Learning on Aggregate Outputs with Kernels

Dino Sejdinovic

Department of Statistics
University of Oxford

2nd Kermes Workshop, Madrid
25/02/2019
Gaussian Processes

Consider function values $f = (f(x_1), \ldots, f(x_n))^\top$ at a set of inputs, and observations $y = (y_1, \ldots, y_n)$. GP regression model is given by

$$f \sim \mathcal{N}(0, K)$$
$$y|f \sim \mathcal{N}(f, \sigma^2 I)$$

where $K$ is the (covariance) kernel matrix on inputs.

- **Posterior distribution:**
  $$f|y \sim \mathcal{N}(K(K + \sigma^2 I)^{-1}y, K - K(K + \sigma^2 I)^{-1}K)$$

- **Posterior predictive distribution:** Suppose $x'$ is an unseen test set. We can extend our model to include the function values $f'$ at the test set:
  $$\begin{pmatrix} y \\ f' \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_{xx} + \sigma^2 I & K_{xx'} \\ K_{x'x} & K_{x'x'} \end{pmatrix} \right)$$

  where $K_{xx'}$ is matrix with $(i, j)$-th entry $k(x_i, x'_j)$.

- Basic Gaussian conditioning gives:
  $$f'|y \sim \mathcal{N}\left( K_{x'x}(K_{xx} + \sigma^2 I)^{-1}y, K_{x'x'} - K_{x'x}(K_{xx} + \sigma^2 I)^{-1}K_{xx'} \right).$$
Non-Gaussian observation models

Consider function values \( \mathbf{f} = (f(x_1), \ldots, f(x_n)) \) at a set of inputs, and observations \( \mathbf{y} = (y_1, \ldots, y_n) \), with a general observation model

\[
\mathbf{f} \sim \mathcal{N}(0, \mathbf{K})
\]

\[
\mathbf{y} | \mathbf{f} \sim p(\mathbf{y} | \mathbf{f}) = \prod_{i=1}^{n} p(y_i | f(x_i)) .
\]

- Posterior distribution \( p(\mathbf{f}|\mathbf{y}) \) is no longer tractable.
- **Variational approximation**: write \( q(\mathbf{f}) = \mathcal{N}(f|\mu, \Sigma) \) and learn \( \mu, \Sigma \) by optimizing the evidence lower bound (ELBO):

\[
\mathcal{L}(\mu, \Sigma) = \mathbb{E}_q \log p(\mathbf{y}|\mathbf{f}) - KL (q(\mathbf{f})\|p(\mathbf{f}))
\]

\(\text{tractable}\)

- **Inducing points / landmarks**: often coupled with a scalable GP approximation, taking \( m \ll n \) inducing inputs \( z_1, \ldots, z_m \) and respective values \( \mathbf{u} = (f(z_1), \ldots, f(z_m)) \), with a joint variational posterior \( q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \), so that only variational parameters of \( q(\mathbf{u}) \) need to be inferred.
GP regression and Kernel Ridge Regression

If KRR and GPR use the same kernel and if the regularization parameter $\lambda$ equals the noise variance $\sigma^2$, KRR estimate of the function coincides with the GPR posterior mean/mode. Indeed, recall that in KRR we are solving empirical risk minimization

$$\min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \sigma^2 \|f\|_{\mathcal{H}_k}^2,$$

and are fitting a function of the form $f(x) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$. Closed form solution is given by $\alpha = (K_{xx} + \sigma^2 I)^{-1} y$. But then if we wish to predict function values at a new set $x' = \{x'_j\}_{j=1}^{m}$ of input vectors, we have

$$f(x'_j) = \sum_{i=1}^{n} \alpha_i k(x'_j, x_i) = [k(x'_j, x_1), \ldots, k(x'_j, x_n)] (K_{xx} + \sigma^2 I)^{-1} y,$$

and $[k(x'_j, x_1), \ldots, k(x'_j, x_n)]$ is the $j$-th row of $K_{x'x}$.

More generally, GP posterior mode for any likelihood model lies in the RKHS (essentially the same proof as the representer theorem).
GPs and RKHSs: shared mathematical foundations

- The same notion of a (positive definite) kernel, but conceptual gaps between communities.
- Orthogonal projection in RKHS $\Leftrightarrow$ Conditioning in GPs.
- Beware! 0/1 laws: GP sample paths with (infinite-dimensional) covariance kernel $k$ almost surely fall outside of $\mathcal{H}_k$.
  - But the space of sample paths is only slightly larger than $\mathcal{H}_k$ (outer shell).
  - It is typically also an RKHS (with another kernel).
- Worst-case in RKHS $\Leftrightarrow$ Average-case in GPs.

$$\text{MMD}^2(P, Q; \mathcal{H}_k) = \left( \sup_{\|f\|_{\mathcal{H}_k} \leq 1} (Pf - Qf) \right)^2 = \mathbb{E}_{f \sim \mathcal{GP}(0, k)} \left[ (Pf - Qf)^2 \right].$$

Radford Neal, 1998: “prior beliefs regarding the true function being modeled and expectations regarding the properties of the best predictor for this function [...] need not be at all similar.”

Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences
M. Kanagawa, P. Hennig, DS, and B. K. Sriperumbudur

ArXiv e-prints:1807.02582
https://arxiv.org/abs/1807.02582
Learning on Aggregate Outputs
Motivation- Disease modelling

Suppose you have a country with $n$ regions and data on:

- number of malaria incidences per region (low resolution)
- many covariates per region (high resolution)

**Goal:** Predict malaria incidences at a higher resolution, given low resolution label data and high definition covariate data.

![Figure: Log incidence rate of malaria](image)

Low definition labels (Data)  

High definition covariates (Data)

Figure: Log incidence rate of malaria
Motivation - Disease modelling

Suppose you have a country with \( n \) regions and data on:

- number of malaria incidences per region (low resolution)
- many covariates per region (high resolution)

**Goal:** Predict malaria incidences at a higher resolution, given low resolution label data and high definition covariate data.

**Figure:** Log incidence rate of malaria
Setup

Formally, we have $n$ regions, and for each region $a$, we have $N_a$ pixels (1km by 1km regions):

Data: $(\{x_{1i}^a\}_{i=1}^{N_1}, y^1), \ldots, (\{x_{ni}^a\}_{i=1}^{N_n}, y^n)$

- $x_{ai}^a$ is the covariates for the $i^{th}$ pixel of region $a$
- $y^a$ is the total observed incidence for region $a$

Goal: Predict $y_{ai}^a$, the unknown incidence for pixel $i$ in region $a$. 
Learning on Aggregates

- **Supervised learning**: obtaining inputs has a lower cost than obtaining outputs/labels, hence we build a (predictive) functional relationship or a conditional probabilistic model of outputs given inputs.

- **Semisupervised learning**: because of the lower cost, there is much more unlabelled than labelled inputs.

- **Weakly supervised learning on aggregates**: because of the lower cost, inputs are at a much higher resolution than outputs.

**Figure:** left: Malaria incidences reported per administrative unit; centre: land surface temperature at night; centre: topographic wetness index
Distribution regression: train on bags, predict on bags

- Represents input bags using the (empirical) kernel mean embeddings 
  \( \hat{\mu}_\alpha = \sum_{i=1}^{N_\alpha} k(\cdot, x^\alpha_i) \).
- Individual labels need not exist - the label is a function of the whole population, e.g. a function of the parameters of the \( \alpha \)-th sampling distribution.
- Individual predictions using RKHS-valued features of individual inputs? Covariate shift.
Output disaggregation: train on bags, predict on individuals

- Weakly supervised ML problem. Classification instance widely studied in ML (learning with label proportions) [Quadrianto et al, 2009; Yu et al, 2013], but little work on regression / other observation likelihoods.
- This work: scalable variational GP machinery + general aggregation model.
Weakly supervised ML problem. Classification instance widely studied in ML (learning with label proportions) [Quadrianto et al, 2009; Yu et al, 2013], but little work on regression / other observation likelihoods.


This work: scalable variational GP machinery + general aggregation model.
An exponential family model \( p(y|\eta) \) for output \( y \in \mathcal{Y} \), with mean parameter \( \eta = \eta(x) \) depending on the individual input \( x \in \mathcal{X} \).

Given a fixed set of points \( x^a_i \in \mathcal{X} \) such that \( \mathbf{x}^a = \{x^a_1, \ldots, x^a_{N_a}\} \), i.e. a bag of points with \( N_a \) individuals.

Observe the aggregate outputs for each of the bags: training data \( (\{x^1_i\}_{i=1}^{N_1}, y^1), \ldots (\{x^n_i\}_{i=1}^{N_n}, y^n) \).

However, we wish to estimate the regression value \( \eta(x^a_i) \) for each individual (in-sample or out-of-sample), not for new bags.

No restrictions on the collection of the individuals, with the bagging process possibly dependent on covariates \( x^a_i \).

To relate the aggregate \( y^a \) and the bag \( \mathbf{x}^a = (x^a_i)_{i=1}^{N_a} \), we use the following bag observation model:

\[
y^a | \mathbf{x}^a \sim p(y | \eta^a), \quad \eta^a = \sum_{i=1}^{N_a} p^a_i \eta(x^a_i),
\]

where \( p^a_i \) is an optional fixed non-negative weight used to adjust the scales.
The total observed incidence (of region $a$) $y^a$ is assumed to follow

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

where $p_i^a$ is the population for pixel $i$ for region $a$ and $\lambda(x_i^a)$ is a model on covariates and is the intended goal.

This model also implies that the unobserved pixel incidences follows:

$$y_i^a \sim \text{Poisson}(y_i^a|p_i^a \lambda(x_i^a))$$

Potential model formulation for $\lambda(\cdot)$:

- Neural network
- **Gaussian process**
Gaussian process

For $\lambda(\cdot)$, we use a Gaussian Process (GP):

$$
\lambda(x^a_i) = \Psi(f(x^a_i)), \quad f \sim GP(\mu, k)
$$

where $\Psi(\cdot)$ is a non-negative valued function taken to be $f^2$ or $\exp(f)$. Some features of Gaussian process here:

- Bayesian non-parametric model, i.e. flexible model
- Provides uncertainty on predictions
- Intractable posterior
- Complexity $O(n^3)$

**Main contribution**
For scalable inference, we derived a scheme based on variational inference, with new proposed bounds and approximations.
Poisson Bag Model

\[ y^a | x^a \sim \text{Poisson} \left( \sum_{i=1}^{N_a} p_i^a \lambda_i^a \right), \quad \lambda_i^a = \Psi(f(x_i^a)), \quad f \sim \text{GP}(\mu, k) \]

Nonnegative link functions: \( \Psi(f) = f^2 \) and \( \Psi(f) = e^f \).

Standard variational bound using inducing points \( u = [f(w_1), \ldots, f(w_m)]^\top \) and a multivariate normal variational posterior \( q(u) \)

\[
\log p(y|\Theta) = \log \int \int p(y, f, u|X, W, \Theta) df du \\
\geq \int \int \log \left\{ p(y|f, \Theta) \frac{p(u)}{q(u)} \right\} p(f|u, \Theta) q(u) df du \quad \text{(Jensen’s inequality)}
\]

\[
= \sum_a y^a \int \log \left( \sum_{i=1}^{N_a} p_i^a \Psi(f(x_i^a)) \right) q(f) df - \sum_a \sum_{i=1}^{N_a} \int p_i^a \Psi(f(x_i^a)) q(f) df \\
- \sum_a \log(y^a!) - KL(q(u)||p(u)) =: \mathcal{L}(q, \Theta),
\]

is still intractable due to aggregation. Needs a further lower bound or an approximation.
Log-sum Lemma

Lemma

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

\[
\int \log \left( \sum_{i=1}^{N} w_i \Psi(v_i) \right) q(v) dv \geq \log \left( \sum_{i=1}^{N} w_i e^{\xi_i} \right),
\]

where

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

Additionally, a Taylor approximation can be used for \( \Psi(f) = f^2 \) (where intractable term essentially becomes \( E \log \|V\|^2 \) where \( V \) is a multivariate normal) – note that log-sum lemma still gives a lower bound in terms of special functions in that case (problematic for backpropagation!)
Inference

For scalability and tractability, we use variational inference \cite{2} with approximating distribution \( q(u) \sim \mathcal{N}(\eta_u, \Sigma_u) \). This leads us to:

1. \( \Psi(f) = f^2 \), an additional approximation using Taylor expansion is applied.

\[
L_1^s := \sum_{a=1}^{n} y^a \zeta^a - \sum_{a=1}^{n} \sum_{i=1}^{N_a} \left\{ (m_i^a)^2 + S_{ii}^a/2 \right\} - KL(q(u)||p(u|W))
\]

2. \( \Psi(f) = \exp(f) \), an additional lower bound is taken.

\[
L_1^e := \sum_{a=1}^{n} y^a \log\left(\sum_{i=1}^{N_a} e^{m_i^a}\right) - \sum_{a=1}^{n} \sum_{i=1}^{N_a} e^{m_i^a+S_{ii}^a/2} - KL(q(u)||p(u|W)).
\]

We can optimise these w.r.t variational parameters \( \{\eta_u, \Sigma_u\} \), kernel parameters, using stochastic gradient descent (SGD).
**Goal:** Predict the underlying incidence rate (represented by the colour)

**Data:** Simulated data, where covariates $x_i^a$ are locations on the swiss roll, and bags are constructed through moving along the $z$-axis.

Tensorflow implementation: [https://github.com/hclllaw/VBAgg](https://github.com/hclllaw/VBAgg)
**Results on Malaria data**

**Goal:** Predict the underlying malaria incidence rate in each 1km by 1km region (pixel)

**Data:** Aggregated incidence of malaria $y^{a}$ at 957 regions, where $N_a$ ranges from 13 to 6,667, with a total of 1,044,683 pixels. Covariates $x^{a}_i \in \mathbb{R}^{18}$, collected by remote sensing.
Malaria results

Data (constant)  Predicted

- Log incidence rate of malaria per 1000.
- Triangle denotes approximate start and end of river location, a widely reported association with malaria.
- Crosses denotes non-train set bags.
Learning on aggregates: the responses are only available at the coarse level. Statistical modelling can be brought to bear in tandem with performant machine learning models.

Increasing confluence between statistics and ML: making use of the well engineered (deep) learning infrastructure, while carefully considering appropriate statistical models for the problem at hand.

Flexibility of the RKHS framework and Gaussian processes as a common ground between machine learning and statistical inference.