIST

Figure 3: Sample reconstructions using the A-S-GPLVM. Each image shows the posterior predictive mean of the reconstructed observation. The model is able to produce a reasonable quality of reconstruction, although the difference in quality to the A-MF-GPLVM (see Figure 14, Appendix K) is not immediately obvious.

on the RFF approximation. Later work in this line by Li et al. [2024] and Yang et al. [2025] applied a VI approach to approximating the RFLVM posterior. A different approach by Xu et al. [2025] used the same sparse GP formulation of the mapping from the latent to observed space, but used Annealed Importance Sampling (AIS) to tighten the lower bound of the GPLVM. As far as we are aware, no prior work has proposed a method for making the latent space conditional on the GP mapping as we do here.

5 Experiments

We compare our extension of the GPLVM framework, the Amortized-Structured-GPLVM (A-S-GPLVM), which applies SSVI with amortization of the latent variables, to variants of the mean field approach. On the smaller scale datasets we compare against two inference strategies: the MF-GPLVM with SVI, which directly approximates the MF-GPLVM variational posterior by optimizing the parameters of the latent variable distribution $\{\mu_n, \Sigma_n\}_{n=1}^N$ and the inducing point variational distribution $\{\mu_p, \Sigma_p\}_{p=1}^P$, and the Amortized-MF-GPLVM (A-MF-GPLVM), where the latent variables are also amortized. We base our implementation of these on Lalchand et al. [2022]. The initialization is the same for all models, and is reported in Appendix D. A Bayesian diagram for the A-MF-GPLVM and A-S-GPLVM is given in Figure 5, in Appendix D. For larger datasets, we do not include the non-amortized MF-GPLVM.

We additionally compare against an amortized importance-weighted variational GPLVM (IWVI-GPLVM), based on Salimbeni et al. [2019] and discussed in Section 4. This baseline preserves the same underlying sparse GPLVM model but improves the variational objective through importance

weighting rather than introducing the explicit posterior dependence used by A-S-GPLVM. It therefore allows us to assess whether the benefits of structured inference persist when compared with an alternative approach to tightening the variational bound. We use the same model and optimization hyperparameters as for the other amortized models, with the additional IWVI-specific parameter, the number of importance samples, set to 5.

Data setup. The experiments are conducted on four standard datasets: two small-scale ones and two larger datasets. Our small-scale datasets are the Oilflow dataset, consisting of 1000 twelve-dimensional data points belonging to three classes corresponding to different phases of oil flow in a pipeline [Bishop and James, 1993], and the qPCR dataset, consisting of 48-dimensional single-cell data obtained from mice with 10 classes [Guo et al., 2010]. For our larger-scale experiments, we use the MNIST [LeCun et al., 1998] and Fashion MNIST [Xiao et al., 2017], which we subsample to contain 1500 examples from each of the 10 classes. Code to recreate the Oilflow experiment using the A-S-GPLVM model is provided in the accompanying repository (see https://github.com/jamesacodgers/conditional_overlaping_gaussian_processes).

Evaluation Metrics. To evaluate the quality of our posterior approximation, we use five metrics. First, we use the negative ELBO from our model, which should give a lower value if our posterior is able to achieve a value closer to the true posterior. As the ELBO is also influenced by hyperparameters, which can differ between models, we also use additional metrics to evaluate the reconstruction quality of our learned data manifold. We define our variational reconstruction distribution with

$$q(y_n) = \int p(y_n|f_n)q(f_n, x_n, U)df_n dx_n dU, \tag{15}$$

for each of our data points. We report two values calculated from this: the Negative Log Predictive Density (NLPD) of $q(y_n)$ evaluated at the true data point, and the Root Mean Squared Error (RMSE) between the true observation y_n and the mean of $q(y_n)$.

To evaluate predictive uncertainty more directly, we additionally report the Continuous Ranked Probability Score (CRPS) and empirical coverage of marginal 95% posterior predictive intervals. CRPS evaluates the full predictive distribution and rewards both calibration and sharpness, with lower values indicating better predictions. Empirical coverage measures the fraction of observed values contained in the corresponding 95% predictive intervals; values closer to 0.95 indicate better calibration.

We also split our data into training and test sets. The training set is the data used to train the GP hyperparameters, the inducing point parameters, and either the values of the latent variables or the amortization network for the latent variables. The test data were not seen during training, and we use this data to evaluate if our model can generalize well to unseen observations. For the smaller datasets, we use a 90/10 train–test split, while for the MNIST and FMNIST datasets, we used 1000 training points and 500 test points.

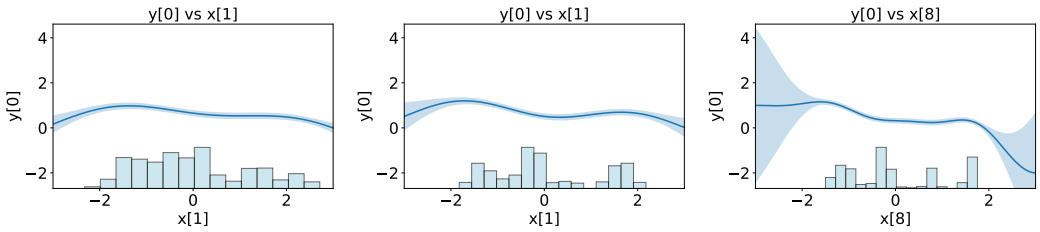
We additionally compare A-MF-GPLVM, IWVI-GPLVM, and A-S-GPLVM on the full MNIST dataset, using the same model and optimization configurations as in the subsampled experiments, described in Appendix D. Consistent with the subsampled experiments, the A-S-GPLVM outperforms both baselines in terms of train and test RMSE, NLPD, and CRPS, demonstrating that these improvements persist on the full dataset. Further details and results are provided in Appendix J.

Data pre-processing. For all datasets, we apply mean-centering as a preprocessing step. Specifically, each observed dimension is centered by subtracting its empirical mean, computed over the training set.

5.1 Results.

Table 1 reports the train and test RMSE and NLPD, as well as the final negative ELBO where applicable. Compared with the A-MF-GPLVM, the A-S-GPLVM achieves lower RMSE and NLPD on all four datasets. Two-sided Welch's t-tests show that these improvements are statistically significant, with the exception of the train and test RMSE improvements on Oilflow.

Compared with IWVI-GPLVM, the A-S-GPLVM consistently achieves significantly lower train and test NLPD across all four datasets. The improvements in RMSE are more dependent on the dataset.



(a) MF-GPLVM (b) A-MF-GPLVM (c) A-S-GPLVM

Figure 4: Latent slice through the most informative latent dimension and a histogram of the learned posterior distribution of the training data along the horizontal axis. Curves show predictive means; shaded areas denote predictive uncertainty for each output. The epistemic uncertainty at points far from the training data clearly increases more rapidly with the A-S-GPLVM than with the MF models.

On MNIST and FMNIST, the A-S-GPLVM achieves significantly lower train and test RMSE, whereas on Oilflow, IWVI-GPLVM performs better. On qPCR, the RMSE differences are not statistically significant. These results suggest that the main advantage of structured inference lies in improving the predictive distribution rather than consistently reducing point prediction errors.

The CRPS results in Table 2 provide further evidence of this improvement. Compared with the A-MF-GPLVM, the A-S-GPLVM achieves significantly lower train and test CRPS across all four datasets. The A-S-GPLVM also significantly outperforms IWVI-GPLVM on MNIST and FMNIST, while the differences on Oilflow and qPCR are not statistically significant. The empirical coverage results are more mixed, with no method consistently achieving coverage closest to the nominal 95% level. Overall, these results support the benefit of structured variational inference for predictive distribution quality, with consistent improvements in NLPD and CRPS despite the more mixed coverage results.

The poor performance of the non-amortized MF-GPLVM on test data is because $q(x_n)$ needs to be numerically optimized for each test point, which is a non-convex optimization and can converge to poor local optima. Figure 1 shows the convergence of the training NLPD across the four datasets. The A-S-GPLVM seems to converge at very similar rates to the A-MF-GPLVM, but to a better final value. In Appendix K, Figures 13-16 show similar plots for the other metrics we tested. The strong performance of the A-S-GPLVM compared to the MF variants suggests that our approach is using the greater flexibility of the variational family to achieve a better approximation of the exact posterior, and hence better performance.

We show two-dimensional projections of the latent space onto the two most significant dimensions of the GP, as in Titsias and Lawrence [2010]. For A-S-GPLVM, latent samples are obtained by first sampling the inducing variables U from their variational posterior $q(U)$, finding $q(X \mid U)$, and sampling from that latent distribution. For the MF-GPLVM and the A-MF-GPLVM, latent samples are generated by sampling $q(X)$ directly.

For the FMNIST dataset in Figure 2, the A-S-GPLVM produces tighter clusters than the A-MF-GPLVM. The A-S-GPLVM latent space has taken on a crescent shape, with one end containing footwear (sandal, sneaker, ankle boot), the other end containing clothing items (trouser, dress), and a clear gap containing no data between them. The A-MF-GPLVM also seems to capture a similar shape, but the separation between the two prongs is much less pronounced. This shows that, for this dataset, our method more effectively captures meaningful semantic information from the images. Additional latent space plots for the Oilflow, qPCR, and MNIST datasets are provided in Appendix K.

Figure 3 shows the mean of $q(y_n)$ from the A-S-GPLVM for eight data points, demonstrating reasonable reconstruction quality. We provide similar plots for the MF-GPLVM reconstructions in Figure 14, Appendix K, which are visually of similar quality. While the improvements of the A-S-GPLVM reconstructions over the A-MF-GPLVM are mostly statistically significant, as shown in Table 3, the visual similarity suggests that further work improving other aspects of the GPLVM model would be needed to make it a truly effective generative model.

While the reconstructions are an important measure of latent space estimation quality, a specific advantage of GPs is their ability to increase epistemic uncertainty away from the data. We visualize

this in Figure 4, which shows a slice of the prediction of the GP from the latent to one of the observed dimensions along the single most informative latent coordinate, *i.e.*, the coordinate with the shortest GP length scale, with other coordinates fixed at zero. Figure 4 shows that the MF-GPLVMs have a slow increase in the epistemic uncertainty of the GP mapping between the region where there is a significant probability of observing a latent variable, roughly corresponding to $[-2, 2]$ on the x axis, and the region where there is a low probability of observing a latent variable. In contrast, the uncertainty of the GP mapping learned by the A-S-GPLVM increases much more rapidly away from the high-probability inputs. We show similar plots for four output dimensions of the Oilflow dataset in Figure 11, Appendix K. We do not expect the A-S-GPLVM to achieve a perfect estimation of the exact epistemic uncertainty of the posterior due to the approximation on the GP mapping [Wenger et al., 2022, 2024] and other restrictions of the variational posterior. However, the A-S-GPLVM shows a clear improvement in the epistemic uncertainty estimation and, as all of the methods presented here share the same GP approximation, these results indicate that the coupling of the variational posteriors for the GP and the latent space is directly responsible for this difference in epistemic uncertainty estimation.

6 Limitations

Computational overhead. Compared with the amortized mean-field GPLVM, the A-S-GPLVM introduces additional computational cost because the structured objective requires Monte Carlo samples $U_k \sim q(U)$, as in Equation 12. For each sampled global state, the encoder must evaluate a conditional local posterior $q(x_n \mid U_k)$ and the expected likelihood must be evaluated. The K sampled inducing-variable states can be processed in parallel, so the dominant practical cost is often increased memory usage rather than a fully sequential increase in wall-clock time. We empirically quantify this in Appendix H, where we show that the A-S-GPLVM suffers a slight wall-clock slowdown, but a significant increase in memory usage. Our additional ablation in Figure 9 shows that increasing K generally improves predictive performance, particularly NLPD, at the cost of higher GPU memory consumption.

Restrictive variational family. A second limitation is that, while our variational posterior is more expressive than the MF-GPLVM alternative, we still make strong assumptions on the form of our variational posterior. For example, we assume $q(x_n \mid U)$ is a normal distribution with a diagonal covariance matrix and $q(U) = \prod_{p=1}^{P} q(u_p)$.

Under the mean-field approximation, Damianou et al. [2016] showed that the optimal variational distribution over the inducing variables is a factorized Gaussian, $q^*(U) = \prod_{p=1}^{P} \mathcal{N}(u_p \mid \mu_p, \Sigma_p)$. However, this result relies on the independence between the latent variables and the inducing variables. Since our structured variational family introduces the dependence $q(X \mid U)$, the same optimality argument no longer applies. We therefore expect that allowing more flexible distributions for $q(U)$ could further improve the variational approximation.

7 Conclusion

This work introduces SSVI for GPLVMs by coupling the variational posterior of the latent variables to the inducing variables. Across Oilflow, qPCR, MNIST, and FMNIST, the A-S-GPLVM achieves statistically significant improvements in train and test CRPS over the A-MF-GPLVM. Its RMSE and NLPD improvements are also statistically significant, except for RMSE on Oilflow.

Compared with IWVI-GPLVM, the A-S-GPLVM achieves significantly lower train and test NLPD on all four datasets, and significantly lower RMSE and CRPS on MNIST and FMNIST. Results on the smaller datasets are more mixed, and empirical 95% coverage does not show uniform improvement. Together, these findings provide quantitative evidence that structured posterior dependence improves the quality of the learned predictive distribution. The qualitative uncertainty plots additionally show a more rapid increase in uncertainty away from the observed data (Figure 4), consistent with our motivation to better capture epistemic uncertainty in the learned data manifold.

Future work will explore how to use the improved uncertainty of the data manifold. For example, as a method of improving the sharing of information and uncertainty estimates in multiobjective Bayesian optimization, as in Sedgwick et al. [2025].

8 Acknowledgments

The authors acknowledge financial support from Eli Lilly and Co. and the Engineering and Physical Sciences Research Council of the UK via Prosperity Partnership (grant number EP/T005556/1). RM holds concurrent appointments as a Professor at Imperial College London and as an Amazon Scholar. This paper describes work performed at Imperial College London and is not associated with Amazon. VF was supported by the Branco Weiss Fellowship.

References

Sumon Ahmed, Magnus Rattray, and Alexis Boukouvalas. Grandprix: scaling up the bayesian gplvm for single-cell data. *Bioinformatics*, 35(1):47–54, January 2019. doi: 10.1093/bioinformatics/bty533.

Christopher M Bishop and Gwilym D James. Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 327(2-3):580–593, 1993. doi: 10.1016/0168-9002(93)90728-Z.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Andreas C. Damianou, Michalis K. Titsias, and Neil D. Lawrence. Variational inference for latent variables and uncertain inputs in gaussian processes. *Journal of Machine Learning Research*, 17 (42):1–62, 2016. URL https://www.jmlr.org/papers/v17/damianou16a.html.

Anish Dhir, Samuel Power, and Mark Van Der Wilk. Bivariate causal discovery using Bayesian model selection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10710–10735. PMLR, 2024. URL https://proceedings.mlr.press/v235/dhir24a.html.

Anish Dhir, Ruby Sedgwick, Avinash Kori, Ben Glocker, and Mark Van Der Wilk. Continuous Bayesian model selection for multivariate causal discovery. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 13486–13512. PMLR, 2025. URL https://proceedings.mlr.press/v267/dhir25a.html.

Gregory Gundersen, Michael Zhang, and Barbara Engelhardt. Latent variable modeling with random features. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1333–1341. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/gundersen21a.html.

Guoji Guo, Mikael Huss, Guang Q. Tong, Chong Wang, Liang L. Sun, Neil D. Clarke, and Paul Robson. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell*, 18(4):675–685, 2010. doi: 10.1016/j.devcel.2010.02.012.

James Hensman, Nicolo Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, volume R11 of *Proceedings of Machine Learning Research*, pages 432–440. PMLR, 2013. URL https://proceedings.mlr.press/r11/hensman13a.html.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Matthew D. Hoffman and David M. Blei. Structured stochastic variational inference. In Francis Bach and David Blei, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *Proceedings of Machine Learning Research*, pages 361–369, San Diego, California, USA, 2015. JMLR Workshop and Conference Proceedings. URL https://proceedings.mlr.press/v38/hoffman15.html.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(40):1303–1347, 2013. URL https://www.jmlr.org/papers/v14/hoffman13a.html.

Alessandro Davide Ialongo, Mark Van Der Wilk, James Hensman, and Carl Edward Rasmussen. Overcoming mean-field approximations in recurrent Gaussian process models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2931–2940. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ialongo19a.html.

Noémie Jaquier, Leonel Rozo, Miguel González-Duque, Viacheslav Borovitskiy, and Tamim Asfour. Bringing motion taxonomies to continuous domains via GPLVM on hyperbolic manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21373–21409. PMLR, 2024. URL https://proceedings.mlr.press/v235/jaquier24a.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, Banff, AB, Canada, April 14-16 2014. URL https://arxiv.org/abs/1312.6114. Conference Track Proceedings.

Vidhi Lalchand, Aditya Ravuri, and Neil D. Lawrence. Generalised GPLVM with stochastic variational inference. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7841–7864. PMLR, 2022. URL https://proceedings.mlr.press/v151/lalchand22a.html.

Neil D. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(60):1783–1816, 2005. URL https://www.jmlr.org/papers/v6/lawrence05a.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. ISSN 0018-9219. doi: 10.1109/5.726791.

Ying Li, Zhidi Lin, Feng Yin, and Michael Minyi Zhang. Preventing model collapse in Gaussian process latent variable models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28278–28308. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24as.html.

Thomas Minka. Divergence measures and message passing, December 2005. URL https://tminka.github.io/papers/message-passing/.

James A. C. Odgers, Ruby Sedgwick, Chrysoula Dimitra Kappatou, Ruth Misener, and Sarah Lucie Filippi. Weighted Sum of Gaussian Process Latent Variable Models. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pages 3610–3618. PMLR, 2025. URL https://proceedings.mlr.press/v258/odgers25a.html.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. ISBN 978-0-262-18253-9. URL https://gaussianprocess.org/gpml/.

John E. Reid and Lorenz Wernisch. Pseudotime estimation: deconfounding single cell time series. *Bioinformatics*, 32(19):2973–2980, October 2016. doi: 10.1093/bioinformatics/btw372.

Leonel Rozo, Miguel González-Duque, Noémie Jaquier, and Søren Hauberg. Riemann^2: Learning riemannian submanifolds from riemannian data. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3097–3105. PMLR, 2025. URL https://proceedings.mlr.press/v258/rozo25a.html.

Hugh Salimbeni, Vincent Dutordoir, James Hensman, and Marc Deisenroth. Deep gaussian processes with importance-weighted variational inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5589–5598. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/salimbeni19a.html.

Ruby Sedgwick, John P. Goertz, Molly M. Stevens, Ruth Misener, and Mark van der Wilk. Transfer learning Bayesian optimization for competitor DNA molecule design for use in diagnostic assays. *Biotechnology and Bioengineering*, 122(1):189–210, 2025. ISSN 1097-0290. doi: 10.1002/bit.28854.

Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, September 1999. ISSN 1369-7412. doi: 10.1111/1467-9868.00196.

Michalis Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 567–574. PMLR, April 2009. URL https://proceedings.mlr.press/v5/titsias09a.html.

Michalis Titsias and Neil D. Lawrence. Bayesian gaussian process latent variable model. In Yee Whye Teh and D. Michael Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *Proceedings of Machine Learning Research*, pages 844–851, Chia Laguna Resort, Sardinia, Italy, 2010. JMLR Workshop and Conference Proceedings. URL https://proceedings.mlr.press/v9/titsias10a.html.

Jonathan Wenger, Geoff Pleiss, Marvin Pförtner, Philipp Hennig, and John Cunningham. Posterior and computational uncertainty in gaussian processes. In *Advances in Neural Information Processing Systems*, volume 35, pages 10876–10890. Curran Associates, Inc., 2022. doi: 10.52202/068431-0790. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4683beb6bab325650db13afd05d1a14a-Paper-Conference.pdf.

Jonathan Wenger, Kaiwen Wu, Philipp Hennig, Jacob R. Gardner, Geoff Pleiss, and John P. Cunningham. Computation-aware gaussian processes: Model selection and linear-time inference. In *Advances in Neural Information Processing Systems*, volume 37, pages 31316–31349. Curran Associates, Inc., 2024. doi: 10.52202/079017-0984. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/379ea6eb0faad176b570c2e26d58ff2b-Paper-Conference.pdf.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL https://arxiv.org/abs/1708.07747.

Jian Xu, Shian Du, Junmei Yang, Qianli Ma, Delu Zeng, and John Paisley. Variational learning of gaussian process latent variable models through stochastic gradient annealed importance sampling. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 4663–4680. PMLR, 2025. URL https://proceedings.mlr.press/v286/xu25a.html.

Zi Yang, Ying Li, Zhidi Lin, Michael Minyi Zhang, and Pablo Olmos. Multi-view oriented gplvm: Expressiveness and efficiency. In *Advances in Neural Information Processing Systems*, volume 38, pages 65283–65327. Curran Associates, Inc., 2025. doi: 10.52202/085713-2190. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5e50b663324972bb8cc7b5c06a059438-Paper-Conference.pdf.

A Graphical Models

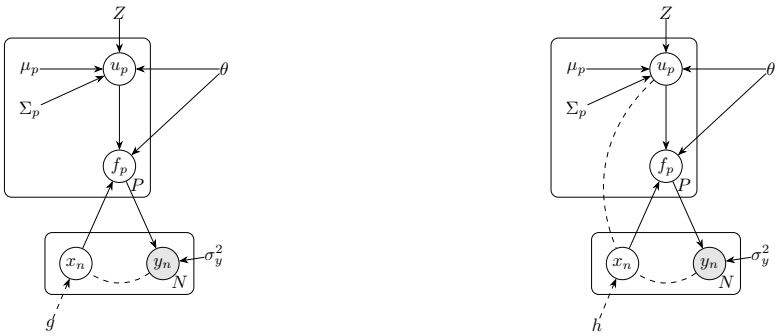


Figure 5: Graphical models comparing the GPLVM variants. Left: Amortized mean-field (A-MF-GPLVM), where the variational posteriors of the latent variables, x_n, are independent of the inducing point values, u_p. Right: Amortized Structured (A-S-GPLVM, *ours*), which conditions the value of the latent variable variational posterior on the global inducing variables, capturing dependencies between local and global variational factors. The dashed line indicates a variable used for amortization.

B SSVI ELBO derivation details

As stated in Section 3, we approximate the ELBO using K Monte Carlo samples $U_k \sim q(U)$, yielding

$$\mathcal{L}_S \approx \frac{1}{K} \sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\mathcal{L}}_n(U_k) - D_{\mathrm{KL}}\big(q(U) \,\|\, p(U)\big). \tag{16}$$

To calculate the local ELBO, it is helpful to explicitly split the calculation across the P output dimensions. Recalling the conditional relationships from Equation (6), we can see that for one latent variable x_n and inducing point sample for a single output $u_{pk} \in \mathbb{R}^M$, then the function output for a single observation and output, $f_{npk} \in \mathbb{R}$, the conditional distribution is given by

$$p(f_{npk} \mid u_{pk}, x_n) = \mathcal{N}(f_{npk} \mid \mu^f_{npk}, \sigma^f_{npk}) \tag{17}$$

where $\mu^f_{npk}, \sigma^f_{npk} \in \mathbb{R}$ are given by

$$\mu^f_{npk} = k_\theta(x_n, Z)^\top K_{ZZ}^{-1} u_{pk}, \tag{18}$$

$$\sigma^f_{npk} = k_\theta(x_n, x_n) - k_\theta(x_n, Z) K_{ZZ}^{-1} k_\theta(x_n, Z)^\top. \tag{19}$$

Note that σ^f_{npk} is the variance of the distribution, but we have not included the squared term to keep the notation cleaner. Additionally, $p(y_n \mid f_n, x_n, U_k) = p(y_n | f_n)$ so we can further simplify our notation here by dropping x_n and U_k from the notation of the likelihood.

The local ELBO can then be written as a sum of the P expected log-likelihood terms

$$\mathcal{F}_{npk} = \mathbb{E}_{p(f_{npk}|x_n,U_k)\, q(x_n|U_k)}[\log p(y_{np} \mid f_{npk})] \tag{20}$$

and a KL divergence to give

$$\hat{\mathcal{L}}_n(U_k) = \sum_{p=1}^{P} \mathcal{F}_{npk} - D_{\mathrm{KL}}(q(x_n \mid U_k) \,\|\, p(x_n)). \tag{21}$$

If the variational posterior is constrained to a parametric form, as it is in our case, the KL term $D_{KL}(q(x_n \mid U) \| p(x_n))$ is available in closed form.

Proceeding in a very similar way to the MF-GPLVM, we can take the expectation of the log-likelihood over $p(f_{npk}|x_n, U_k)$, giving

$$\mathcal{F}_{npk} = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\,\mathbb{E}_{q(x_n|U_k)}\Big[(y_{np} - \mu^f_{npk})^2 + \sigma^f_{npk}\Big]. \tag{22}$$

We then define the so-called ψ-statistics, which are the required kernel expectations:

$$\begin{aligned}\psi_{0n} &= \mathbb{E}_{q(x_n)}[k_\theta(x_n, x_n)] \in \mathbb{R}, \\ \psi_{1n} &= \mathbb{E}_{q(x_n)}[k_\theta(x_n, Z)] \in \mathbb{R}^{1\times M}, \\ \psi_{2n} &= \mathbb{E}_{q(x_n)}[k_\theta(x_n, Z)^\top k_\theta(x_n, Z)] \in \mathbb{R}^{M\times M}.\end{aligned} \tag{23}$$

We can separate the two terms in the latent variable expectation in Equation (22). The expectation of σ^f_{npk} with respect to $q(x_n|U_k)$ is

$$\hat{\sigma}^f_{npk} = \psi_{0n} - \operatorname{tr}\big(K_{ZZ}^{-1}\psi_{2n}\big). \tag{24}$$

Expanding the quadratic expectation term gives

$$\mathbb{E}_{q(x_n|U_k)}\big[(y_{np} - \mu^f_{npk})^2\big] = y^2_{np} - 2y_{np}\,\mathbb{E}_{q(x_n|U_k)}[\mu^f_{npk}] + \mathbb{E}_{q(x_n|U_k)}\big[(\mu^f_{npk})^2\big]. \tag{25}$$

For the ARD kernel we use, the expectations have closed-form solutions [Titsias and Lawrence, 2010].

Substituting the kernel expectations

$$\begin{aligned}\mathbb{E}_{q(x_n|U_k)}\big[\mu^f_{npk}\big] &= \psi_{1n}K_{ZZ}^{-1}u_{pk}, \\ \mathbb{E}_{q(x_n|U_k)}\big[(\mu^f_{npk})^2\big] &= u_{pk}^\top K_{ZZ}^{-1}\psi_{2n}K_{ZZ}^{-1}u_{pk}.\end{aligned} \tag{26}$$

For notational convenience, we use

$$\hat{\mu}^f_{npk} = \mathbb{E}_{q(x_n|U_k)}[\mu^f_{npk}] \tag{27}$$

The expectations above were done for the pth output dimension of y_n. As a result, for all dimensions of y_n we obtain:

$$\sum_{p=1}^{P}\mathcal{F}_{npk} = -\frac{P}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{p=1}^{P}\Big[\hat{\sigma}^f_{npk} + y^2_{np} - 2y_{np}\,\hat{\mu}^f_{npk} + u_{pk}^\top K_{ZZ}^{-1}\psi_{2n}K_{ZZ}^{-1}u_{pk}\Big]. \tag{28}$$

C Amortized Encoder

As noted by Hoffman and Blei [2015], the function $h(\cdot,\cdot)$ can be defined implicitly as $h(y_n, U) = \arg\max_{m_n,S_n}\hat{\mathcal{L}}_n(U)$, however, this requires solving a new optimization problem for $q(x_n|U)$ at each training step and for each sample of U. To avoid this expensive step, amortized inference can be used. In our testing, we found several design choices improved performance, which we detail below.

Encoder architecture. In our experiments, we used two separate two-hidden-layer multilayer perceptrons (MLPs), one to encode the mean and one to encode the log-variance. The rationale for this was that the magnitudes of variation in optimal means and variances differ significantly.

Reparameterization of inducing variables. Both of these networks received the same inputs. The goal of these networks is to couple the information from the observation to the values of the inducing variables to improve the representation of the latent space. The naive way to do this would be to pass in the tuple (U_k, y_n) for each observation and sample. While this approach is possible, we found that at late stages of training, when the variational distribution of the inducing variables, $q(U)$, had a high degree of certainty, all of the samples of the inducing variables U_k lie very close to one another. In order to give a better-conditioned input to the network, while preserving the information required to couple the inducing points to the observations, we use the reparameterization trick to sample U_k.

For each sample of $u_{pk} \in \mathbb{R}^M$, we sample a standard normal variable $\xi_{pk} \sim \mathcal{N}(0, I)$ which we transform to be a sample from the variational posterior, $q(u_p) = \mathcal{N}(\mu_p, \Sigma_p)$ using

$$u_{pk} = \mu_p + L_p\xi_{pk}, \tag{29}$$

where L_p is the Cholesky factor of Σ_p. The main benefit of the reparameterization trick is that it allows gradients to propagate through samples of the global inducing variables U. However, as an added benefit, the reparameterized input Ξ_k is well spread out over the input space, while maintaining all of the sample-specific information required to couple the inducing sample to the output and condition the latent variable on the inducing point samples. Therefore, we pass the variable $\Xi_k = \{\xi_{pk}\}_{p=1}^P$.

Local variational posterior. We additionally apply a soft-plus transform to ensure that the variance of the variational posterior is strictly positive, so the local variational posterior becomes

$$\mu_{nk} = h_\mu(y_n, \Xi_k), \tag{30}$$

$$\log \sigma_{nk} = h_\sigma(y_n, \Xi_k), \tag{31}$$

$$\Sigma_{nk} = \text{diag}(\text{softplus}(\log \sigma_{nk}) + \epsilon_{\min}), \tag{32}$$

where $\epsilon_{\min}$ is a small positive constant ensuring numerical stability. These parameters define the local variational distribution $q(x_n \mid U_k) = \mathcal{N}(\mu_{nk}, \text{diag}(\sigma_{nk}^2))$.

Annealed prior regularization. When training on larger datasets, we found that the mean encoder network was slow to converge to making predictions with variance approximately equal to one. To speed up this convergence, we adapted our local variable by scaling the KL regularization term up by a factor of β larger than one at the start of training, similar to the β-VAE [Higgins et al., 2017], so the local ELBO looks like

$$\mathcal{L}_S \approx \frac{1}{K} \sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\mathcal{L}}_n(U_k) - \beta D_{\text{KL}}\big(q(U) \,\|\, p(U)\big). \tag{33}$$

Over the course of training, we decay this variable from an initial value $\beta > 1$ to $\beta = 1$ at the end of training. We applied this schedule only to the A-S-GPLVM on the MNIST and FMNIST datasets, with an initial value of $\beta = 10$ and a quadratic decay to $\beta = 1$. We did not apply it to the mean-field baselines, as they did not exhibit the same optimization issue, and our preliminary experiments showed no improvement in the MF-GPLVM ELBO from using this schedule. It is important to emphasize that this variable does not change the objective at the end of training. Its only function is to speed up convergence to the optimal variational posterior.

Comparison to mean field amortization. In amortized mean field GPLVMs, the encoder takes only y_n as input and produces a variational posterior $q(x_n)$ independent of the inducing variables.

D Model Configuration

For the smaller datasets (oil flow, qPCR), the hyperparameters are:

- Latent Dimensionality (Q): 10 dimensions
- Inducing Points (M): 32 per output dimension
- Training Iterations: 100,000 for all four models
- Batch Size: 32
- Random Seeds: 0,1,2,3,4
- Initial lengthscales of 1.0 and observation noise variance of 1.0
- Number of samples from the inducing variables (K): 64
- Learning rate: $1e^{-3}$
- MLP hidden layer dims: $(128, 64)$

For the larger datasets (MNIST, FMNIST), the hyperparameters are:

- Latent Dimensionality (Q): 10 dimensions
- Inducing Points (M): 64 per output dimension
- Training Iterations: 200,000 for all three models
- Batch Size: 32
- Random Seeds: 0,1,2,3,4
- Initial lengthscales of 1.0 and observation noise variance of 1.0
- Number of samples from the inducing variables (K): 4
- Learning rate: $1e^{-4}$
- MLP hidden layer dims: $(512, 256)$

E Computational setup

All experiments reported for Oilflow, qPCR, MNIST, and FMNIST using the hyperparameter configurations described above were run on similar GPU nodes in a high-performance computing cluster. The precise hardware specifications varied slightly across runs. Detailed measurements of the runtime and memory overhead of the proposed structured approximation, measured on a representative node, are provided in Appendix H.

F Optimal ELBO Guarantee

In this section, we formally show that the mean-field variational family is a special case of our structured variational family. This establishes a theoretical guarantee that the optimal ELBO achievable by A-S-GPLVM is at least as high as that of A-MF-GPLVM.

Proposition. Let $\mathcal{Q}_{\mathrm{MF}}$ denote the set of variational posteriors of the form of Equation (7), and $\mathcal{Q}_{\mathrm{S}}$ those of the form of Equation (9), where both use the same parametrisation $q(u_p) = \mathcal{N}(\mu_p, \Sigma_p)$ and the same class of encoder networks. Then $\mathcal{Q}_{\mathrm{MF}} \subseteq \mathcal{Q}_{\mathrm{S}}$, and for any fixed setting of the GP hyperparameters,

$$\sup_{q \in \mathcal{Q}_{\mathrm{S}}} \mathcal{L}_{\mathrm{S}}(q) \geq \sup_{q \in \mathcal{Q}_{\mathrm{MF}}} \mathcal{L}_{\mathrm{MF}}(q). \tag{34}$$

Proof. Any $q \in \mathcal{Q}_{\mathrm{MF}}$ is recovered from Equation (9) by taking $h(y_n, U) = g(y_n)\ \forall U$, i.e. an encoder which is constant in U, so $\mathcal{Q}_{\mathrm{MF}} \subseteq \mathcal{Q}_{\mathrm{S}}$. On such a member the two objectives agree: the local KL term in Equation (11) is constant in U and can therefore be pulled out of the expectation over $q(U)$, leaving the remaining terms exactly as in the mean-field ELBO of Equation (8). Since $\mathcal{L}_{\mathrm{S}}$ and $\mathcal{L}_{\mathrm{MF}}$ are therefore the same bound on $\log p(Y)$ evaluated over nested families, the supremum over the larger family cannot be smaller. □

This result shows that, provided that the encoder can model functions which are constant with respect to some dimensions, the global optimum ELBO of an A-S-GPLVM must be at least as good as the optimum of the A-MF-GPLVM ELBO. Of course, in most cases, we will not achieve the global optimum, but we empirically show that the A-S-GPLVM outperforms its MF counterpart.

Table 3: Two-sided Welch's t-test p-values over 5 independent runs comparing A-MF-GPLVM and A-S-GPLVM.

Metric	Oilflow	qPCR	MNIST	FMNIST
RMSE (train)	0.12	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-4}$
RMSE (test)	0.55	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-4}$
NLPD (train)	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-4}$
NLPD (test)	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-4}$
CRPS (train)	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
CRPS (test)	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
Coverage (train)	0.071	0.013	0.357	$< 10^{-3}$
Coverage (test)	0.022	0.031	0.703	0.045
ELBO	$< 10^{-4}$	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-4}$

G Statistical significance of the observed improvements

To assess whether the gains from the structured variational family are robust across random initialization and stochastic optimization noise, we performed a two-sided Welch's t-test over 5 independent runs, comparing A-MF-GPLVM and A-S-GPLVM on each dataset. The resulting p-values are reported in Table 3.

Overall, the A-S-GPLVM achieves statistically significant improvements in NLPD and CRPS across all four datasets, on both the training and test sets. In particular, all CRPS improvements are significant

at the 10^{-4} level, while the NLPD improvements are significant at the 10^{-3} level or smaller. The RMSE improvements are also statistically significant on qPCR, MNIST, and FMNIST. On Oilflow, however, the RMSE differences are not statistically significant ($p = 0.12$ on the training set and $p = 0.55$ on the test set). The empirical coverage results are more mixed, with some statistically significant differences between the methods but no consistent improvement towards the nominal 95% level.

This distinction is important for interpreting the effect of the proposed method. In particular, the Oilflow results show that the structured posterior can significantly improve the quality of the predictive distribution, as measured by NLPD and CRPS, even when the point-estimate accuracy measured by RMSE does not change significantly. This supports our claim that the main benefit of the structured variational family lies in improving the posterior approximation and predictive uncertainty, rather than merely shifting the predictive mean.

Table 4: Empirical runtime and peak GPU memory over 100 training iterations on Oilflow, using $B = 32$, $M = 32$, $Q = 10$, $K = 32$ and float64 precision.

Metric	A-MF-GPLVM	A-S-GPLVM
Wall-clock time (s)	5.03	8.36
Peak GPU memory (MB)	29.0	598.7

H Runtime and memory overhead

The proposed structured approximation introduces two main sources of overhead relative to A-MF-GPLVM. First, the encoder input becomes larger, since the sampled inducing-variable state is concatenated with the observation rather than processing the observation alone. Second, each ELBO evaluation uses K Monte Carlo replicas. Both operations are parallelizable, so the primary practical trade-off is expected to be increased memory use rather than an inherently sequential increase in computation.

To quantify this overhead empirically, we measured runtime and peak GPU memory on the Oilflow dataset for 100 training iterations using a minibatch size of $B = 32$, $M = 32$ inducing points, latent dimension $Q = 10$, 12 observed dimensions, $K = 32$ Monte Carlo samples, and float64 precision throughout. The experiments were run on an NVIDIA RTX 2080Ti GPU (22 GB VRAM) with an AMD EPYC 7282 CPU and CUDA 13.0. The results are shown in Table 4.

In this configuration, A-S-GPLVM increases the wall-clock time from $5.03\,\mathrm{s}$ to $8.36\,\mathrm{s}$ over 100 iterations, corresponding to an approximate $1.66\times$ increase. In terms of memory, peak GPU usage rises from $29.0\,\mathrm{MB}$ to $598.7\,\mathrm{MB}$. This is consistent with the fact that the K inducing-variable samples can be processed in parallel, so the dominant cost of the structured estimator is storing the replicated computation graph and intermediate activations.

We emphasize that these measurements characterize the present implementation rather than an unavoidable algorithmic constant. In our current codebase, the additional wall-clock overhead is partly influenced by implementation details in a shared layered design across models. The more stable conclusion is that the proposed method preserves the same sparse, minibatched stochastic variational structure as amortized mean-field GPLVMs, but pays a memory cost in exchange for a richer posterior approximation.

I Ablation studies

To better isolate the source of the improvement in A-S-GPLVM, we carried out ablations on the encoder architecture, the number of inducing points, and the number of sampled inducing-variable states. The corresponding plots are shown in Figures 6, 7, and 8.

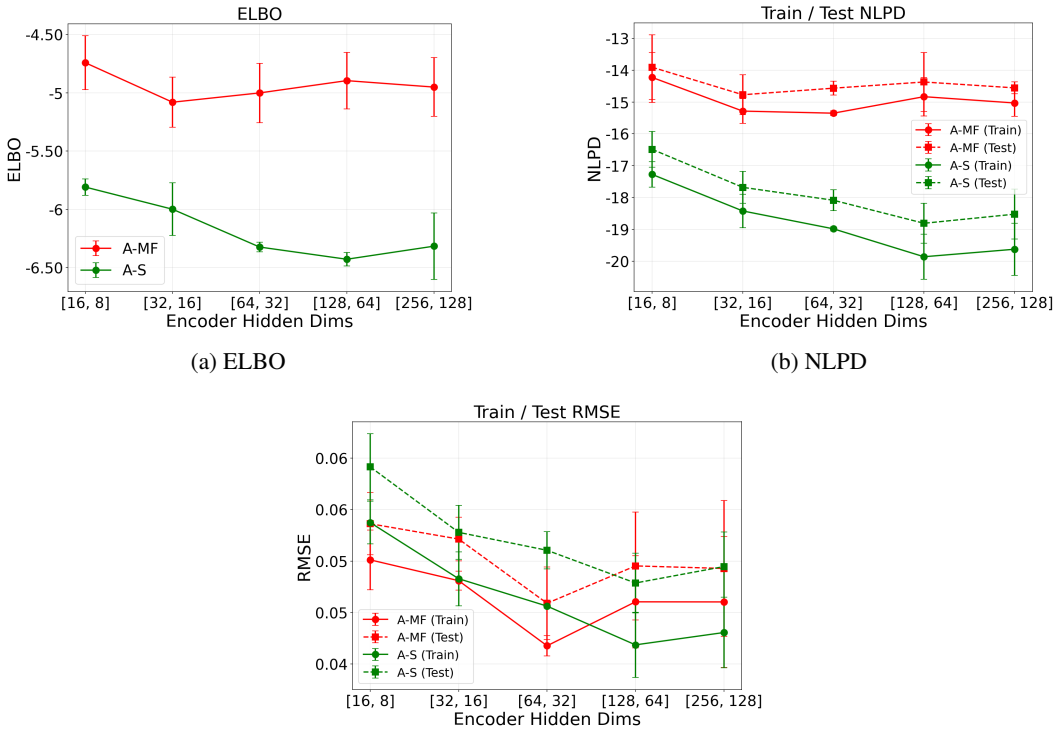
Encoder architecture. We first varied the widths of the two-hidden-layer MLP encoder across the settings $[16, 8]$, $[32, 16]$, $[64, 32]$, $[128, 64]$, and $[256, 128]$ for both A-MF-GPLVM and A-S-GPLVM. Across this range, the total encoder size did not have a strong effect on performance, while A-S-GPLVM consistently outperformed A-MF-GPLVM on the probabilistic metrics (ELBO and NLPD). This supports our claim that the main benefit does not come from increased encoder capacity.

Increasing encoder size in the mean-field model only makes the deterministic map $y_n \mapsto q(x_n)$ more flexible, whereas the structured model can adapt the local posterior to the sampled global GP state through $q(x_n \mid U)$. In particular, even a small encoder with SSVI remains better than a larger encoder with MFVI.

Number of inducing points. We next fit the Oilflow dataset using 16, 32, and 64 inducing points. Across all tested values, A-S-GPLVM performed better than A-MF-GPLVM on ELBO and NLPD, and often performed comparably to or better than A-MF-GPLVM with twice as many inducing points. We also observe that the margin between the two models becomes smaller with larger numbers of inducing points. In our current implementation, we suspect this happens because the encoder input dimension in A-S-GPLVM grows with the number of inducing points and becomes too large for the small MLP used in this ablation.

Number of inducing-variable samples. Finally, we varied the number of sampled inducing-variable states K. In the ablation plots, we compare $K = 4$ and $K = 64$, and also show the effect of using learning rates 10^{-3} and 3×10^{-4}. Larger values of K lead to better results, which is consistent with the reduced variance of the gradient estimator when more samples are used. In memory-constrained settings, however, increasing K may not be possible. The ablation also shows that for small K, lowering the learning rate can recover performance by countering the increased gradient variance, at the cost of slower convergence.

Taken together, these ablations support the interpretation that the main gains come from the structured variational family rather than from increased encoder capacity, while also clarifying the dependence of performance on the number of inducing points and inducing-variable samples.

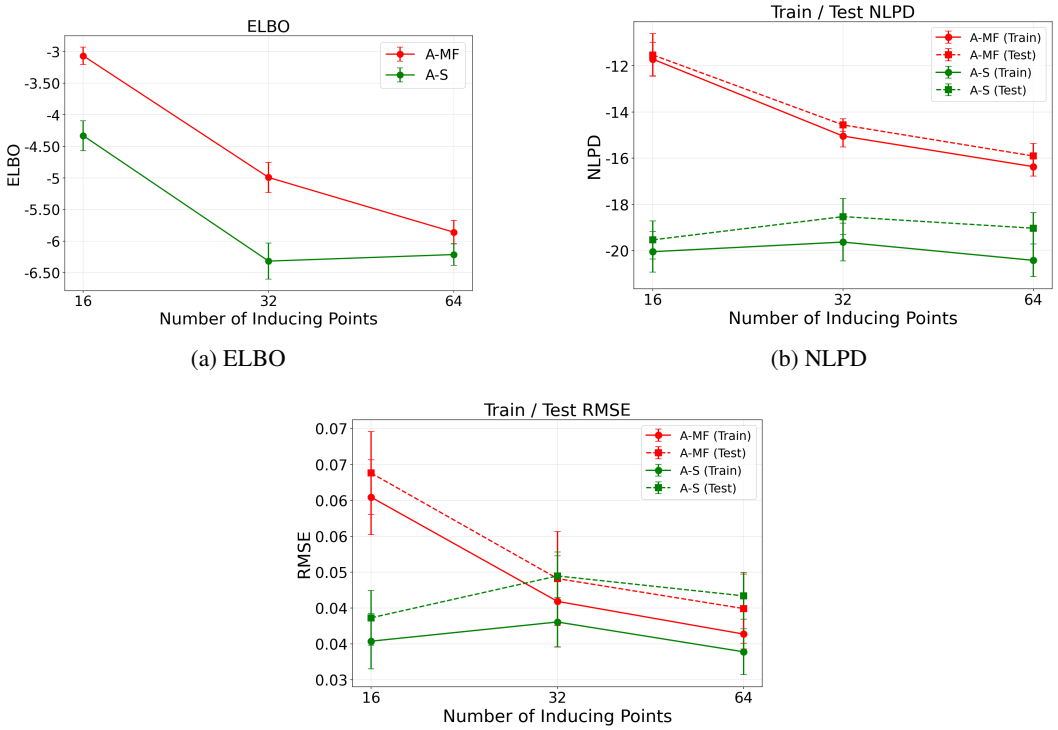


(a) ELBO

(b) NLPD

(c) RMSE

Figure 6: Ablation over encoder architecture on the Oilflow dataset. We compare A-MF-GPLVM and A-S-GPLVM across different two-hidden-layer MLP encoder widths.

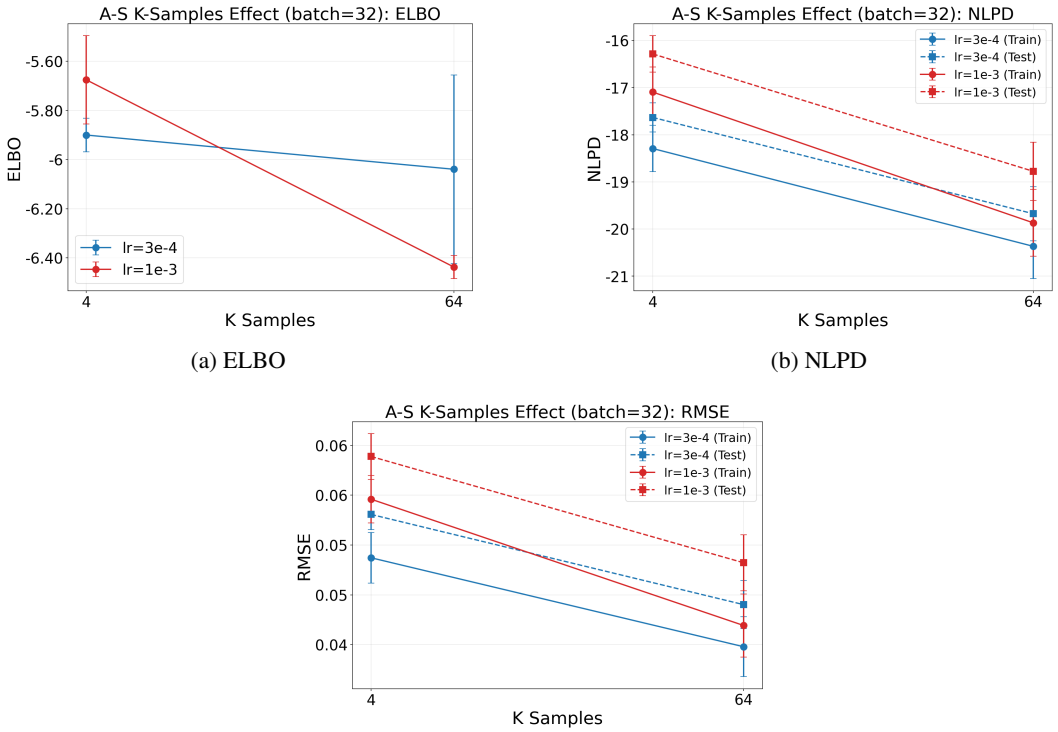


(a) ELBO

(b) NLPD

(c) RMSE

Figure 7: Ablation over the number of inducing points on the Oilflow dataset. We compare A-MF-GPLVM and A-S-GPLVM for different values of the number of inducing points.



(a) ELBO

(b) NLPD

(c) RMSE

Figure 8: Ablation over the number of sampled inducing-variable states on the Oilflow dataset.

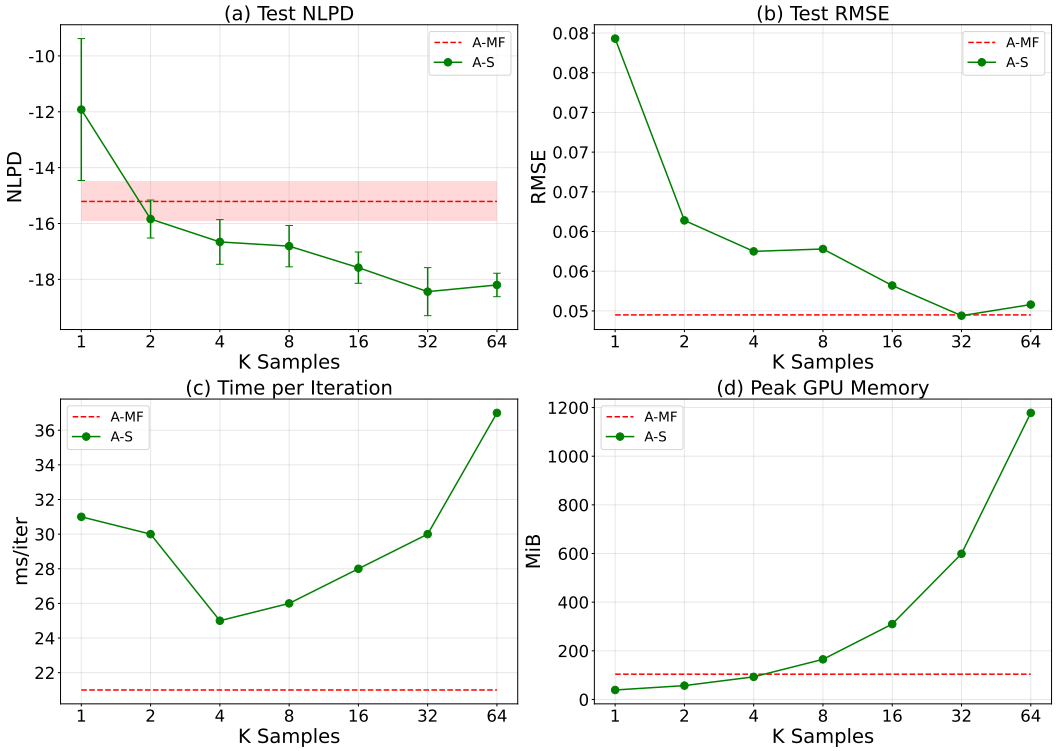


Figure 9: Ablation over the number of Monte Carlo samples K on the Oilflow dataset. The A-S-GPLVM results are shown in green, while the dashed red lines indicate the A-MF-GPLVM baseline. Error bars represent the standard deviation of test NLPD, with the shaded region indicating the corresponding uncertainty for A-MF-GPLVM.

Predictive Performance and Computational Cost Across Monte Carlo Sample Counts We further investigate the effect of the number of Monte Carlo samples K on predictive performance and computational cost on the Oilflow dataset. We consider $K \in \{1, 2, 4, 8, 16, 32, 64\}$ and compare A-S-GPLVM against A-MF-GPLVM.

As shown in Figure 9, increasing the number of Monte Carlo samples generally improves predictive performance, particularly in terms of test NLPD. The best observed results are achieved at $K = 32$, with a test NLPD of -18.44 ± 0.86 and a test RMSE of 0.0494, compared to -15.21 ± 0.70 and 0.0495, respectively, for A-MF-GPLVM.

However, these improvements come at an increased computational cost. At $K = 32$, A-S-GPLVM requires 30 ms per iteration and 599 MiB of peak GPU memory, compared to 21 ms and 104 MiB for A-MF-GPLVM.

Increasing K further to 64 does not improve predictive performance, while increasing peak GPU memory to 1178 MiB and iteration time to 37 ms. These results highlight the trade-off between predictive performance and computational cost, with $K = 32$ providing the best observed predictive performance among the evaluated configurations.

J Full-scale MNIST Experiment

To further evaluate the scalability of our approach, we conducted additional experiments on the full MNIST dataset, extending the subsampled MNIST experiments reported in Section 5. We used the same model and optimization configurations described in Appendix D, but trained on the full MNIST dataset instead of the subsampled version. We evaluated all three amortized GPLVM variants (A-MF-GPLVM, IWVI-GPLVM, and A-S-GPLVM) over five independent runs with different random seeds.

Table 5: Reconstruction performance on the full MNIST dataset. Values are reported as mean $\pm$ standard deviation over five runs. The best result for each metric is shown in bold. The final negative ELBO is not reported for IWVI-GPLVM because it optimizes a different importance-weighted objective.

Metric	A-MF	IWVI-GPLVM	A-S (Ours)
Train RMSE $\downarrow$	0.1562 ± 0.0003	0.1551 ± 0.0019	$\mathbf{0.1528 \pm 0.0005}$
Test RMSE $\downarrow$	0.1552 ± 0.0003	0.1553 ± 0.0017	$\mathbf{0.1517 \pm 0.0003}$
Train NLPD $\downarrow$	-336.85 ± 1.52	-342.32 ± 9.98	$\mathbf{-354.86 \pm 1.16}$
Test NLPD $\downarrow$	-341.99 ± 1.71	-341.18 ± 8.83	$\mathbf{-360.32 \pm 2.67}$
Final neg. ELBO $\downarrow$	-302.28 ± 1.26	–	$\mathbf{-313.00 \pm 2.47}$

Table 6: Predictive uncertainty evaluation on the full MNIST dataset. Values are reported as mean $\pm$ standard deviation over five runs. The best result for each metric is shown in bold, with coverage evaluated relative to its nominal level.

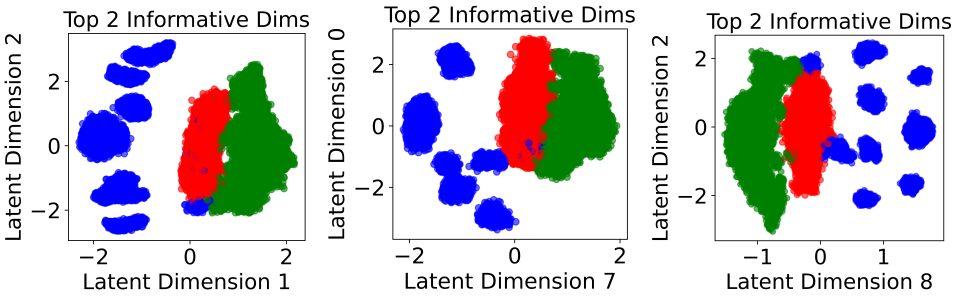
Metric	A-MF	IWVI-GPLVM	A-S (Ours)
Train CRPS $\downarrow$	0.1010 ± 0.0002	0.1007 ± 0.0017	$\mathbf{0.0985 \pm 0.0003}$
Test CRPS $\downarrow$	0.1006 ± 0.0002	0.1008 ± 0.0016	$\mathbf{0.0981 \pm 0.0002}$
Train coverage (95%)	0.9260 ± 0.0002	$\mathbf{0.9297 \pm 0.0008}$	0.9274 ± 0.0003
Test coverage (95%)	0.9270 ± 0.0002	$\mathbf{0.9295 \pm 0.0011}$	0.9284 ± 0.0007

The results are reported in Tables 5 and 6. Consistent with our findings on the subsampled MNIST dataset, A-S-GPLVM achieves the lowest train and test RMSE, NLPD, and CRPS among all three methods. Two-sided Welch's t-tests confirm that A-S-GPLVM significantly outperforms IWVI-GPLVM on all three test metrics (RMSE, NLPD, and CRPS; $p < 0.05$).

These results are consistent with our findings on the subsampled MNIST dataset, demonstrating that the improvements achieved by structured variational inference persist when scaling to the full dataset.

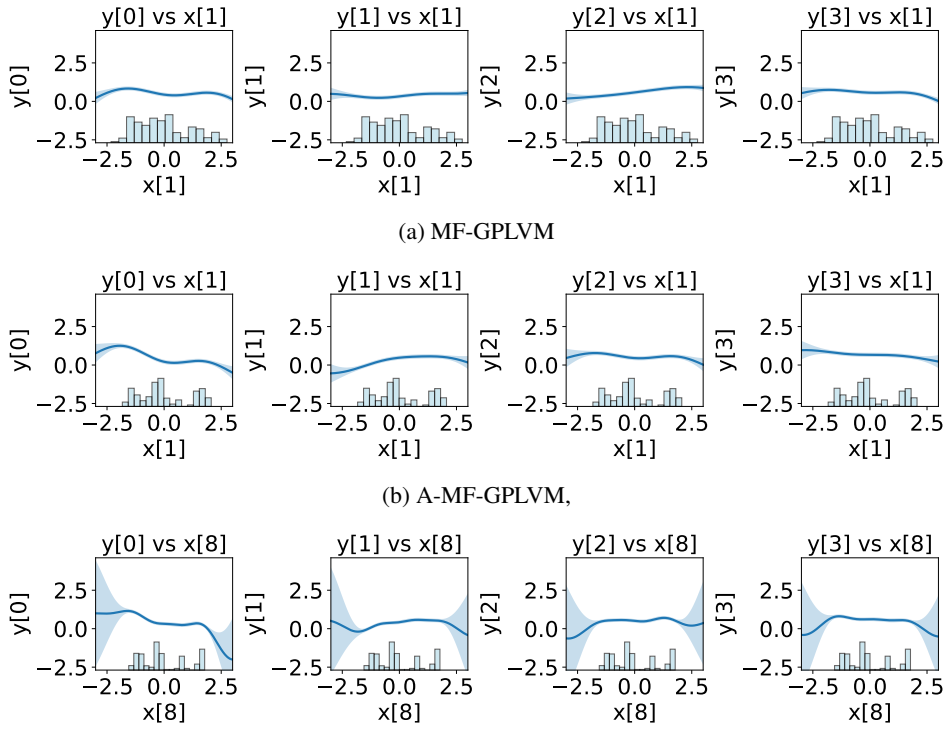
K Additional Results Plots

We have included additional plots relating to the experiments described in Section 5 here.



(a) MF-GPLVM (b) A-MF-GPLVM (c) A-S-GPLVM

Figure 10: Latent spaces for the Oilflow dataset.

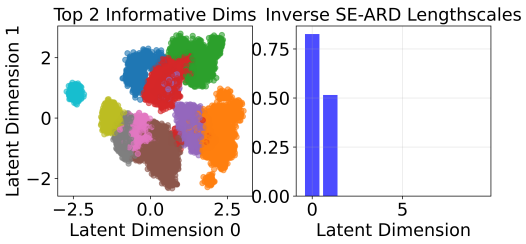


(a) MF-GPLVM

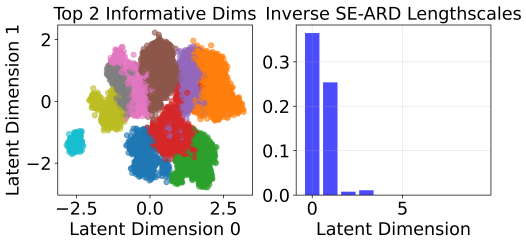
(b) A-MF-GPLVM,

(c) A-S-GPLVM

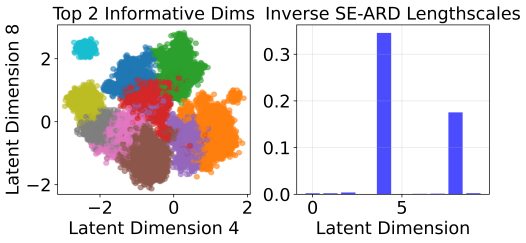
Figure 11: Latent slice through the most informative latent dimension. Curves show predictive means; shaded areas denote predictive uncertainty for each output. The A-S-GPLVM GP mapping has significantly larger predictive uncertainty away from the data.



(a) MF-GPLVM



(b) A-MF-GPLVM



(c) A-S-GPLVM

Figure 12: Performance comparison on the qPCR dataset, for (a) MF-GPLVM, (b) A-MF-GPLVM, (c) A-S-GPLVM.

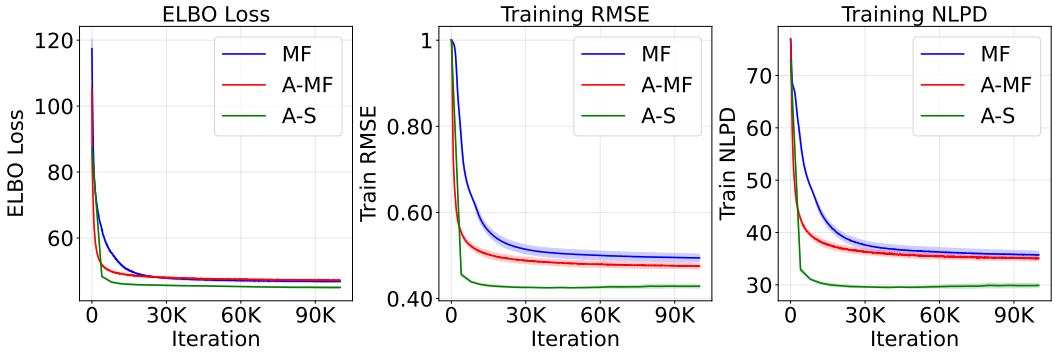
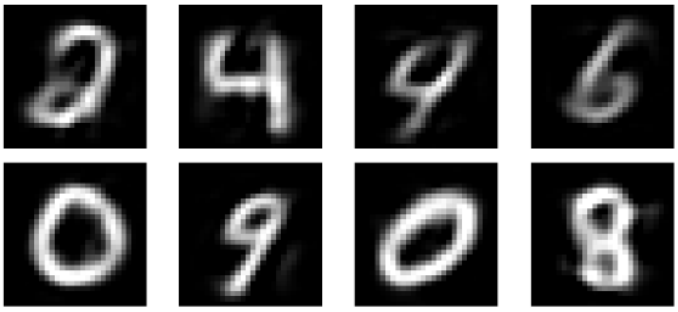
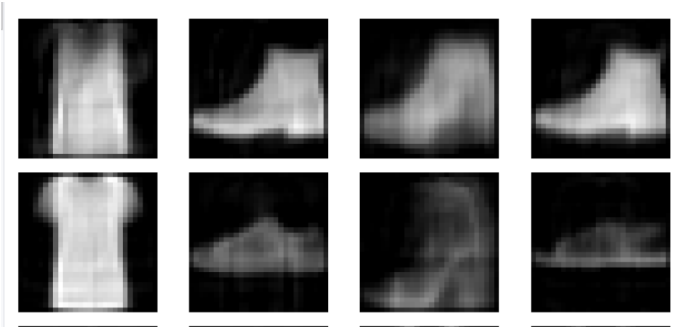


Figure 13: Training dynamics comparison across inference methods on the qPCR dataset. The panels show the ELBO, RMSE, and NLPD trajectories over training iterations. A-S-GPLVM exhibits faster convergence and lower final NLPD, indicating improved uncertainty calibration and optimization stability compared to MF-GPLVM and A-MF-GPLVM.

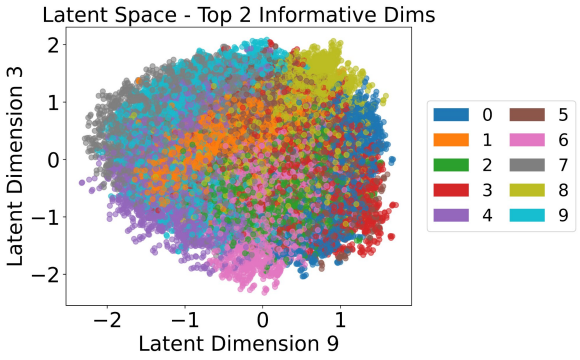


(a) MNIST

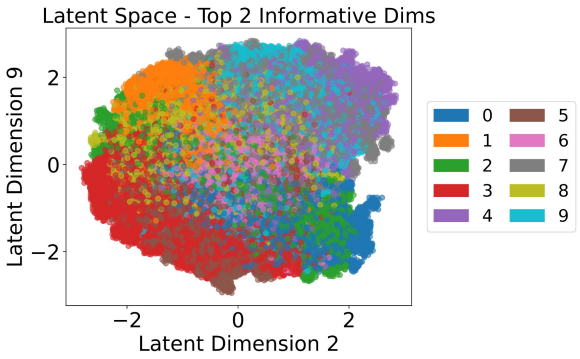


(b) FMNIST

Figure 14: Sample reconstructions using the A-MF-GPLVM. Each image shows the posterior predictive mean of the reconstructed observation.



(a) A-S-GPLVM



(b) A-MF-GPLVM

Figure 15: Latent spaces for the MNIST dataset.

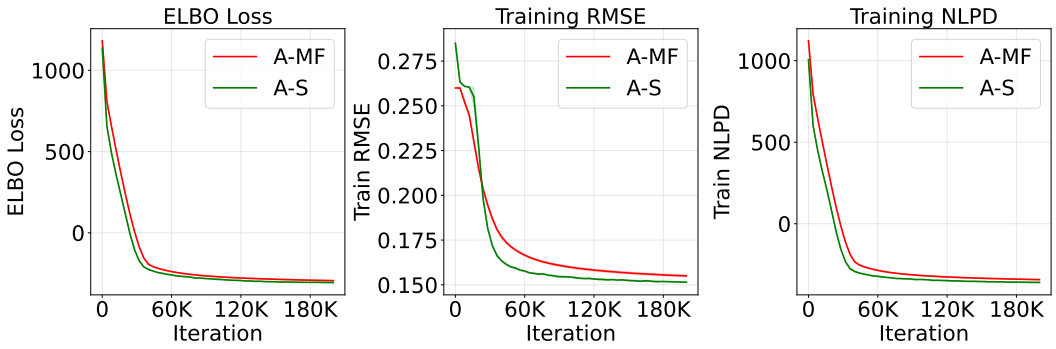


Figure 16: Training dynamics comparison across inference methods on the MNIST dataset. The panels show the ELBO, RMSE, and NLPD trajectories over training iterations.

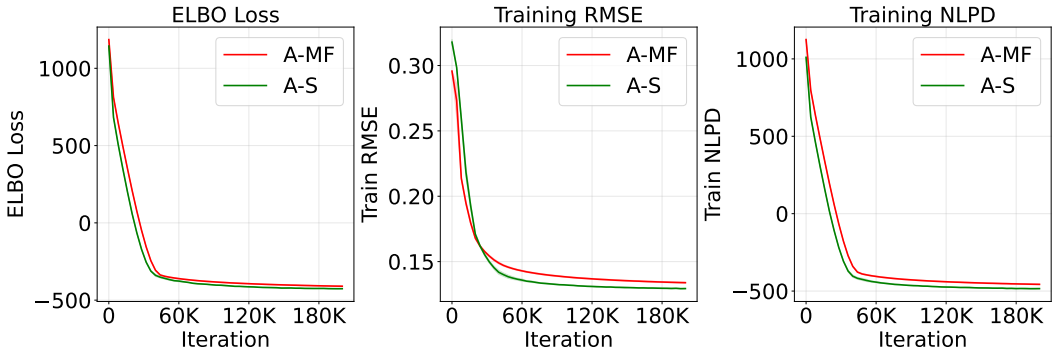


Figure 17: Training dynamics comparison across inference methods on the Fashion-MNIST dataset. The panels show the ELBO, RMSE, and NLPD trajectories over training iterations.

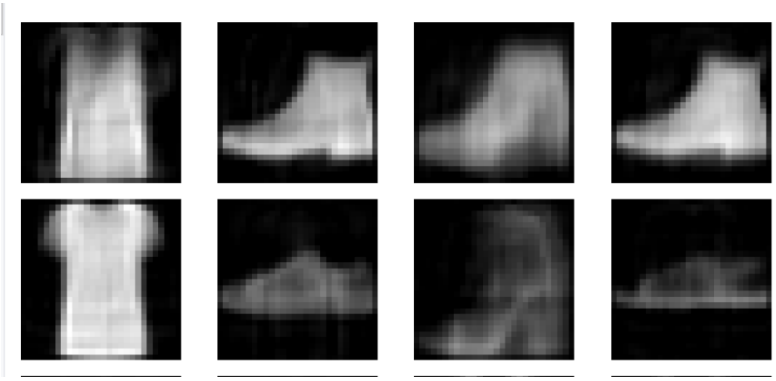


Figure 18: Sample reconstructions on Fashion-MNIST using the A-MF-GPLVM. Each image shows the posterior predictive mean of the reconstructed observation.

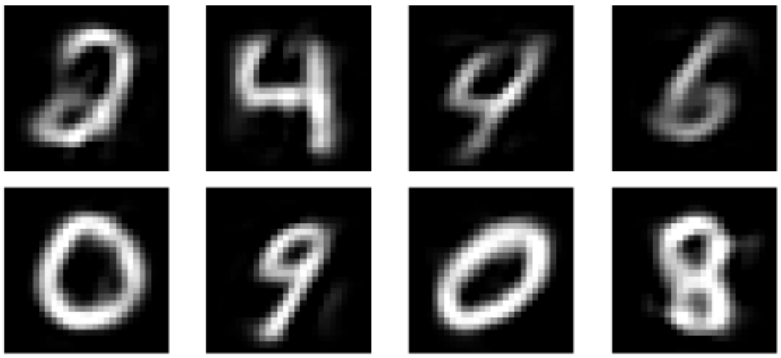


Figure 19: Sample reconstructions on MNIST using the A-MF-GPLVM. Each image shows the posterior predictive mean of the reconstructed observation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state that the paper introduces an amortized structured variational posterior for GPLVMs by coupling latent variables with inducing variables. The experimental claims are supported by the results in Section 5, Table 1, and the qualitative uncertainty analysis in Figure 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated limitations section in Section 6. It discusses the computational overhead of A-S-GPLVM, with runtime and memory measurements reported in Appendix H and Table 4. It also discusses remaining restrictions in the variational posterior our work has not addressed.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main mathematical content consists of ELBO derivations and variational factorization arguments, which are given including all relevant assumptions in Section 2 and Section 3. The paper does not present formal theorems or proofs as theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the datasets, preprocessing, train-test splits, baselines, evaluation metrics, random seeds, and main hyperparameters in Section 5 and Appendix D. These details are sufficient to reproduce the main experimental comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides an anonymized repository link to reproduce the experiments. Section 5 and Appendix D specify the datasets, preprocessing, train–test splits, baselines, evaluation metrics, random seeds, and main hyperparameters, which are sufficient to reproduce the main experimental comparisons.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides an anonymized repository link to reproduce the experiments. Section 5 and Appendix D specify the datasets, preprocessing, train–test splits, baselines, evaluation metrics, random seeds, and main hyperparameters. Additional implementation details, including the specific optimizers and related training choices, are provided in the accompanying code, making the main experimental comparisons reproducible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 1 reports means and standard deviations over independent runs. Appendix G additionally reports two-sided Welch's t-test p-values comparing A-MF-GPLVM and A-S-GPLVM.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E and Appendix H report runtime and peak GPU memory usage, and execution times for representative benchmarks on an RTX 2080Ti GPU with an AMD EPYC 7282 CPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is methodological and uses standard public benchmark datasets. To the best of our knowledge, the research conforms to the NeurIPS Code of Ethics and does not involve deception, unsafe deployment, or non-consensual human-subject data collection by the authors.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The current paper does not include a dedicated broader impacts discussion. The work is foundational methodology for probabilistic latent-variable modeling, with potential positive impact through improved uncertainty estimation, but no direct deployment setting or specific societal risk is analyzed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, pretrained generative models, or systems with obvious dual-use misuse risks. Therefore, special release safeguards are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original sources for the datasets used, including Oilflow, qPCR, MNIST, and Fashion MNIST.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces code for the proposed A-S-GPLVM method and provides an anonymized repository link for reproducing at least the Oilflow experiment. The implementation details and model configuration are documented in Section 5 and Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or new research with human subjects. All experiments use existing benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve new human-subject experiments conducted by the authors. Therefore, IRB approval or an equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology does not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.