Shared Stochastic Gaussian process Decoders: A Probabilistic Generative model for Quasar Spectra

Vidhi Lalchand [1] **Anna-Christina Eilers** [2]

Abstract

This work proposes a scalable probabilistic latent variable model based on Gaussian processes (Lawrence, 2004) in the context of multiple observation spaces. We focus on an application in astrophysics where it is typical for data sets to contain both observed spectral features as well as scientific properties of astrophysical objects such as galaxies or exoplanets. In our application, we study the spectra of very luminous galaxies known as quasars, and their properties, such as the mass of their central supermassive black hole, their accretion rate and their luminosity, and hence, there can be multiple observation spaces. A single data point is then characterised by different classes of observations which may have different likelihoods. Our proposed model extends the baseline stochastic variational Gaussian process latent variable model (GPLVM) (Lalchand et al., 2022) to this setting, proposing a seamless generative model where the quasar spectra and the scientific labels can be generated *simultaneously* when modelled with a shared latent space acting as input to different sets of Gaussian process decoders, one for each observation space. Further, this framework allows training in the missing data setting where a large number of dimensions per data point may be unobserved. We demonstrate high-fidelity reconstructions of the spectra and the scientific labels during test-time inference and briefly discuss the scientific interpretations of the results along with the significance of such a generative model.

[*]Equal contribution [1] University of Cambridge, Cambridge, UK [2]MIT Kavli Institute for Astrophysics and Space Research. Correspondence to: Vidhi Lalchand <vr308@cam.ac.uk>.

ICML 2023 Workshop on Machine Learning for Astrophysics, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

Many challenges in the contemporary physical sciences arise from the analysis of large scale, noisy and high-dimensional datasets (Clarke et al., 2016). Generative latent variable models of late have supplanted traditional dimensionality reduction techniques as they offer the simultaneous benefits of a probabilistic interpretation and data generation while learning a faithful embedding of the high-dimensional training data in low-dimensional latent space. A generative probabilistic framework like the GPLVM (Lawrence, 2004) works by optimising the parameters of a Gaussian process *decoder* from low dimensional latent space ($Z \in \mathbb{R}^{N\times Q}$) to high-dimensional data space ($X \in \mathbb{R}^{N\times D}$) such that $Q << D$ and points close in latent space are nearby in data space. Since the decoder is a non-parametric Gaussian process, the kernel function controls the inductive biases of the function mapping like smoothness and periodicity. There typically is no encoder mapping hence, these models are also called Gaussian process decoders.

This work proposes a novel formulation of the GPLVM based on the idea of a shared latent space. The earlier work by Ek (2009) was the first to propose the idea of a shared data generation process but precluded truly scalable inference due to the standard $\mathcal{O}(N^3)$ scaling. We extend this framework in two important ways. First, we show that the shared GPLVM is compatible with stochastic variational inference (SVI) (Hoffman et al., 2013) where we derive a joint evidence lower bound which factorises across multiple observation spaces due to conditional independence but share predictive strength though inducing locations and latent variables. Secondly, we train the entire model in the presence of missing dimensions in one or both of the observation spaces. Crucially, we demonstrate that it is possible to share predictive strength by learning a common latent variable space Z across multiple-outputs (X, Y) where $Y \in \mathbb{R}^{N\times L}$ is an additional observation space with L dimensions. In this way we indirectly model the relationships and correlation structure between the different observation spaces.

We demonstrate this scalable model in an astrophysical application using data of quasars. Quasars are the most luminous galaxies in the universe, powered by accretion onto a central supermassive black hole (SMBH) with millions to

billions of solar masses in size. Understanding the formation, growth, and evolution across cosmic time of quasars and their SMBHs is one of the major goals of observational cosmology today. To this end, precise measurements of the physical properties of quasars are crucial, but usually require very expensive and extremely time-intensive observations, since multiple epochs of observations are required to accurately determine the quasar's SMBH mass for instance. The high-dimensional data used in this work contains 20,000 quasars with their spectral information (along 590 dimensions/pixels) along with 4 scientific labels per quasar: their black hole mass, luminosity, redshift and so-called Eddington ratio – a measure of the quasar's accretion rate. By modelling the spectra and scientific labels through a generative model acting on a shared latent space we aim to reason about the physical properties of the quasars just through its "single-epoch" spectral information, thus circumventing the time-intensive multi-epoch observations. Earlier work on applying probabilistic generative modelling to high-dimensional quasar spectra using Gaussian processes (Eilers et al., 2022) have been constrained on scalability and examine less than 50 astronomical objects. This is because the authors use exact GPs and construct the full data marginal likelihood for both spectra and labels, this set-up scaled cubically in the number of objects. By using inducing point based sparse GPs (Titsias, 2009) and stochastic variational inference we demonstrate our framework on datasets $400\times$ bigger.

2. Stochastic Variational GPLVM with a Shared latent space

The fundamental contribution of this work is to develop an inference scheme to show that a shared latent space with a joint evidence lower bound enables highly scalable inference through SVI. We summarise that in the section below:

2.1. SV-GPLVM: Stochastic Variational GPLVM

In the traditional formulation underlying GPLVMs we have a training set comprising of N D-dimensional real valued observations $X \equiv \{\boldsymbol{x}_n\}_{n=1}^N \in \mathbb{R}^{N\times D}$. These data are associated with N Q-dimensional latent variables, $Z \equiv \{\boldsymbol{z}_n\}_{n=1}^N \in \mathbb{R}^{N\times Q}$ where $Q << D$ provides dimensionality reduction (Lawrence, 2004). The forward mapping ($Z \longrightarrow X$) is governed by GPs independently defined across dimensions D. The sparse GP formulation describing the data is as follows:

$$p(Z) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{z}_n; \mathbf{0}, \mathbb{I}_Q),$$

$$p(F|U, Z, \theta) = \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{f}_d; K_{nm}K_{mm}^{-1}\boldsymbol{u}_d, Q_{nn}) \quad (1)$$

$$p(X|F, Z) = \prod_{n=1}^{N}\prod_{d=1}^{D} \mathcal{N}(x_{n,d}; f_d(\boldsymbol{z}_n), \sigma_x^2),$$

where $Q_{nn} = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$, $F \equiv \{\boldsymbol{f}_d\}_{d=1}^D$, $U \equiv \{\boldsymbol{u}_d\}_{d=1}^D$ and $\boldsymbol{x_d}$ is the d^{th} column of X. K_{nn} is the covariance matrix corresponding to a user chosen positive-definite kernel function $k_\theta(x, x')$ evaluated on latent points $\{\boldsymbol{x}_n\}_{n=1}^N$ and parameterised by hyperparameters θ shared across dimensions. The inducing variables per dimension $\{\boldsymbol{u}_d\}_{d=1}^D$ are distributed with a GP prior $\boldsymbol{u}_d|\tilde{Z} \sim \mathcal{N}(\boldsymbol{u}_d; \mathbf{0}, K_{mm})$ computed on inducing input locations $\tilde{Z} \in \mathbb{R}^{M\times Q}$ which live in latent space with Z and have dimensionality Q (matching $\boldsymbol{z}_n$).

The crux of SVI applied to sparse variational GPs as proposed in the seminal work of Hensman et al. (2013) is that we can variationally integrate out $\boldsymbol{u}_d$ by learning their variational distributions $q(\boldsymbol{u}_d) \sim \mathcal{N}(\boldsymbol{m}_d, S_d)$ numerically using stochastic gradient methods. Essentially, by keeping the representation of $\boldsymbol{u}_d$ uncollapsed. While (Hensman et al., 2013) proposed SVI for GP regression, (Lalchand et al., 2022) extended this work to GPLVMs where the inputs Z to the GPs are unobserved and each dimension of the high-dimensional output space $\boldsymbol{x}_d$ is modelled by an independent GP $\boldsymbol{f}_d$ but with shared kernel hyperparameters. If we choose to optimize the latent variables Z as point estimates rather than variationally integrate them out, one can bound the intractable log-marginal likelihood $p(X|\theta)$ in the model formulation above with the following evidence lower-bound,

$$p(X|\theta) \geq \int p(F|U, Z)q(U) \log \frac{p(X|F, Z)p(U|\tilde{Z})p(Z)}{q(U)} dF dU = \mathcal{L}_x \quad (2)$$

$$= \sum_{n,d} \langle \log p(x_{n,d}|\boldsymbol{f}_d, z_n, \sigma_x^2)\rangle_{q(.)} - \text{KL}(q(U)||p(U|\tilde{Z})) + \log p(Z)$$

where the variational distribution $q(.)$ is given by,

$$p(F, U|X) = \Big[\prod_{d=1}^{D} p(\boldsymbol{f}_d|\boldsymbol{u}_d, Z)q(\boldsymbol{u}_d)\Big] \approx q(F, U) \quad (3)$$

Latent point estimates of Z can be learnt along $\boldsymbol{\theta}$ and variational parameters $(\tilde{Z}, \boldsymbol{m}_d, S_d)$ by taking gradients of the ELBO above. An important constraint however is that this formulation assumes a single kernel matrix (single set of kernel hyperparameters) underlying all the D independent GPs. In the next section, we introduce the idea of an additional observation space with L dimensions and how they can be modelled by their own stack of independent GPs $\boldsymbol{f}_l$ and learn their own set of hyperparameters for additional flexibility but share the latent embedding Z and inducing inputs $\tilde{Z}$ to model correlations between the different output spaces.

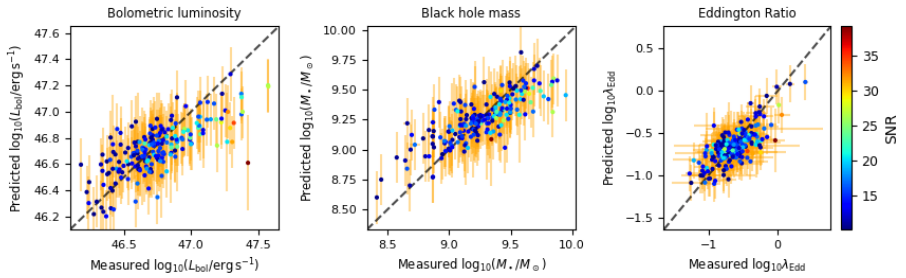

Figure 1. Scientific label prediction based on unseen (X^*, Y^*). The scatter are colored by SNR which is a measured quantity available as part of the dataset but is not used during training. The dashed black line (- - -) denotes a 45° line to aid visualisation of reconstruction accuracy. The vertical and horizontal orange lines (——) denotes posterior predictive standard deviation and the recorded measurement uncertainty for each object (data point) and dimension.

2.2. Shared joint variational lower bound

In the astrophysical application we focus on in this work we have two observation spaces corresponding to N quasars. We denote the quasar spectra (pixels) with the matrix $X \in \mathbb{R}^{N\times D}$ and the scientific labels corresponding to the N objects with $Y \in \mathbb{R}^{N\times L}$. The GPLVM construction models each column (pixel dimension and label dimension) with an independent GP, with the GPs corresponding to the pixel dimensions $\{f_d\}_{d=1}^D$ and label dimensions $\{f_l\}_{l=1}^L$ modelled with their own independent kernels and kernel hyperparameters, θ_x and θ_y. Within each observation space the kernel hyperparameters are shared, so we learn 2 sets of hyperparameters corresponding to two observation spaces.

$$f_d \sim \mathcal{GP}(0, k_{\theta_x}) \qquad f_l \sim \mathcal{GP}(0, k_{\theta_y}) \tag{4}$$

The priors over finite function values are given by,

$$p(\boldsymbol{f}_d|\theta_x) = \mathcal{N}(\mathbf{0}, K_{nn}^{(d)}) \quad p(\boldsymbol{f}_l|\theta_y) \quad = \mathcal{N}(\mathbf{0}, K_{nn}^{(l)}) \tag{5}$$

where $K_{nn}^{(d)}$ and $K_{nn}^{(l)}$ denote the $N \times N$ kernel matrices which rely on their own set of hyperparameters. The two observation spaces also yield two data likelihoods given by,

$$p(X|f_{1:D}, Z) = \prod_{n=1}^{N}\prod_{d=1}^{D}\mathcal{N}(x_{n,d}; f_d(\boldsymbol{z}_n), \sigma_x^2) \tag{6}$$

$$p(Y|f_{1:L}, Z) = \prod_{n=1}^{N}\prod_{l=1}^{L}\mathcal{N}(y_{n,l}; f_l(\boldsymbol{z}_n), \sigma_y^2) \tag{7}$$

In the absence of sparsity the log-marginal likelihood of the joint model compartmentalises nicely due to the assumed factorisation in the likelihoods. We marginalise out the latent function values $f_{1:D}$ and $f_{1:L}$ per dimension,[1]

$$\begin{aligned}
&p(X, Y|\theta_x, \theta_y, Z) \\
&= \int\int p(X|f_{1:D}, Z)p(Y|f_{1:L}, Z)p(\boldsymbol{f}_d|\theta_x)p(\boldsymbol{f}_l|\theta_y)d\boldsymbol{f}_{1:D}d\boldsymbol{f}_{1:L} \\
&= \int_{1:D} p(X|f_{1:D})p(\boldsymbol{f}_d|\theta_x)d\boldsymbol{f}_{1:D}\int_{1:L} p(Y|f_{1:L})p(\boldsymbol{f}_l|\theta_y)d\boldsymbol{f}_{1:L} \\
&= \prod_{d=1}^{D} p(\boldsymbol{x}_d|\theta_x)\prod_{l=1}^{L} p(\boldsymbol{y}_l|\theta_y) = \prod_{d=1}^{D}\mathcal{N}(\mathbf{0}, K_{nn}^{(d)} + \sigma_x^2) \\
&\qquad\qquad \times \prod_{l=1}^{L}\mathcal{N}(\mathbf{0}, K_{nn}^{(l)} + \sigma_y^2)
\end{aligned}$$

where $\boldsymbol{x}_d$ and $\boldsymbol{y}_l$ denote a single column/dimension of the observation spaces X and Y. The log marginal likelihood objective is then given by the following,

$$\log p(X, Y|\theta_x, \theta_y, Z) = \sum_{d=1}^{D}\log p(\boldsymbol{x}_d|\theta_x, Z) + \sum_{l=1}^{L}\log p(\boldsymbol{y}_l|\theta_y, Z) \tag{8}$$

With sparse GPs each of the terms in the decomposition above can be bounded by $\mathcal{L}_x$, while the inducing points $\tilde{Z}$ can be shared between the terms yielding the joint evidence lower bound.

$$\begin{aligned}
\log p(X, Y|\theta_x, \theta_y, Z) \geq &\sum_{n,d}\langle\log p(x_{n,d}|\boldsymbol{f}_d, z_n, \sigma_x^2)\rangle_{q(.)} \\
-\sum_d \mathrm{KL}(q(\boldsymbol{u}_d)||p(\boldsymbol{u}_d|\tilde{Z})) &+ \sum_{n,l}\langle\log p(y_{n,l}|\boldsymbol{f}_l, z_n, \sigma_y^2)\rangle_{q(.)} \\
&- \sum_l \mathrm{KL}(q(\boldsymbol{u}_l)||p(\boldsymbol{u}_l|\tilde{Z})) + \log p(Z)
\end{aligned} \tag{9}$$

The joint variational lower bound in eq. 9 is optimised for the shared latent embedding Z (local variational parameters), kernel hyperparameters and global variational parameters.We include the full training algorithm and detail the prediction framework in the appendix.

3. Experiments

In this section we demonstrate experiments aimed at assessing the reconstruction quality of unseen quasar spectra and scientific attributes. The data used in this work are quasar

[1]Note that there are $D + L$ dimensions in total.

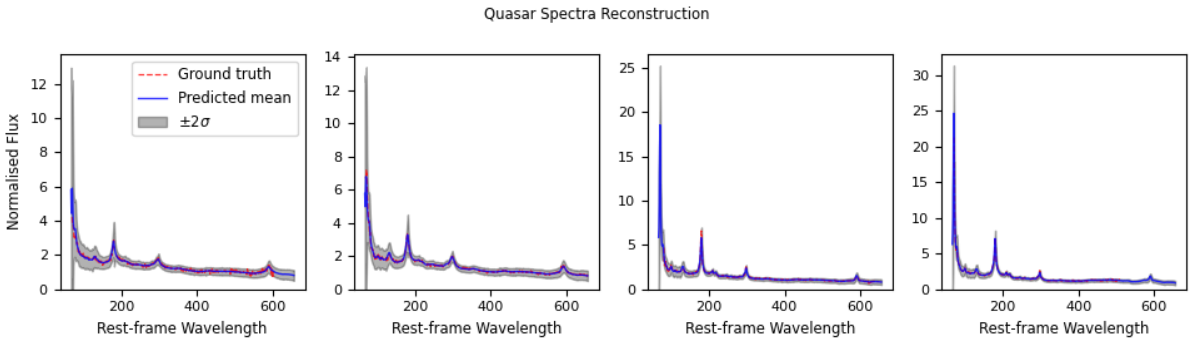

Figure 2. Reconstruction plots of unseen quasar spectra with $\pm 2\sigma$ predictions intervals. Note that the rightmost plot demonstrates extrapolation at test-time where pixel dimensions [504-576] were missing. The blue curve denotes the posterior predictive mean at each dimension.

Metrics ($\longrightarrow$)	RMSE	
Models ($\longrightarrow$)	Baseline	Shared (ours)
Spectra	0.181 ± 0.005	$\mathbf{0.1416 \pm 0.008}$
Blackhole Mass	0.298 ± 0.011	$\mathbf{0.233 \pm 0.014}$
Luminosity	0.261 ± 0.025	$\mathbf{0.228 \pm 0.022}$
Eddington Ratio	0.206 ± 0.006	$\mathbf{0.182 \pm 0.009}$

Table 1. Summary of test-time reconstruction abilities for the spectra and the scientific labels. Square root mean-squared error (RMSE) on un-normalised data ($\pm$ standard error of mean) evaluated on average of 10 splits with 80% of the data used for training.

spectra (X space) observed as part of the Sloan Digital Sky Survey (SDSS) DR16 (Lyke et al., 2020). We chose all quasars with spectra that have a signal-to-noise ratio (SNR) per pixel > 10. The four scientific labels for these quasars are (1) their SMBH mass, (2) their bolometric luminosity, i.e. the total power output across all electromagnetic wavelengths, (3) their redshift which denotes the factor by which the emitted wavelengths have been "stretched" due to the expansion of the universe, and (4) their Eddington ratio, which is a measure of the accretion and growth rate of the SMBH. All measurements were previously uniformly determined by Wu and Shen (2022). The baseline model in the experiments refers to the canonical stochastic variational GPLVM (Lalchand et al., 2022) which treats multiple observation spaces using $D + L$ independent GPs with the same kernel and optimising a single set of kernel hyperparameters. In terms of the lower bound, the baseline model has the same structure of the shared lower bound (eq. 9) without independent terms (involving l) for the scientific labels.

It may be important to note that the approach in Eilers et al. (2022) does not scale to $20,000$ objects making a direct comparison infeasible; further smaller datasets (under 1000) objects do not necessitate stochastic variational training which is mainly motivated by the need to train on much bigger datasets.

3.1. Reconstructing Quasar spectra

We assess the quality of our probabilistic generative model in reconstructing unseen quasar spectra. At test-time we deal with unseen spectra and scientific labels stacked row-wise and denoted by (X^*_{gt}, Y^*_{gt}). The 2-step prediction learns low-dimensional shared latent variables Z^* (point estimate per data point), followed by a forward pass through the GP decoder corresponding to the spectra $Z^* \rightarrow X^*_{est.}$. Note that the ground truth spectra contains several missing pixels (dimensions) and the probabilistic decoder provides a reasonable reconstruction at those locations. In fig. 2 we visualise the reconstruction (posterior predictive mean) of 4 test quasars along with ground-truth measurements and 2σ uncertainty intervals. We achieve a remarkably good reconstruction even when a significant chunk of the ground-truth spectra are missing (rightmost plot). Further, the prediction intervals capture the the ground spectra providing robust coverage at peaks and extrapolated regions.

3.2. Predicting unseen scientific labels Y^*

The L dimensions corresponding to the scientific labels in the dataset governed by their own GP decoders $\{f_l\}_{l=1}^{L}$ are a critical prediction quantity. The ability to reconstruct these quantities from learnt latent variables underscores the generalisation abilities of our model. In fig. 1 we demonstrate the accuracy of our reconstructions by plotting each of the dimensions against ground truth held-out data. We show reconstructions for 200 test points in fig. 1 sampled randomly from the full test set containing 2000 quasars. section 3 reports the RMSE on the full test set. Each point on the scatter denotes a quasar and the x-axis denotes the ground-truth measurement. The orange vertical error bars denote 2σ intervals computed by extracting the diagonals from the GP posterior predictive for each dimension. We can observe a high-degree of prediction accuracy across the three scientific labels and the ground-truth data is incorporated within the 2σ interval for over 90% of the points in the test set. Further, we can observe that the reconstruction qual-

ity is robust and independent of the signal-to-noise (SNR) ratio as there is no strong pattern of correlation between prediction quality and SNR. table 2 summarises the NLLs (quality of uncertainty) for two modes of inference 1) where the scientific attributes for an unseen test quasar (with fully observed spectra and labels) were reconstructed from the latent estimate Z^* and 2) where the scientific attributes were reconstructed only from the spectra. The model exhibits reasonable behaviour reflecting higher uncertainty (higher NLL) in the latter case with marginal to no degradation in the accuracy of the estimate as seen from the scatter plots in fig. 7 relative to fig. 1.

Scientific Labels (⟶)	Black hole Mass	Luminosity	Eddington Ratio
Fully observed test-point	-0.1251	-0.2339	-0.1814
Only spectra observed	-0.1209	-0.2235	-0.1626

Table 2. Summary of test-time uncertainty quantification under the full and split reconstruction framework. Negative log-likelihood (lower is better) on un-normalised observed data.

4. Significance

Our new generative model allows us to *simultaneously* model the spectral properties of quasars as well as their scientific labels, thus opening up novel possibilities to study the evolution of quasars across cosmic time and the formation and growth of SMBHs. Additionally, we have shown that our model is able to predict other physical properties of quasars such as their bolometric luminosities (see Fig. 1). This implies that we can obtain a measurement of the quasars' absolute luminosities from their spectra alone, which enables us to use quasars as so-called "standard candles". Standard candles are incredibly valuable for astronomy, as knowing the luminosity of an object allows one to determine its distance. Previously, supernovae have been famously used as standard candles, which lead to the Nobel Prize winning discovery of the expansion of our universe and the existence of dark energy (Riess et al., 1998). Our new model allows us to use quasars as standard candles and can be probed to larger distances due to their on average much higher luminosities, enabling new constraints on the dark energy content of our universe in the future.

References

P. Clarke, P. V. Coveney, A. F. Heavens, J. Jäykkä, B. Joachimi, A. Karastergiou, N. Konstantinidis, A. Korn, R. G. Mann, J. D. McEwen, S. de Ridder, S. Roberts, T. Scanlon, E. P. S. Shellard, and J. A. Yates. Big data in the physical sciences: challenges and opportunities. *ATI Scoping Report*, 2016.

A.-C. Eilers, D. W. Hogg, B. Schölkopf, D. Foreman-Mackey, F. B. Davies, and J.-T. Schindler. A generative model for quasar spectra. *The Astrophysical Journal*, 938 (1):17, 2022.

C. H. Ek. *Shared Gaussian process latent variable models*. PhD thesis, Citeseer, 2009.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI2013)*, 2013.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013. URL http://jmlr.org/papers/v14/hoffman13a.html.

V. Lalchand, A. Ravuri, and N. D. Lawrence. Generalised GPLVM with Stochastic Variational Inference. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7841–7864. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/lalchand22a.html.

N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in neural information processing systems*, pages 329–336, 2004.

B. W. Lyke, A. N. Higley, J. N. McLane, D. P. Schurhammer, A. D. Myers, A. J. Ross, K. Dawson, S. Chabanier, P. Martini, N. G. Busca, H. d. Mas des Bourboux, M. Salvato, A. Streblyanska, P. Zarrouk, E. Burtin, S. F. Anderson, J. Bautista, D. Bizyaev, W. N. Brandt, J. Brinkmann, J. R. Brownstein, J. Comparat, P. Green, A. de la Macorra, A. Muñoz Gutiérrez, J. Hou, J. A. Newman, N. Palanque-Delabrouille, I. Pâris, W. J. Percival, P. Petitjean, J. Rich, G. Rossi, D. P. Schneider, A. Smith, M. Vivek, and B. A. Weaver. The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release. *The Astrophysical Journal Supplement Series*, 250(1):8, Sept. 2020. doi: 10.3847/1538-4365/aba623.

A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan,

S. Jha, R. P. Kirshner, B. Leibundgut, M. M. Phillips, D. Reiss, B. P. Schmidt, R. A. Schommer, R. C. Smith, J. Spyromilio, C. Stubbs, N. B. Suntzeff, and J. Tonry. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *The astronomical journal*, 116(3):1009–1038, Sept. 1998. doi: 10.1086/300499.

M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Q. Wu and Y. Shen. A Catalog of Quasar Properties from Sloan Digital Sky Survey Data Release 16. *The Astrophysical Journal Supplement Series*, 263(2):42, Dec. 2022. doi: 10.3847/1538-4365/ac9ead.

A. Schematic and Graphical Model

In fig. 3 we present a schematic of the model architecture with double observation spaces (X, Y), the corresponding stacks of individual GPs $\{f_d\}$ and $\{f_l\}$ which model the individual columns of the spectra X and scientific attributes Y respectively and the low-dimensional latent space Z. The dimensionality of the latent and observation spaces are denoted by Q, D, L respectively and N denotes the number of objects / data points (quasars). Note that the correlation between the two observation spaces are not explicitly but implicitly modelled through a shared latent space. Generating a single data point $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ (a row across X and Y) entails a forward pass through the GPs, where $\boldsymbol{x}_n = [\ldots, x_{nd}, \ldots]$ is generated as $[f_1(\boldsymbol{z}_n), f_2(\boldsymbol{z}_n), \ldots, f_D(\boldsymbol{z}_n)]$ and $\boldsymbol{y}_n = [\ldots, y_{nl}, \ldots]$ is generated as $[f_1(\boldsymbol{z}_n), \ldots, f_L(\boldsymbol{z}_n)]$.

B. Algorithm

Below we enclose the pseudo-code in Algorithm 1 for stochastic variational inference in the context of the shared model for clarity. Let $\mathcal{L}_x$ and $\mathcal{L}_y$ denote the ELBO's for each of the observation spaces and let $\mathcal{L}_x^{(B)}$ and $\mathcal{L}_y^{(B)}$ denote the ELBOs formed with a randomly drawn mini-batch of the data (across all dimensions). For a mini-batch (subset) of the data $X_B \subset X$, the mini-batch ELBO is given by,

$$\mathcal{L}_x \simeq \mathcal{L}_x^{(B)} = \frac{N}{B}\left(\sum_{b,d}\langle \log p(x_{b,d}|\boldsymbol{f}_d, z_b, \sigma_x^2)\rangle_{q(.)} + \sum_b \log p(\boldsymbol{z}_b)\right) - \sum_d \mathrm{KL}(q(\boldsymbol{u}_d)||p(\boldsymbol{u}_d|\tilde{Z})) \quad (10)$$

where the scaling term is important for the mini-batch ELBO to be an estimator of the full-dataset ELBO.

B.1. Computational Cost

The training cost of the canonical stochastic variational GPLVM is dominated by the number of inducing points $\mathcal{O}(M^3 D)$ (free of N) where $M << N$ and D is the data-dimensionality (we have D GP mappings f_d, one per output dimension). The practical algorithm is made further scalable with the use of mini-batched learning. In our shared model with 2 sets of GPs the dynamics of the training cost are the same except that they go up linearly in the number of additional dimensions (L), making the cost $\mathcal{O}(M^3(D+L))$. The number of global variational parameters to be updated in each step (parameters of $q(U)$) is $MQ + M(D+L) + M^2(D+L)$, where MQ are the M Q-dimensional inducing inputs $\tilde{Z}$ (shared), $M(D+L)$ is the size of the mean parameters of the inducing variables $\boldsymbol{u}_d$, $\boldsymbol{u}_l$ and $M^2(D+L)$ are the full-rank covariances of the inducing variables. The local variational parameters Z (the latent embedding shared across GPs) are of size NQ and model hyperparameters (kernel hyperparameters) are of size $2Q + 4$, which account for Q input lengthscales, a scalar signal variance and noise variance per GP group $\{f_d\}$ and $\{f_l\}$. We use the squared exponential kernel with automatic relevance determination across both sets of GPs.

C. Experimental set-up

In this section we detail the configuration of the experiments in section 3 of the main paper. We conduct experiments across two data sets with 1K and 20K points. For each of the datasets we repeat every experiment with 10 random seeds yielding different splits of the 80% training data. The baseline model in the experiments refers to the canonical stochastic variational GPLVM (Lalchand et al., 2022) which treats multiple observation spaces using the same set of independent GPs learning a single set of kernel hyperparameters. The attributes of the data and sparse GP set-up are given in table 3. We used a learning rate of 0.005 across all parameters and ran the mini-batch loop with a batch size of 100 for 10,000 iterations on an Intel Core i7 processor with a GeForce RTX 3070 GPU with 8GB RAM memory. In order to give an estimate of the scale of the model for the 20k dataset we enclose a summary snapshot of the number of trainable parameters in our shared model.

D. Additional Experimental Results

D.0.1. Reconstructing missing spectra

In this experiment we test the generative models ability to learn from massively missing chunks of the spectra at test-time. We observe a partial window of the spectra in each plot (given by the shaded region), hence the latent variables corresponding to these points are only informed by the observed region. We then reconstruct the whole spectra

Algorithm 1: Shared Stochastic GPLVM for Quasar Spectra

TRAINING FRAMEWORK

Input: ELBO objective $\mathcal{L} = \mathcal{L}_x + \mathcal{L}_y$, gradient based optimiser `optim()`, observation spaces X (spectra) and Y (scientific labels)

Initial model params:

$\theta = (\theta_x, \theta_y)$ (covariance hyperparameters for GP mappings $f_{1:D}, f_{1:L}$),

$\sigma^2 = (\sigma_x^2, \sigma_y^2)$ (variance of the noise model for each likelihood),

$Z \equiv \{\boldsymbol{z}_n\}_{n=1}^N$ (point estimates for latent embedding)

Initial variational params:

$\tilde{Z} \in \mathbb{R}^{M \times Q}$ (inducing locations),

$\lambda = \{m_h, S_h\}_{h=1}^{D+L}$ (global variational params for inducing variables per dimension $\boldsymbol{u}_h$),

while *not converged* **do**

- Choose a random mini-batch of the data from both the observation spaces $X_B \subset X, Y_B \subset Y$.
- Form a mini-batch estimate of the ELBO: $\mathcal{L}_x^{(B)} + \mathcal{L}_y^{(B)}$
- Gradient step for global parameters $\boldsymbol{g} \leftarrow \nabla_{\theta,\sigma^2,\tilde{Z},\lambda}(\mathcal{L}_x^{(B)} + \mathcal{L}_y^{(B)})$
- Gradient step for local parameters $\boldsymbol{l} \leftarrow \nabla_{Z_B}(\mathcal{L}_x^{(B)} + \mathcal{L}_y^{(B)})$ (where Z_B are the latent embeddings corresponding to points in the mini-batch)
- Update all parameters $\tilde{Z}, \boldsymbol{\theta}, \sigma^2, \lambda, Z_B \equiv \{\boldsymbol{z}_b\}_{n=1}^B \longleftarrow$ `optim()` using gradients $\boldsymbol{g}, \boldsymbol{l}$

end

return $\theta, \sigma^2, \tilde{Z}, \lambda, Z$

PREDICTION FRAMEWORK

(Predict Z^* corresponding to unseen X^*, Y^*)

Input: Trained global and local parameters $\theta, \sigma^2, \tilde{Z}, \lambda, Z$, unseen observation spaces X^* (spectra) and Y^* (scientific labels).

1. Initialise latent embedding $Z^* \equiv \{\boldsymbol{z}_{n^*}\}_{n^*=1}^{N^*}$ corresponding to unseen points.
2. Extend the joint ELBO to include terms corresponding to the N^* additional data points.

$\mathcal{L}_x^* \leftarrow \mathcal{L}_x + \sum_{n^*,d} \langle \log p(x_{n^*,d} | \boldsymbol{f}_d, z_{n^*}, \sigma_x^2) \rangle_{q(.)} + \sum_{n^*} \log p(\boldsymbol{z}_{n^*})$

$\mathcal{L}_y^* \leftarrow \mathcal{L}_y + \sum_{n^*,l} \langle \log p(y_{n^*,l} | \boldsymbol{f}_l, z_{n^*}, \sigma_y^2) \rangle_{q(.)} + \sum_{n^*} \log p(\boldsymbol{z}_{n^*})$

$\mathcal{L}^* = \mathcal{L}_x^* + \mathcal{L}_y^*$

3. Freeze all global and local parameters except for Z^*

while *not converged* **do**

- Gradient step for Z^*: $\boldsymbol{l}^* \leftarrow \nabla_{Z^*} \mathcal{L}^*$
- Update $Z^* \leftarrow$ `optim()` using gradients $\boldsymbol{l}^*$.

end

return Z^*

(Note that the gradients of $\mathcal{L}_x$ and $\mathcal{L}_y$ with respect to Z^* are 0 and the only terms that are optimised are the additional terms corresponding to the new data points.)

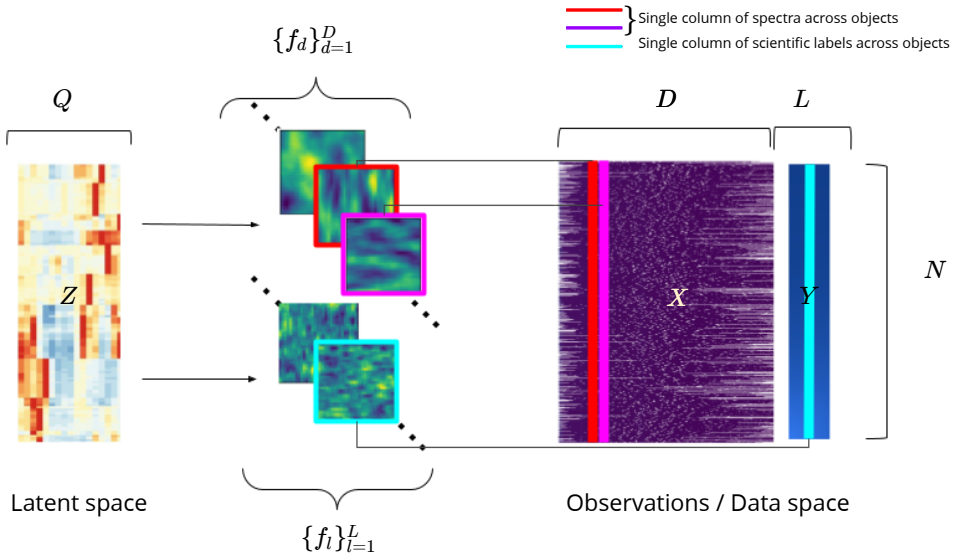

Figure 3. Shared GPLVM with multiple observation spaces. The blocks on the right-hand side denote the double observation spaces (X, Y) of quasar spectra and scientific labels respectively. In the center are two stacks of GPs, one for each observation space which control the data generation process through the shared latent space. In the figure above we assume $Q = 2$ (for ease of visualisation) since we denote the GPs are two dimensional surfaces, however, typically Q can be higher than 2 corresponding to higher dimensional GPs.

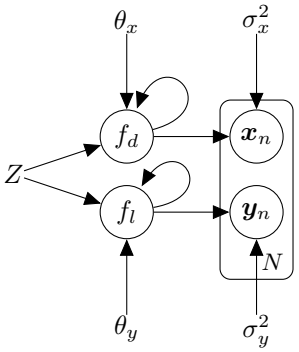

Figure 4. The graphical model of the shared GPLVM with two sets of independent GPs and their respective hyperparameter sets.

Dataset	N	D	L	Num inducing M	Latent dim. Q
20K	22844	657	4	250	10

Table 3. Experimental configuration to reproduce experiments in section 3 of the main paper.

from the latent variables informed by the partial spectra. We enclose our results in fig. 6. The reconstruction entails the inference steps: $X^*_{partial} \rightarrow Z^* \rightarrow X^*_{full}$. We note that both the quality of the mean prediction and the coverage of the uncertainty intervals deteriorate compared to the fully observed test point predictions. However, the coverage of the prediction intervals is only weaker as we move away from the shaded observed regions. Reconstruction quality at the observed regions is much higher.

D.0.2. Predicting scientific labels **only** from spectra X^*

Very often astronomers want to reason about the scientific attributes of quasars just by analysing their spectra. In this experiment we demonstrate precisely this use case where the latent variables Z^* are informed only by the spectra X^*, computing the prediction entails the inference steps: $X^* \rightarrow Z^* \rightarrow Y^*$. We can note from fig. 7 that we can predict the scientific labels with a high-degree of accuracy with a marginal to no degradation in prediction quality, in terms of both the mean estimate and the uncertainty intervals. Compared to fig. 1 we do note that the uncertainty intervals around each data point is slightly higher in this experiment. This behaviour is perfectly reasonable and captures the higher epistemic uncertainty in this set-up. This manifests in the marginally weaker log-likelihoods we show in table 2.

D.1. Cross-validating M and Q

The two main parameters of our shared framework which need to be fixed at the outset are: the number of inducing points M and the latent space dimensionality Q. We set M to be 250 in all the 20k experiments after extensive cross-validation upto $M = 1000$ and found that $M = 250$ gave the best possible trade-off in terms of speed and accuracy. The reconstruction results with $M = 1000$ were only marginally better than with $M = 250$ inducing points but significantly increased compute due to the cubic scaling in inducing points.

Modules	Parameters
inducing_inputs	2500
Z.Z	208440
model_spectra.variational_strategy._variational_distribution.variational_mean	164250
model_spectra.variational_strategy._variational_distribution.chol_variational_covar	41062500
model_spectra.mean_module.constant	657
model_spectra.covar_module.raw_outputscale	1
model_spectra.covar_module.base_kernel.raw_lengthscale	10
model_labels.variational_strategy._variational_distribution.variational_mean	1000
model_labels.variational_strategy._variational_distribution.chol_variational_covar	250000
model_labels.mean_module.constant	4
model_labels.covar_module.raw_outputscale	1
model_labels.covar_module.base_kernel.raw_lengthscale	10

Total Trainable Params: 41689373

Figure 5. Number of trainable parameters for the 20k model where `model_spectra` refers to the GPs corresponding to X and `model_labels` refers to the GPs corresponding to Y observation space.

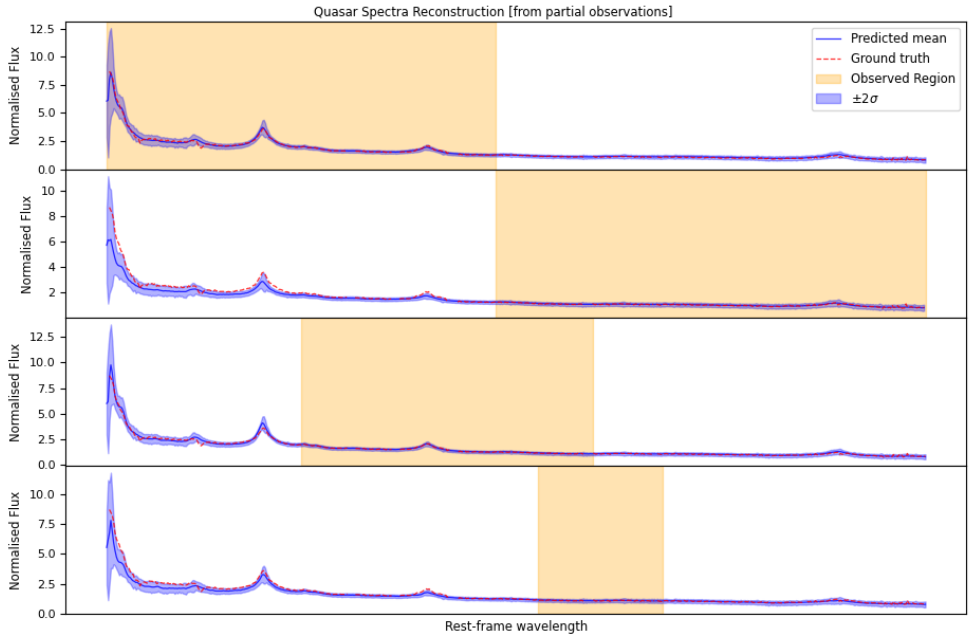

Figure 6. Reconstruction of a single spectra by masking out large wavelength chunks. The shaded orange regions denote the observed wavelengths. Note in the 2nd and 4th plots the 2σ prediction intervals underestimate the ground-truth at the initial wavelengths (left) as their observed regions are further out. On the contrary, the 1st and 3rd plots where the observed regions are closer to the initial wavelengths tend to reconstruct those dimensions better.

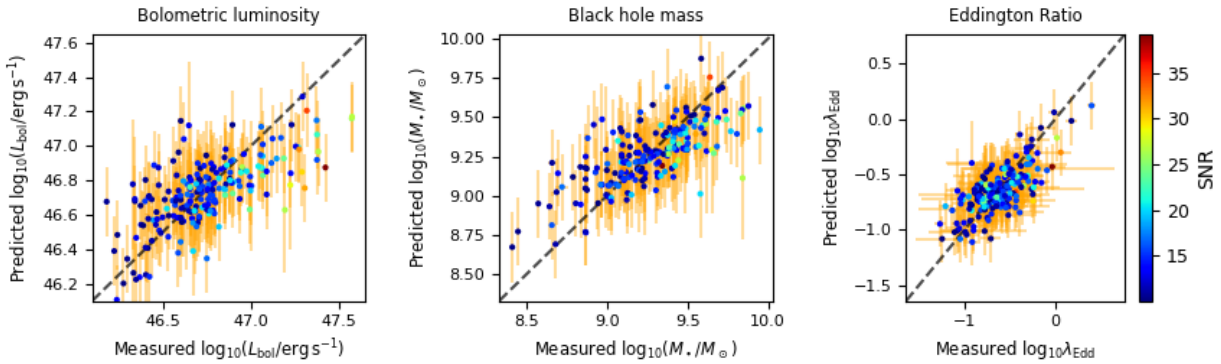

Figure 7. Scientific label prediction based on based on unseen X^* **only**. The dashed black line (- - -) denotes a 45° line to aid visualisation of reconstruction accuracy. The vertical and horizontal orange lines (——) denotes posterior predictive standard deviation and the recorded measurement uncertainty for each object (data point) and dimension.

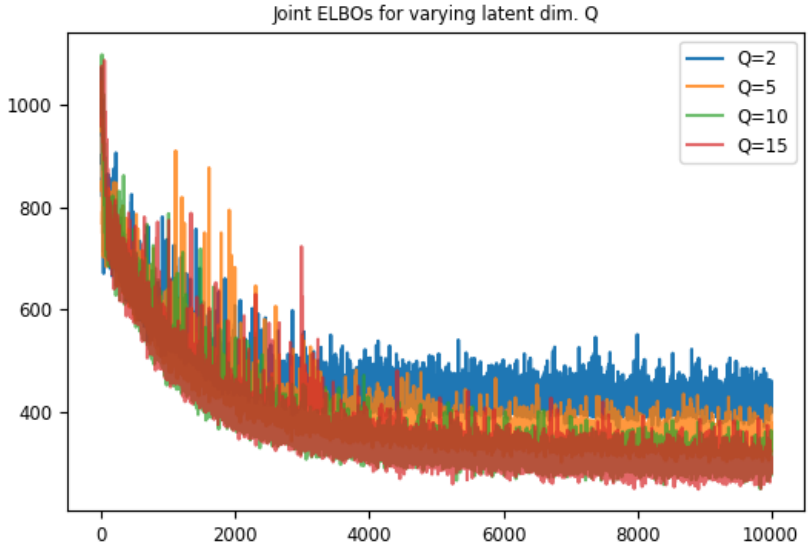

Figure 8. Sensitivity to Q: The negative ELBO objective for varying latent space dimensionality (lower is better)

In fig. 8 we visualise the evolution of the ELBOs across varying latent dimensionality. We notice a meaningful improvement in increasing the dimensionality from $Q = 2$ but very marginal gains beyond $Q = 10$; we use this setting in experiments. It may be important to highlight that due to automatic relevance determination of the squared exponential kernel, setting a high latent dimensionality should not degrade results as the model automatically prunes redundant dimensions by driving the corresponding inverse lengthscales to 0. However, they do increase the compute cost, hence, it is important to set Q at a reasonable value which is flexible enough for structure discovery and not too constrained, while simultaneously minimising the computational burden.