Approximate Kernel Methods and Learning on Aggregates

Dino Sejdinovic
joint work with Leon Law, Seth Flaxman, Dougal Sutherland, Kenji Fukumizu, Ewan Cameron, Tim Lucas, Katherine Battle (and many others)

Department of Statistics
University of Oxford

GPSS Workshop on Advances in Kernel Methods, Sheffield
06/09/2018

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.

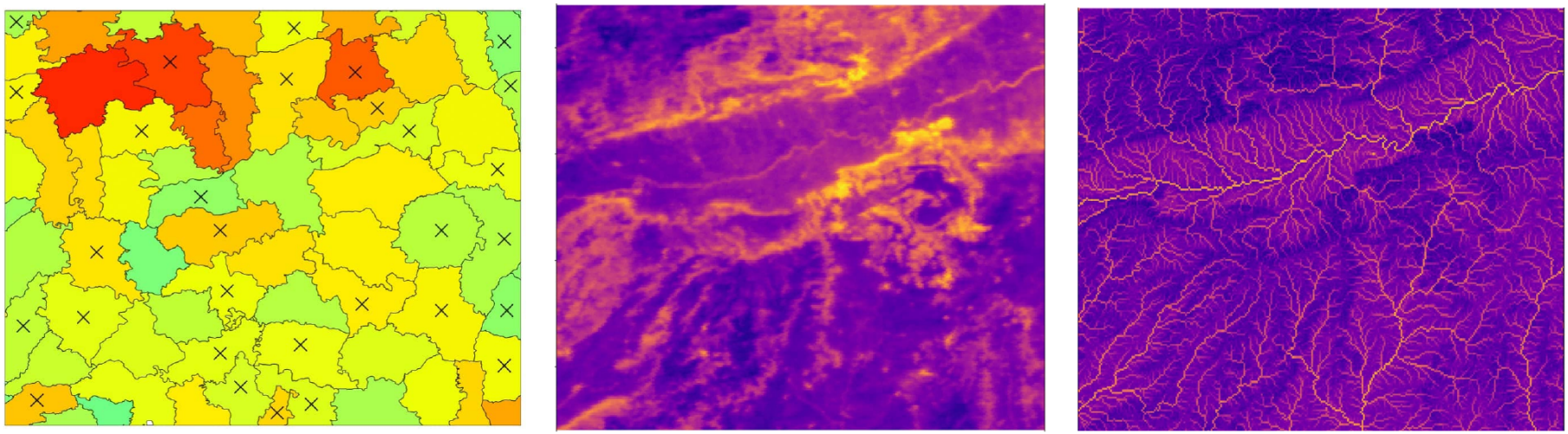


Figure: **left**: Malaria incidences reported per administrative unit; **centre**: land surface temperature at night; **centre**: topographic wetness index

Outline

1. Preliminaries on Kernels and GPs
2. Bayesian Approaches to Distribution Regression
3. Variational Learning on Aggregates with GPs

Outline

Reproducing Kernel Hilbert Space (RKHS)

Definition ([Aronszajn, 1950; Berlinet & Thomas-Agnan, 2004])

Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$, and
2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space*.

Equivalent to the notion of kernel as an *inner product of features*: any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$.

In particular, for any $x, y \in \mathcal{X}$, $k(x, y) = \langle k(\cdot, y), k(\cdot, x) \rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}$.
Thus $\mathcal{H}$ servers as a canonical *feature space* with feature map $x \mapsto k(\cdot, x)$.

- Equivalently, all evaluation functionals $f \mapsto f(x)$ are continuous (norm convergence implies pointwise convergence).
- Moore-Aronszajn Theorem: every positive semidefinite $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel and has a unique RKHS $\mathcal{H}_k$.

Reproducing Kernel Hilbert Space (RKHS)

Definition ([Aronszajn, 1950; Berlinet & Thomas-Agnan, 2004])

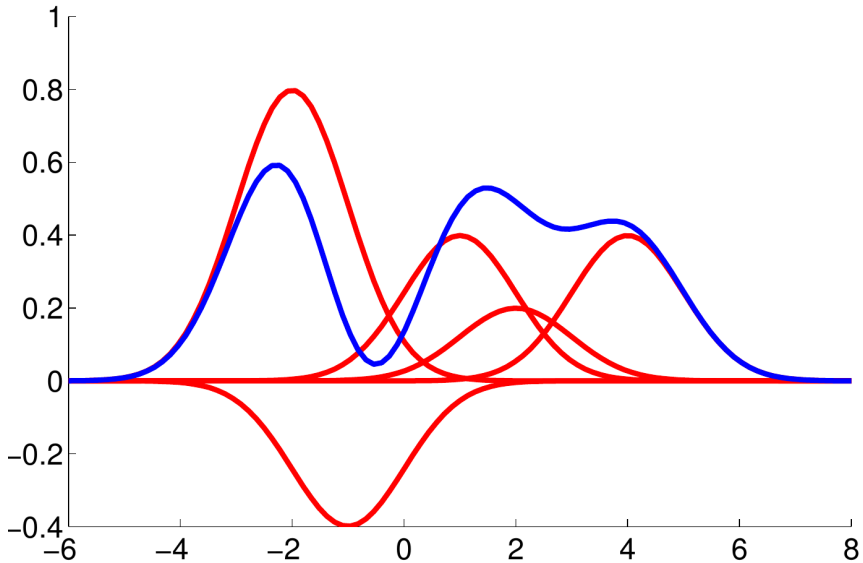
Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$, and
2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space*.

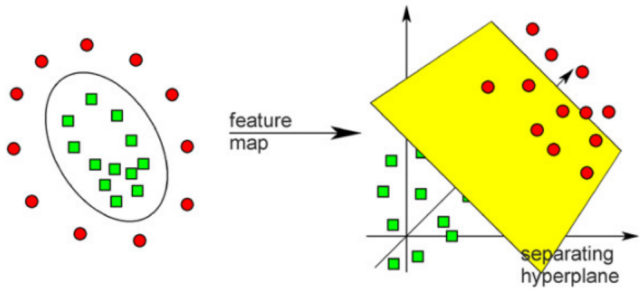
Gaussian RBF kernel $k(x, x') = \exp\left(-\frac{1}{2\gamma^2} \|x - x'\|^2\right)$ has an infinite-dimensional $\mathcal{H}$ with elements $h(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$ and their limits which give completion with respect to the inner product

$$\left\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \sum_{j=1}^{m} \beta_j k(y_j, \cdot) \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j).$$



Kernel Trick and Kernel Mean Trick

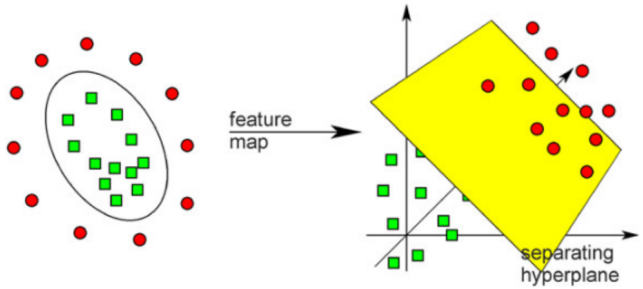
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

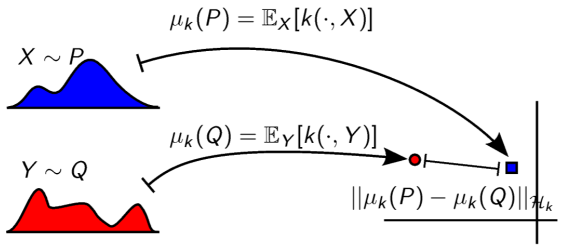
Kernel Trick and Kernel Mean Trick

- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

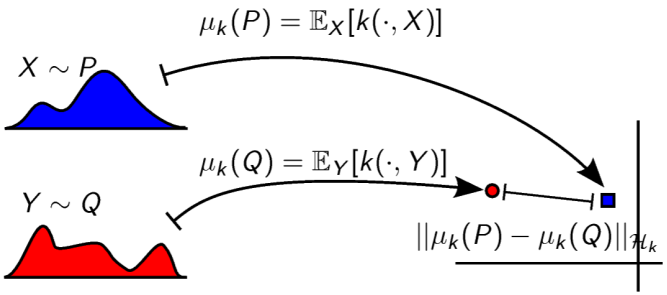
- **RKHS embedding**: implicit feature mean
  [Smola et al, 2007; Sriperumbudur et al, 2010; Muandet et al, 2017]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\phi_1(X), \ldots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y)$
  inner products easy to estimate
  - nonparametric two-sample, independence, conditional independence, interaction testing, learning on distributions

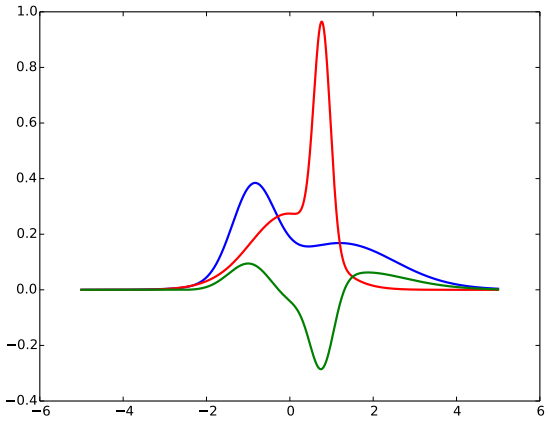


[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS et al, 2013; Muandet et al, 2012; Szabo et al, 2015]

Maximum Mean Discrepancy

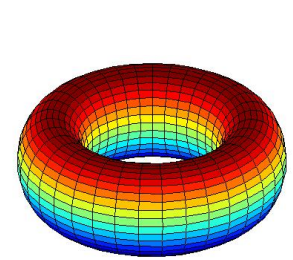
- **Maximum Mean Discrepancy (MMD)** [Borgwardt et al, 2006; Gretton et al, 2007] between P and Q:

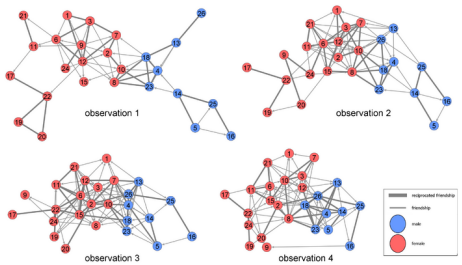




$$\mathrm{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E} f(X) - \mathbb{E} f(Y)|$$

- Characteristic kernels: $\mathrm{MMD}_k(P, Q) = 0$ iff $P = Q$ (also metrizes weak* [Sriperumbudur,2010]).
  - Gaussian RBF $\exp(-\frac{1}{2\sigma^2}\|x - x'\|_2^2)$, Matérn family, inverse multiquadrics.
- Can encode structural properties in the data: kernels on non-Euclidean domains, networks, images, text...





Lorem ipsum dolor sit amet, consecte adipiscing elit. posuere tortor vitae elit. Sed vitae metus a elit bibendum malesuada cras pulvinar. Quisque pellentesque nibh in sem. Curabitur ligula. Suspendisse potenti. Duis sit amet augue eu arcu ultrices auctor. Suspendisse elementum, nunc ut molestie elementum, neque augue vulputate elit, eu blandit enim velit vitae nulla. Duis sed.

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.

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

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

Some uses of MMD

within-sample average similarity

–

between-sample average similarity

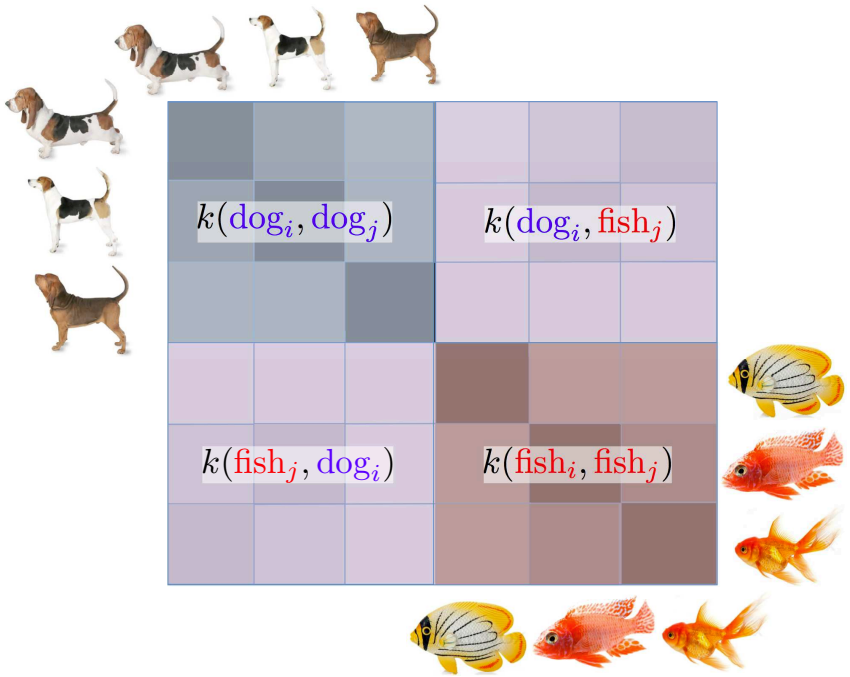


Figure by Arthur Gretton

MMD has been applied to:

- two-sample tests and independence tests (on graphs, text, audio...) [Gretton et al, 2009, Gretton et al, 2012]
- model criticism and interpretability [Lloyd & Ghahramani, 2015; Kim, Khanna & Koyejo, 2016]
- analysis of Bayesian quadrature [Briol et al, 2018]
- ABC summary statistics [Park, Jitkrittum & DS, 2015; Mitrovic, DS & Teh, 2016]
- summarising streaming data [Paige, DS & Wood, 2016]
- traversal of manifolds learned by convolutional nets [Gardner et al, 2015]
- MMD-GAN: training deep generative models [Dziugaite, Roy & Ghahramani, 2015; Sutherland et al, 2017; Li et al, 2017]

$$\text{MMD}_k^2(P, Q) = \mathbb{E}_{X,X' \overset{i.i.d.}{\sim} P} k(X, X') + \mathbb{E}_{Y,Y' \overset{i.i.d.}{\sim} Q} k(Y, Y') - 2\mathbb{E}_{X \sim P, Y \sim Q} k(X, Y).$$

Some uses of MMD

within-sample average similarity

–

between-sample average similarity

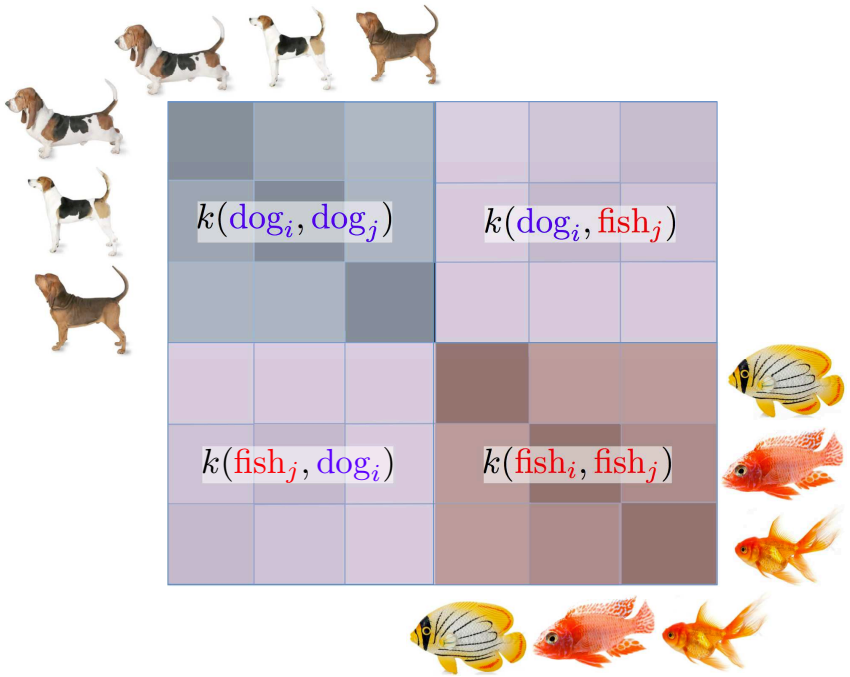


Figure by Arthur Gretton

MMD has been applied to:

- two-sample tests and independence tests (on graphs, text, audio...) [Gretton et al, 2009, Gretton et al, 2012]
- model criticism and interpretability [Lloyd & Ghahramani, 2015; Kim, Khanna & Koyejo, 2016]
- analysis of Bayesian quadrature [Briol et al, 2018]
- ABC summary statistics [Park, Jitkrittum & DS, 2015; Mitrovic, DS & Teh, 2016]
- summarising streaming data [Paige, DS & Wood, 2016]
- traversal of manifolds learned by convolutional nets [Gardner et al, 2015]
- MMD-GAN: training deep generative models [Dziugaite, Roy & Ghahramani, 2015; Sutherland et al, 2017; Li et al, 2017]

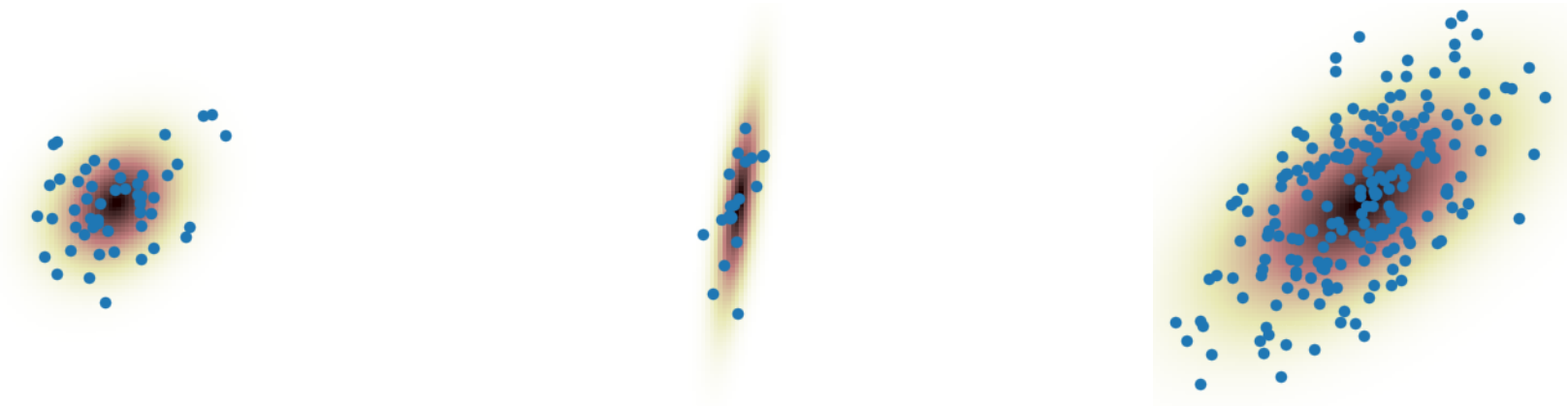
$$\widehat{\text{MMD}}_k^2(P, Q) = \frac{1}{n_x(n_x-1)} \sum_{i \neq j} k(X_i, X_j) + \frac{1}{n_y(n_y-1)} \sum_{i \neq j} k(Y_i, Y_j) - \frac{2}{n_x n_y} \sum_{i,j} k(X_i, Y_j).$$

Kernel Embeddings for Distribution Regression

-0.856 0.562 1.39

- Labels $y_i = f(P_i)$ but observe only $\{x_i^j\}_{j=1}^{N_i} \sim P_i$.
- The goal: build a predictive model $\hat{y}_\star = f(\{x_\star^j\}_{j=1}^{N_\star})$ for a new sample $\{x_\star^j\}_{j=1}^{N_\star} \sim P_\star$.
- Represent each sample with the empirical mean embedding $\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} k(\cdot, x_i^j) \in \mathcal{H}_k$.
- Now can use the induced inner product structure on empirical measures to build a regression model:
  - Linear kernel on the RKHS: $K\left(\hat{\mu}_i, \hat{\mu}_j\right) = \langle \hat{\mu}_i, \hat{\mu}_j \rangle_{\mathcal{H}_k} = \frac{1}{N_i N_j} \sum_{r,s} k(x_i^r, x_j^s)$
  - Gaussian kernel on the RKHS: $K\left(\hat{\mu}_i, \hat{\mu}_j\right) = \exp(-\gamma \|\hat{\mu}_i - \hat{\mu}_j\|_{\mathcal{H}_k}^2) = \exp\left(-\gamma \widehat{\mathrm{MMD}}_k^2\left(P_i, P_j\right)\right)$

Kernel Embeddings for Distribution Regression



- Labels $y_i = f(P_i)$ but observe only $\{x_i^j\}_{j=1}^{N_i} \sim P_i$.
- The goal: build a predictive model $\hat{y}_\star = f(\{x_\star^j\}_{j=1}^{N_\star})$ for a new sample $\{x_\star^j\}_{j=1}^{N_\star} \sim P_\star$.
- Represent each sample with the empirical mean embedding $\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} k(\cdot, x_i^j) \in \mathcal{H}_k$.
- Now can use the induced inner product structure on empirical measures to build a regression model:
  - Linear kernel on the RKHS: $K\left(\hat{\mu}_i, \hat{\mu}_j\right) = \langle \hat{\mu}_i, \hat{\mu}_j \rangle_{\mathcal{H}_k} = \frac{1}{N_i N_j} \sum_{r,s} k(x_i^r, x_j^s)$
  - Gaussian kernel on the RKHS: $K\left(\hat{\mu}_i, \hat{\mu}_j\right) = \exp(-\gamma \|\hat{\mu}_i - \hat{\mu}_j\|_{\mathcal{H}_k}^2) = \exp\left(-\gamma \widehat{\mathrm{MMD}}_k^2\left(P_i, P_j\right)\right)$

Kernel Embeddings for Distribution Regression

- supervised learning where labels are available at the group, rather than at the individual level.

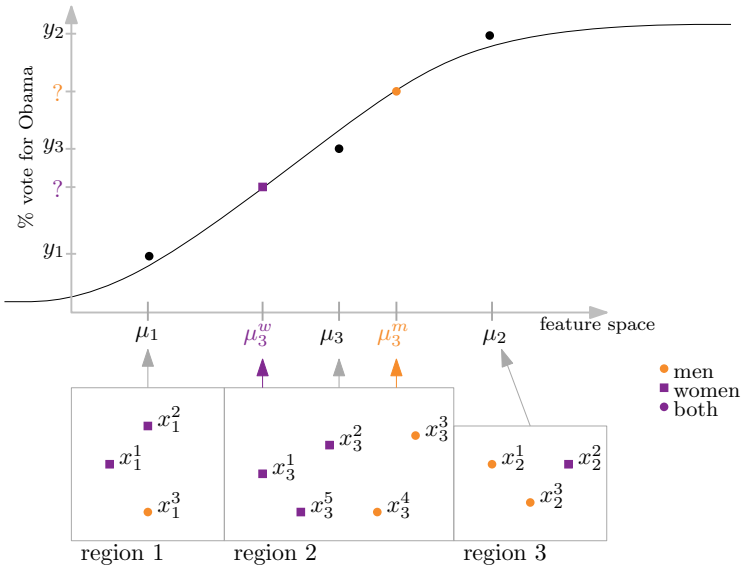


Figure from Flaxman et al, 2015

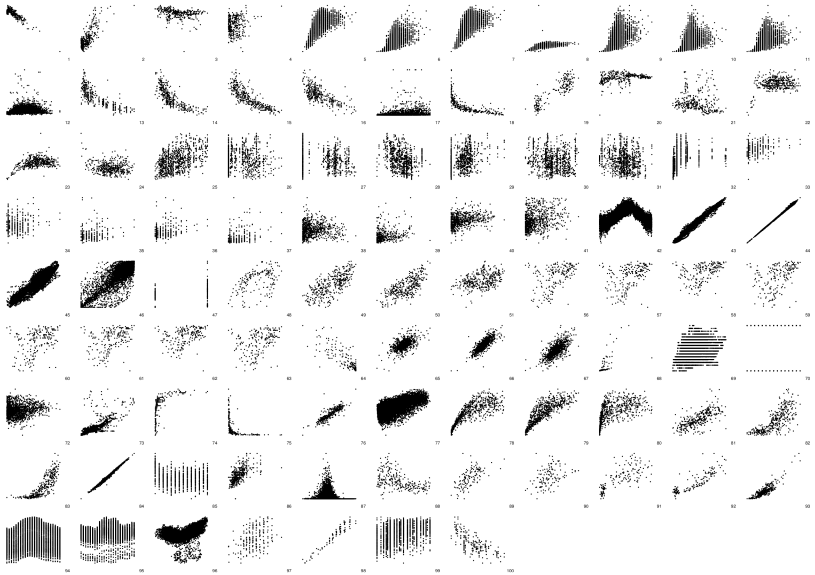


Figure from Mooij et al, 2014

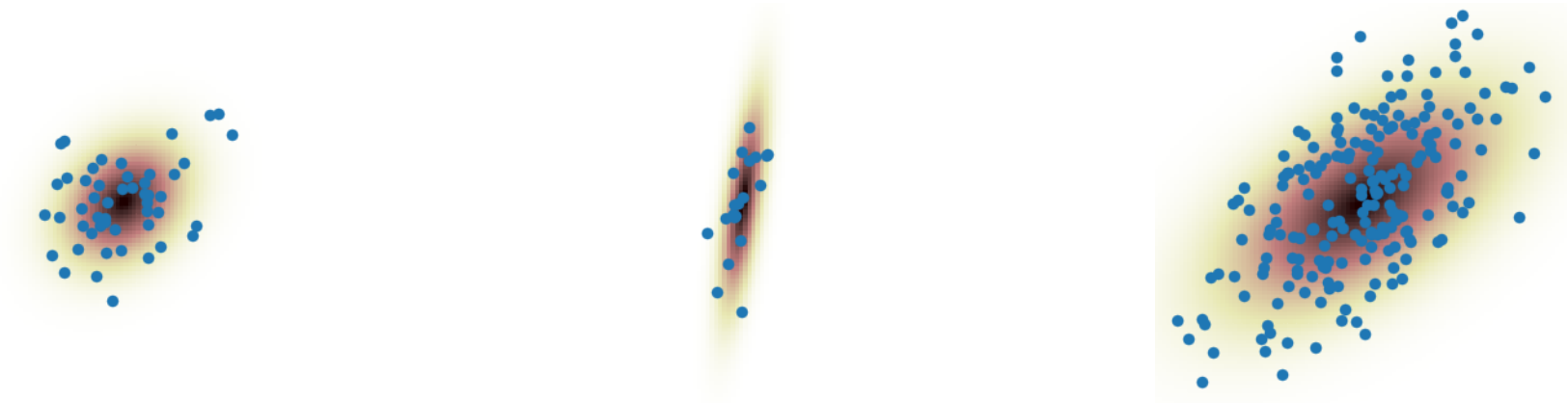
- classifying text based on word features [Yoshikawa et al, 2014; Kusner et al, 2015]
- aggregate voting behaviour of demographic groups [Flaxman et al, 2015; 2016]
- image labels based on a distribution of small patches [Szabo et al, 2016]
- "traditional" parametric statistical inference by learning a function from sets of samples to parameters: ABC [Mitrovic et al, 2016], EP [Jitkrittum et al, 2015]
- identify the cause-effect direction between a pair of variables from a joint sample [Lopez-Paz et al,2015]

Next:

- How to model uncertainty of kernel embeddings when learning on aggregates?
  - A simple Bayesian (GP) model for kernel mean embeddings leads to shrinkage estimators with better predictive performance in high noise regimes.
- How to predict on individual inputs when only aggregate count data is available?
  - Variational bounds leading to improved prediction accuracy and scalability to large datasets, while explicitly taking uncertainty into account.

Outline

Uncertainty in Bag Sizes



-0.856 0.562 1.39

- Recall: we represent each sample with the empirical mean embedding $\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} k(\cdot, x_i^j) \in \mathcal{H}_k$.
- Empirical mean in infinite-dimensional space? Stein's phenomenon? Shrinkage estimators can be better behaved [Muandet et al, 2013]
- These inputs (with or without shrinkage) are *noisy* - we do not observe the true embedding μ_i. Moreover, bags with small N_i are noisier - can this uncertainty be included in the predictive model?

Bayesian Approaches to Distribution Regression
Ho Chung Leon Law, Dougal Sutherland, DS, and Seth Flaxman
AISTATS 2018
http://proceedings.mlr.press/v84/law18a.html

Uncertainty in Mean Embeddings

- The empirical mean embedding is $\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} k(\cdot, x_i^j) \in \mathcal{H}_k$
- Bayesian model for kernel mean embeddings [Flaxman,DS,Cunningham & Filippi, UAI 2016]:
  - Place prior on the RKHS $\mu_i \sim GP\left(m_0(\cdot), r(\cdot,\cdot)\right)$ (requires care due to 0/1 laws [Kallianpur, 1970; Wahba, 1990; Steinwart, 2014+])
  - Posit normal likelihood for the *evaluations of the embedding* at a set of points $\mathbf{u}$:

$$\hat{\mu}_i(\mathbf{u}) | \mu_i(\mathbf{u}) \sim \mathcal{N}(\mu_i(\mathbf{u}), \Sigma_i / N_i)$$

  - Leads to a closed-form GP posterior $\mu_i | \{x_i^j\}$:

$$\mu_i(\mathbf{z}) | \{x_i^j\} \sim \mathcal{N}\Big(R_{\mathbf{zu}}(R_{\mathbf{uu}} + \Sigma_i/N_i)^{-1}(\hat{\mu}_i - m_0) + m_0, \\ R_{\mathbf{zz}} - R_{\mathbf{zu}}(R_{\mathbf{uu}} + \Sigma_i/N_i)^{-1} R_{\mathbf{uz}} \Big)$$

  - Recovers frequentist shrinkage estimator of mean embeddings [Muandet et al, 2013] (but with r instead of k), similar to James-Stein estimator.

Distribution Regression Model

- Model label as a function of the "true" kernel mean embedding:

$$y_i = f(\mu_i) + \epsilon, \qquad \mu_i = \mathbb{E}_{X \sim P_i} k(\cdot, X)$$

- Linear model on the evaluation of kernel mean embedding at a set of "landmark points" $\mathbf{z}$:

$$f(\mu_i) = \beta^\top \mu_i(\mathbf{z})$$

- Can model uncertainty in β (BLR) or in μ_i (shrinkage) or in both (BDR, which requires MCMC due to non-conjugacy).
- **Shrinkage**: Integrate likelihood $y_i \sim \mathcal{N}(f(\mu_i), \sigma^2)$ through the posterior $\mu_i | \{x_i^j\}$ to obtain

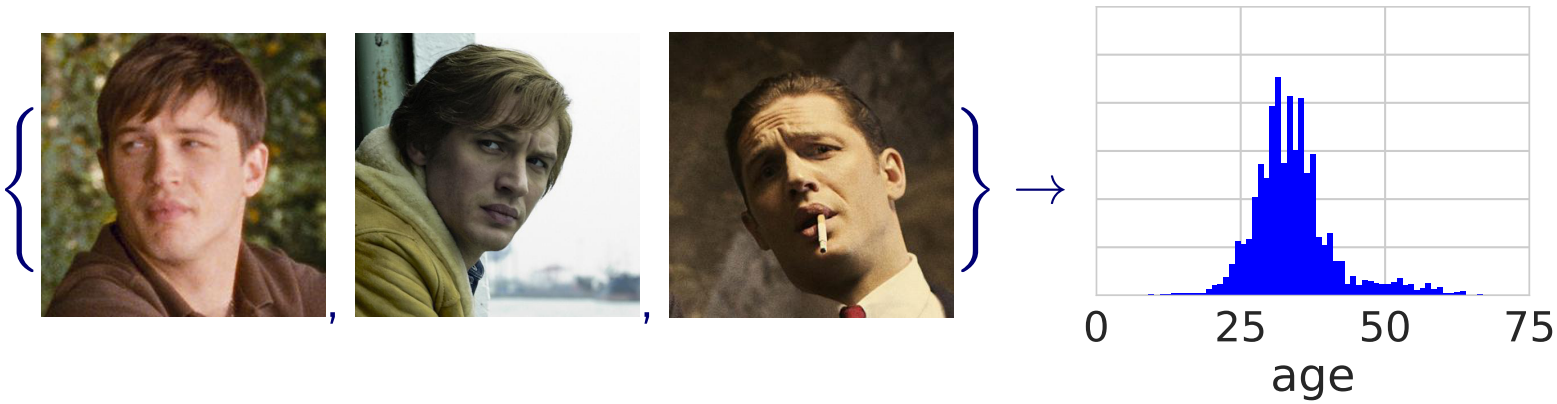
$$y_i \mid \{x_i^j\}, \beta \sim \mathcal{N}(\xi_i^\beta, \nu_i^\beta)$$

$$\xi_i^\beta = \beta^\top R_{\mathbf{z}\mathbf{x}_i} \left(R_{\mathbf{x}_i\mathbf{x}_i} + \frac{\Sigma_i}{N_i} \right)^{-1} (\hat{\mu}_i - m_0) + \beta^\top m_0$$

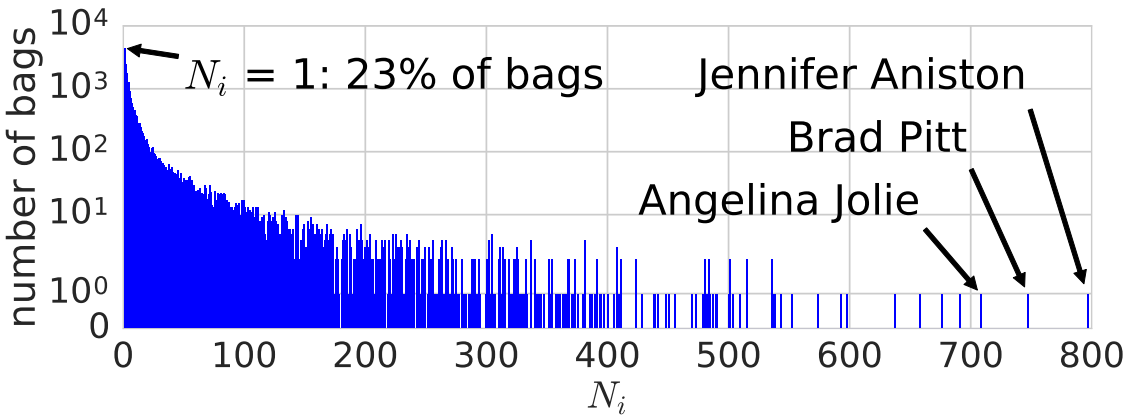
$$\nu_i^\beta = \beta^\top \left(R_{\mathbf{z}\mathbf{z}} - R_{\mathbf{z}\mathbf{x}_i} \left(R_{\mathbf{x}_i\mathbf{x}_i} + \frac{\Sigma_i}{N_i} \right)^{-1} R_{\mathbf{x}_i\mathbf{z}}^\top \right) \beta + \sigma^2.$$

- Can be optimized to find MAP of β, σ^2, kernel parameters, locations of landmark points, ...

Age prediction from images



- IMDb-Wiki database of images with age labels
  - *Very* noisy labels in the dataset
- Distribution regression: group pictures of actors, predict *mean age*
- Image features: last hidden layer from a convolutional neural network by [Rothe et al, IJCV 2016]
- *Lots* of variation in N_i:



Age prediction from images

Propagating uncertainty using shrinkage helps!

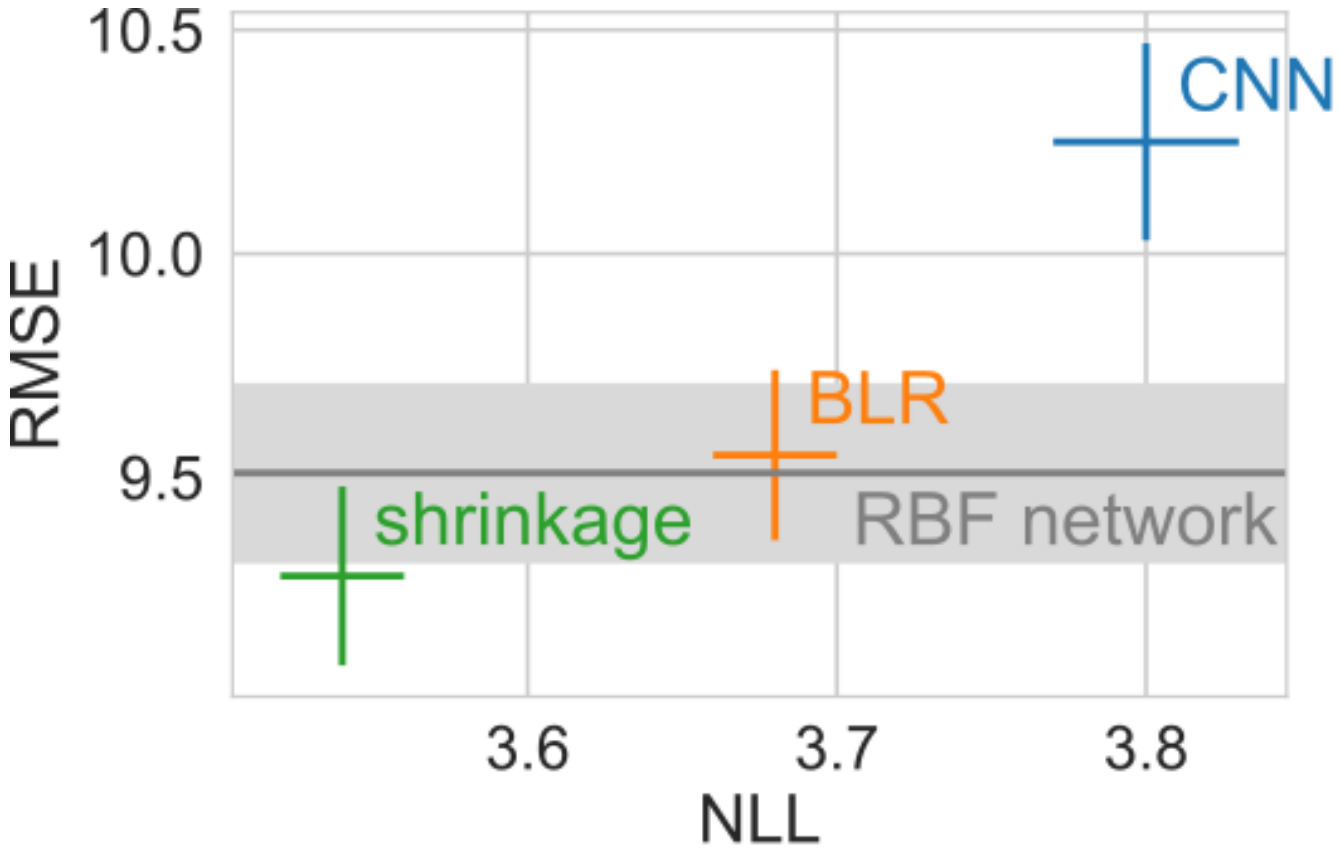
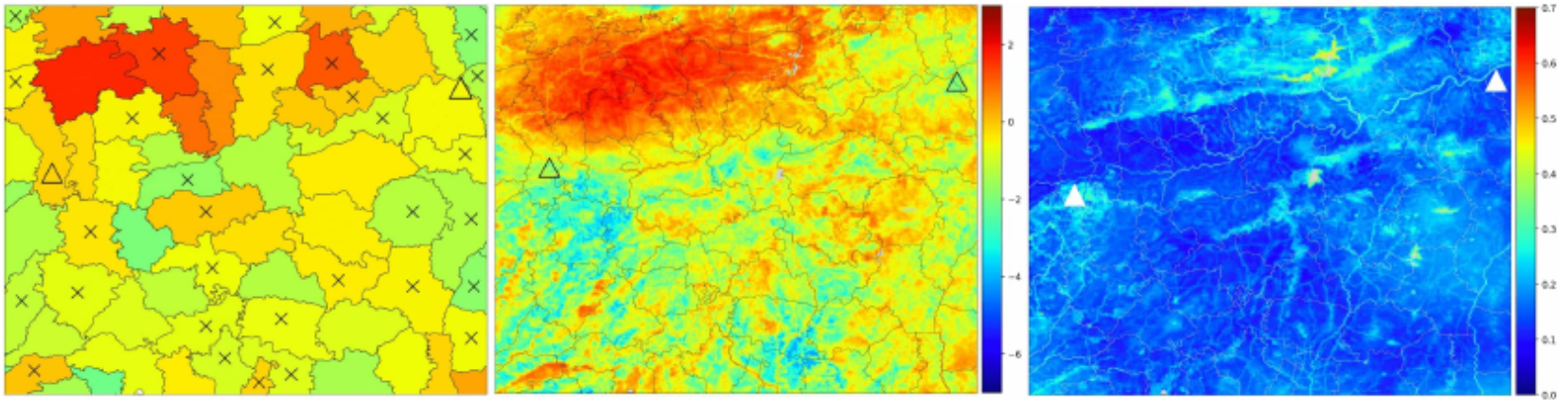


Figure: Results across 10 data splits (means and standard deviations). RBF net is tuned for RMSE, other methods for NLL. CNN takes the mean of the predictive distributions of [Rothe, 2016] for each point in the bag.

Tensorflow implementation: https://github.com/hcllaw/bdr

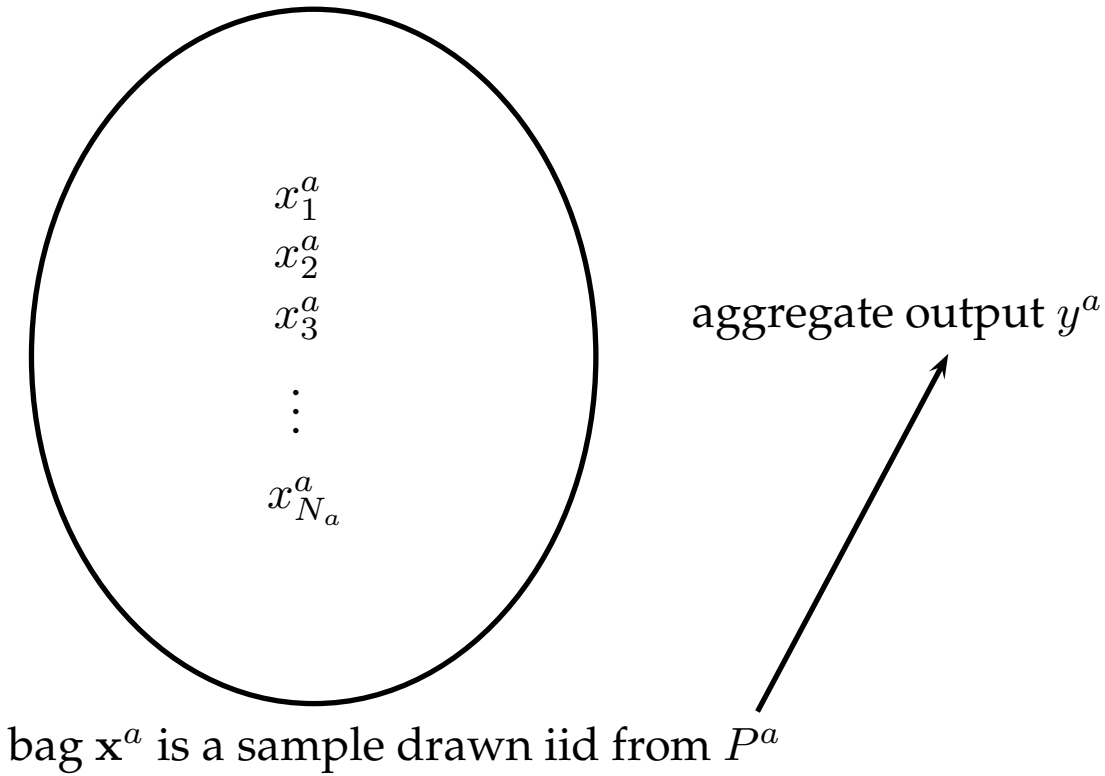
Outline

Disaggregating Aggregate Outputs



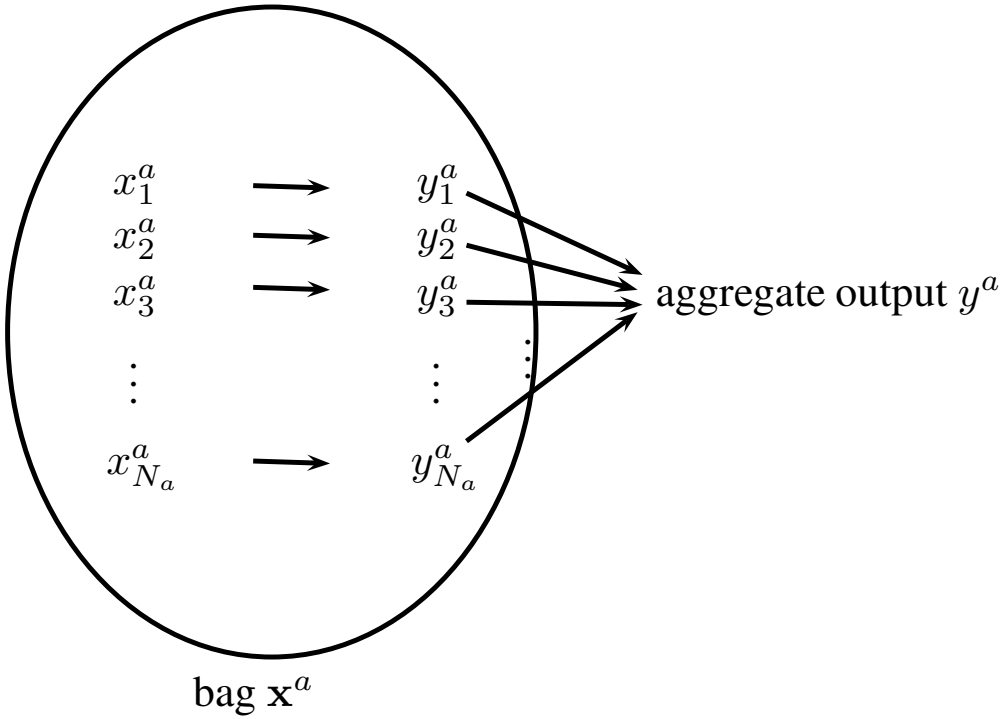
Variational Learning on Aggregate Outputs with Gaussian Processes
H. C. L. Law, DS, E. Cameron, T. C. D. Lucas, S. Flaxman, K. Battle, and K. Fukumizu
to appear in **NIPS 2018**
https://arxiv.org/abs/1805.08463

Distribution regression: train on bags, predict on bags



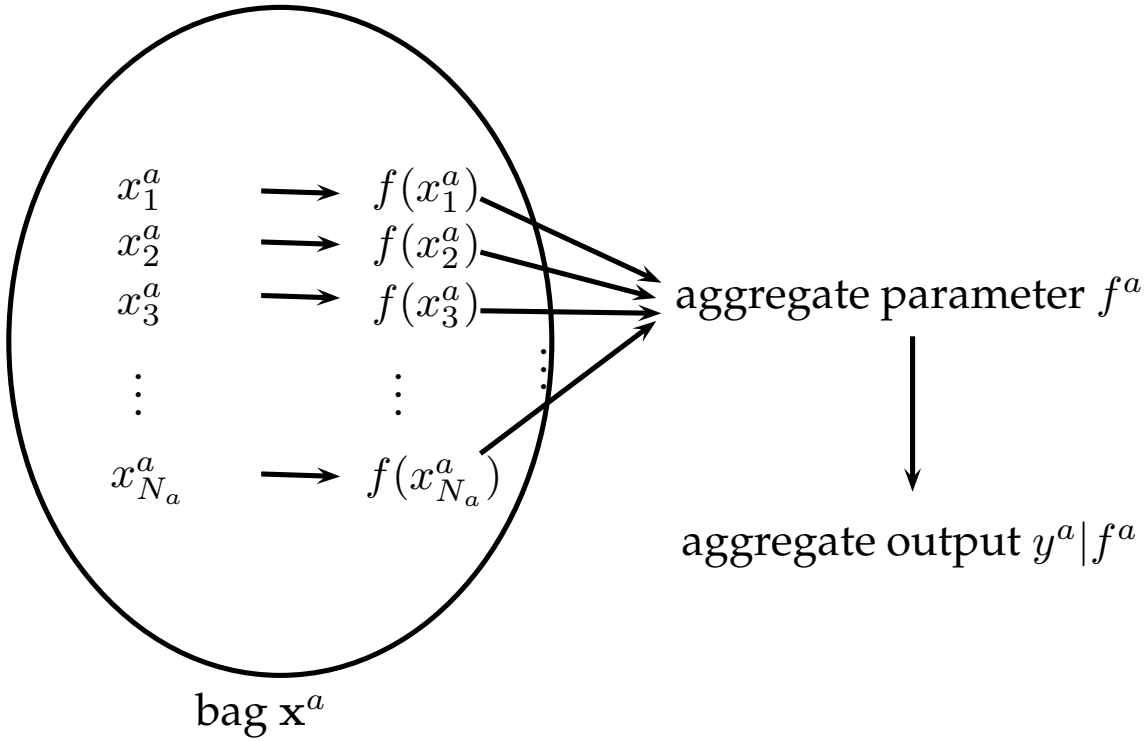
- Individual labels need not exist - the label is a function of the whole population.

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.
- Spatial statistics: 'down-scaling', 'fine-scale modelling' or 'spatial disaggregation' in the analysis of disease mapping, agricultural data, and species distribution modelling, but mostly simple linear models.
- This work: scalable variational GP machinery + general aggregation model.

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.
- Spatial statistics: 'down-scaling', 'fine-scale modelling' or 'spatial disaggregation' in the analysis of disease mapping, agricultural data, and species distribution modelling, but mostly simple linear models.
- This work: scalable variational GP machinery + general aggregation model.

Bag Observation Model: Aggregation in Mean Parameters

- 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 $x_i^a \in \mathcal{X}$ such that $\mathbf{x}^a = \{x_1^a, \ldots, x_{N_a}^a\}$, i.e. a *bag* of points with N_a *individuals*
- Observe the *aggregate outputs* for each of the bags: training data $(\{x_i^1\}_{i=1}^{N_1}, y^1), \ldots (\{x_i^n\}_{i=1}^{N_n}, y^n)$.
- However, we wish to estimate the regression value $\eta(x_i^a)$ 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_i^a.

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*:

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

where p_i^a is an optional fixed non-negative weight used to adjust the scales. .

Poisson Bag Model

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

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 $q(u)$

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

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* $\Psi(v)$*,*

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

where

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

Additionally, a Taylor approximation can be used for $\Psi(f) = f^2$ (where intractable term essentially becomes $\mathbb{E} \log \|V\|^2$ 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!)

Results

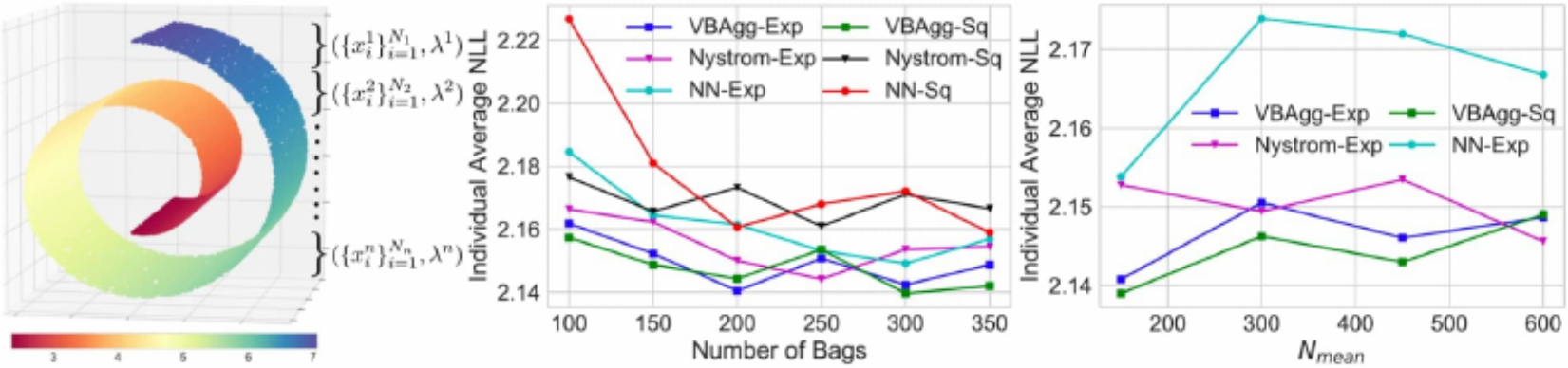


Figure 1: **Left**: Random samples on the Swiss roll manifold. **Middle, Right**: Individual Average NLL on train set for varying number of training bags n and increasing N_{mean}, over 5 repetitions. Constant prediction within bag gives a NLL of 2.22. bag-pixel model gives NLL above 2.4 for the varying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

Tensorflow implementation: https://github.com/hcllaw/VBAgg

Results

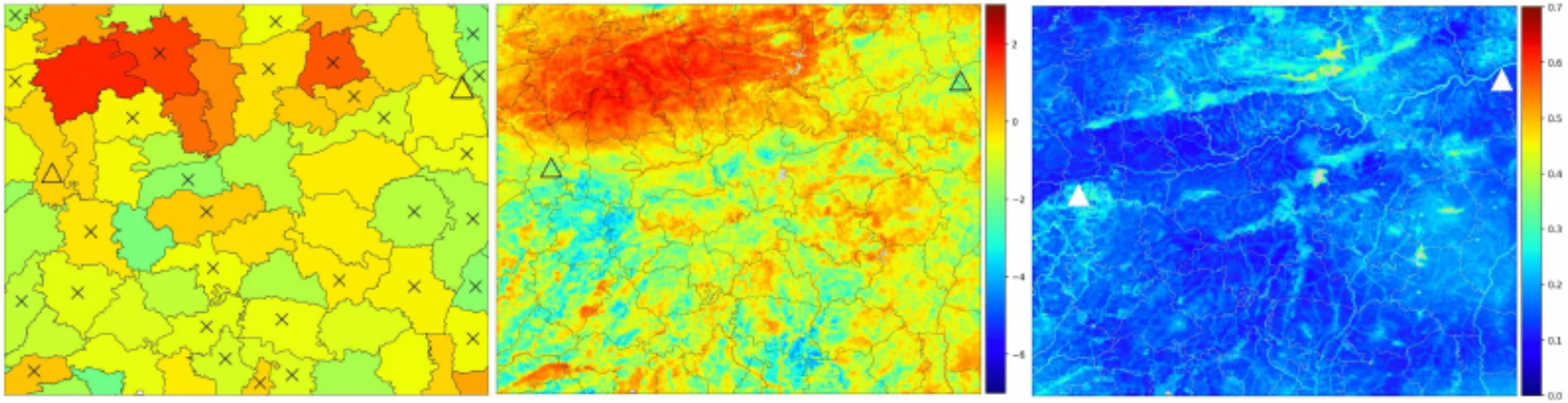


Figure 2: Triangle denotes approximate start and end of river location, crosses denotes non-train set bags. Malaria incidence rate λ_i^a is per 1000 people. **Left, Middle**: $\log(\hat{\lambda}_i^a)$, with constant model (Left), and VBAgg-Obj-Sq (tuned on $\mathcal{L}_1^s$) (Middle). **Right**: Standard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

Tensorflow implementation: https://github.com/hcllaw/VBAgg

Summary

- Both contributions study learning on aggregates, i.e. where the responses are available at the group level, and demonstrate how statistical modelling can be brought to bear.
- Increasing confluence between statistical modelling and machine learning – making use of the well engineered deep learning (black-box) infrastructure, while carefully considering appropriate statistical models.
- Flexibility of the RKHS framework and Gaussian processes as a common ground between deep learning and statistical inference.

References

- Ho Chung Leon Law, Dougal J. Sutherland, DS, and Seth Flaxman, Bayesian Approaches to Distribution Regression, in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018, PMLR 84:1167-1176.
- Ho Chung Leon Law, DS, Ewan Cameron, Tim Lucas, Seth Flaxman, Katherine Battle, and Kenji Fukumizu, Variational Learning on Aggregate Outputs with Gaussian Processes, in *Advances in Neural Information Processing Systems (NIPS)*, 2018, to appear. *ArXiv e-prints:1805.08463*, 2018.