Variational Learning on Aggregate Outputs with Gaussian Processes

Ho Chung Leon Law[*] University of Oxford
Dino Sejdinovic[*] University of Oxford
Ewan Cameron[†] University of Oxford
Tim CD Lucas[†] University of Oxford

Seth Flaxman[‡] Imperial College London
Katherine Battle[†] University Of Oxford
Kenji Fukumizu[§] Institute of Statistical Mathematics

Abstract

While a typical supervised learning framework assumes that the inputs and the outputs are measured at the same levels of granularity, many applications, including global mapping of disease, only have access to outputs at a much coarser level than that of the inputs. Aggregation of outputs makes generalization to new inputs much more difficult. We consider an approach to this problem based on variational learning with a model of output aggregation and Gaussian processes, where aggregation leads to intractability of the standard evidence lower bounds. We propose new bounds and tractable approximations, leading to improved prediction accuracy and scalability to large datasets, while explicitly taking uncertainty into account. We develop a framework which extends to several types of likelihoods, including the Poisson model for aggregated count data. We apply our framework to a challenging and important problem, the fine-scale spatial modelling of malaria incidence, with over 1 million observations.

1 Introduction

A typical supervised learning setup assumes existence of a set of input-output examples $\{(x_\ell, y_\ell)\}_\ell$ from which a functional relationship or a conditional probabilistic model of outputs given inputs can be learned. A prototypical use-case is the situation where obtaining outputs $y_\star$ for new, previously unseen, inputs $x_\star$ is costly, i.e., labelling is expensive and requires human intervention, but measurements of inputs are cheap and automated. Similarly, in many applications, due to a much greater cost in acquiring labels, they are only available at a much coarser resolution than the level at which the inputs are available and at which we wish to make predictions. This is the problem of *weakly supervised* learning on aggregate outputs [14, 20], which has been studied in the literature in a variety of forms, with classification and regression notably being developed separately and without any unified treatment which can allow more flexible observation models. In this contribution, we consider a framework of observation models of aggregated outputs given bagged inputs, which reside in exponential families. While we develop a more general treatment, the main focus in the paper is on the Poisson likelihood for count data, which is motivated by the applications in spatial statistics.

In particular, we consider the important problem of fine-scale mapping of diseases. High resolution maps of infectious disease risk can offer a powerful tool for developing National Strategic Plans,

[*]Department of Statistics, Oxford, UK. <ho.law@stats.ox.ac.uk, dino.sejdinovic@stats.ox.ac.uk>
[†]Big Data Institute, Oxford, UK. <dr.ewan.cameron@gmail.com, timcdlucas@gmail.com, katherine.battle@bdi.ox.ac.uk>
[‡]Department of Mathematics and Data Science Institute, London, UK. <s.flaxman@imperial.ac.uk>
[§]Tachikawa, Japan. <fukumizu@ism.ac.jp>

Preprint.

allowing accurate stratification of intervention types to areas of greatest impact [5]. In low resource settings these maps must be constructed through probabilistic models linking the limited observational data to a suite of spatial covariates (often from remote-sensing images) describing social, economic, and environmental factors thought to influence exposure to the relevant infectious pathways. In this paper, we apply our method to the incidence of clinical malaria cases. Point incidence data of malaria is typically available at a high temporal frequency (weekly or monthly), but lacks spatial precision, being aggregated by administrative district or by health facility catchment. The challenge for risk modelling is to produce fine-scale predictions from these coarse incidence data, leveraging the remote-sensing covariates and appropriate regularity conditions to ensure a well-behaved problem.

Methodologically, the Poisson distribution is a popular choice for modelling count data. In the mapping setting, the intensity of the Poisson distribution is modelled as a function of spatial and other covariates. We use Gaussian processes (GPs) as a flexible model for the intensity. GPs are a widely used approach in spatial modelling but also one of the pillars of Bayesian machine learning, enabling predictive models which explicitly quantify their uncertainty. Recently, we have seen many advances in variational GP posterior approximations, allowing them to couple with more complex observation likelihoods (e.g. binary or Poisson data [21, 17]) as well as a number of effective scalable GP approaches [24, 30, 8, 9], extending the applicability of GPs to dataset sizes previously deemed prohibitive.

Contribution Our contributions can be summarised as follows. A general framework is developed for *aggregated observation models* using exponential families and Gaussian processes. This is novel, as previous work on aggregation or bag models focuses on specific types of output models such as binary classification. Tractable and scalable variational inference methods are proposed for several instances of the aggregated observation models, making use of novel lower bounds on the model evidence. In experiments, it is demonstrated that the proposed methods can scale to dataset sizes of more than 1 million observations. We thoroughly investigate an application of the developed methodology to disease mapping from coarse measurements, where the observation model is Poisson, giving encouraging results. Uncertainty quantification, which is explicit in our models, is essential for this application.

Related Work The framework of learning from aggregate data was believed to have been first introduced in [20], which considers the two regimes of classification and regression. However, while the task of classification of individuals from aggregate data (also known as *learning from label proportions*) has been explored widely in the literature [23, 22, 13, 18, 35, 34, 14], there has been little literature on the analogous regression regime in the machine learning community. Perhaps the closest literature available is [13], who considers a general framework for learning from aggregate data, but also only considers the classification case for experiments. In this work, we will appropriately adjust the framework in [13] and take this to be our baseline. A related problem arises in the spatial statistics community under the name of 'down-scaling', 'fine-scale modelling' or 'spatial disaggregation' [11, 10], in the analysis of disease mapping, agricultural data, and species distribution modelling, with a variety of proposed methodologies (cf. [33] and references therein), including kriging [6]. However, to the best of our knowledge, approaches making use of recent advances in scalable variational inference for GPs are not considered.

Another closely related topic is *multiple instance learning* (MIL), concerned with classification with max-aggregation over labels in a bag, i.e. a bag is positively labeled if at least one individual is positive, and it is otherwise negatively labelled. While the task in MIL is typically to predict labels of new unobserved *bags*, [7] demonstrates that individual labels of a GP classifier can also be inferred in MIL setting with variational inference. Our work parallels that approach, considering bag observation models in exponential families and deriving new approximation bounds for some common generalized linear models. In deriving these bounds, we have taken an approach similar to [17], who considers the problem of Gaussian process-modulated Poisson process estimation using variational inference. However, our problem is made more complicated by the aggregation of labels. Other related research topics include distribution regression and set regression, as in [28, 15, 16] and [36]. In these regression problems, while the input data for learning is the same as the current setup, the goal is to learn a function at the bag level, rather than the individual level, the application of these methods in our setting, naively treating single individuals as "distributions", may lead to suboptimal

performance. An overview of some other approaches for classification using bags of instances is given in [4].

2 Bag observation model: aggregation in mean parameters

Suppose we have a statistical model $p(y|\eta)$ for output $y \in \mathcal{Y}$, with parameter η given by a function of input $x \in \mathcal{X}$, i.e., $\eta = \eta(x)$. Although one can formulate $p(y|\eta)$ in an arbitrary fashion, practitioners often only focus on interpretable simple models, hence we restrict our attention to $p(y|\eta)$ arising from exponential families. We assume that η is the mean parameter of the exponential family.

Assume that we have a fixed set of points $x_i^a \in \mathcal{X}$ such that $\mathbf{x}^a = \{x_1^a, \ldots, x_{N_a}^a\}$ is a *bag* of points with N_a *individuals*, and we wish to estimate the regression value $\eta(x_i^a)$ for each individual. However, instead of the typical setup where we have a paired sample $\{(x_\ell, y_\ell)\}_\ell$ of individuals and their outputs to use as a training set, we observe only *aggregate outputs* y^a for each of the bags. Hence, our training data is of the form

$$(\{x_i^1\}_{i=1}^{N_1}, y^1), \ldots (\{x_i^n\}_{i=1}^{N_n}, y^n), \tag{1}$$

and the goal is to estimate parameters $\eta(x_i^a)$ corresponding to individuals. To relate the aggregate y^a and the bag $\mathbf{x}^a = (x_i^a)_{i=1}^{N_a}$, we use the following *bag observation model*:

$$y^a|\mathbf{x}^a \sim p(y|\eta^a), \qquad \eta^a = \sum_{i=1}^{N_a} w_i^a \eta(x_i^a), \tag{2}$$

where w_i^a is an optional fixed non-negative weight used to adjust the scales (see Section 3 for an example). Note that the aggregation in the bag observation model is on the mean parameters for individuals, not necessarily on the individual responses y_i^a. This implies that each individual contributes to the mean bag response and that the observation model for bags belongs to the same parametric form as the one for individuals. For tractable and scalable estimation, we will use variational methods, as the aggregated observation model leads to intractable posteriors. We consider the Poisson, normal, and exponential distributions, but devote a special focus to the Poisson model in this paper, and refer readers to Appendix A for other cases and experimental results for the Normal model in Appendix H.2.

It is also worth noting that we place no restrictions on the collection of the individuals, with the bagging process possibly dependent on covariates x_i^a or any unseen factors. The bags can also be of different sizes, with potentially the same individuals appearing in multiple bags. After we obtain our individual model $\eta(x)$, we can use it for prediction of in-bag individuals, as well as out-of-bag individuals.

3 Poisson bag model: Modelling aggregate counts

The Poisson distribution $p(y|\lambda) = \lambda^y e^{-\lambda}/(y!)$ is considered for count observations, and this paper discusses the Poisson regression with intensity $\lambda(x_i^a)$ multiplied by a 'population' p_i^a, which is a constant assumed to be known for each individual (or 'sub-bag') in the bag. The population for a bag a is given by $p^a = \sum_i p_i^a$. An observed bag count y^a is assumed to follow

$$y^a|\mathbf{x}^a \sim \text{Poisson}(p^a \lambda^a), \quad \lambda^a := \sum_{i=1}^{N_a} \frac{p_i^a}{p^a} \lambda(x_i^a).$$

Note that, by introducing unobserved counts $y_i^a \sim \text{Poisson}(y_i^a | p_i^a \lambda(x_i^a))$, the bag observation y^a has the same distribution as $\sum_{i=1}^{N_a} y_i^a$ since the Poisson distribution is closed under convolutions. If a bag and its individuals correspond to an area and its partition in geostatistical applications, as in the malaria example in Section 4.2, the population in the above bag model can be regarded as the population of an area or a sub-area. With this formulation, the goal is to estimate the basic intensity function $\lambda(x)$ from the aggregated observations (1). Assuming independence given $\{\mathbf{x}^a\}_a$, the negative log-likelihood (NLL) ℓ_0 across bags is

$$-\log[\Pi_{a=1}^n p(y^a|\mathbf{x}^a)] \stackrel{c}{=} \sum_{a=1}^n p^a \lambda^a - y^a \log(p^a \lambda^a) \stackrel{c}{=} \sum_{a=1}^n \left[\sum_{i=1}^{N_a} p_i^a \lambda(x_i^a) - y^a \log \left(\sum_{i=1}^{N_a} p_i^a \lambda(x_i^a) \right) \right], \tag{3}$$

where $\stackrel{c}{=}$ denotes an equality up to additive constant. During training, this term will pass information from the bag level observations $\{y^a\}$ to the individual basic intensity $\lambda(x_i^a)$. It is noted that once we have trained an appropriate model for $\lambda(x_i^a)$, we will be able to make individual level predictions, and also bag level predictions if desired. We will consider baselines with (3) using penalized likelihoods inspired by manifold regularization in semi-supervised learning [2] – presented in Appendix B. In the next section, we propose a model for λ based on GPs.

3.1 VBAgg-Poisson: Gaussian processes for aggregate counts

Suppose now we model f as a Gaussian process (GP), then we have:

$$y^a|\mathbf{x}^a \sim \text{Poisson}\left(\sum_{i=1}^{N_a} p_i^a \lambda_i^a\right), \qquad \lambda_i^a = \Psi(f(x_i^a)), \qquad f \sim GP(\mu, k) \tag{4}$$

where μ and k are some appropriate mean function and covariance kernel $k(x, y)$. (For implementation, we consider a constant mean function.) Since the intensity is always non-negative, in all models, we will need to use a transformation $\lambda(x) = \Psi(f(x))$, where Ψ is a non-negative valued function. We will consider cases $\Psi(f) = f^2$ and $\Psi(f) = e^f$. A discussion of various choices of this link function in the context of Poisson intensities modulated by GPs is given in [17]. Modelling f with a GP allows us to propagate uncertainty on the predictions to λ_i^a, which is especially important in this weakly supervised problem setting, where we do not directly observe any individual output y_i^a. Since the total number of individuals in our target application of disease mapping is typically in the millions (see Section 4.2), we will approximate the posterior over $\lambda_i^a := \lambda(x_i^a)$ using variational inference, with details found in Appendix E.

For scalability of the GP method, as in previous literature [7, 17], we use a set of inducing points $\{u_\ell\}_{\ell=1}^m$, which are given by the function evaluations of the Gaussian process f at landmark points $W = \{w_1, \ldots, w_m\}$; i.e., $u_\ell = f(w_\ell)$. The distribution $p(u|W)$ is thus given by

$$u \sim N(\mu_W, K_{WW}), \qquad \mu_W = (\mu(w_\ell))_\ell, \quad K_{WW} = (k(w_s, w_t))_{s,t}. \tag{5}$$

The joint likelihood is given by:

$$p(y, f, u|X, W, \Theta) = \prod_{a=1}^{n} \prod_{i=1}^{N_a} \text{Poisson}(y^a|p^a\lambda^a) p(f|u)p(u|W), \text{ with } f|u \sim GP(\tilde{\mu}_u, \tilde{K}), \tag{6}$$

$$\tilde{\mu}(z) = \mu_z + \mathbf{k}_{zW} K_{WW}^{-1}(u - \mu_W), \quad \tilde{K}(z, z') = k(z, z') - \mathbf{k}_{zW} K_{WW}^{-1} \mathbf{k}_{Wz'} \tag{7}$$

where $\mathbf{k}_{zW} = (k(z, w_1), \ldots, k(z, w_\ell))^T$, with μ_W, μ_z denoting their respective evaluations of the mean function μ and Θ being parameters of the mean and kernel functions of the GP. Proceeding similarly to [17], which discusses (non-bag) Poisson regression with GP, we obtain a lower bound of the marginal log-likelihood $\log p(y|\Theta)$:

$$\begin{aligned}
\log p(y|\Theta) &= \log \int\int p(y, f, u|X, W, \Theta) df du \\
&\geq \int\int \log\Big\{p(y|f, \Theta)\frac{p(u|W)}{q(u)}\Big\} p(f|u, \Theta) q(u) df du \quad \text{(Jensen's inequality)} \\
&= \sum_a \int\int \Big\{y^a \log\Big(\sum_{i=1}^{N_a} p_i^a \Psi(f(x_i^a)\Big) - \Big(\sum_{i=1}^{N_a} p_i^a \Psi(f(x_i^a))\Big)\Big\} p(f|u)q(u) df du \\
&\quad - \sum_a \log(y^a!) - KL(q(u)||p(u|W)) =: \mathcal{L}(q, \Theta),
\end{aligned} \tag{8}$$

where $q(u)$ is a variational distribution to be optimized. The general solution to the maximization over q of the evidence lower bound $\mathcal{L}(q, \Theta)$ above is given by the posterior of the inducing points $p(u|y)$, which is intractable. We introduce a restriction to the class of $q(u)$ to approximate the posterior $p(u|y)$. Suppose that the variational distribution $q(u)$ is Gaussian, $q(u) = N(\eta_u, \Sigma_u)$. We then need to maximize the lower bound $\mathcal{L}(q, \Theta)$ over the variational parameters η_u and Σ_u. The resulting $q(u)$ gives an approximation to the posterior $p(u|y)$ which also leads to a Gaussian approximation $q(f) = \int p(f|u)q(u)du$ to the posterior $p(f|y)$, which we finally then transform through Ψ to obtain

the desired approximate posterior on each $\lambda(x_a^i)$ (which is either log-normal or non-central χ^2 depending on the form of Ψ). The approximate posterior on λ will then allow us to make predictions for individuals while, crucially, taking into account the uncertainties in f (note that even the posterior predictive mean of λ will depend on the predictive variance in f due to the nonlinearity Ψ). We also want to emphasis the use of inducing variables is essential for scalability in our model: we cannot directly obtain approximations to the posterior of $\lambda(x_i^a)$ for all individuals, since this is often large in our problem setting (Section 4.2).

As the $p(u|W)$ and $q(u)$ are both Gaussian, the last term (KL-divergence) of (8) can be computed explicitly with exact form found in Appendix E.3. To consider the first two terms, let $q^a(v^a)$ be the marginal normal distribution of $v^a = (f(x_1^a), \ldots, f(x_{N_a}^a))$, where f follows the variational posterior $q(f)$. The distribution of v^a is then $N(m^a, S^a)$, using (7) :

$$m^a = \mu_{\mathbf{x}^a} + K_{\mathbf{x}^a W} K_{WW}^{-1}(\eta_u - \mu_W),\ S^a = K_{\mathbf{x}^a,\mathbf{x}^a} - K_{\mathbf{x}^a W}\left(K_{WW}^{-1} - K_{WW}^{-1}\Sigma_u K_{WW}^{-1}\right) K_{W\mathbf{x}^a} \tag{9}$$

In the first term of (8), each summand is of the form

$$y^a \int \log\Big(\sum_{i=1}^{N_a} p_i^a \Psi\left(v_i^a\right)\Big) q^a(v^a) dv^a - \sum_{i=1}^{N_a} p_i^a \int \Psi\left(v_i^a\right) q^a(v^a) dv^a, \tag{10}$$

in which the second term is tractable for both of $\Psi(f) = f^2$ and $\Psi(f) = e^f$. The integral of the first term, however with q^a Gaussian is not tractable. To solve this, we take different approaches for $\Psi(f) = f^2$ and $\Psi(f) = e^f$; for the former, approximation by Taylor expansion is applied, while for the latter, further lower bound is taken.

First consider the case $\Psi(f) = f^2$, and rewrite the first term of (8) as:

$$y^a \mathbb{E} \log \|V^a\|^2 \quad \text{, where } V^a \sim N(\tilde{m}^a, \tilde{S}^a),$$

with $P^a = diag\left(p_1^a, \ldots, p_{N_a}^a\right), \tilde{m}^a = P^{a\,1/2} m^a$ and $\tilde{S}^a = P^{a\,1/2} S^a P^{a\,1/2}$. By a Taylor series approximation for $\mathbb{E} \log \|V^a\|^2$ (similar to [29]) around $\mathbb{E}\|V^a\|^2 = \|\tilde{m}^a\|^2 + tr\tilde{S}^a$, we obtain

$$\begin{aligned}
&\int \log\Big(\sum_{i=1}^{N_a} p_i^a (v_i^a)^2\Big) q^a(v^a) dv^a \\
&\qquad \approx \log\left(m^{a\top} P^a m^a + tr(S^a P^a)\right) - \frac{2 m^{a\top} P^a S^a P^a m^a + tr\Big((S^a P^a)^2\Big)}{\left(m^{a\top} P^a m^a + tr(S^a P^a)\right)^2} =: \zeta^a.
\end{aligned} \tag{11}$$

with details are in Appendix E.4. An alternative approach which we use for the case $\Psi(f) = e^f$ is to take a further lower bound, which is applicable to a general class of Ψ (we provide further details for the analogous approach for $\Psi(v) = v^2$ in Appendix E.2). We use the following Lemma (proof found in Appendix E.1):

Lemma 1. *Let $v = [v_1, \ldots, v_N]^\top$ be a random vector with probability density $q(v)$ with marginal densities $q_i(v)$, and let $w_i \geq 0$, $i = 1, \ldots, N$. Then, for any non-negative valued function $\Psi(v)$,*

$$\int \log\Big(\sum_{i=1}^{N} w_i \Psi(v_i)\Big) q(v) dv \geq \log\Big(\sum_{i=1}^{N} w_i e^{\xi_i}\Big), \quad where \quad \xi_i := \int \log \Psi(v_i) q_i(v_i) dv_i.$$

Hence we obtain that

$$\int \log\Big(\sum_{i=1}^{N_a} p_i^a e^{v_i^a}\Big) q^a(v^a) dv^a \geq \log\Big(\sum_{i=1}^{N_a} p_i^a e^{m_i^a}\Big), \tag{12}$$

Using the above two approximation schemes, our objective (up to constant terms) can be formulated as: 1) $\Psi(v) = v^2$

$$\mathcal{L}_1^s(\Theta, \eta_u, \Sigma_u, W) := \sum_{a=1}^{n} y^a \zeta^a - \sum_{a=1}^{n} \sum_{i=1}^{N_a} \left\{(m_i^a)^2 + S_{ii}^a/2\right\} - KL(q(u)||p(u|W)), \tag{13}$$

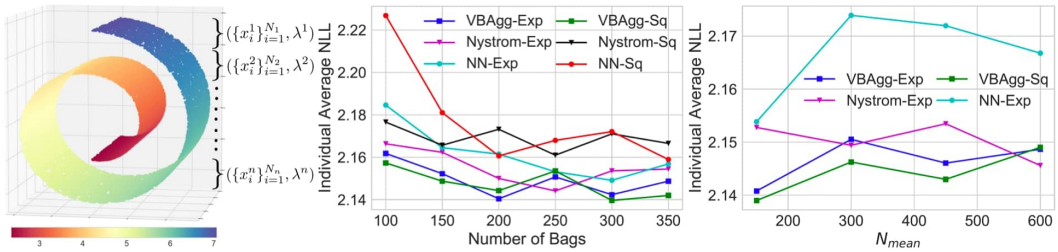

Figure 1: **Left**: Random samples on the Swiss roll manifold. **Middle, Right**: Individual Average NLL on train set for varying number of training bags n and increasing N_{mean}, over 5 repetitions. Constant prediction within bag gives a NLL of 2.22. bag-pixel model gives NLL above 2.4 for the varying number of bags experiment.

2) $\Psi(v) = e^v$

$$\mathcal{L}_1^e(\Theta, \eta_u, \Sigma_u, W) := \sum_{a=1}^{n} y^a \log\Big(\sum_{i=1}^{N_a} e^{m_i^a}\Big) - \sum_{j=1}^{n}\sum_{i=1}^{N_a} e^{m_i^a + S_{ii}^a/2} - KL(q(u)||p(u|W)). \quad (14)$$

Given these objectives, we can now optimise these lower bounds with respect to variational parameters $\{\eta_u, \Sigma_u\}$, parameters Θ of the mean and kernel functions, using stochastic gradient descent (SGD) on bags. Additionally, we might also learn W, locations for the landmark points. In this form, we can also see that the bound for $\Psi(v) = e^v$ has the added computational advantage of not requiring the full computation of the matrix S^a, but only its diagonals, while for $\Psi(v) = v^2$ computation of ζ^a involves full S^a, which may be problematic for extremely large bag sizes.

4 Experiments

We will now demonstrate various approaches: Variational Bayes with Gaussian Process (VBAgg), a MAP estimator of Bayesian Poisson regression with explicit feature maps (Nyström) and a neural network (NN) – the latter two employing manifold regularisation with RBF kernel (unless stated otherwise). For additional baselines, we consider a constant within bag model (constant), i.e. $\hat{\lambda_i^a} = \frac{y^a}{p^a}$ and also consider creating 'individual' covariates by aggregation of the covariates within a bag (bag-pixel). For details of all these approaches, see Appendix B. We also denote $\Psi(v) = e^v$ and v^2 as Exp and Sq respectively.

We implement our models in *TensorFlow*[5] and use SGD with Adam [12] to optimise their respective objectives, and we split the dataset into 4 parts, namely train, early-stop, validation and test set. Here the early-stop set is used for early stopping for the Nyström, NN and bag-pixel models, while the VBAgg approach ignores this partition as it optimises the lower bound to the marginal likelihood. The validation set is used for parameter tuning of any regularisation scaling, as well as learning rate, layer size and multiple initialisations. Throughout, VBAgg and Nyström have access to the same set of landmarks for fair comparison. It is also important to highlight that we perform early stopping and tuning based on *bag* level performance on NLL only, as this is the only information available to us.

For the VBAgg model, there are two approaches to tuning, one approach is to choose parameters based on NLL on the validation bag sets, another approach is to select all parameters based on the training objective $\mathcal{L}_1$, the lower bound to the marginal likelihood. We denote the latter approach VBAgg-Obj and report its toy experimental results in Appendix H.1.1 for presentation purposes. In general, the results are relatively *insensitive* to this choice, especially when $\Psi(v) = v^2$. To make predictions, we use the mean of our approximated posterior (provided by a log-normal and non-central χ^2 distribution for Exp and Sq). As an additional evaluation, we report mean square error (MSE) and bag performance results in Appendix H.

[5]Code will be available for use.

4.1 Poisson Model: Swiss Roll

We first demonstrate our method on the swiss roll dataset[6], illustrated in Figure 1 (left). To make this an aggregate learning problem, we first construct n bags with sizes drawn from a negative binomial distribution $N_a \sim NB(N_{mean}, N_{std})$, where N_{mean} and N_{std} represents the respective mean and standard deviation of N_a. We then randomly select $\sum_{a=1}^{n} N_a$ points from the swiss roll manifold to be the locations, giving us a set of colored locations in $\mathbb{R}^3$. Ordering these random locations by their z-axis coordinate, we group them, filling up each bag in turn as we move along the z-axis. The aim of this is to simulate that in real life the partitioning of locations into bags is often not independent of covariates. Taking the colour of each location as the underlying rate λ_i^a at that location, we simulate $y_i^a \sim Poisson(\lambda_i^a)$, and take our observed outputs to be $y^a = \sum_{i=1}^{N_a} y_i^a \sim Poisson(\lambda^a)$, where $\lambda^a = \sum_{i=1}^{N_a} \lambda_i^a$. Our goal is then to predict the underlying individual rate parameter λ_i^a, given only bag-level observations y^a. To make this problem even more challenging, we embed the data manifold into $\mathbb{R}^{18}$ by rotating it with a random orthogonal matrix. For the choice of k for VBAgg and Nyström, we use the RBF kernel, with the bandwidth parameter learnt. For landmark locations, we use the K-means++ algorithm, so that landmark points lie evenly across the data manifold.

Varying number of Bags: n To see the effect of increasing number of bags available for training, we fix $N_{mean} = 150$ and $N_{std} = 50$, and vary the number of bags n for the training set from 100 to 350 with the same number of bags for early stopping and validation. Each experiment is repeated for 5 runs, and results are shown in Figure 1 for individual NLL on the train set. Again we emphasise that the individual labels are not used in training. We see that all versions of VBAgg outperform all other models, in terms of MSE and NLL, with statistical significance confirmed by a signed rank permutation test (see Appendix H.1.1). We also observe that the bag-pixel model has poor performance, as a result of losing individual level covariate information in training by simply aggregating them.

Varying number of individuals per bag: N_{mean} To study the effect of increasing bag sizes (with larger bag sizes, we expect "disaggregation" to be more difficult), we fix the number of training bags to be 600 with early stopping and validation set to be 150 bags, while varying the number of individuals per bag through N_{mean} and N_{std} in the negative binomial distribution. To keep the relative scales between N_{mean} and N_{std} the same, we take $N_{std} = N_{mean}/2$. The results are shown in Figure 1, focusing on the best performing methods in the previous experiment. Here, we observe that VBAgg models again perform better than the Nyström and NN models with statistical significance as reported in Appendix H.1.1, with performance stable as N_{mean} increases.

Discussion To gain more insight into the VBAgg model, we look at the calibration of our two different Bayesian models: VBAgg-Exp and VBAgg-Square. We compute their respective posterior quantiles and observe the ratio of times the true λ_i^a lie in these quantiles. We present these in Appendix H.1.1. The calibration plots reveal an interesting nature about using the two different approximations for using e^v versus v^2 for $\Psi(v)$. While experiments showed that the two model perform similarly in terms of NLL, the calibration of the models is very different. While the VBAgg-Square is well calibrated in general, the VBAgg-Exp suffers from poor calibration. This is not surprising, as VBAgg-Exp uses an additional lower bound on model evidence. Thus, uncertainty estimates given by VBAgg-Exp should be treated with care.

4.2 Malaria Incidence Prediction

We now demonstrate the proposed methodology on an important real life malaria prediction problem for an endemic country from the Malaria Atlas Project database[7]. In this problem, we would like to predict the underlying malaria incidence rate in each 1km by 1km region (referred to as a pixel), while having only observed aggregated incidences of malaria y^a at much larger regional levels, which are treated as bags of pixels. These bags are non-overlapping administrative units, with N_a pixels per bag ranging from 13 to 6,667, with a total of 1,044,683 pixels. In total, data is available for 957 bags[8].

[6]The swiss roll manifold function (for sampling) can be found on the Python *scikit-learn* package.

[7]Due to confidentiality reasons, we do not report country or plot the full map of our results.

[8]We consider 576 bags for train, 95 bags each for validation and early-stop, with 191 bags for testing, with different splits across different trials, selecting them to ensure distributions of labels are similar across sets.

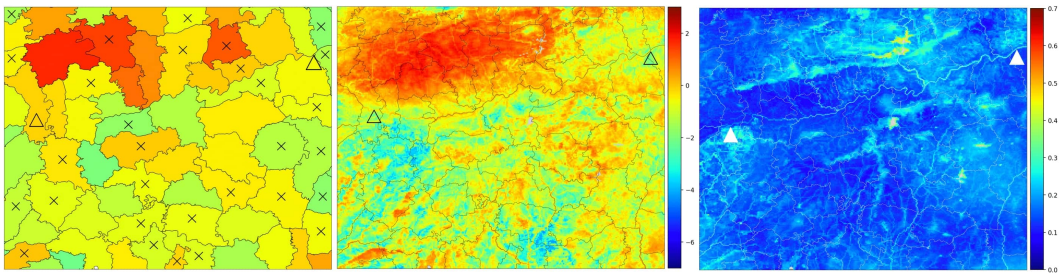

Figure 2: Triangle denotes approximate start and end of river location, crosses denotes non-train set bags. Malaria incidence rate λ_i^a is per 1000 people. **Left, Middle**: $\log(\hat{\lambda}_i^a)$, with constant model (Left), and VBAgg-Obj-Sq (tuned on $\mathcal{L}_1^s$) (Middle). **Right**: Standard deviation of the posterior v in (9) with VBAgg-Obj-Sq.

Along with these pixels, we also have population estimates p_i^a (per 1000 people) for pixel i in bag a, spatial coordinates given by s_i^a, as well as covariates $x_i^a \in \mathbb{R}^{18}$, collected by remote sensing. Some examples of covariates includes accessibility, distance to water, mean of land surface temperature and stable night lights. It is clear that rather than expecting malaria incidence rate to be constant throughout the entire bag (as in Figure 2), we expect pixel incidence rate to vary, depending on social, economic and environmental factors [32]. Our goal is therefore to build models that can predict malaria incidence rates at a *pixel* level.

We assume a Poisson model on each individual pixel, i.e. $y^a \sim Poisson(\sum_i p_i^a \lambda_i^a)$, where λ_i^a is the underlying pixel incidence rate of malaria per 1000 people that we are interested in predicting. We consider the VBAgg, Nyström and NN as prediction models and use a kernel given as a sum of an ARD (automatic relevance determination) kernel on covariates and a Matérn kernel on spatial locations for the VBAgg and Nyström methods, learning all kernel parameters (the kernel expression is provided in Appendix G). We use the same kernel for manifold regularisation in the NN model. This kernel choice incorporates spatial information, while allowing feature selection amongst other covariates. For choice of landmarks, we ensure landmarks are placed evenly throughout space by using one landmark point per training bag (selected by k-means++). This is so that the uncertainty estimates we obtain are not too sensitive to the choice of landmarks. In this problem, no individual-level labels are available, so we report Bag NLL and MSE (on observed incidences) on the test bags in Appendix G over 10 different re-splits of the data. Although we can see that Nyström is the best performing method, the improvement over VBAgg models is not statistically significant. On the other hand, both VBAgg and Nyström models statistically significantly outperform NN, which also has some instability in its predictions, as discussed in Appendix G.1. However, a caution should be exercised when using the measure of performance at the bag level as a surrogate for the measure of performance at the individual level: in order to perform well at the bag level, one can simply utilise spatial coordinates and ignore other covariates, as malaria intensity appears to smoothly vary between the bags (Left of Figure 2). However, we do not expect this to be true at the individual level.

To further investigate this, we consider a particular region, and look at the predicted individual malaria incidence rate, with results found in Figure 2 and in Appendix G.1 across 3 different data splits, where the behaviours of each of these models can be observed. While Nyström and VBAgg methods both provide good bag-level performance, Nyström and VBAgg-Exp can sometimes provide overly-smooth spatial patterns, which does not seem to be the case for the VBAgg-Sq method (recall that VBAgg-Sq performed best in both prediction and calibration for the toy experiments). In particular, VBAgg-Sq consistently predicts higher intensity along rivers (a known factor [31]; indicated by triangles in Figure 2) using only coarse aggregated intensities, demonstrating that prediction of (unobserved) pixel-level intensities is possible using fine-scale environmental covariates, especially ones known to be relevant such as covariates indicated by the Topographic Wetness Index, a measure of wetness, see Appendix G.2 for more details.

In summary, by optimising the lower bound to the marginal likelihood, the proposed variational methods are able to learn useful relations between the covariates and pixel level intensities, while avoiding the issue of overfitting to spatial coordinates. Furthermore, they also give uncertainty estimates (Figure 2, right), which are essential for problems like these, where validation of predictions is difficult, but they may guide policy and planning.

5 Conclusion

Motivated by the vitally important problem of malaria, which is the direct cause of around 187 million clinical cases [3] and 631,000 deaths [5] each year in sub-Saharan Africa, we have proposed a general framework of *aggregated observation models* using Gaussian processes, along with scalable variational methods for inference in those models, making them applicable to large datasets. The proposed method allows learning in situations where outputs of interest are available at a much coarser level than that of the inputs, while explicitly quantifying uncertainty of predictions. The recent uptake of digital health information systems offers a wealth of new data which is abstracted to the aggregate or regional levels to preserve patient anonymity. The volume of this data, as well as the availability of much more granular covariates provided by remote sensing and other geospatially tagged data sources, allows to probabilistically disaggregate outputs of interest for finer risk stratification, e.g. assisting public health agencies to plan the delivery of disease interventions. This task demands new high-performance machine learning methods and we see those that we have developed here as an important step in this direction.

Acknowledgement

We thank Kaspar Martens for useful discussions, and Dougal Sutherland for providing the code base in which this work was based on. HCLL is supported by the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). HCLL and KF are supported by JSPS KAKENHI 26280009. EC and KB are supported by OPP1152978, TL by OPP1132415 and the MAP database by OPP1106023. DS is supported in part by the ERC (FP7/617071) and by The Alan Turing Institute (EP/N510129/1). The data were provided by the Malaria Atlas Project supported by the Bill and Melinda Gates Foundation.

References

[1] LU Ancarani and G Gasaneo. Derivatives of any order of the confluent hypergeometric function f 1 1 (a, b, z) with respect to the parameter a or b. *Journal of Mathematical Physics*, 49(6):063508, 2008.

[2] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[3] Samir Bhatt, DJ Weiss, E Cameron, D Bisanzio, B Mappin, U Dalrymple, KE Battle, CL Moyes, A Henry, PA Eckhoff, et al. The effect of malaria control on plasmodium falciparum in africa between 2000 and 2015. *Nature*, 526(7572):207, 2015.

[4] Veronika Cheplygina, David M.J. Tax, and Marco Loog. On classification with bags, groups and sets. *Pattern Recognition Letters*, 59:11 – 17, 2015.

[5] Peter W Gething, Daniel C Casey, Daniel J Weiss, Donal Bisanzio, Samir Bhatt, Ewan Cameron, Katherine E Battle, Ursula Dalrymple, Jennifer Rozier, Puja C Rao, et al. Mapping plasmodium falciparum mortality in africa between 1990 and 2015. *New England Journal of Medicine*, 375(25):2435–2445, 2016.

[6] Pierre Goovaerts. Combining areal and point data in geostatistical interpolation: Applications to soil science and medical geography. *Mathematical Geosciences*, 42(5):535–554, Jul 2010.

[7] Manuel Haußmann, Fred A Hamprecht, and Melih Kandemir. Variational bayesian multiple instance learning with gaussian processes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6570–6579, 2017.

[8] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. 2013.

[9] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable Variational Gaussian Process Classification. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360, San Diego, California, USA, 09–12 May 2015. PMLR.

[10] Richard Howitt and Arnaud Reynaud. Spatial disaggregation of agricultural production data using maximum entropy. *European Review of Agricultural Economics*, 30(3):359–387, 2003.

[11] Petr Keil, Jonathan Belmaker, Adam M Wilson, Philip Unitt, and Walter Jetz. Downscaling of species distribution models: a hierarchical approach. *Methods in Ecology and Evolution*, 4(1):82–94, 2013.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597–606. ACM, 2015.

[14] H. Kueck and N. de Freitas. Learning about individuals from group statistics. In *UAI*, pages 332–339, 2005.

[15] H. C. L. Law, C. Yau, and D. Sejdinovic. Testing and learning on distributions with symmetric noise invariance. In *NIPS*, 2017.

[16] Ho Chung Leon Law, Dougal Sutherland, Dino Sejdinovic, and Seth Flaxman. Bayesian approaches to distribution regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1167–1176, 2018.

[17] Chris Lloyd, Tom Gunter, Michael Osborne, and Stephen Roberts. Variational inference for gaussian process modulated poisson processes. In *International Conference on Machine Learning*, pages 1814–1822, 2015.

[18] Vitalik Melnikov and Eyke Hüllermeier. Learning to aggregate using uninorms. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 756–771. Springer, 2016.

[19] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyonds. *arXiv preprint arXiv:1605.09522*, 2016.

[20] David R Musicant, Janara M Christensen, and Jamie F Olson. Supervised learning by training on aggregate outputs. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 252–261. IEEE, 2007.

[21] H. Nickisch and CE. Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, October 2008.

[22] Giorgio Patrini, Richard Nock, Tiberio Caetano, and Paul Rivera. (Almost) no label no cry. In *NIPS*. 2014.

[23] Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. *JMLR*, 10:2349–2374, 2009.

[24] Joaquin Quiñonero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959, December 2005.

[25] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2007.

[26] Carl Edward Rasmussen and Christopher KI Williams. Gaussian processes for machine learning, 2006.

[27] Alex J Smola and Peter L Bartlett. Sparse greedy gaussian process regression. In *Advances in neural information processing systems*, pages 619–625, 2001.

[28] Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

[29] Yee W Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360, 2007.

[30] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[31] DA Warrel, T Cox, J Firth, and Jr E Benz. Oxford textbook of medicine, 2017.

[32] Daniel J Weiss, Bonnie Mappin, Ursula Dalrymple, Samir Bhatt, Ewan Cameron, Simon I Hay, and Peter W Gething. Re-examining environmental correlates of plasmodium falciparum malaria endemicity: a data-intensive variable selection approach. *Malaria journal*, 14(1):68, 2015.

[33] António Xavier, Maria de Belém Costa Freitas, Maria do Socorro Rosário, and Rui Fragoso. Disaggregating statistical data at the field level: An entropy approach. *Spatial Statistics*, 23:91 – 108, 2018.

[34] Felix X Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014.

[35] Felix X Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. $propto$ svm for learning with label proportions. *arXiv preprint arXiv:1306.0886*, 2013.

[36] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *NIPS*, 2017.

A Aggregated Exponential Family Models

Consider an observation model of the form

$$p(y|\theta) = \exp\left(\frac{y\theta - c(\theta)}{\tau}\right) h(y, \tau), \tag{15}$$

where response y is one-dimensional, θ is a natural parameter corresponding to the statistic y, τ is a dispersion parameter, and $h(y, \tau)$ is base measure. For simplicity, we will assume that natural parameters corresponding to the other parts of the sufficient statistic are fixed and folded into the base measure. Let η be the corresponding mean parameter, i.e.

$$\eta = \mathbb{E}_\theta y = \int y p(y|\theta) dy$$

and $\theta = F(\eta)$ be the link function mapping from mean to the natural parameters and $G(\theta)$ its inverse. We wish to model the mean parameter $\eta = \eta(x)$ using a Gaussian process on a domain $\mathcal{X}$ together with a function Ψ which transforms the GP value to the natural parameter space, i.e.

$$\eta(x) = \Psi(f(x)), \qquad f \sim \mathcal{GP}(\mu, k). \tag{16}$$

For example, the mean parameter for some models is restricted to the positive part of the real line, while the GP values cover the whole real line. We will consider the following examples:

- **Normal** (with fixed variance). $F = G = idenity$ and Ψ can be identity, too, as there are no restrictions on the mean parameter space.
- **Poisson**. $F(\eta) = \log \eta$, $G(\theta) = e^\theta$. Ψ should take a positive value, so we consider $\Psi(v) = e^v$ or $\Psi(v) = v^2$.
- **Exponential**. $p(y|\eta) = \exp(-y/\eta)/\eta$ and $\theta = -\eta$, $F(\eta) = -1/\eta$, $G(\theta) = -1/\theta$. Ψ should take a positive value, so we consider $\Psi(v) = e^v$ or $\Psi(v) = v^2$

Note that the link function F is concave for all the examples above.

A.1 Bag model

We will consider the aggregation in the mean parameter space. Namely, let $y^1, \ldots, y^n$ be n independent aggregate responses for each of the n bags of covariates $\mathbf{x}^a = \{x_1^a, \ldots, x_{N_a}^a\}$, $a = 1, \ldots, n$. We assume the following aggregation model:

$$y^a \sim p(y|\eta_a), \quad \eta^a = \sum_{i=1}^{N_a} w_i^a \eta_i^a = \sum_{i=1}^{N_a} w_i^a \Psi(f(x_i^a)), \quad a = 1, \ldots, n. \tag{17}$$

where w_i^a are fixed weights to adjust the scales among the individuals and the bag (e.g., adjusting for population size).

We also can model individual (unobserved) variables y_i^a ($i = 1, \ldots, N_a$), which follow:

$$y_i^a \sim p(y|\eta_i^a), \qquad \eta_i^a = \Psi(f(x_i^a)), \quad i = 1, \ldots, N_a,\ a = 1, \ldots, n. \tag{18}$$

Note that we consider aggregation in mean parameters of responses, not in the responses themselves. If we consider a case where underlying individual responses y_i^a aggregate to y^a as a weighted sum, the form of the bag likelihood and individual likelihood would be different unless we restrict attention to distribution families which are closed under both scaling and convolution. However, when aggregation occurs in the mean parameter space, the form of the bag likelihood and individual likelihood is always the same. This corresponds to the following measurement process:

- Each individual has a mean parameter η_i^a - if it were possible to sample a response for that particular individual, we would obtain a sample $y_i^a \sim p(\cdot|\eta_i^a)$
- However, we cannot sample the individual and we can only observe a bag response. But in that case, only a single bag response is taken and depends on all individuals simultaneously. Each individual contributes in terms of an increase in a mean bag response, but this measurement process is different from the two-stage procedure by which we aggregate individual responses.

A.2 Marginal likelihood and ELBO

Let $Y = (y^1, \ldots, y^n)$ (bag observations). With the inducing points $u = f(W)$, the marginal likelihood is

$$p(Y) = \int \int \prod_{a=1}^{n} p(y^a|\eta^a) p(f|u) p(u) du df. \tag{19}$$

The evidence lower bound can be derived as

$$\begin{aligned}
\log p(Y) &= \log \int \int \Big\{ \prod_{a=1}^{n} p(y^a|\eta^a) \frac{p(u)}{q(u)} \Big\} p(f|u) q(u) du df \\
&\geq \int \int \log \Big\{ \prod_{a=1}^{n} p(y^a|\eta^a) \frac{p(u)}{q(u)} \Big\} p(f|u) q(u) du df \\
&= \sum_{a=1}^{n} \frac{y^a}{\tau} \int F\Big(\sum_i w_i^a \Psi(f(x_i^a)) \Big) q(f) df - \int c\Big(F\Big(\sum_i w_i^a \Psi(f(x_i^a)) \Big) \Big) q(f) df \\
&\quad - \int q(u) \log \frac{q(u)}{p(u)} du,
\end{aligned} \tag{20}$$

where $q(f) = \int p(f|u) q(u) du$.

By setting the variational distribution $q(u)$ as Gaussian, the third term is tractable. The first and second terms are however tractable only in limited cases. The cases we develop are the Poisson bag model, described in the main text, as well as the normal bag model and the exponential bag model, described below.

A.3 Normal bag model

F is identity and $c(\theta) = \theta^2/2$, which makes both the first and the second terms tractable with the choice of $\Psi(v) = v$. Moreover, the viewpoints of aggregating in the mean parameters and in the individual responses are equivalent for this model and we can also allow different variance parameters for different bags (and individuals).

Consider a bag a of items $\{x_i^a\}_{i=1}^{N_a}$. Each item x_i^a is assumed to have a weight w_i^a. At the individual level, we model the (unobserved) responses y_i^a as

$$y_i^a | x_i^a \sim \mathcal{N}\left(w_i^a \mu_i^a, (w_i^a)^2 \tau_i^a \right) \tag{21}$$

where $\mu_i^a = \mu(x_i^a)$, thus μ_i^a is a *mean parameter per unit weight* corresponding to the item x_i^a and it is assumed to be a function of both x_i^a. Similarly, τ_i^a is a variance parameter per unit weight. At the bag level, we consider the following model for the observed aggregate response y^a, assuming conditional independence of individual responses given covariates $\mathbf{x}^a = \{x_1^a, \ldots, x_{N_a}^a\}$:

$$y^a = \sum_{i=1}^{N_a} y_i^a, \text{ i.e. } y^a | \mathbf{x}^a \sim \mathcal{N}(w^a \mu^a, (w^a)^2 \tau^a), \qquad \mu^a = \sum_{i=1}^{N_a} \frac{w_i^a}{w^a} \mu_i^a, \tau^a = \frac{\sum_{i=1}^{N_a} (w_i^a)^2 \tau_i^a}{(w^a)^2} \tag{22}$$

where μ^a and τ^a are the mean and variance parameters per unit weight of the whole bag a and $w^a = \sum_{i=1}^{N_a} w_i^a$ is the *total weight* of bag a. Although we can take τ_i^a to also be a function of the covariates, here for simplicity, we take $\tau_i^a = \tau^a$ to be constant per bag (note the abuse of notation). We can now compute the negative log-likelihood (NLL) across bags (assuming conditional independence given the $\mathbf{x}^a$):

$$\ell_0 = -\log \left[\Pi_{a=1}^{n} p(y^a|\mathbf{x}^a) \right] = \frac{1}{2} \sum_{a=1}^{n} \left\{ \log \left(2\pi\tau^a \sum_{i=1}^{N_a} (w_i^a)^2 \right) + \frac{\left(y^a - \sum_{i=1}^{N_a} w_i^a \mu_i^a \right)^2}{\sum_{i=1}^{N_a} (w_i^a)^2 \tau^a} \right\} \tag{23}$$

where $\mu_i^a = f(x_i^a)$ is the function we are interested in, and τ^a are the variance parameters to be learnt.

We can now consider the lower bound to the marginal likelihood as below (assuming $w_i^a = 1$ here to simplify notation, while the analogous expression with non-uniform weights is straightforward):

$$\begin{aligned}
\log p(y|\Theta) &= \log \int\int p(y, f, u|X, W, \Theta) df du \\
&= \log \int\int \left(\prod_{a=1}^{n} \frac{1}{\sqrt{2\pi N_a \tau^a}} \exp\left(-\frac{(y^a - \sum_{i=1}^{N_a} f(x_i^a))^2}{2N_a\tau^a} \right) \right) \frac{p(u|W)}{q(u)} p(f|u)q(u) df du \\
&\geq \int\int \log \left\{ \prod_{a=1}^{n} \frac{1}{\sqrt{2\pi N_a \tau^a}} \exp\left(-\frac{(y^a - \sum_{i=1}^{N_a} f(x_i^a))^2}{2N_a\tau^a} \right) \frac{p(u|W)}{q(u)} \right\} p(f|u)q(u) df du \\
&= -\frac{1}{2} \sum_a \int\int \left\{ \frac{(y^a)^2 - 2y^a \sum_{i=1}^{N_a} f(x_i^a) + \left(\sum_{i=1}^{N_a} f(x_i^a) \right)^2}{N_a \tau^a} \right\} p(f|u)q(u) df du \\
&\quad - \frac{1}{2} \sum_a \log(2\pi N_a \tau^a) - \int q(u) \log \frac{q(u)}{p(u|W)} du. \qquad (24)
\end{aligned}$$

Using again a Gaussian distribution for $q(u)$, we have $q(f) = \int p(f|u)q(u)du$, which is a normal distribution and let $q^a(f^a)$ be its marginal normal distribution of $f^a = (f(x_1^a), \ldots, f(x_{N_a}^a))$ with mean and covariance given by m^a and S^a as before in (9).

Then all expectations with respect to $q(f)$ are tractable and the ELBO is simply

$$\begin{aligned}
\mathcal{L}(q, \theta) = -\frac{1}{2} \sum_{a=1}^{n} \left\{ \frac{(y^a)^2 - 2y^a \mathbf{1}^\top m^a + \mathbf{1}^\top \left(S^a + m^a (m^a)^\top \right) \mathbf{1}}{N_a \tau^a} \right\} - \frac{1}{2} \sum_a \log(2\pi N_a \tau^a) \\
- KL(q(u)||p(u|W)). \qquad (25)
\end{aligned}$$

A.4 Exponential bag model

In this case, we have $F(\eta) = -1/\eta$. We can apply the similar argument as in Lemma 1. For any $\alpha_i > 0$ with $\sum_i \alpha_i = 1$, by the concavity of F,

$$\begin{aligned}
\int F\left(\sum_i w_i \Psi(v_i) \right) q(v_i) dv_i &= \int F\left(\sum_i \alpha_i w_i / \alpha_i \Psi(v_i) \right) q(v_i) dv_i \\
&\geq \int \sum_i \alpha_i F\left(w_i / \alpha_i \Psi(v_i) \right) q(v_i) dv_i \\
&= \sum_i \alpha_i \int F\left(w_i / \alpha_i \Psi(v_i) \right) q(v_i) dv_i.
\end{aligned}$$

For $F(\eta) = -1/\eta$, the last line is equal to

$$\sum_i \frac{\alpha_i^2}{w_i} \int \frac{1}{\Psi(v_i)} q(v_i) dv_i.$$

When using a normal q, this is tractable for several choices of Ψ including e^v and v^2. If we let $\xi_i := \int \frac{1}{\Psi(v_i)} q(v_i) dv_i$, and maximize

$$\sum_i \alpha_i^2 \frac{\xi_i}{w_i}$$

under the constraint $\sum_i \alpha_i = 1$, we obtain

$$\alpha_i = \frac{(w_i/\xi_i)}{\sum_\ell (w_i/\xi_i)}.$$

Finally, we have a lower bound

$$\int F\left(\sum_i w^i \Psi(v_i)\right) q(v_i) dv_i \geq -\frac{\sum_i (w_i/\xi_i)}{\sum_i (w_i/\xi_i)^2} \tag{26}$$

where

$$\xi_i = \int \frac{1}{\Psi(v_i)} q(v_i) dv_i.$$

which is tractable for a Gaussian variational family. Also with an explicit form of Ψ, it is easy to take the derivatives of the resulting lower bound with respect to the variational parameters in $q(v)$.

B Alternative approaches

Constant For the Poisson model, we can take $\lambda_i^a = \lambda_c^a$, a constant rate across the bag, then:

$$\hat{\lambda}_c^a = \frac{y^a}{p^a}$$

hence the individual level predictive distribution is the form $y_i^a \sim Poisson(\hat{\lambda}_c^a)$, and for unseen bag r, $\hat{\lambda}_c^{\text{bag}} = \frac{1}{\sum_{a=1}^n p^a} \sum_{a=1}^n y^a$, with predictive distribution given by $y^r \sim Poisson(p^r \hat{\lambda}_c^{\text{bag}})$.

bag-pixel: Bag as Individual Another baseline is to train a model from the weighted average of the covariates, given by $x^a = \sum_{i=1}^{N_a} \frac{p_i^a}{p^a} x_i^a$ in the Poisson case, and $x^a = \sum_{i=1}^{N_a} \frac{w_i^a}{w^a} x_i^a$ in the normal case. The purpose of this baseline is to demonstrate that modelling at the individual level is important during training. Since we now have labels and covariates at the bag level, we can consider the following model:

$$y^a | x^a \sim Poisson(p^a \lambda(x^a))$$

with $\lambda(x^a) = \Psi(f(x^a))$ for the Poisson model. For the normal model, we have:

$$y^a | x^a \sim Normal(w^a \mu(x^a), (w^a)^2 \tau)$$

where $\mu(x^a) = f(x^a)$ and τ is a parameter to be learnt (assuming constant across bags). Now we observe that these models are identical to the individual model, except for a difference in indexing. Hence, after learning the function f at the bag level, we can transfer the model to the individual level. Essentially here we have created fake individual level instances by aggregation of individual covariates inside a bag.

Nyström: Bayesian MAP for Poisson regression on explicit feature maps Instead of the posterior based on the model (6), we can also consider an explicit feature map in order to directly construct a MAP estimator. While this method does not provide posterior uncertainty over λ_i^a, it does provide an interesting connection to the settings we have considered and also manifold-regularized neural networks, as discussed below. Let K_{zz} be the covariance function defined on covariates $\{z_1, \dots z_n\}$, and consider its low rank approximation $K_{zz} \approx \mathbf{k}_{zW} K_{WW}^{-1} \mathbf{k}_{Wz}$ with landmark points $W = \{w_\ell\}_{\ell=1}^m$ and $\mathbf{k}_{zW} = (k(z, w_1), \dots, k(z, w_\ell))^T$. By using landmark points W, we have avoided computation of the full kernel matrix, reducing computational complexity. Under this setup, we have that $K_{zz} \approx \Phi_z \Phi_z^\top$, with $\Phi_z = \mathbf{k}_{zW} K_{WW}^{-\frac{1}{2}}$ being the explicit (Nyström) feature map. Using this explicit feature map Φ, we have the following model:

$$f_i^a = \phi_i^a \beta, \qquad \beta \sim \mathcal{N}(0, \gamma^2 I)$$

$$y^a | \mathbf{x}^a \sim \text{Poisson}\left(\sum_{i=1}^{N_a} p_i^a \lambda(x_i^a)\right), \qquad \lambda(x_i^a) = \Psi(f_i^a),$$

where γ is a prior parameter and ϕ_i^a is the corresponding i^{th} row of $\Phi_{\mathbf{x}^a}$. We can then consider a MAP estimator of the model coefficients β:

$$\hat{\beta} = \text{argmax}_\beta \log[\Pi_{a=1}^n p(y^a | \beta, \mathbf{x}^a)] + \log p(\beta). \tag{27}$$

This essentially recovers the same model as in (3) with the standard l_2 loss regularising the complexity of the function. This model can be thought of in several different ways, for example as a weight space view of the GP ([26] for an overview), or as a MAP of the Subset of Regressors (SoR) approximation [27] of the GP when $\sigma = 1$. Additional we may include manifold regulariser as part of the prior, see discussion below about neural network.

NN: Manifold-regularized neural networks The next approach we consider is a parametric model for f as in [13], and search the best parameter to minimize negative log-likelihood ℓ_0 across bags. This paper considers a neural network with parameters θ for the model f, and uses the back-propagation to learn θ and hence individual level model f. However, since we only have aggregated observations at the bag level, but lots of individual covariate information, it is useful to incorporate this information also, by enforcing smoothness on the data manifold given by the unlabelled data. To do this, following [13] and [22], we pursue a semisupervised view of the problem and include an additional manifold regularisation term [2] (rescaling with N_{total}^2 during implementation):

$$\ell_1 = \sum_{w=1}^{N_{\text{total}}} \sum_{u=1}^{N_{\text{total}}} (f_u - f_w)^2 k_L(x_u, x_w) = \mathrm{f}^\top \mathrm{L}\, \mathrm{f} \tag{28}$$

where we have suppressed the bag index, N_{total} represents the total number of individuals, $k_L(\cdot,\cdot)$ is some user-specified kernel[9], $\mathrm{f} = [f_1, \ldots, f_{N_{\text{total}}}]^\top$, L is the Laplacian defined as $\mathrm{L} = diag(\mathrm{K_L}\mathbb{1}^\top) - \mathrm{K_L}$, where $\mathbb{1}$ is just $[1, \ldots, 1]$ and $\mathrm{K_L}$ is a kernel matrix. Although this term involves calculation of a kernel matrix across individuals, in practice we consider stochastic gradient descent (SGD) and also random Fourier features [25] or Nyström approximation (see Appendix C), with scale parameter λ_1 to control the strength of the regularisation. Similarly, one can also consider manifold regularisation at the bag level, if bag-level covariates/embeddings are available, for further details, see Appendix D.

In fact, the same regularisation can be applied to the MAP estimation with the explicit feature maps. This is equivalent to having a prior $\beta \sim \mathcal{N}(0, \sigma^2 I + (\lambda_1 \Phi^\top \mathrm{L} \Phi)^{-1})$ that is data dependent and incorporates the structure of the manifold [10].

For implementation, we consider a one hidden layer neural network, with also an output layer, for a fair comparison to the Nyström approach. For activation function, we consider the Rectified Linear Unit (ReLU).

MAP estimation of GP We introduce $p(f,u) = p(f|u)p(u|W)$ and consider the posterior given by $p(u|f,y,w,\theta)$, where here the conditional distribution $f|u$ is given by:

$$f|u \sim GP(\tilde{\mu}_u, \tilde{K}), \tag{29}$$

$$\tilde{\mu}(z) = \mu_z + \mathbf{k}_{zW} K_{WW}^{-1}(u - \mu_W), \quad \tilde{K}(z,z') = k(z,z') - \mathbf{k}_{zW} K_{WW}^{-1} \mathbf{k}_{\mathbf{W}\mathbf{z}'}$$

where $\mathbf{k}_{zW} = (k(z,W_1), \ldots, k(z,W_\ell))^T$. Using Bayes rule, we obtain:

$$\begin{aligned}
\log[p(u|f,y,w)] &= \log[p(y|f,u)p(f,u|X,W)] \\
&= \log[p(y|f)p(f|u,X)p(u|W)] \\
&= \sum_{a=1}^{n} y^a \log(p^a \lambda^a) + \sum_{a=1}^{n} p^a \lambda^a - \sum_{a=1}^{n} \log(y^a!) + \log(p(f|u,X)) + \log(p(u|W))
\end{aligned}$$

where $p(f|u,X) \sim \mathcal{N}(\tilde{\mu}_u, \tilde{K})$ given by above, and $p(u|W) \sim \mathcal{N}(\mu_W, \Sigma_{WW})$, i.e.

$$\log p(f|u,X) + \log p(u|W) = -\frac{1}{2}(\log(|\tilde{K}||\Sigma_{WW}|) + (f-\tilde{\mu}_u)^\top \tilde{K}^{-1}(f-\tilde{\mu}_u) + (u-\mu_W)^\top \Sigma_{WW}^{-1}(u-\mu_W) \tag{30}$$

Here, we can not perform SGD, as the latter terms does not decompose into a sum over the data. More importantly, here we require the computation of $\tilde{K}$, which contains the kernel matrix K, even after the use of landmarks. This direct approach is not feasible for large number of individuals, which is true in our target application, and hence we do not pursue this method, and consider Nyström and NN as baselines.

C Random Fourier Features on Laplacian

Here we discuss using random Fourier features [25] to reduce computational cost in calculation of the Laplacian defined as $\mathrm{L} = diag(\mathrm{K}\mathbb{1}^\top) - \mathrm{K}$, where $\mathbb{1}$ is just $[1, \ldots, 1]$ and K. Suppose the kernel

[9] In practice, this does not have to be a positive semi-definite kernel, it can be derived from any notion of similarity between observations, including k-nearest neighbours.

[10] In order to guarantee positive definiteness of Laplacian, one can add ϵI, where $\epsilon > 0$.

is stationary i.e. $k_w(x-y) = k(x,y)$ (some examples include the gaussian and matern kernel), then using random Fourier features, we obtain $\mathrm{K} \approx \Phi\Phi^\top$, where $\Phi \in \mathbb{R}^{b_N \times m}$, b_N denotes the total number of individuals in the batch and m denotes the number of frequencies. Now we have:

$$\mathrm{f}^\top \mathrm{Lf} \approx \mathrm{f}^\top diag(\Phi\Phi^\top \mathbb{1}^\top)\mathrm{f} - \mathrm{f}^\top \Phi\Phi^\top \mathrm{f} = \mathrm{f}^\top diag(\Phi\Phi^\top \mathbb{1}^\top)\mathrm{f} - ||\Phi^\top f||_2^2 \tag{31}$$

In both terms, we can avoid computing the kernel matrix, by carefully selecting the order of computation. Note another option is to consider Nyström approximation with landmark points $\{z_1, \dots z_m\}$, then $K \approx K_{nm}K_{mm}^{-1}K_{mn}$, where K_{mm} denotes the kernel matrix on landmark points, while K_{nm} is the kernel matrix between landmark and data. Then $\Phi = K_{nm}K_{mm}^{-\frac{1}{2}}$.

D Bag Manifold regularisation

Suppose we have bag covariates s^a (note these are for the entire bag), and also some summary statistics of a bag, e.g. mean embeddings [19] given by $H^a = \frac{1}{N_a}\sum_{i=1}^{N_a} h(x_i^a)$, with some user-defined h. Then similarly to individual level manifold regularisation, we can consider manifold regularisation at the bag level (assuming a seperable kernel for simplicity), i.e.

$$\ell_2 = \sum_{l=1}^{n}\sum_{m=1}^{n}(F^l - F^m)^2 k_s(s^l, s^m)k_h(H^l, H^m) = \mathrm{F}^\top \mathrm{L}_{\text{bag}}\mathrm{F} \tag{32}$$

where $F^a = \frac{1}{N_l}\sum_{i=1}^{N_a} f_i^a$, k_s is a kernel on bag covariates s^a, k_μ is a kernel on H^a, L_{bag} is the bag level Laplacian with the corresponding kernel, and $F = [F^1, \dots, F^n]^\top$. Combining all these terms, we have the following loss function to minimise:

$$\ell = \frac{1}{b}\ell_0 + \frac{\lambda_1}{b_N^2}\ell_1 + \frac{\lambda_2}{b_N^2}\ell_2 \tag{33}$$

where b is the mini-batch size in SGD, B_N is the total number of individuals in each mini-batch, λ_1 and λ_2 are parameters controlling the strength of the respective regularisation.

E Additional details for Poisson variational derivation

E.1 Log-sum lemma

Lemma 2. *Let $v = [v_1, \dots, v_N]^\top$ be a random vector with probability density $q(v)$, and let $w_i \geq 0$, $i = 1, \dots, N$. Then, for any non-negative valued function $\Psi(v)$,*

$$\int \log\big(\sum_{i=1}^{N} w_i \Psi(v_i)\big) q(v)dv \geq \log\Big(\sum_{i=1}^{N} w_i e^{\xi_i}\Big),$$

where

$$\xi_i := \int \log \Psi(v_i) q_i(v_i) dv_i.$$

Proof. Let $\alpha_1, \dots, \alpha_N$ be non-negative numbers with $\sum_{i=1}^{N}\alpha_i = 1$. It follows from Jensen's inequality that

$$\begin{aligned}
&\int \log\big(\sum_{i=1}^{N} w_i \Psi(v_i)\big) q(v)dv = \\
&\int \log\Big(\sum_{i=1}^{N} \alpha_i \tfrac{w_i}{\alpha_i} \Psi(v_i)\Big) q(v)dv \geq \\
&\sum_{i=1}^{N} \alpha_i \left[\int \log\Big(\Psi(v_i)\Big) q(v_i)dv_i + \log\frac{w_i}{\alpha_i}\right] = \\
&\sum_{i=1}^{N} \alpha_i \xi_i + \sum_{i=1}^{N} \alpha_i \log\frac{w_i}{\alpha_i}.
\end{aligned} \tag{34}$$

By Lagrange multiplier method, maximizing the last line with respect to α gives

$$\alpha_i = \frac{w_i e^{\xi_i}}{\sum_{j=1}^{N} w_j e^{\xi_j}}.$$

Plugging this to (34) completes the proof. □

E.2 A lower bound of marginal likelihood for $\Psi(f) = e^f$ and $\Psi(f) = f^2$

Using Lemma 2, we obtain that

$$\int \log\big(\sum_{i=1}^{N} p_i^a \Psi(v_i^a)\big) q(v^a) dv^a \geq \log\Big(\sum_{i=1}^{N} p_i^a \Psi(\xi_i^a)\Big), \tag{35}$$

where

$$\xi_i^a = \int \log \Psi(v_i^a) q_i^a(v_i^a) dv_i^a.$$

The above lower bound is tractable for the popular functions $\Psi(v) = v^2$ and $\Psi(v) = e^v$ under the normal variational distributions $q^a(v^a) \sim \mathcal{N}(m^a, S^a)$. In particular,

$$\begin{aligned}
\Psi(v) = e^v: \quad \xi_i^a &= \int v_i^a q_i^a(v_i^a) dv_i^a = m_i^a, \\
\Psi(v) = v^2: \quad \xi_i^a &= \int \log(v_i^a)^2 q_i^a(v_i^a) dv_i^a = -G\left(-\frac{m_i^a}{2S_{ii}^a}\right) + \log\left(\frac{S_{ii}^a}{2}\right) - \gamma,
\end{aligned}$$

where γ is the Euler constant and

$$G(t) = 2t \sum_{j=0}^{\infty} \frac{j!}{(2)_j (3/2)_j} t^j$$

is the partial derivative of the confluent hypergeometric function [17, 1]. However, in this work we focus on the Taylor series approximation for $\Psi(v) = v^2$, as implementation of the above bound uses a large look-up table and involves linear interpolation. Furthermore, it is suggested in experiments that the secondary lower bound proposed above in Lemma 2 can lead to poor calibration, for more details, refer to Section 4.

E.3 KL Term

Since $q(u)$ and $p(u|W)$ are both normal distribution, the KL divergence is tractable:

$$KL(q(u)||p(u|W)) = \frac{1}{2}\Big\{Tr[K_{WW}^{-1}\Sigma_u] + \log\frac{|K_{WW}|}{|\Sigma_u|} - m + (\mu_W - \eta_u)^T K_{WW}^{-1}(\mu_W - \eta_u)\Big\} \tag{36}$$

E.4 Taylor series approximation in the variational method

We consider the integral

$$\int \log\big(\sum_{i=1}^{N} p_i^a (v_i^a)^2\big) q^a(v^a) dv^a$$

where q^a is $\mathcal{N}(m^a, S^a)$. We note that this can be written as $\mathbb{E} \log \|V^a\|^2$, where $V^a \sim N(\tilde{m}^a, \tilde{S}^a)$, with $P^a = diag\left(p_1^a, \ldots, p_{N_a}^a\right), \tilde{m}^a = P^{a\,1/2} m^a$ and $\tilde{S}^a = P^{a\,1/2} S^a P^{a\,1/2}$. Note that $\|V^a\|^2$ follows a non-central chi-squared distribution. We now resort to a Taylor series approximation for

$\mathbb{E}\log\|V^a\|^2$ (similar to [29]) around $\mathbb{E}\|V^a\|^2 = \|\tilde{m}^a\|^2 + tr\tilde{S}^a$, resulting in

$$\begin{aligned}\mathbb{E}\log\left(\|V^a\|^2\right) &= \log\left(\mathbb{E}\|V^a\|^2\right) \\ &\quad +\mathbb{E}\left[\frac{\|V^a\|^2 - \mathbb{E}\|V^a\|^2}{\mathbb{E}\|V^a\|^2} - \frac{\left(\|V^a\|^2 - \mathbb{E}\|V^a\|^2\right)^2}{2\left(\mathbb{E}\|V^a\|^2\right)^2} + \mathcal{O}\left(\left(\|V^a\|^2 - \mathbb{E}\|V^a\|^2\right)^3\right)\right] \\ &\approx \log\left(\|\tilde{m}^a\|^2 + tr\tilde{S}^a\right) - \frac{2\tilde{m}^{a\top}\tilde{S}^a\tilde{m}^a + tr\left(\left(\tilde{S}^a\right)^2\right)}{\left(\|\tilde{m}^a\|^2 + tr\tilde{S}^a\right)^2}.\end{aligned}$$

As commented in [29], approximation is very accurate when $\mathbb{E}\|V^a\|^2$ is large, but the caveat is that the Taylor series converges only for $\|V\|^2 \in (0, 2\mathbb{E}\|V\|^2)$ so this approach effectively ignores the tail of the non-central chi-squared.

F Code

All of our models were implemented in TensorFlow, and code will be published and available for use.

G Additional Malaria Experimental Results

Here we provide additional experimental results for the malaria dataset. In table 1, we provide results for bag level performance for NLL and MSE with 10 different test sets (after retrial of the experiments, splitting the data across train, early-stop, validation and testing). Statistical significance was not establish for the best performing Nyström method versus the VBAgg methods, this is shown in Table 2. We further provide additional prediction/uncertainty patches for 3 different splits to highlight the general behaviour of the trained models, with further explanation and details below.

It is also noted in all cases λ_i^a is the incidence rate per 1000 people. For VBAgg and Nyström, we use an additive kernel, between an ARD kernel and a Matern kernel:

$$k((x,s_x),(y,s_y)) = \gamma_1 \exp\left(-\frac{1}{2}\sum_{k=1}^{18}\frac{1}{\ell_k}(x_k - y_k)^2\right) + \gamma_2\left(1 + \frac{\sqrt{3}||s_x - s_y||_2}{\rho}\right)\exp\left(-\frac{\sqrt{3}||s_x - s_y||_2}{\rho}\right) \tag{37}$$

where x, y are covariates, and s_x, s_y are their respective spatial location. Here, we learn any scale parameters and weights during training. For the NN, we also use this kernel as part of manifold regularisation, however we use an RBF kernel instead of an ARD kernel, due to parameter tuning reasons (we can no longer learn these scales).

For constant model, bag rate predictions are computed by, $p^a\hat{\lambda}_c^{\text{bag}}$,where $\hat{\lambda}_c^{\text{bag}} = \frac{1}{\sum_{a=1}^n p^a}\sum_{a=1}^n y^a$. This essentially takes into account of population.

Table 1: Results for the Poisson Model on the malaria dataset with 10 different re-splits of train, early-stopping, validation and test. Approximately, 191 bags are used for test set. Bag performance is measured on a test set, with MSE computed between $\log(y^a)$ and $\log(\sum_{i=1}^{N_a} p_i^a\hat{\lambda}_i^a)$. Brackets include standard deviation.

	Bag NLL	Bag MSE (Log)
Constant	173.1 (31.2)	4.08 (0.13)
Nyström-Exp	88.1 (25.1)	1.31 (0.15)
VBAgg-Sq-Obj	94.1 (34.0)	1.21 (0.05)
VBAgg-Exp-Obj	97.2 (39.6)	1.04 (0.11)
VBAgg-Sq	97.6 (39.0)	1.38 (0.18)
VBAgg-Exp	99.2 (39.8)	1.21 (0.19)
NN-Exp	164.4 (127.8)	1.82 (0.29)

Table 2: p-values from a Wilcoxon signed-rank test for Nyström-Exp versus the methods below for Bag NLL and MSE for the malaria dataset. The null hypothesis is Nyström-Exp performs equal or worse than the considered method on the test bag performance.

	NLL	MSE
Constant	0.0009766	0.0009766
NN-Exp	0.00293	0.0009766
VBAgg-Sq-Obj	0.1162	0.958
VBAgg-Sq	0.1377	0.1611
VBAgg-Exp-Obj	0.08008	1.0
VBAgg-Exp	0.09668	0.958

Table 3: p-values from a Wilcoxon signed-rank test for VBAgg-Sq versus the methods below for Bag NLL and MSE for the malaria dataset. The null hypothesis is VBAgg-Sq performs equal or worse than the considered method on the test bag performance.

	NLL	MSE
Constant	0.0009766	0.0009766
NN-Exp	0.01855	0.001953
VBAgg-Sq-Obj	0.6234	0.9861
Nyström-Exp	0.8838	0.8623
VBAgg-Exp-Obj	0.6875	1.0
VBAgg-Exp	0.3477	0.9346

G.1 Predicted log malaria incidence rate for various models

Constant: Bag level observed incidences This is the baseline with $\hat{\lambda}_i^a$ being constant throughout the bag, as shown in Figure 3. For training, we only use 60% of the data.

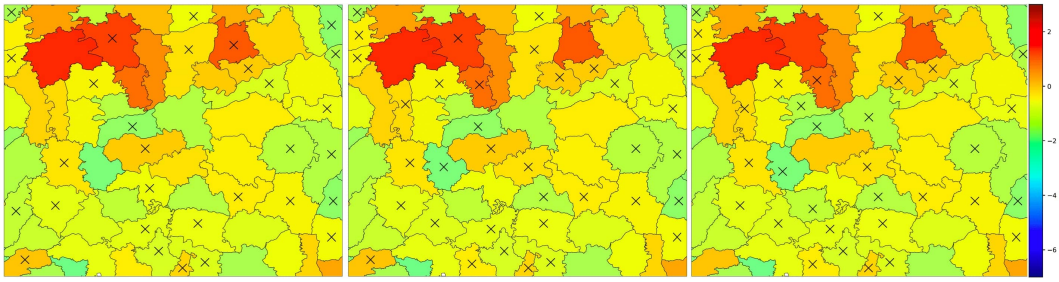

Figure 3: Predicted $\hat{\lambda}_i^a$ on log scale using constant model, for 3 different re-splits of the data.$\times$ denote non-train set bags.

VBAgg-Sq-Obj This is the VBAgg model with $\Psi(v) = v^2$ and tuning of hyperparameters is performed based on training objective, the lower bound to the marginal likelihood, we ignore early-stop and validation set here. The uncertainty of the model seems reasonable, and we also observe that in general the areas that are not in the training set have higher uncertainties. Furthermore, in all cases, malaria incidence was predicted to be higher near the river, as discussed in Section 4.2.

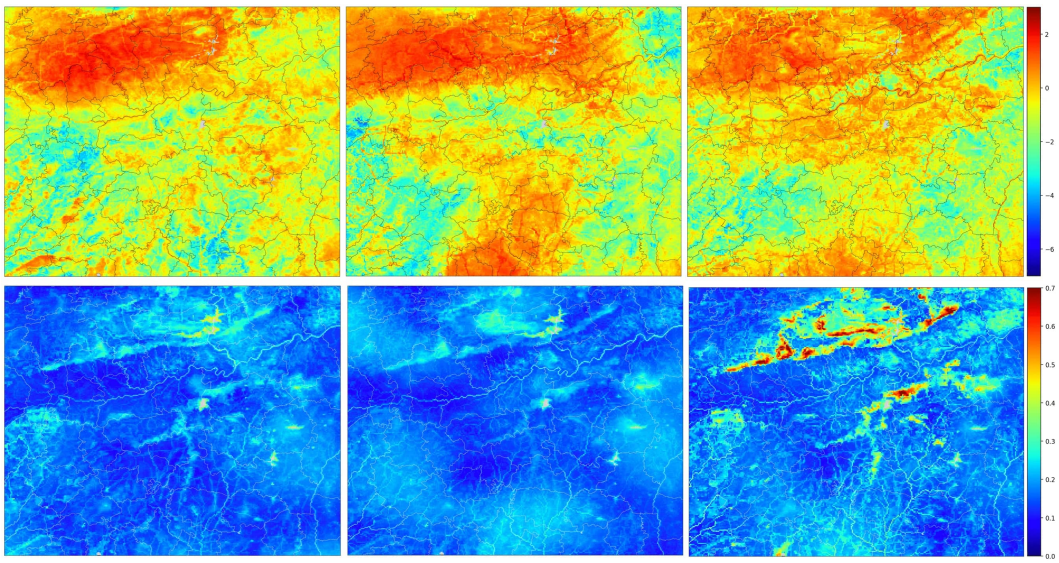

Figure 4: **Top:** Predicted $\hat{\lambda}_i^a$ on log scale for VBAgg-Sq-Obj. **Bottom:** Standard deviation of the posterior v in (9) with VBAgg-Sq-Obj.

VBAgg-Sq This is the VBAgg model with $\Psi(v) = v^2$ and tuning of hyperparameters is performed based on NLL at the bag level. Predicted incidence are similar to the VBAgg-Sq-Obj model. The uncertainty of the model is less reasonable here, this is expected behaviour, as we are tuning hyperparameters based on NLL here. In the first patch, the same parameters was chosen as VBAgg-Sq-Obj.

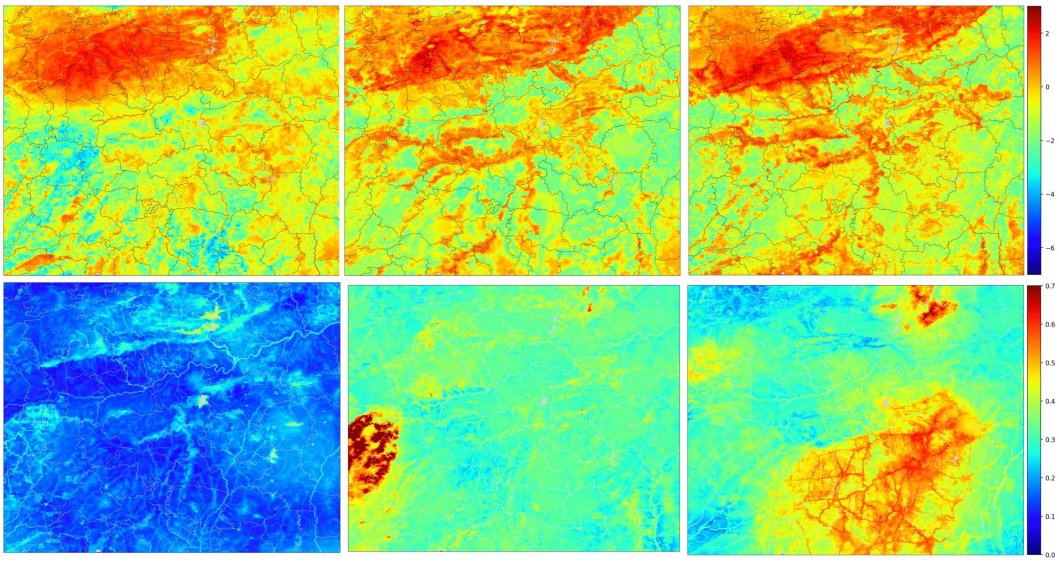

Figure 5: **Top:** Predicted $\hat{\lambda}_i^a$ on log scale for VBAgg-Sq. **Bottom:** Standard deviation of the posterior v in (9) with VBAgg-Sq.

VBAgg-Exp-Obj This is the VBAgg model with $\Psi(v) = e^v$ and tuning of hyperparameters is performed based on training objective, the lower bound to the marginal likelihood, we ignore early-stop and validation set here. Predicted incidence seem to be stable in general, though some smoothness is observed. The uncertainty of the model is also not very reasonably here, but this behaviour was observed in the Toy experiments, and likely due to an additional lower bound.

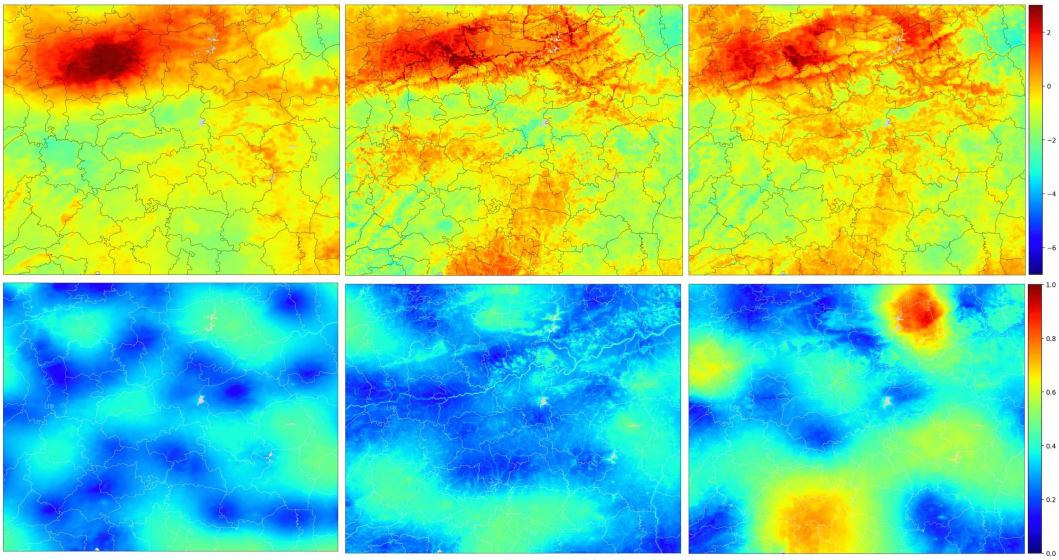

Figure 6: **Top:** Predicted $\hat{\lambda}_i^a$ on log scale for VBAgg-Exp-Obj.**Bottom:** Standard deviation of the posterior v in (9) with VBAgg-Exp-Obj.

VBAgg-Exp This is the VBAgg model with $\Psi(v) = e^v$ and tuning of hyperparameters is performed based on NLL. For details, see discussion above for the VBAgg-Exp-Obj model.

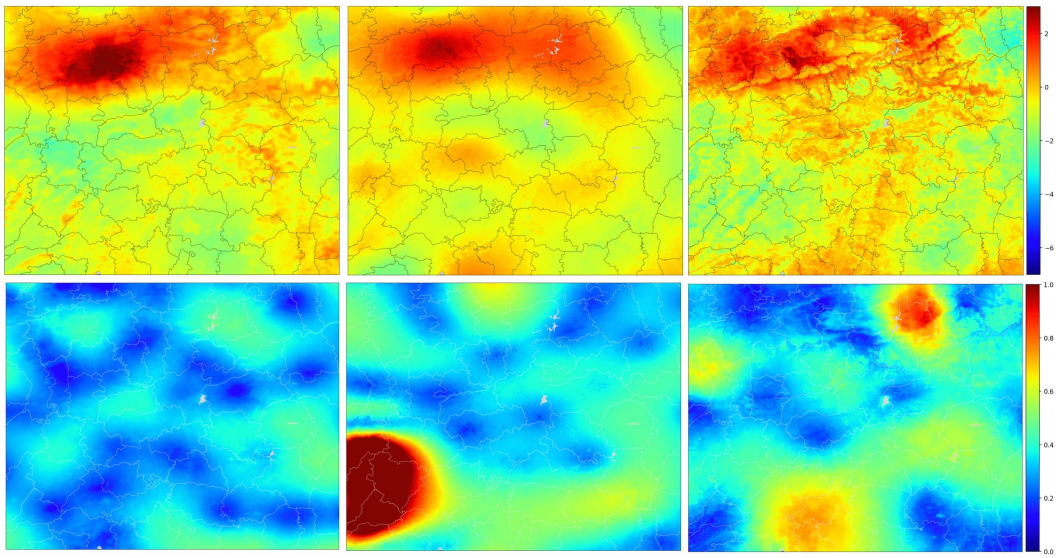

Figure 7: **Top:** Predicted $\hat{\lambda}_i^a$ on log scale for VBAgg-Exp. **Bottom:** Standard deviation of the posterior v in (9) with VBAgg-Exp.

Nyström-Exp This is the Nyström-Exp model, it is clear that while it performs best in terms of bag NLL, sometimes prediction are too smooth in the pixel space, this is because it optimises directly bag NLL. This pattern might be seen to be unrealistic, and may cause useful covariates to be neglected.

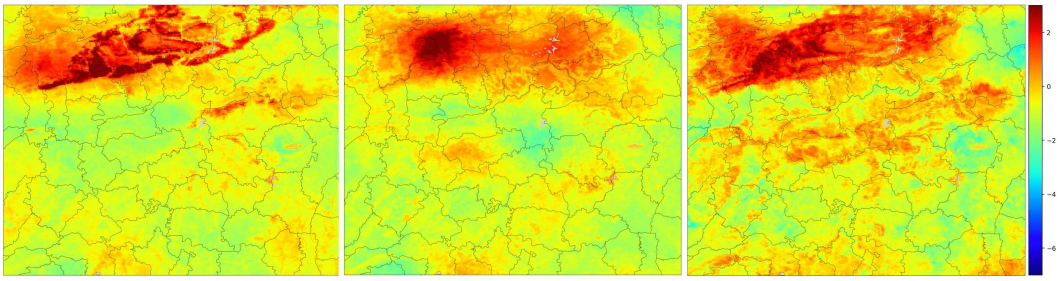

Figure 8: Predicted $\hat{\lambda}_i^a$ on log scale for Nyström-Exp.

NN-Exp We can see that the model is not very stable, this can be potentially due to the model does not have an inbuilt spatial smoothness function unlike other methods. It only uses manifold regularisation for training. Also, the maximum predicted pixel level intensity rate $\hat{\lambda}_i^a$ is over 1000 in some cases, this is clearly physically impossible given λ_i^a is rate per 1000 people.

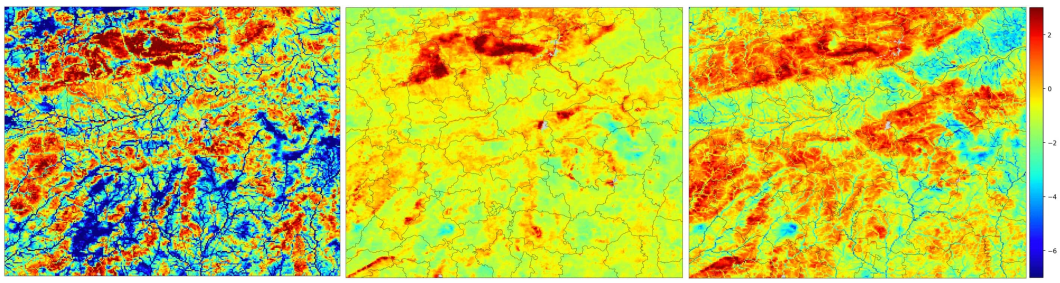

Figure 9: Predicted $\hat{\lambda}_i^a$ on log scale for NN-Exp.

G.2 Remote Sensing covariates that provide the existence of a river

Here, we provide figures for some covariates that give information that there is a river as indicated by the triangles in Figure 2.

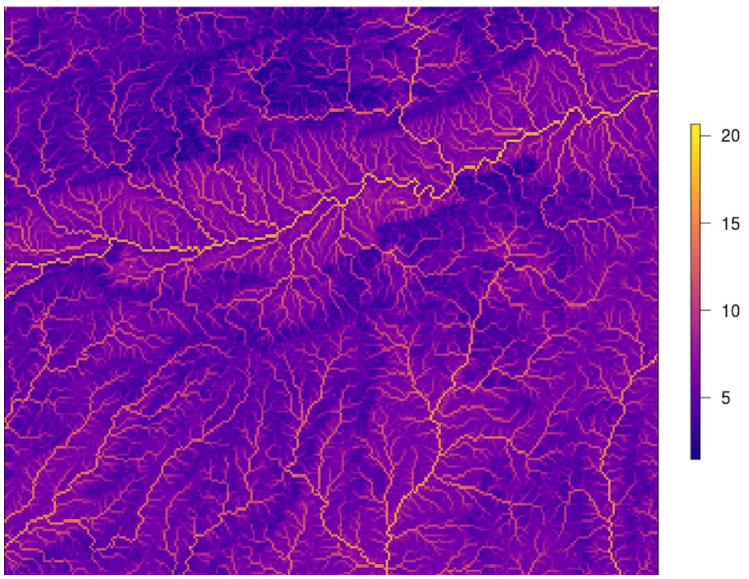

Figure 10: Topographic wetness index, measures the wetness of an area, rivers are wetter than others, as clearly highlighted.

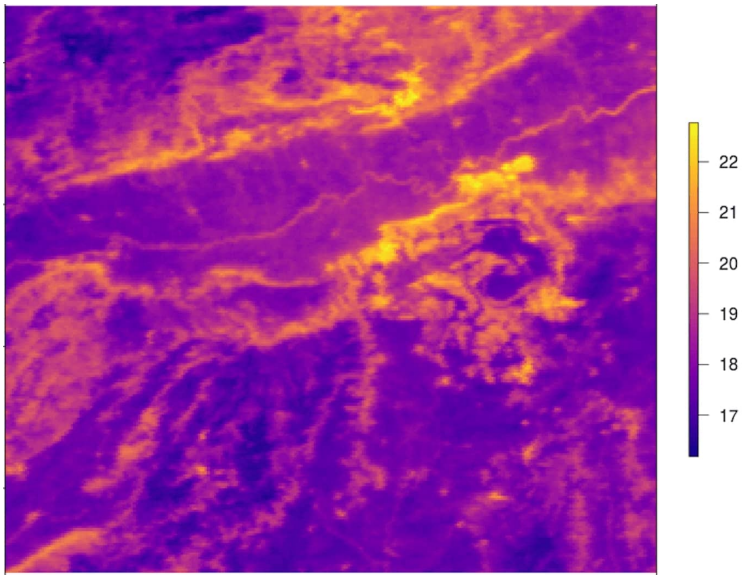

Figure 11: Land Surface Temperature at night, river is hotter at night, due to river being able to retain heat better.

H Additional Toy Experimental Results

In this section, we provide additional experimental results for the Normal and Poisson model. In particular, we provide results on test bag level performance, and provide also prediction, calibration and uncertainty plots.

For the VBAgg model, during the tuning process, it is possible to choose tuning parameters (e.g. learning rate, multiple-initialisations, landmark choices) based on NLL with an additional validation set or on the objective $\mathcal{L}_1$ on the training set. To compare the difference, we denote the model tuned on NLL as VBAgg and the model tuned on $\mathcal{L}_1$ as VBAgg-Obj. Intuitively, as VBAgg-Obj attempts to obtain as tight a bound to the marginal likelihood, we would expect better performance in calibration, i.e. more accurate uncertainties.

For calibration plots, we compute the α quantiles of the approximated posterior distribution and consider the ratio of times the underlying rate parameter λ_i^a (or μ_i^a for the normal model) appear inside the quantiles of the posterior distribution. If the model provides good uncertainties/calibration, we should expect to see the quantiles to match with the observed ratio.

In the case of $\Psi(v) = v^2$, the approximated posterior distribution is simply a non-central χ^2 distribution, while for $\Psi(v) = e^v$, this is a log-normal distribution. For the Normal Model, it is simply a normal distribution, as we do not have any transformations. Calibration plots can be found in Figure 20 and Figure 21 for the Normal Model, with Figure 14 and Figure 15 for the Poisson Model.

For uncertainty plots, we plot the standard deviation of the posterior of $v \sim \mathcal{N}(m^a, S^a)$ (i.e. before transformation through Ψ), as this provides better interpretability. Uncertainty plots can be found in Figure 17 and 23

To demonstrate statistical significance of our result, we aggregate the repetitions in each experiment for each method and consider a one sided rank permutation test (Wilcoxon signed-rank test) to see whether VBAgg is statistically significant better than other approaches for individual NLL and MSE.

H.1 Poisson Model

H.1.1 Swiss Roll Dataset

We provide additional results here for the experimental settings that we consider.

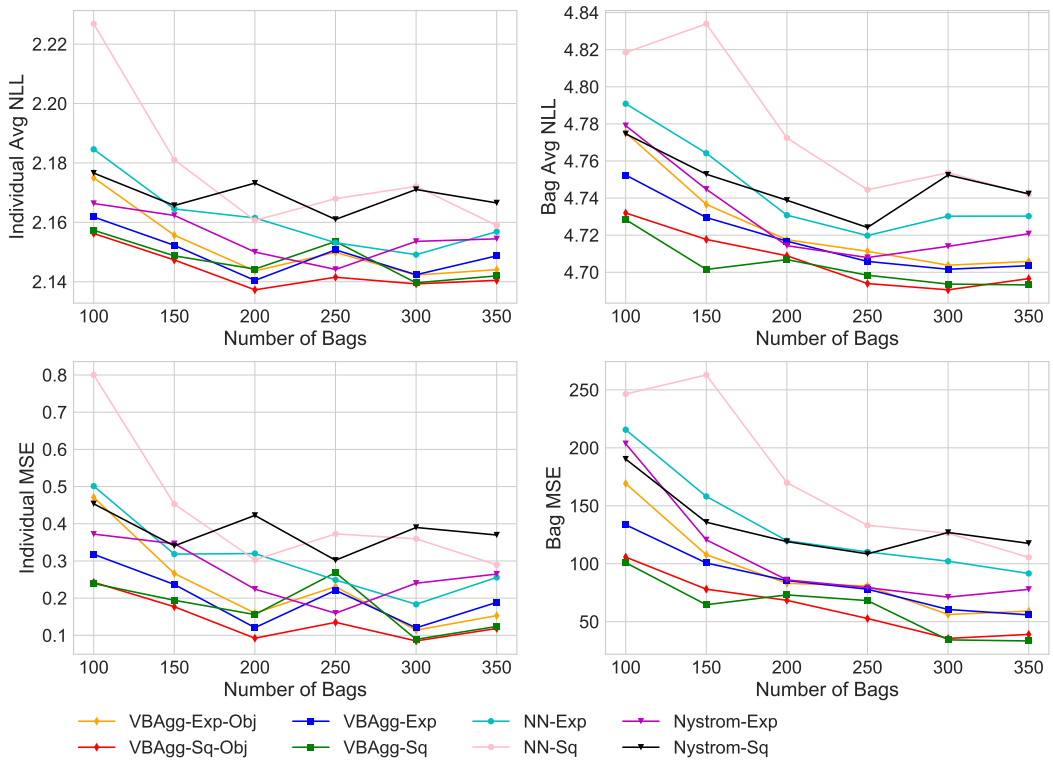

Figure 12: Varying number of bags over 5 repetitions.**Left Column:** Individual average NLL and MSE on train set. **Right Column:** Bag average NLL and MSE on test set (of size 500). Constant prediction NLL and MSE is 2.23 and 0.85 respectively. bag-pixel model prediction NLL is above 2.4 and MSE is above 3.0, hence not shown on graph.

The varying number of bags experimental results is found in Figure 12, with the corresponding table of p-values in Table 4, 5 demonstrating statistical significance of the VBAgg-Exp and VBAgg-Sq method. Similarly, the varying number of individuals per bag through N_{mean} experimental result can be found in Figure 13, with the corresponding table of p-values in Table 6, 7. The comparison between VBAgg-Exp and VBAgg-Sq was found to be non-significant.

Table 4: p-values from a Wilcoxon signed-rank test for VBAgg-Sq versus the methods below for the varying number of bags experiment for the Poisson model. The null hypothesis is VBAgg-Sq performs equal or worse than NN or Nyström in terms of individual NLL or MSE on the train set.

	NLL	MSE
NN-Exp	6.98e−06	0.00025
Nyström-Exp	0.00048	0.00015

Table 5: p-values from a Wilcoxon signed-rank test for VBAgg-Exp versus the methods below for the varying number of bags experiment for the Poisson model. The null hypothesis is VBAgg-Exp performs equal or worse than NN or Nyström in terms of individual NLL or MSE on the train set.

	NLL	MSE
NN-Exp	2.48e−06	2.48e−05
Nyström-Exp	0.0005	0.00025

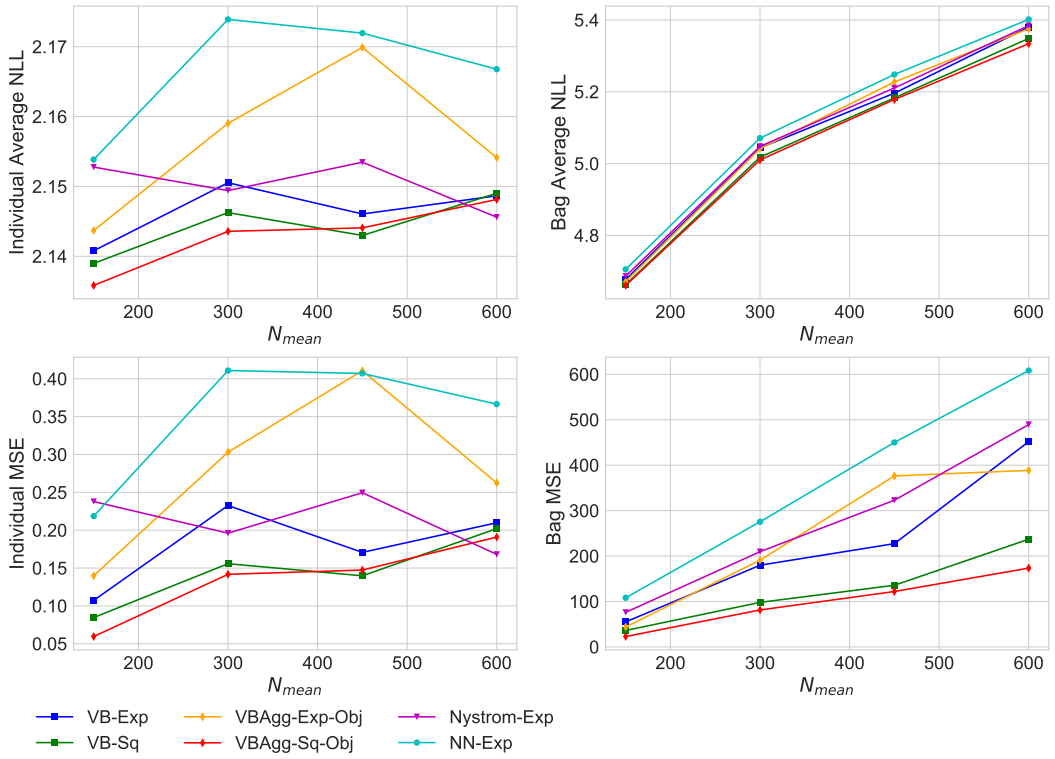

Figure 13: Varying number of individuals per bag N_{mean} over 5 repetitions.**Left Column:** Individual average NLL and MSE on train set. **Right Column:** Bag average NLL and MSE on test set (of size 500). Constant prediction NLL and MSE is 2.23 and 0.85 respectively.

Table 6: p-values from a Wilcoxon signed-rank test for VBAgg-Sq versus the methods below for the varying number of individuals per bag experiment for the Poisson model. The null hypothesis is VBAgg-Sq performs equal or worse than NN or Nyström in terms of individual NLL or MSE on the train set.

	NLL	MSE
NN-Exp	1.81e−05	9.53e−06
Nyström-Exp	0.062	0.041

Table 7: p-values from a Wilcoxon signed-rank test for VBAgg-Exp versus the methods below for the varying number of individuals per bag experiment for the Poisson model. The null hypothesis is VBAgg-Exp performs worse than NN or Nyström in terms of individual NLL or MSE on the train set.

	NLL	MSE
NN-Exp	6.68e−05	0.00016
Nyström-Exp	0.049	0.062

Calibration Plots for the Swiss Roll Dataset In Figure 14 and 15, we provide calibration results for both experiments that we have considered. See top of Appendix H for a further details. It is clear that while VBAgg-Sq-Obj and VBAgg-Sq provides good calibration in general, this is not the case for VBAgg-Exp-Obj and VBAgg-Exp. This is not surprising as the VBAgg-Exp methods uses an additional lower bound.

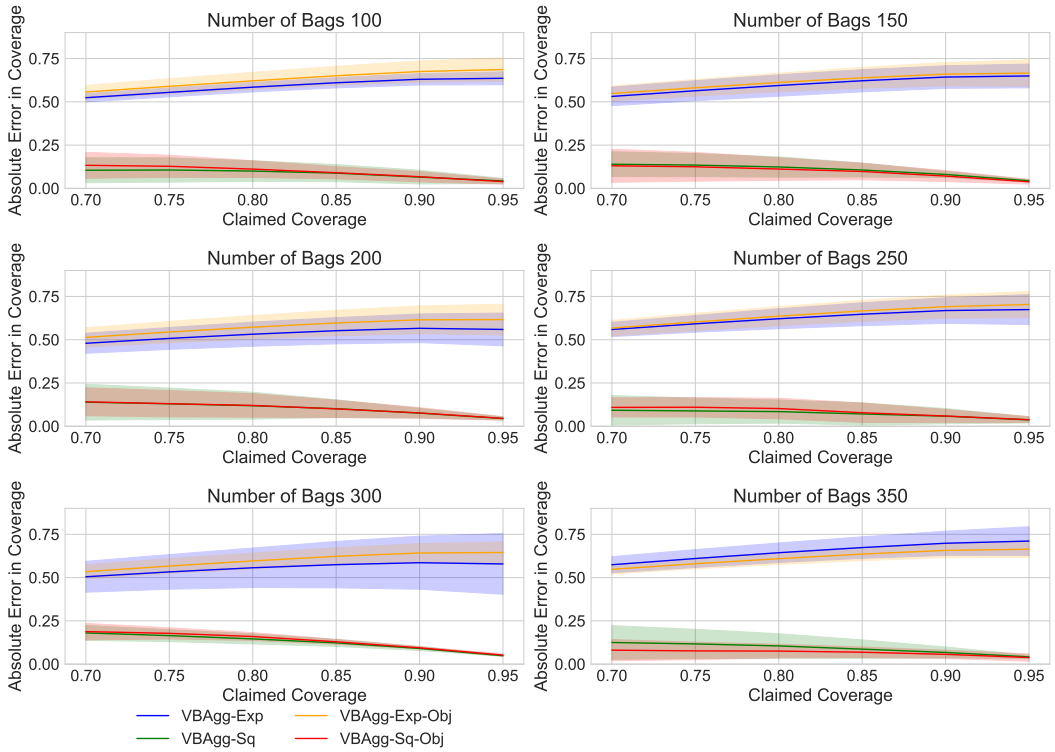

Figure 14: Absolute Error in coverage from 70% to 95% for the increasing number of bags experiment for the Poisson Model. Shaded regions highlight the standard deviation. Perfect coverage would provide a straight line at 0 error.

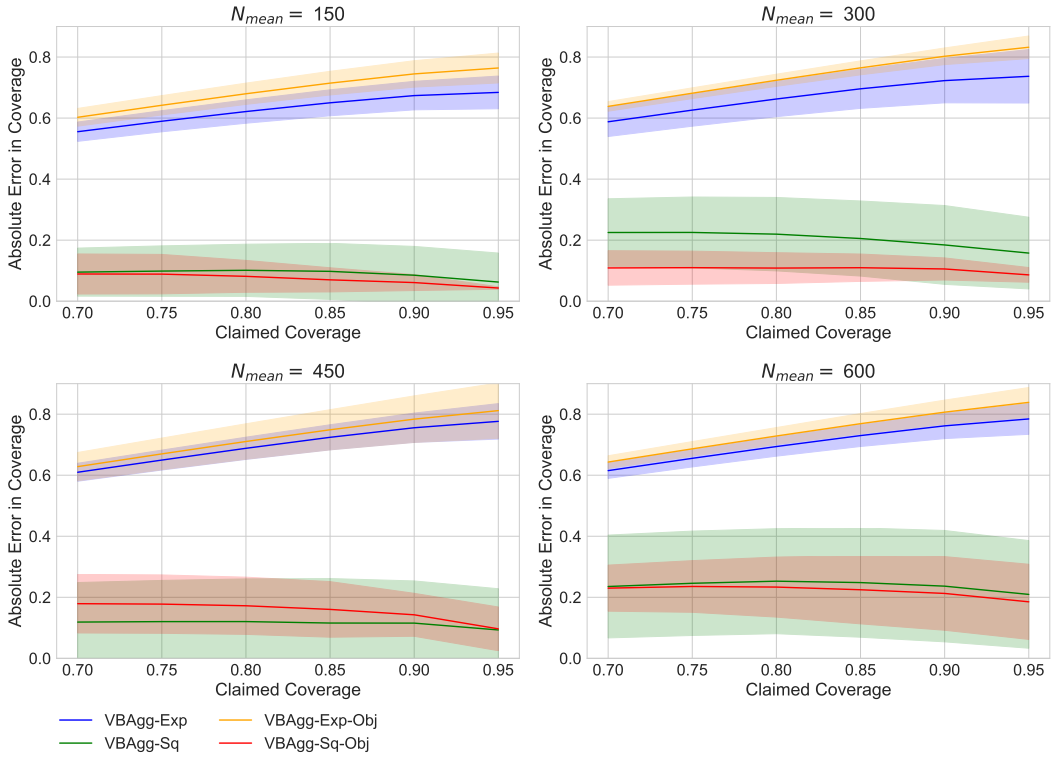

Figure 15: Absolute Error in coverage from 70% to 95% for the increasing number of individuals per bag N_{mean} and N_{std} for the Poisson Model. Shaded regions highlight the standard deviation. Perfect coverage would provide a straight line at 0 error.

Prediction and uncertainty plots In Figure 16 and 17, we provide some prediction plots for different models, and uncertainties for VBAgg models.

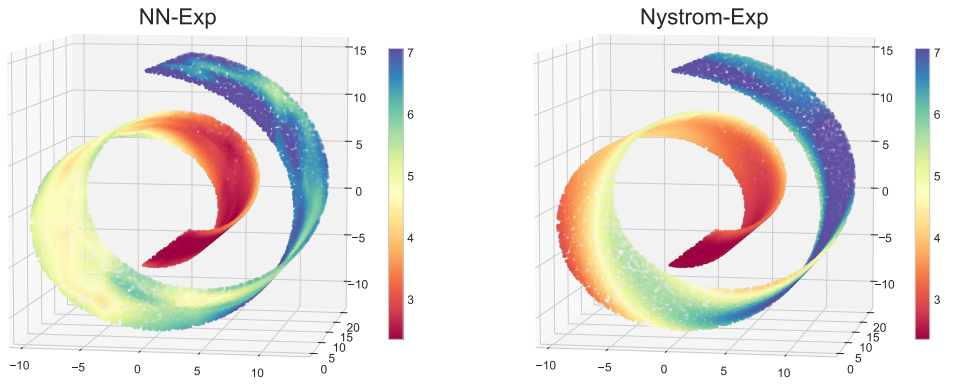

Figure 16: Individual predictions on the train set for the swiss roll dataset with 150 bags for NN and Nyström model. Here $N_{mean} = 150$, with $N_{std} = 50$.

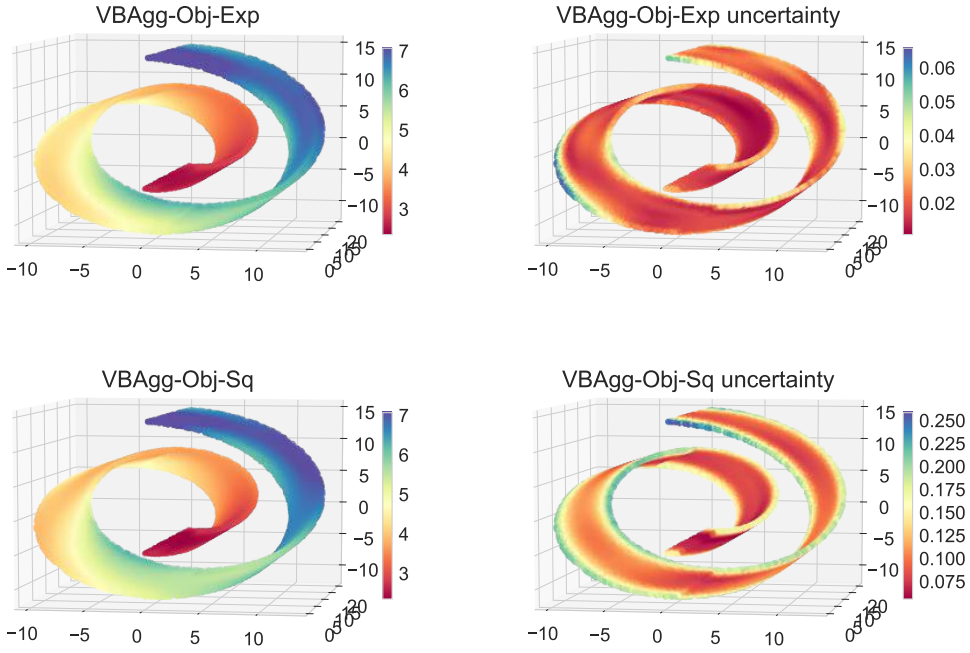

Figure 17: Predictions and uncertainty on the swiss roll dataset with 150 bags for the VBAgg-Obj models. Here $N_{mean} = 150$, with $N_{std} = 50$. For uncertainty, we plot the standard deviation of the posterior of v, coming from $v^a \sim \mathcal{N}(m^a, S^a)$ in (9).

H.2 Normal Model

H.2.1 Swiss Roll Dataset

In this section, we provide some experimental results for the Normal model, where throughout we assume $\tau_i^a = \tau$, same for all individuals.

We consider the same swiss roll dataset as in the Poisson model, here the colour of each point to be the underlying mean μ_i^a. We then consider $y_i^a \sim \mathcal{N}(\mu^a, \tau)$ with $\tau = 0.1$, hence bag observations are given by $y^a = \sum_{i=1}^{N_a} y_i^a \sim \mathcal{N}(\mu^a, N_a\tau)$ with $\mu^a = \sum_{i=1}^{N_a} \mu_i^a$. Here, the goal is to predict μ_i^a and τ, given bag observations y^a only. The results for the experiments are shown below in Figure 18 and Figure 19, which shows the VBAgg outperforming the NN and Nyström model. To show statistical significance, we also report the corresponding table of p-values in Table 8 and Table 9. Furthermore, we would also like to point out that the VBAgg is well calibrated as shown in Figure 20.

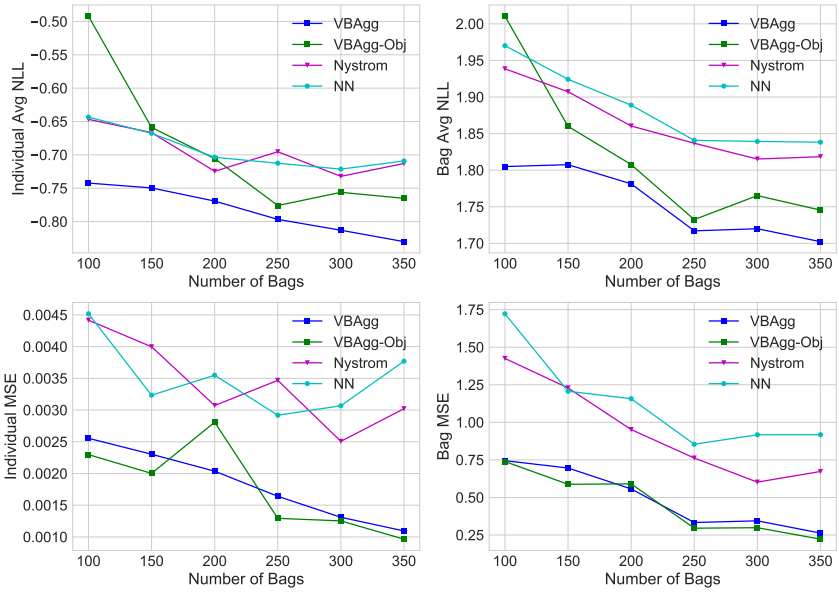

Figure 18: Varying number of bags over 5 repetitions for the Normal model.**Left Column:** Individual average NLL and MSE on train set. **Right Column:** Bag average NLL and MSE on test set (of size 500). Constant model individual MSE is 0.04.

Table 8: p-values from a Wilcoxon signed-rank test for VBAgg versus the methods below for the varying number of bags experiment for the Normal model. The null hypothesis is VBAgg performs equal or worse than NN or Nyström in terms of individual NLL or MSE on the train set.

	NLL	MSE
NN	5.96e−07	4.79e−09
Nyström	4.01e−08	6.52e−09

Table 9: p-values from a Wilcoxon signed-rank test for VBAgg versus the methods below for the varying number of individuals per bag N_{mean} experiment for the Normal nodel. The null hypothesis is VBAgg performs worse than NN or Nyström in terms of individual NLL or MSE on the train set.

	NLL	MSE
NN	4.77e−06	4.77e−06
Nyström	4.77e−06	4.77e−06

Calibration Plots for the Swiss Roll Dataset In Figure 20 and 21, we provide calibration results for both experiments that we have considered. See top of Appendix H for further details. It is clear

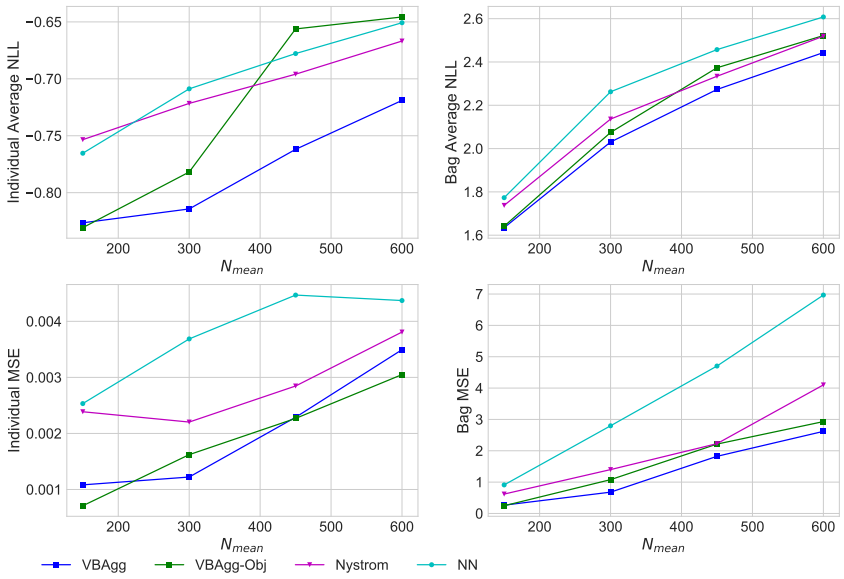

Figure 19: Varying number of individuals per bag N_{mean} over 5 repetitions.**Left Column:** Individual average NLL and MSE on train set. **Right Column:** Bag average NLL and MSE on test set (of size 500). Constant model individual MSE is 0.039.

that VBAgg-Obj has better calibration in general, this is not surprising as it is tuned based on the correct objective, rather than NLL.

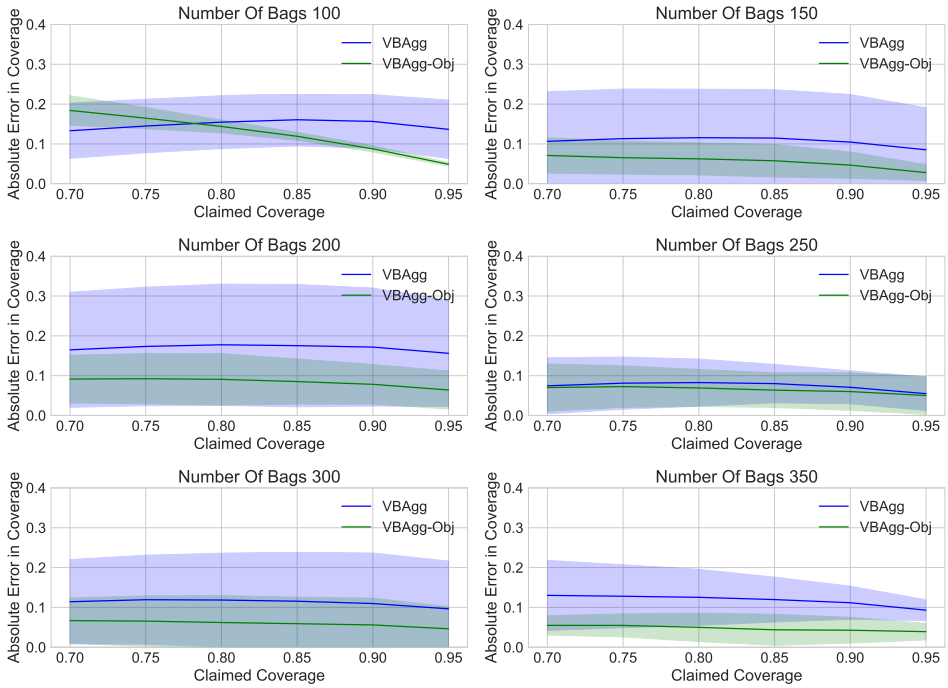

Figure 20: Absolute Error in coverage from 70% to 95% for the increasing number of bags experiment for the Normal Model. Shaded regions highlight the standard deviation. Perfect coverage would provide a straight line at 0 error.

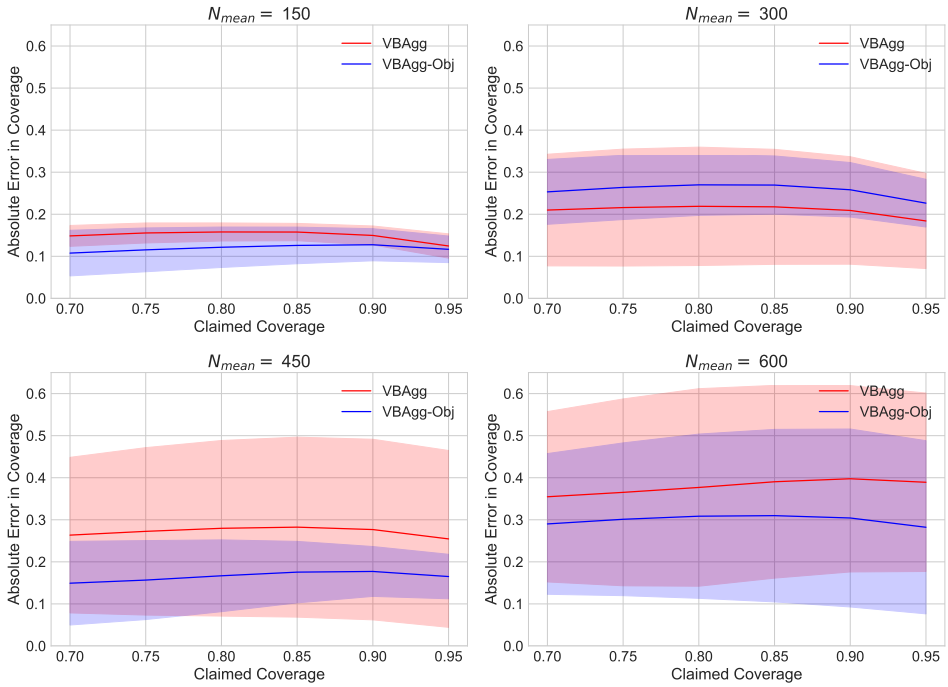

Figure 21: Absolute Error in coverage from 70% to 95% for the increasing number of individuals per bag N_{mean} and N_{std} for the Normal Model. Shaded regions highlight the standard deviation. Perfect coverage would provide a straight line at 0 error.

Prediction and uncertainty plots Here, we provide some prediction plots for different models.

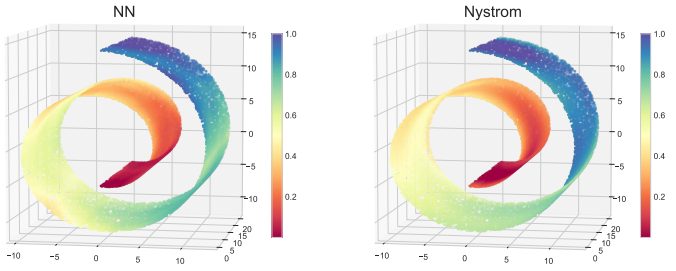

Figure 22: Individual predictions on the train set for the swiss roll dataset with 150 bags for NN and Nyström model. Here $N_{mean} = 150$, with $N_{std} = 50$.

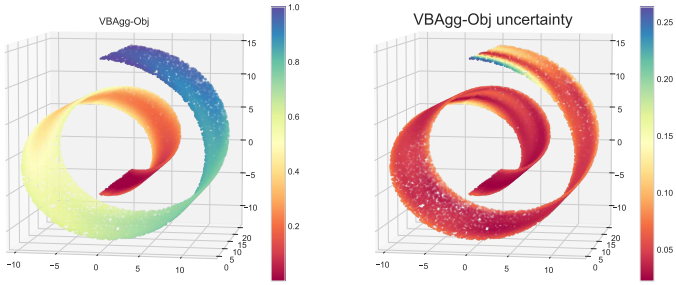

Figure 23: Predictions and uncertainty on the swiss roll dataset with 150 bags for the VBAgg-Obj model. Here $N_{mean} = 150$, with $N_{std} = 50$. For uncertainty, we plot the standard deviation of the posterior of v, coming from $v^a \sim \mathcal{N}(m^a, S^a)$ in (9).

H.2.2 Elevators Dataset

For a real dataset experiment, we consider the elevators dataset[11], which is a large scale regression dataset[12] containing 16599 instances, with each instance $\in \mathbb{R}^{17}$. This dataset is obtained from the task of controlling F16 aircraft, with the label y being a particular action taken on the elevators of the aircraft $\in \mathbb{R}$. For the model formulation we assume each label follows a normal distribution, i.e. $y_l \sim \mathcal{N}(\mu_l, \tau)$, where τ is a fixed quantity to be learnt. In practice, we can imagine the action taken may differ according to the operator.

In order formulate this dataset in an aggregate data setting, we sample bag sizes from a negative binomial distribution as before, with $N_{mean} = 30$ and $N_{std} = 15$, and also take $w_i^a = 1$. To place observations into bags, similar to the swiss roll dataset, we consider a particular covariate, and place instances into bags based on the ordering of the covariate. We now have the bag-level model given by $y^a \sim \mathcal{N}(\mu^a, N_a\tau)$, with individual model $y_i^a \sim \mathcal{N}(\mu_i^a, \tau)$ and it is our goal to predict μ_i^a (and also infer τ), given only y^a. After the bagging process, we obtain approximately 225 bags for training, and 33 bags each for early stopping, validation and testing (for bag level performance). Further, in order to neglect variables that do not provide signal, we use an ARD kernel for the VBAgg and

[11] This dataset is publicly available at http://sci2s.ugr.es/keel/dataset.php?cod=94

[12] We have removed one column that is almost completely sparse.

Table 10: Results for the Normal Model on the elevators dataset with 50 repetitions. Indiv represents individuals on train set here, while bag performance is measured on a test set. Numbers in brackets denotes p-values from a Wilcoxon signed-rank test for VBAgg versus the method. The null hypothesis is VBAgg performs equal or worse than NN or Nyström in terms of individual NLL or MSE on the train set. It is also noted MSE is computed on the observed y_i^a or y^a, rather than the unknown μ_i^a or μ^a.

	Indiv NLL	Bag NLL	Indiv MSE	Bag MSE
Constant	N/A	N/A	0.010	0.366
VBAgg	-1.69	0.003	0.0018	0.052
VBAgg-Obj	-1.71	-0.02	0.0018	0.052
Nyström	$-1.57(1.5\text{e}{-13})$	0.003	0.0024 $(8.9\text{e}{-16})$	0.041
NN	-1.64 (0.0001258)	0.082	0.0021 $(8.8\text{e}{-10})$	0.041

Nyström model, as below:

$$k_{ard}(x, y) = \gamma_{scale} \exp\left(-\frac{1}{2}\sum_{k=1}^{d}\frac{1}{\ell_k}(x_k - y_k)^2\right) \tag{38}$$

and learn kernel parameters γ_{scale} and $\{\ell_k\}_{k=1}^d$. We repeat this process and splitting of the dataset 50 times and report individual NLL results, and also MSE results in Table 10. From the results, we observe that the VBAgg model performs better the Nyström and NN model, with statistical significance.