Kernel Methods, Embeddings and Aggregates

Dino Sejdinovic

Department of Statistics
University of Oxford

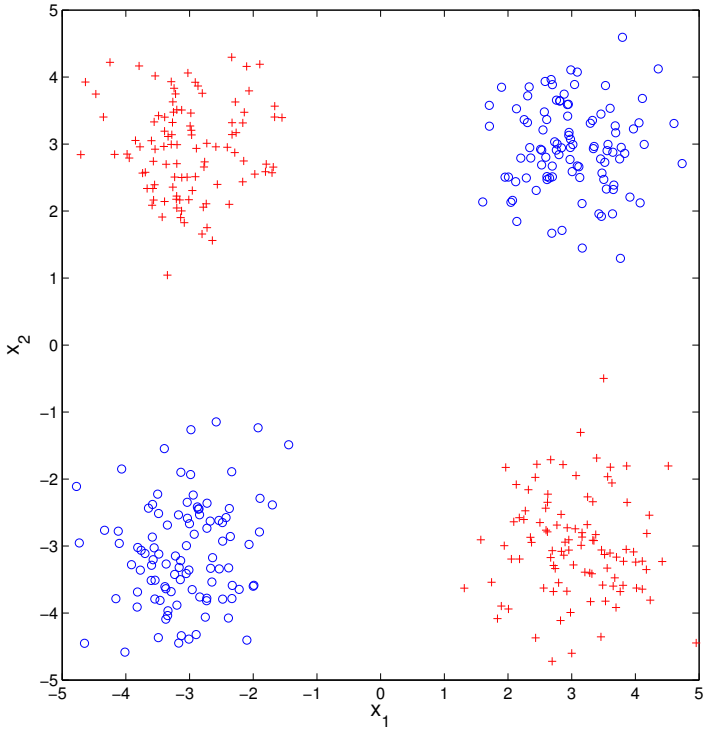
Imperial College London
28/11/2018

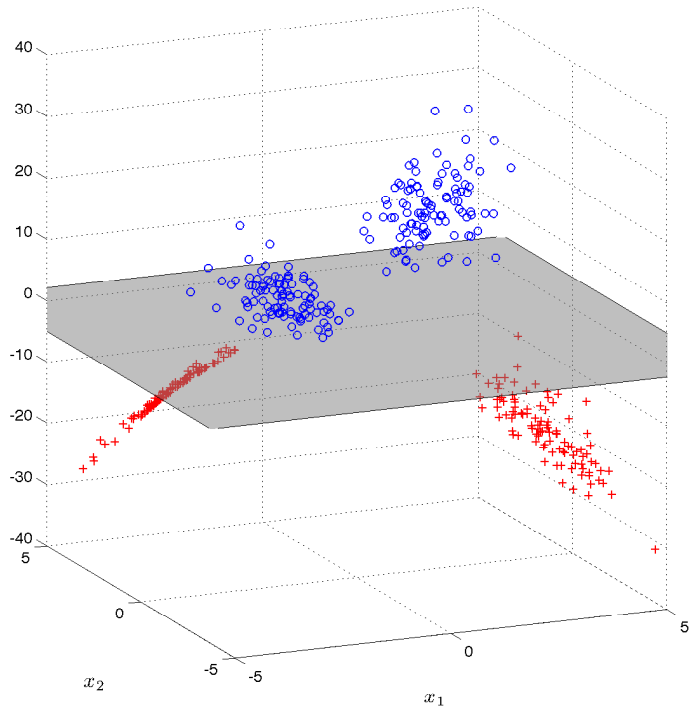
Outline

Outline

Feature maps and feature spaces

Feature maps





- No linear classifier separates red from blue.
- Linear separation after mapping to a **higher dimensional feature space**:

$$\mathbb{R}^2 \ni \begin{pmatrix} x^{(1)} & x^{(2)} \end{pmatrix}^\top = x \quad \mapsto \quad \varphi(x) = \begin{pmatrix} x^{(1)} & x^{(2)} & x^{(1)}x^{(2)} \end{pmatrix}^\top \in \mathbb{R}^3$$

Feature maps and kernel trick

- Kernel methods on a generic domain $\mathcal{X}$ allow constructing nonlinear methods after mapping to a higher dimensional feature space:

$$\varphi : \mathcal{X} \to \mathbb{R}^D$$

- Typically need only inner products $\varphi(x_i)^\top \varphi(x_j)$ are required and the coordinates of the maps $\varphi(x_i) \in \mathbb{R}^D$ need not be computed explicitly - inner product between features can be a simple function (kernel) of x_i and x_j.
- Polynomial kernel $k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j) = (1 + x_i^\top x_j)^q$ on $\mathbb{R}^p$ computes q-order features - never need to compute explicit feature expansion of dimension $D = \binom{p+q}{q}$ where this inner product is defined.
- For example, if $p = 2$ and $q = 2$, we have the feature map with quadratic and mixed non-linearities,

$$\varphi(x) = \left(1, \sqrt{2}x^{(1)}, \sqrt{2}x^{(2)}, \sqrt{2}x^{(1)}x^{(2)}, \left(x^{(1)}\right)^2, \left(x^{(2)}\right)^2\right)^\top \in \mathbb{R}^6$$

Kernel: an inner product between feature maps

Definition (kernel)

Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if there exists a *Hilbert space* and a map $\varphi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} .$$

- Almost no conditions on $\mathcal{X}$ (in particular, $\mathcal{X}$ itself need not have an inner product).
- Think of kernel as a *similarity measure between input features*

A single kernel can correspond to multiple pairs of underlying feature maps and feature spaces. For a simple example, consider $\mathcal{X} := \mathbb{R}^p$:

$$\phi_1(x) = x \qquad \text{and} \qquad \phi_2(x) = \left[\frac{x_1}{\sqrt{2}}, \cdots, \frac{x_p}{\sqrt{2}}, \frac{x_1}{\sqrt{2}}, \cdots, \frac{x_p}{\sqrt{2}} \right]^\top .$$

Both ϕ_1 and ϕ_2 are valid feature maps (with feature spaces $\mathcal{H}_1 = \mathbb{R}^p$ and $\mathcal{H}_2 = \mathbb{R}^{2p}$) of kernel $k(x, x') = x^\top x'$.

Positive semidefinite functions

If we are given a "measure of similarity" with two arguments, $k(x, x')$, how can we determine if it is a valid kernel?

1. Find a feature map?
   - Sometimes not obvious (especially if the feature vector is infinite-dimensional)
2. A simpler direct property of the function: positive semidefiniteness.

Positive semidefinite functions

Definition (Positive semidefinite functions)

A symmetric function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite if $\forall n \geq 1$, $\forall (a_1, \dots a_n) \in \mathbb{R}^n$, $\forall (x_1, \dots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \kappa(x_i, x_j) \geq 0.$$

- Kernel $k(x, y) := \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$ for a Hilbert space $\mathcal{H}$ is positive semidefinite.

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \varphi(x_i), a_j \varphi(x_j) \rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^{n} a_i \varphi(x_i) \right\|_{\mathcal{H}}^2 \geq 0. \end{aligned}$$

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])

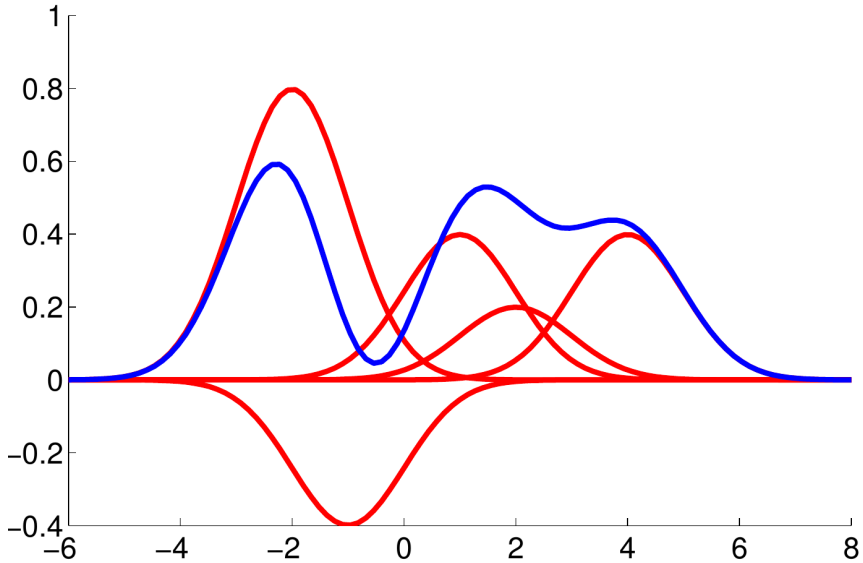
Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

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

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



Representer Theorem

Representer theorem

Standard supervised learning setup: we are given a set of paired observations $(x_1, y_1), \ldots (x_n, y_n)$.
Goal: find the function f^* in the RKHS $\mathcal{H}$ which solves the regularized empirical risk minimization problem.

$$\min_{f \in \mathcal{H}} \hat{R}(f) + \Omega\left(\|f\|_{\mathcal{H}}^2\right),$$

where empirical risk is

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i), x_i),$$

and Ω is a non-decreasing function.

- Classification: L could be a hinge loss $L(y, f(x), x) = (1 - yf(x))_+$ or a logistic loss $L(y, f(x), x) = \log\left(1 + \exp(-yf(x)\right)$.
- Regression: $L(y, f(x), x) = (y - f(x))^2$.

Representer theorem

Theorem (Representer Theorem)

There is a solution to

$$\min_{f \in \mathcal{H}} \hat{R}(f) + \Omega\left(\|f\|_{\mathcal{H}}^2\right)$$

that takes the form

$$f^* = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i).$$

If Ω is strictly increasing, all solutions have this form.

Representer theorem: proof

Proof: Denote f_s projection of f onto the subspace

$$\operatorname{span}\{k(\cdot, x_i) : \ i = 1, \ldots, n\}$$

such that

$$f = f_s + f_\perp,$$

where $f_s = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ and $f_\perp$ is orthogonal to $\operatorname{span}\{k(\cdot, x_i) : \ i = 1, \ldots, n\}$.

Regularizer:

$$\|f\|_{\mathcal{H}}^2 = \|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_s\|_{\mathcal{H}}^2 ,$$

then

$$\Omega\left(\|f\|_{\mathcal{H}}^2\right) \geq \Omega\left(\|f_s\|_{\mathcal{H}}^2\right).$$

Representer theorem: proof

Proof (cont.): Individual terms $f(x_i)$ in the loss:

$$f(x_i) = \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle f_s + f_\perp, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle f_s, k(\cdot, x_i)\rangle_{\mathcal{H}} \,,$$

so

$$L(y_i, f(x_i), x_i) = L(y_i, f_s(x_i), x_i) \forall i \implies \hat{R}(f) = \hat{R}(f_s).$$

Hence

- The empirical risk only depends on the components of f lying in the subspace spanned by canonical features.
- Regularizer $\Omega(\ldots)$ is minimized when $f = f_s$.
- If Ω is strictly non-decreasing, then $\|f_\perp\|_{\mathcal{H}} = 0$ is required at the minimum.

A Simple Kernel Method: Kernel Ridge Regression

Regularised Least Squares

We are given n training points $\{x_i\}_{i=1}^n$ in $\mathbb{R}^p$: Define some $\lambda > 0$. Our goal is:

$$\begin{aligned} w^* &= \arg\min_{w \in \mathbb{R}^p} \left(\sum_{i=1}^n (y_i - x_i^\top w)^2 + \lambda \|w\|^2 \right) \\ &= \arg\min_{w \in \mathbb{R}^p} \left(\|\mathbf{y} - \mathbf{X}w\|^2 + \lambda \|w\|^2 \right), \end{aligned}$$

Solution is:

$$w^* = \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{y},$$

which is the standard regularised least squares solution.

Kernel ridge regression

Use features $\phi(x_i)$ in the place of x_i:

$$w^* \quad = \quad \arg\min_{w \in \mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle w, \phi(x_i) \rangle_{\mathcal{H}} \right)^2 + \lambda \|w\|_{\mathcal{H}}^2 \right).$$

E.g. for finite dimensional feature spaces,

$$\phi_p(x) = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^\ell \end{bmatrix} \qquad \phi_s(x) = \begin{bmatrix} \sin(x) \\ \cos(x) \\ \sin(2x) \\ \vdots \\ \cos\left(\frac{\ell}{2}x\right) \end{bmatrix}$$

In finite dimensions, w is a vector of length ℓ giving weight to each of these features so that learned function is $f_w(x) = w^\top \phi(x)$. Feature vectors can also have *infinite* length.

Kernel ridge regression

Recall that feature maps ϕ and feature spaces $\mathcal{H}$ are not unique, but RKHS $\mathcal{H}_k$ is. Thus, we can identify w with the function f_w (there is an isometry between w and f_w: $\|w\|_{\mathcal{H}} = \|f_w\|_{\mathcal{H}_k}$ regardless of the choice of the feature space $\mathcal{H}$) and write

$$
\begin{aligned}
f^* &= \arg\min_{f \in \mathcal{H}_k} \left(\sum_{i=1}^{n} \left(y_i - \langle f, k(\cdot, x_i) \rangle_{\mathcal{H}} \right)^2 + \lambda \|f\|_{\mathcal{H}_k}^2 \right) \\
&= \arg\min_{f \in \mathcal{H}_k} \left(\sum_{i=1}^{n} \left(y_i - f(x_i) \right)^2 + \lambda \|f\|_{\mathcal{H}_k}^2 \right).
\end{aligned}
$$

Kernel ridge regression

Recall the representer theorem: f is a linear combination of feature space mappings of data points

$$f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i).$$

Then

$$\begin{aligned} \sum_{i=1}^{n} \left(y_i - \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}_k}\right)^2 + \lambda \|f\|^2_{\mathcal{H}_k} &= \|\mathbf{y} - \mathbf{K}\alpha\|^2 + \lambda \alpha^\top \mathbf{K}\alpha \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{K}\alpha + \alpha^\top \left(\mathbf{K}^2 + \lambda \mathbf{K}\right)\alpha \end{aligned}$$

Differentiating wrt α and setting this to zero, we get

$$\alpha^* = (\mathbf{K} + \lambda I_n)^{-1} y.$$

Recall: $\frac{\partial \alpha^\top U \alpha}{\partial \alpha} = (U + U^\top)\alpha, \qquad \frac{\partial v^\top \alpha}{\partial \alpha} = \frac{\partial \alpha^\top v}{\partial \alpha} = v$

Parameter selection for KRR

Given the objective

$$f^* \quad = \quad \arg\min_{f \in \mathcal{H}_k} \left(\sum_{i=1}^{n} \left(y_i - f(x_i) \right)^2 + \lambda \| f \|^2_{\mathcal{H}_k} \right).$$

How do we choose

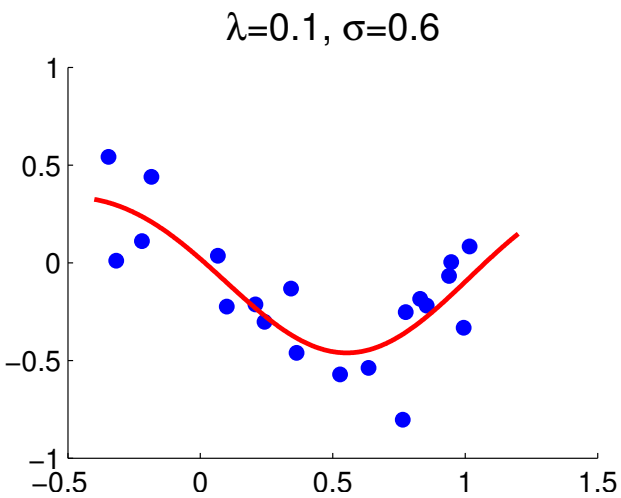
- The regularization parameter λ?
- The kernel parameter: for Gaussian kernel, σ in

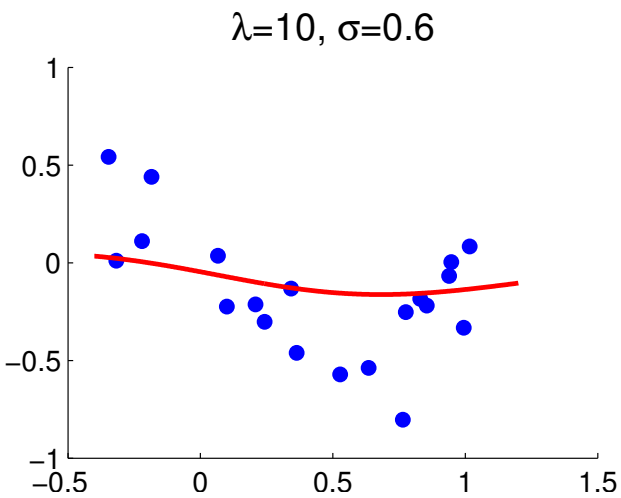
$$k(x, y) = \exp\left(\frac{-\|x - y\|^2}{\sigma} \right).$$

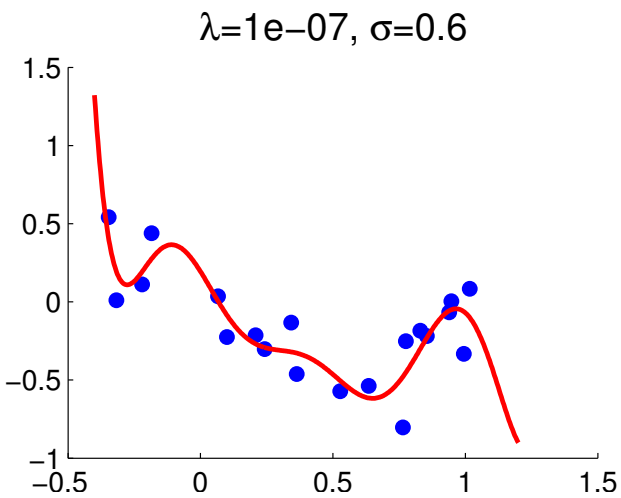
Beware: Gaussian kernel has many different parametrisations in the literature and software packages!

Typically use cross-validation.

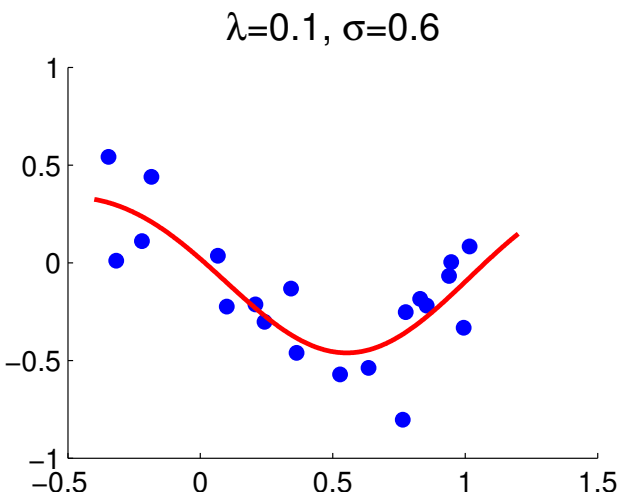
Choice of λ

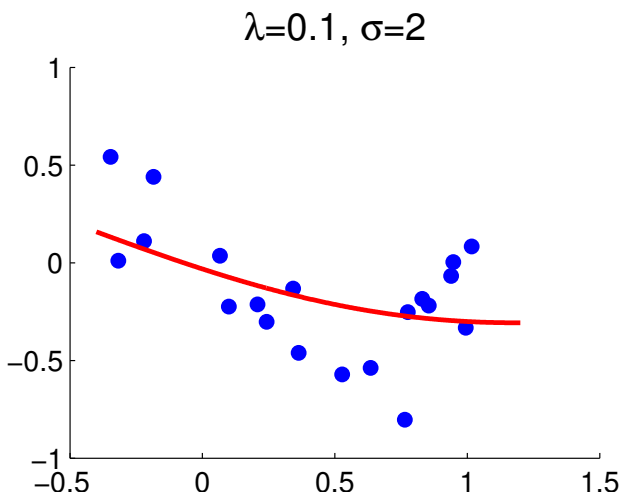


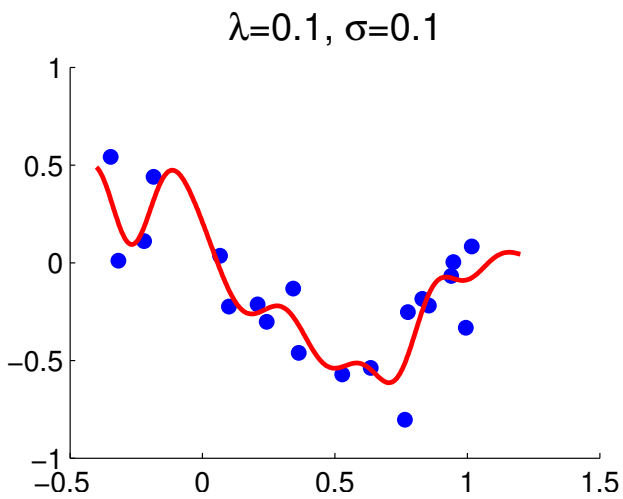




Choice of σ







Kernel families and operations with kernels

Examples of kernels

- *Linear*: $k(x, x') = x^\top x'$.
- *Polynomial*: $k(x, x') = (c + x^\top x')^m$, $c \in \mathbb{R}$, $m \in \mathbb{N}$.
- *Periodic (1d)*: $k(x, x') = \exp\left(-\frac{2\sin^2(\pi|x-x'|/p)}{\gamma^2}\right)$, period p, $\gamma > 0$.
- *Exponential*: $k(x, x') = \exp(\frac{x^\top x'}{\gamma})$, $\gamma > 0$.
- *Gaussian RBF*: $k(x, x') = \exp\left(-\frac{1}{2\gamma^2}\|x - x'\|^2\right)$, $\gamma > 0$.
- *Laplace*: $k(x, x') = \exp\left(-\frac{1}{\gamma}\|x - x'\|\right)$, $\gamma > 0$.
- *Rational quadratic*: $k(x, x') = \left(1 + \frac{\|x-x'\|^2}{2\alpha\gamma^2}\right)^{-\alpha}$, $\alpha, \gamma > 0$.
- *Brownian covariance*: $k(x, x') = \frac{1}{2}\left(\|x\|^\gamma + \|x'\|^\gamma - \|x - x'\|^\gamma\right)$, $\gamma \in [0, 2]$.

all norms are 2-norms unless specified otherwise

Matérn Family

$$k(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{\gamma}\|x-x'\|\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{\gamma}\|x-x'\|\right), \quad \nu>0, \gamma>0,$$

where K_ν is the modified Bessel function of the second kind of order ν.

- $\nu = 1/2$: $k(x,x') = \exp\left(-\frac{1}{\gamma}\|x-x'\|\right)$
- $\nu = 3/2$: $k(x,x') = \left(1+\frac{\sqrt{3}}{\gamma}\|x-x'\|\right)\exp\left(-\frac{\sqrt{3}}{\gamma}\|x-x'\|\right)$
- $\nu = 5/2$:
  $k(x,x') = \left(1+\frac{\sqrt{5}}{\gamma}\|x-x'\|+\frac{5}{3\gamma^2}\|x-x'\|^2\right)\exp\left(-\frac{\sqrt{5}}{\gamma}\|x-x'\|\right)$
- as $\nu\to\infty$, converges to Gaussian RBF $k(x,x') = \exp\left(-\frac{1}{2\gamma^2}\|x-x'\|^2\right)$

Matérn family norms penalize the derivatives of f. In particular, for $\nu = s+1/2$, it penalizes the derivatives up to order $s+1$, e.g. for $\nu = 3/2$ and in one dimension:

$$\|f\|^2_{\mathcal{H}_k} \propto \int f''(x)^2dx + \frac{6}{\gamma^2}\int f'(x)^2dx + \frac{9}{\gamma^4}\int f(x)^2dx$$

New kernels from old: sums, transformations

The great majority of useful kernels are built from simpler kernels.

Lemma (Sums of kernels are kernels)

Given $\alpha > 0$ and k, k_1 and k_2 all kernels on $\mathcal{X}$, then αk and $k_1 + k_2$ are kernels on $\mathcal{X}$.

To prove this, just check inner product definition (features get scaled with $\sqrt{\alpha}$ or concatenated). A difference of kernels need not be a kernel (**why?**)

Lemma (Space transformation)

Let $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ be sets, and consider any map $s : \mathcal{X} \to \widetilde{\mathcal{X}}$. Let $\tilde{k}$ be a kernel on $\widetilde{\mathcal{X}}$. Then $k(x, x') = \tilde{k}(s(x), s(x'))$ is a kernel on $\mathcal{X}$.

Proof: if $\tilde{\varphi}$ is a feature map for $\tilde{k}$, then $\varphi = \tilde{\varphi} \circ s$ is a feature map for k.

New kernels from old: products

Lemma (Products of kernels are kernels)

Given k_1 on $\mathcal{X}_1$ and k_2 on $\mathcal{X}_2$, then $k_1 \times k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$.

The general proof requires technical details about tensor products, but the main idea comes from simple linear algebra. Consider finite-dimensional feature maps $k(x, x') = \varphi(x)^\top \varphi(x')$ and $l(y, y') = \psi(y)^\top \psi(y')$, with $\varphi(x) \in \mathbb{R}^M$, and $\psi(y) \in \mathbb{R}^N$. Note that a valid inner product between matrices $A \in \mathbb{R}^{M \times N}$ and $B \in \mathbb{R}^{M \times N}$ is

$$\langle A, B \rangle_{\mathbb{R}^{M \times N}} = \operatorname{trace}(A^\top B) = \sum_{i=1}^{M} \sum_{j=1}^{N} A_{ij} B_{ij}.$$

Then

$$\begin{aligned} k(x, x')\, l(y, y') &= \varphi(x)^\top \varphi(x') \psi(y')^\top \psi(y) \\ &= \operatorname{tr}(\psi(y) \varphi(x)^\top \varphi(x') \psi(y')^\top) \\ &= \left\langle \varphi(x)\psi(y)^\top, \varphi(x')\psi(y')^\top \right\rangle_{\mathbb{R}^{M \times N}}. \end{aligned}$$

Thus product kernel has (matrix-valued) features $A(x, y) = \varphi(x)\psi(y)^\top$.

More simply, *Kronecker product of positive definite matrices is positive definite*.

More products and Taylor expansions

Lemma (Products of kernels are kernels)

Given kernels k_1 and k_2 on $\mathcal{X}$, $k_1 \times k_2$ is a kernel on $\mathcal{X}$.

Proof: It is certainly a kernel on $\mathcal{X} \times \mathcal{X}$, so just consider space transformation $s : \mathcal{X} \to \mathcal{X} \times \mathcal{X}$ with $s(x) = (x, x)$.

More simply, *Hadamard (entrywise) product of positive definite matrices is positive definite*.

As a corollary:

$$k(x, x') = c + \sum_{j=1}^{d} a_j \langle x, x' \rangle^d \tag{1}$$

is certainly a kernel. Readily extends to

$$k(x, x') = g\left(\langle x, x' \rangle\right) \tag{2}$$

for an analytic function g with nonnegative Taylor coefficients, e.g., $\exp$.

Gaussian RBF is a kernel

As a product of an exponential kernel and a kernel with 1-d feature $x \mapsto \exp\left(-\frac{\|x\|^2}{2\gamma^2}\right)$.

$$\begin{aligned} k(x, x') &= \exp\left(-\frac{1}{2\gamma^2}\|x - x'\|^2\right) \\ &= \exp\left(-\frac{\|x\|^2}{2\gamma^2}\right)\exp\left(-\frac{\|x'\|^2}{2\gamma^2}\right)\exp\left(\frac{1}{\gamma^2}\langle x, x'\rangle\right) \end{aligned}$$

All of the proofs above are constructive: they give a way of constructing new features from old. But the resulting features quickly become very difficult to interpret. There is another, much cleaner way to do this: Mercer's Theorem.

Mercer's theorem

- Assume that $\mathcal{X}$ is a compact metric space, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel and fix a finite measure ν on $\mathcal{X}$ with supp$\nu = \mathcal{X}$.
- To k we can associate a certain operator T_k on $L_2(\mathcal{X}; \nu)$ which is compact, positive and self-adjoint

$$[T_k f](y) = \int f(x)k(x, y)\nu(dx)$$

- There exist an orthonormal set of *continuous* L_2 functions $\{e_j\}_{j \in J}$ and $\{\lambda_j\}_{j \in J}$ (strictly positive eigenvalues with $\lambda_j \to 0$; J at most countable).

Theorem (Mercer's theorem)

$\forall x, y \in \mathcal{X}$ *with convergence uniform on* $\mathcal{X} \times \mathcal{X}$*:*

$$k(x, y) \quad = \quad \sum_{j \in J} \lambda_j e_j(x) e_j(y).$$

Mercer's theorem

$$
\begin{aligned}
k(x,y) &= \sum_{j\in J} \lambda_j e_j(x) e_j(y) \\
&= \left\langle \left\{ \sqrt{\lambda_j} e_j(x) \right\}, \left\{ \sqrt{\lambda_j} e_j(y) \right\} \right\rangle_{\ell^2(J)}
\end{aligned}
$$

Another (Mercer) feature map:

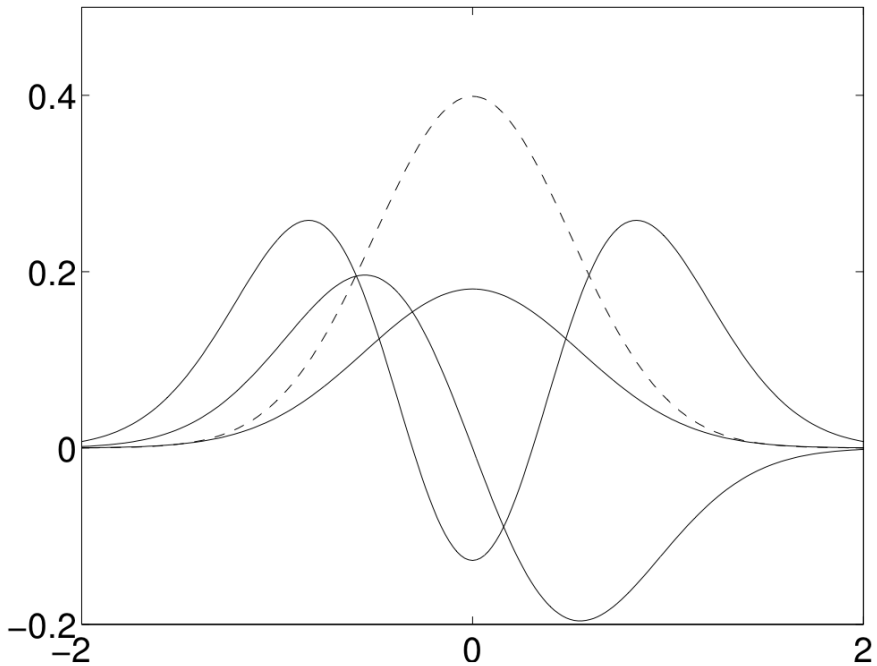
$$
\begin{aligned}
\varphi : \mathcal{X} &\to \ell^2(J) \\
\varphi : x &\mapsto \left\{ \sqrt{\lambda_j} e_j(x) \right\}_{j\in J}
\end{aligned}
$$

Mercer's Theorem and Smoothness

What does $\|f\|_{\mathcal{H}}$ have to do with smoothing? For the Gaussian kernel:

$$f(x) = \sum_{r=1}^{\infty} a_r e_r(x), \qquad \|f\|_{\mathcal{H}}^2 = \sum_{r=1}^{\infty} \frac{a_r^2}{\lambda_r}.$$

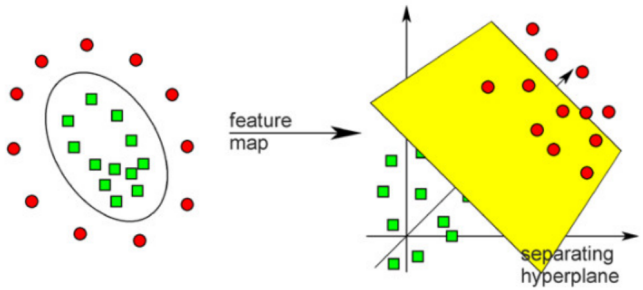
$\lambda_r \sim B^r \to 0$, as $r \to \infty$ for $B \in (0,1)$ and $e_r(x)$ are functions of increasing complexity as r increases (r zero-crossings) – related to r-th order *Hermite polynomials*. Figure from Rasmussen and Williams, 2006



Kernel Embeddings

Kernel Trick and Kernel Mean Trick

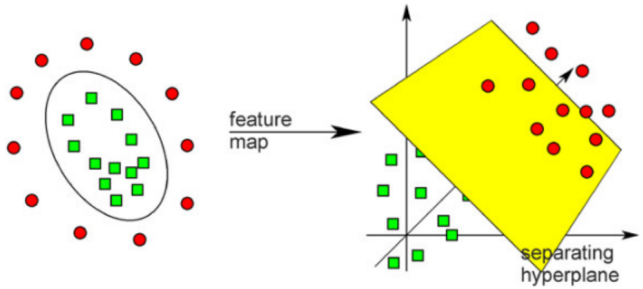
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions,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



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

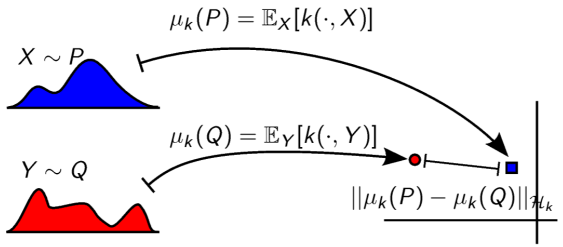
Kernel Trick and Kernel Mean Trick

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



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

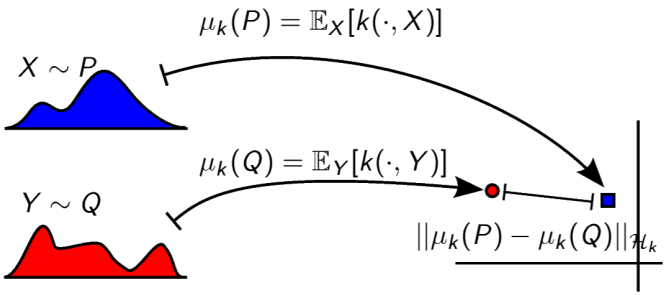
- **RKHS embedding**: implicit feature mean
  [Smola et al, 2007; Sriperumbudur et al, 2010; Muandet et al, 2017]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\phi_1(X), \dots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P),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

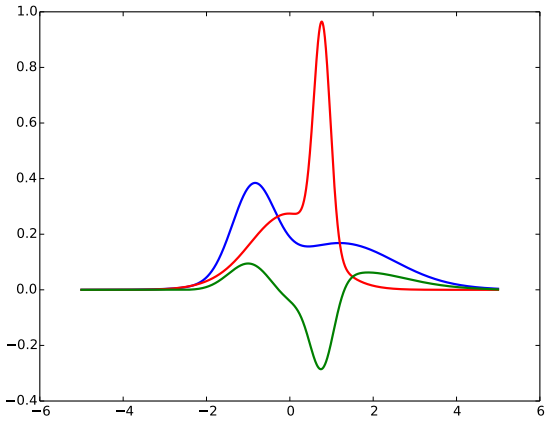


[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS et al, 2013; Muandet et al, 2012; Szabo et al, 2015]

Maximum Mean Discrepancy

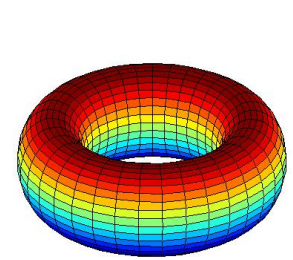
- **Maximum Mean Discrepancy (MMD)** [Borgwardt et al, 2006; Gretton et al, 2007] between P and Q:

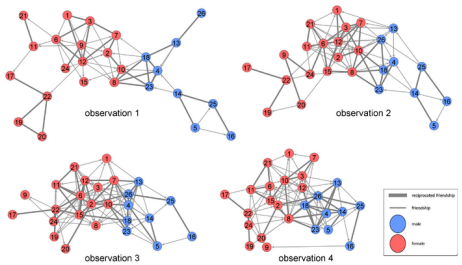




$$\mathrm{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}f(X) - \mathbb{E}f(Y)|$$

- 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, consectetur 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.

Some uses of MMD

within-sample average similarity

–

between-sample average similarity

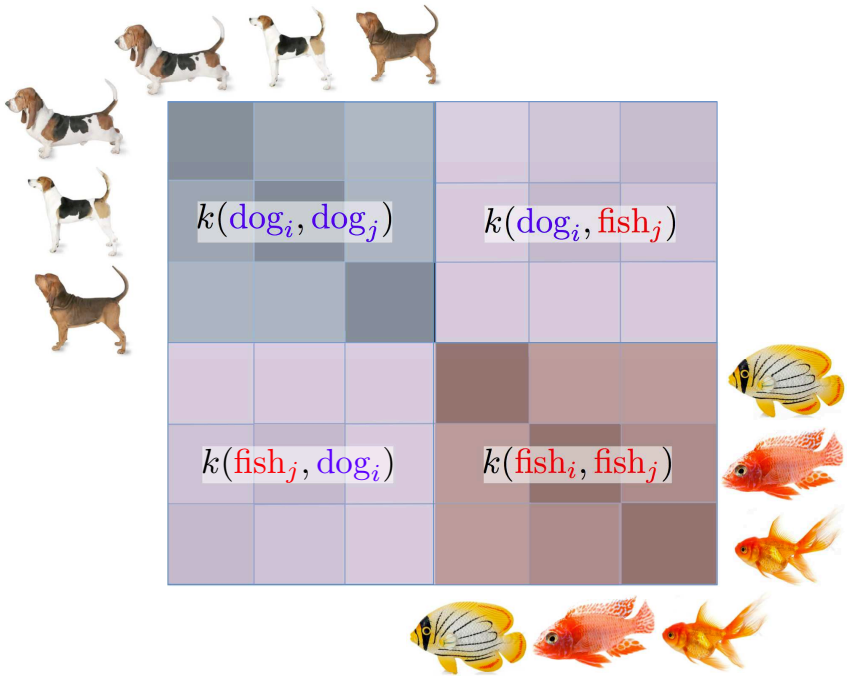


Figure by Arthur Gretton

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

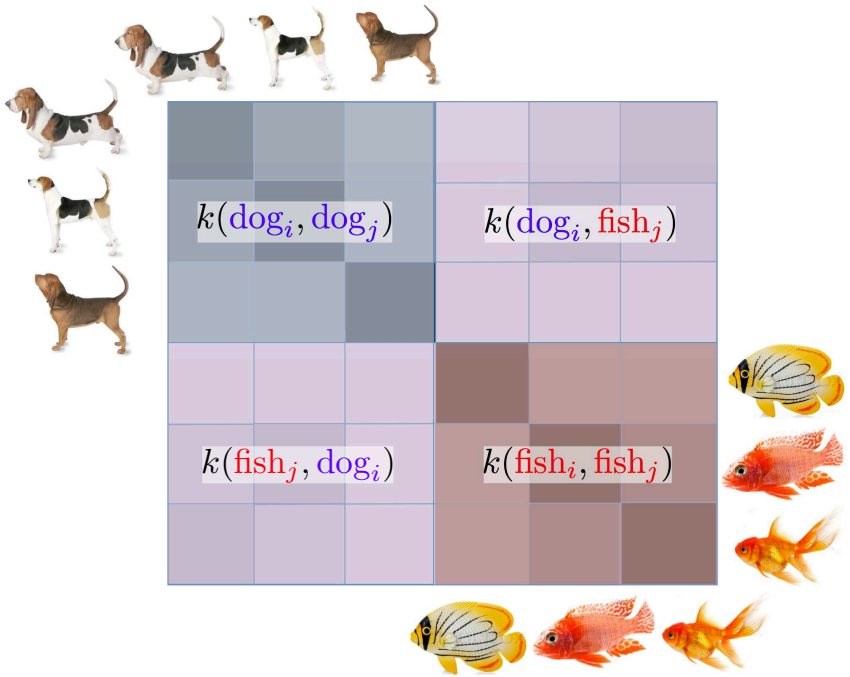


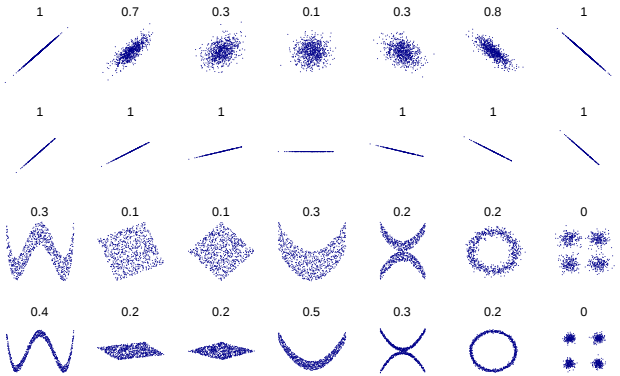
Figure by Arthur Gretton

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 dependence measures: HSIC



cor vs. dcor

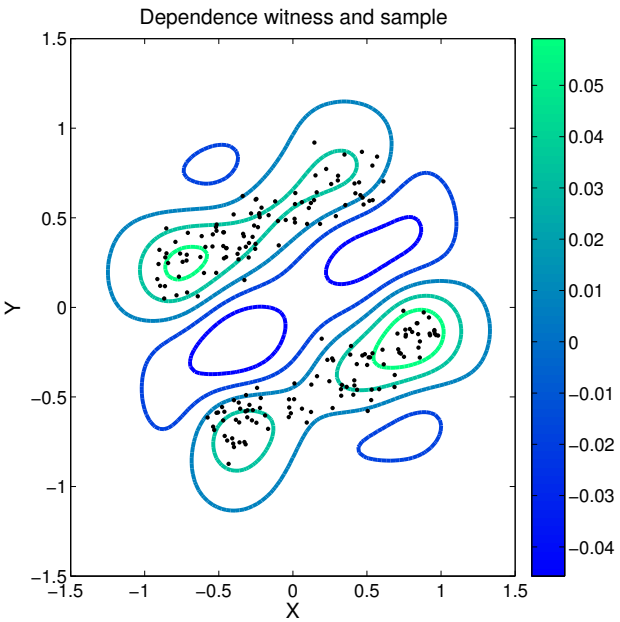
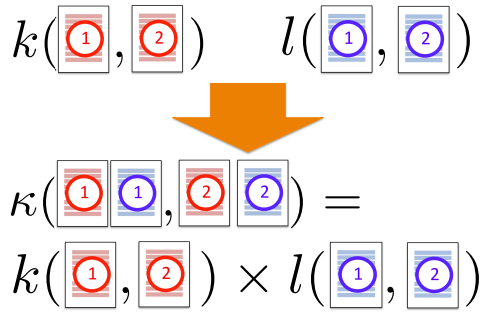


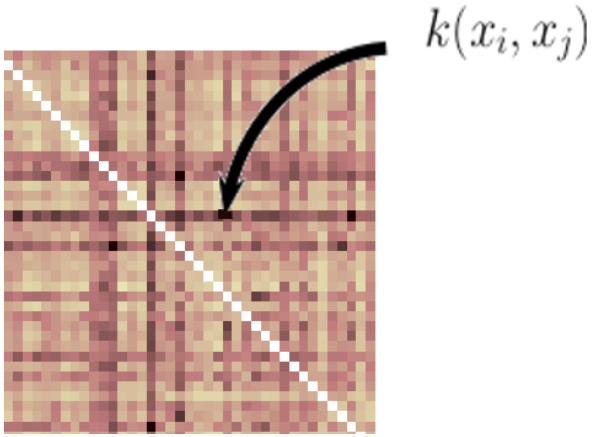
Figure by Arthur Gretton

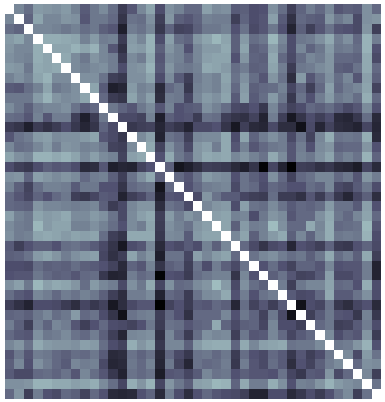
- $HSIC^2(X, Y; \kappa) = \|\mu_\kappa(P_{XY}) - \mu_\kappa(P_X P_Y)\|^2_{\mathcal{H}_\kappa}$
- Hilbert-Schmidt norm of the feature-space cross-covariance [Gretton et al, 2009]
- dependence witness is a smooth function in the RKHS $\mathcal{H}_\kappa$ of functions on $\mathcal{X} \times \mathcal{Y}$



- Independence testing framework that generalises Distance Correlation (dcor) of [Szekely et al, 2007]: HSIC with Brownian motion kernels [DS et al, 2013]
- Extends to multivariate interaction and joint dependence measures [DS et al, 2013; Pfister et al, 2017]

Kernel dependence measures: HSIC (2)

$k($  ,  $) \rightarrow \mathbf{K} =$ 

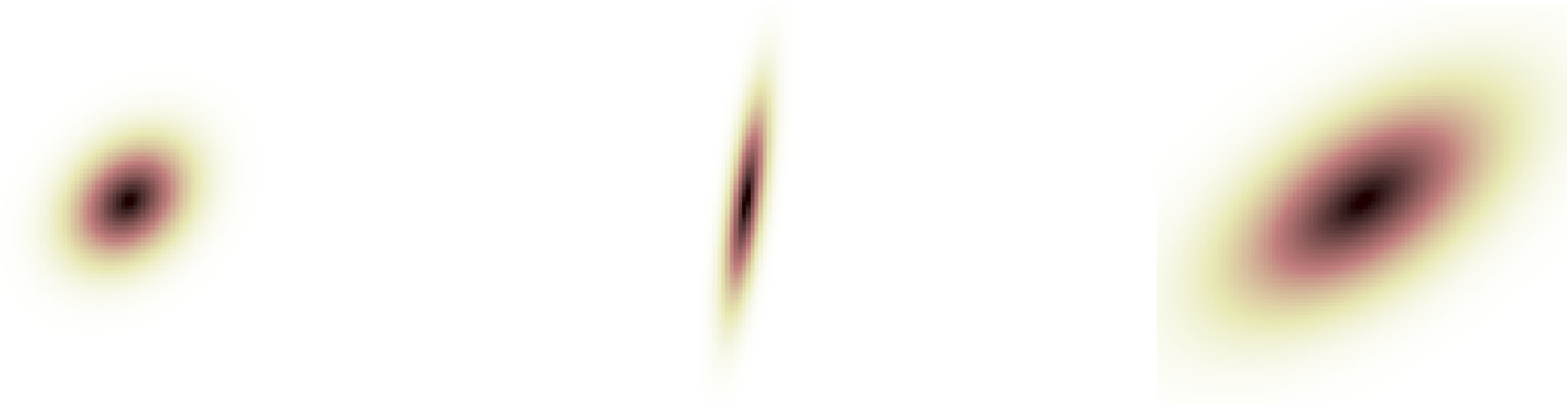
$\ell($ The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps... , Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament... $) \rightarrow \mathbf{L} =$ 

Hilbert-Schmidt Independence Criterion (**HSIC**): similarity between the kernel matrices $\left\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \right\rangle = \boxed{\text{Tr}\left(\tilde{\mathbf{K}}\tilde{\mathbf{L}}\right)}$, where $\tilde{\mathbf{K}} = \mathbf{HKH}$, and $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix.

[Gretton et al, 2008; Fukumizu et al, 2008; Song et al, 2012]

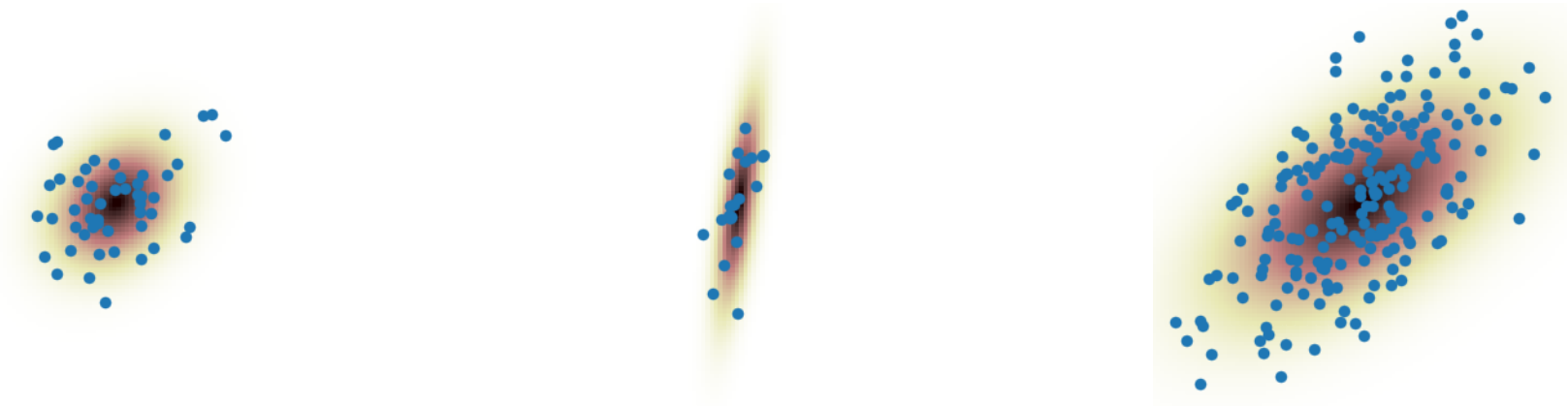
Distribution Regression

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



- 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.

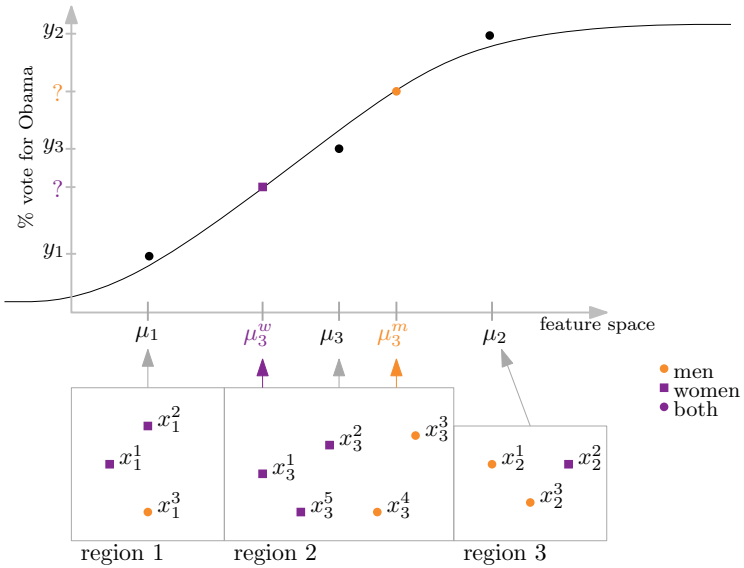


Figure from Flaxman et al, 2015

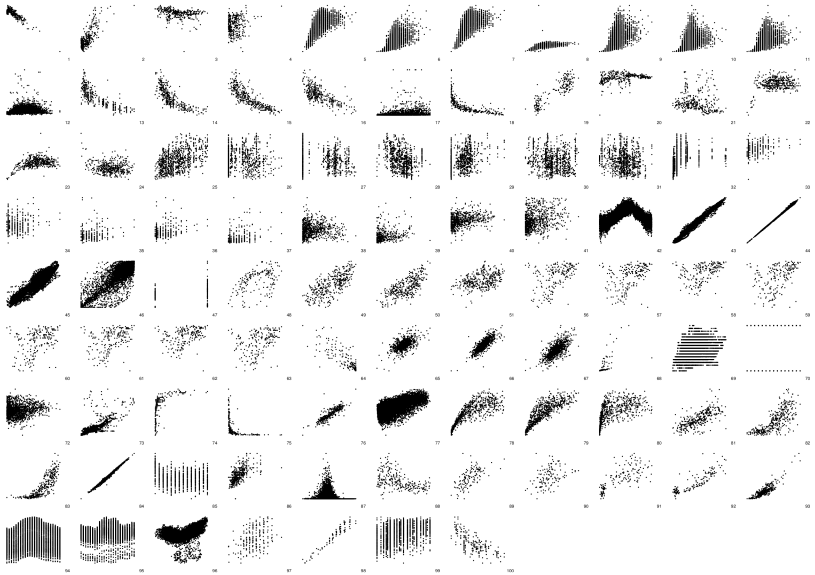


Figure from Mooij et al, 2014

- 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]

Kernel Methods and Gaussian Processes

Different Flavours of Regression

- We can model response y_i as a noisy version of the underlying function f evaluated at input x_i:

$$y_i | f(x_i) \sim \mathcal{N}(f(x_i), \sigma^2)$$

  Appropriate loss: $L(y, f(x)) = (y - f(x))^2$
- *Frequentist Parametric* approach: model f as f_θ for some parameter vector θ. Fit θ by ML / ERM with squared loss (linear regression).
- *Frequentist Nonparametric* approach: model f as the unknown parameter taking values in an infinite-dimensional space of functions. Fit f by regularized ML / ERM with squared loss (kernel ridge regression)
- *Bayesian Parametric* approach: model f as f_θ for some parameter vector θ. Put a prior on θ and compute a posterior $p(\theta|\mathcal{D})$ (Bayesian linear regression).
- *Bayesian Nonparametric* approach: treat f as the random variable taking values in an infinite-dimensional space of functions. Put a prior over functions $f \in \mathcal{F}$, and compute a posterior $p(f|\mathcal{D})$ (Gaussian Process regression).

Priors on function values

- Work with the function values at a set of inputs $\mathbf{f} = (f(x_1), \ldots, f(x_n))^\top$
- What properties of the function can we incorporate?
  - Multivariate normal prior on $\mathbf{f}$:

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

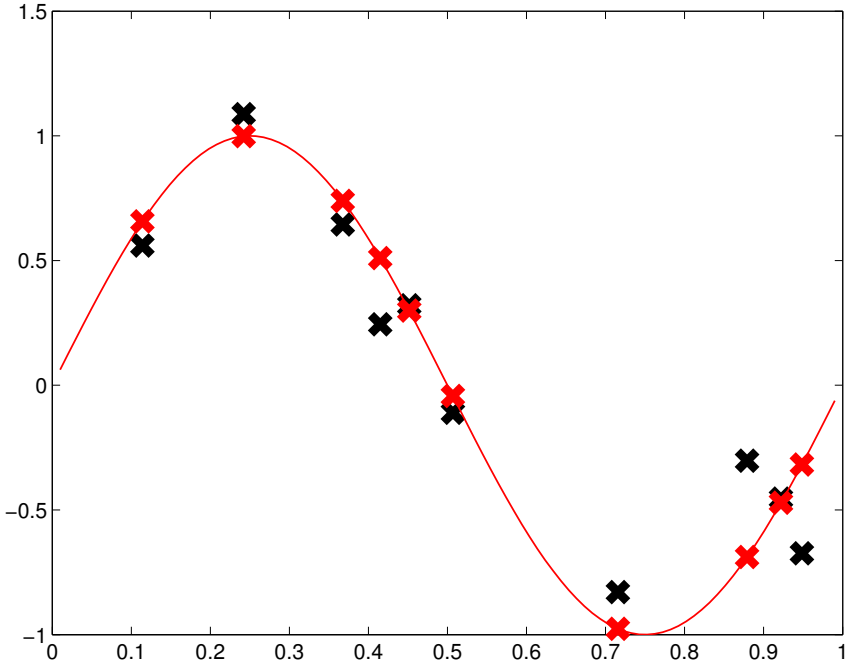
  - Use a (positive definite) covariance function k to define $\mathbf{K}$:

$$\mathbf{K}_{ij} = k(x_i, x_j)$$

  - Expect regression functions to be smooth: If x and x' are close by, then $f(x)$ and $f(x')$ have similar values, i.e. strongly correlated.

$$\begin{pmatrix} f(x) \\ f(x') \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{pmatrix} \right)$$

The prior $p(\mathbf{f})$ encodes our prior knowledge about the function.



- Model:

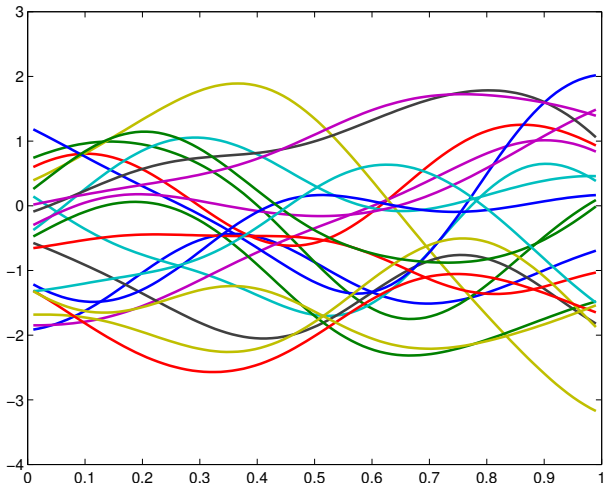
$$\mathbf{f} \sim \mathcal{N}(0, \mathbf{K})$$
$$y_i | f_i \sim \mathcal{N}(f_i, \sigma^2)$$

Gaussian Processes

- What does a multivariate normal prior mean?
- Imagine $\mathbf{x}$ forms an infinitesimally dense grid of data space. Simulate prior draws

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

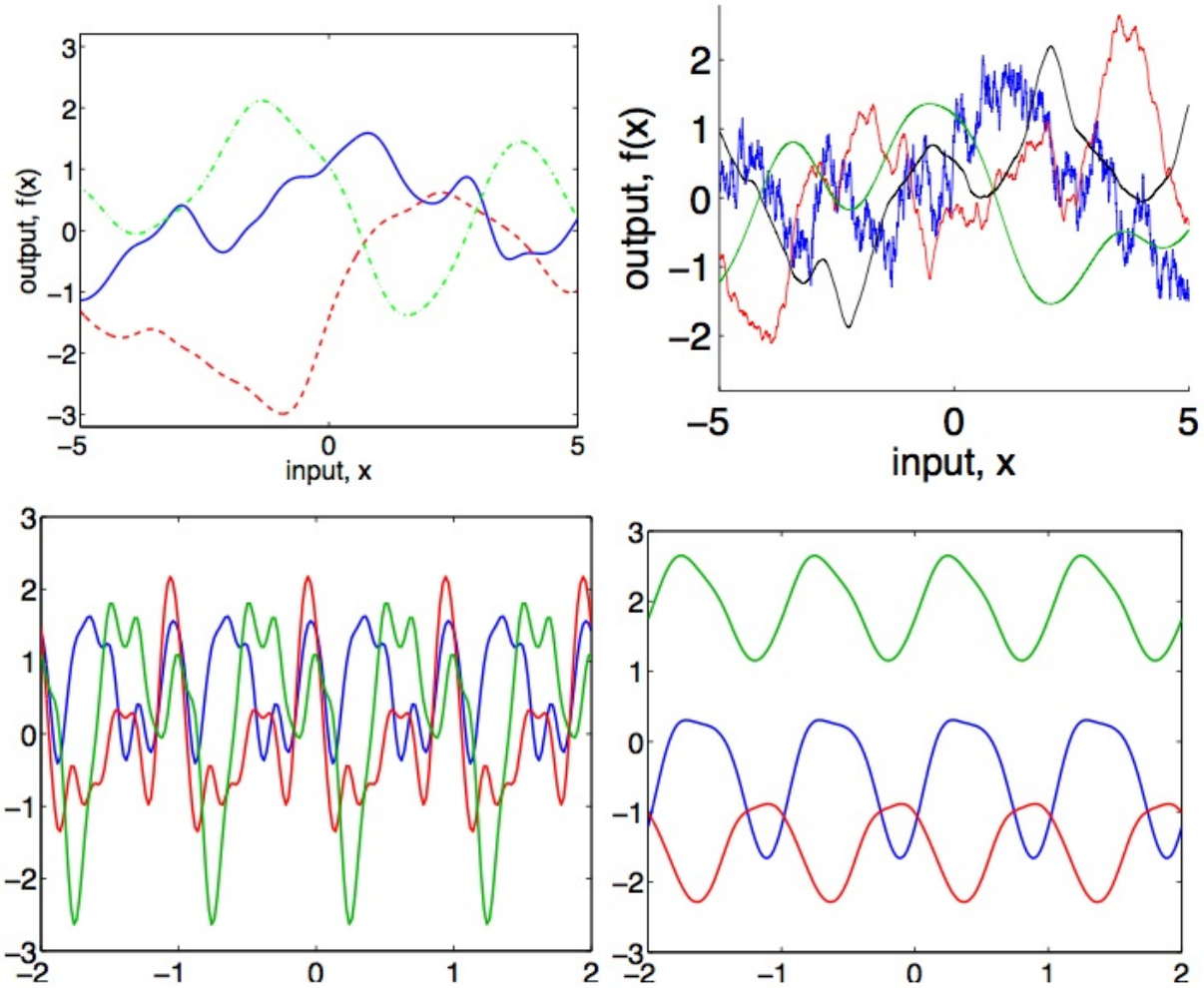
  Plot f_i vs x_i for $i = 1, \ldots, n$.
- The corresponding prior over functions is called a *Gaussian Process* (GP): any finite number of evaluations of which follow a Gaussian distribution.



http://www.gaussianprocess.org/

Gaussian Processes

- Different kernels lead to different function characteristics.



Carl Rasmussen. Tutorial on Gaussian Processes at NIPS 2006.

Gaussian Processes

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

- Posterior distribution:

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

- Posterior predictive distribution: Suppose $\mathbf{x}'$ is a test set. We can extend our model to include the function values $\mathbf{f}'$ at the test set:

$$\begin{pmatrix}\mathbf{f}\\ \mathbf{f}'\end{pmatrix}|\mathbf{x}, \mathbf{x}' \sim \mathcal{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\mathbf{K_{xx}} & \mathbf{K_{xx'}}\\ \mathbf{K_{x'x}} & \mathbf{K_{x'x'}}\end{pmatrix}\right)$$
$$\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2 I)$$

  where $\mathbf{K_{xx'}}$ is matrix with (i,j)-th entry $k(x_i, x'_j)$.
- Some manipulation of multivariate normals gives:

$$\mathbf{f}'|\mathbf{y} \sim \mathcal{N}\left(\mathbf{K_{x'x}}(\mathbf{K_{xx}}+\sigma^2 I)^{-1}\mathbf{y}, \mathbf{K_{x'x'}} - \mathbf{K_{x'x}}(\mathbf{K_{xx}}+\sigma^2 I)^{-1}\mathbf{K_{xx'}}\right)$$

GP regression and Kernel Ridge Regression

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

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

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

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

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

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.

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

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

Large-Scale Kernel Approximations

Kernel methods at scale

- Expressivity of kernel methods comes at a price of $O(n^2)$ or $O(n^3)$ in the number of observations n (due to having to compute, store and often invert the Gram matrix)
- Problematic when we have a lot of observations (and this is exactly when we want to use a rich expressive model with a high-dimensional hypothesis class!)
- Scaling up kernel methods is a very active research area

  [Sonnenburg et al, 2006; Rahimi & Recht, 2007; Le, Sarlos & Smola, 2013; Wilson et al, 2014; Dai et al, 2014; Sriperumbudur & Szabo, 2015; Bach, 2015; Avron et al, 2017; Li et al, 2019].
- Main idea: study the RKHS and construct a (random) low-dimensional space with similar inner product structure for a given data - then undo the kernel trick(!?)

explicit basis functions

$\downarrow$

implicit basis functions

$\downarrow$

explicit random basis functions

Random Fourier features: Inverse Kernel Trick

Bochner's representation: Assume that k is a positive definite translation-invariant kernel on $\mathbb{R}^p$. Then k can be written as

$$\begin{aligned} k(x,y) &= \int_{\mathbb{R}^p} \exp\left(i\omega^\top (x-y)\right) d\Lambda(\omega) \\ &= 2\int_{\mathbb{R}^p} \left\{\cos\left(\omega^\top x\right)\cos\left(\omega^\top y\right) + \sin\left(\omega^\top x\right)\sin\left(\omega^\top y\right)\right\} d\Lambda(\omega) \end{aligned}$$

for some positive measure (w.l.o.g. a probability distribution) Λ.

- Sample m frequencies $\Omega = \{\omega_j\}_{j=1}^m \sim \Lambda$ and use a Monte Carlo estimator of the kernel function instead [Rahimi & Recht, 2007]:

$$\begin{aligned} \hat{k}(x,y) &= \frac{2}{m}\sum_{j=1}^m \left\{\cos\left(\omega_j^\top x\right)\cos\left(\omega_j^\top y\right) + \sin\left(\omega_j^\top x\right)\sin\left(\omega_j^\top y\right)\right\} \\ &= \langle \xi_\Omega(x), \xi_\Omega(y)\rangle_{\mathbb{R}^{2m}}, \end{aligned}$$

  with an explicit set of features $\xi_\Omega : x \mapsto \sqrt{\frac{2}{m}}\left[\cos\left(\omega_1^\top x\right), \sin\left(\omega_1^\top x\right), \ldots\right]^\top$.
- The cost drops: $O(n^3) \to O(m^2 n + m^3)$, $O(n^2) \to O(mn + m^2)$. How fast does m need to grow with n? Often sublinear and can be as low as $\log n$ without sacrificing convergence rates [Bach, 2015; Rudi et al, 2017; Avron et al, 2017; Li et al, 2019].

Inducing variables / Nyström

- Directly approximate the $n \times n$ Gram matrix K_{XX} of a set of inputs $\{x_i\}_{i=1}^n$ with

$$\hat{K}_{XX} = K_{XZ} K_{ZZ}^{-1} K_{ZX}$$

  where K_{ZZ} is $m \times m$ on "inducing" inputs $\{z_i\}_{i=1}^m$.
- Corresponds to explicit feature representation $x \mapsto K_{xZ} K_{ZZ}^{-1/2}$.
- Surrogate kernel $\hat{k}(x, x') = \langle k_|(\cdot, x), k_|(\cdot, x') \rangle$, where $k_|(\cdot, x)$ is a projection of $k(\cdot, x)$ to span $\{k(\cdot, z_1), \ldots, k(\cdot, z_m)\}$
- Often used in regression with Gaussian processes: with the use of Sherman-Morrison-Woodbury identity, reduces $O(n^3)$ cost to $O(nm^2)$.

  [Quiñonero-Candela and Rasmussen, 2005, Snelson and Ghahramani, 2006]
- m can grow much slower than n in regression without sacrificing performance

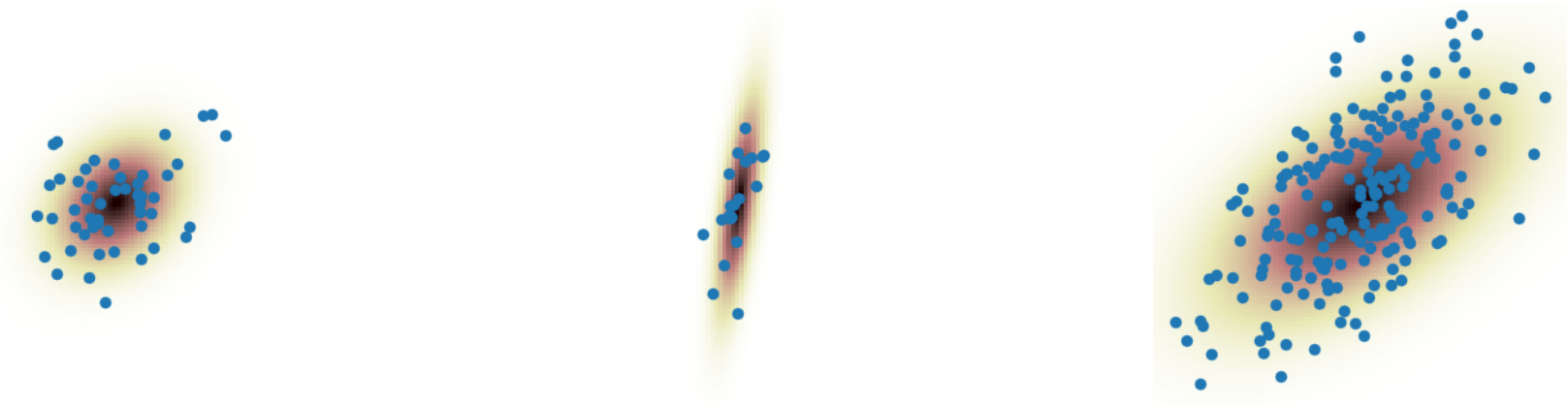
  [Rudi, Camoriano & Rosasco, 2015].

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

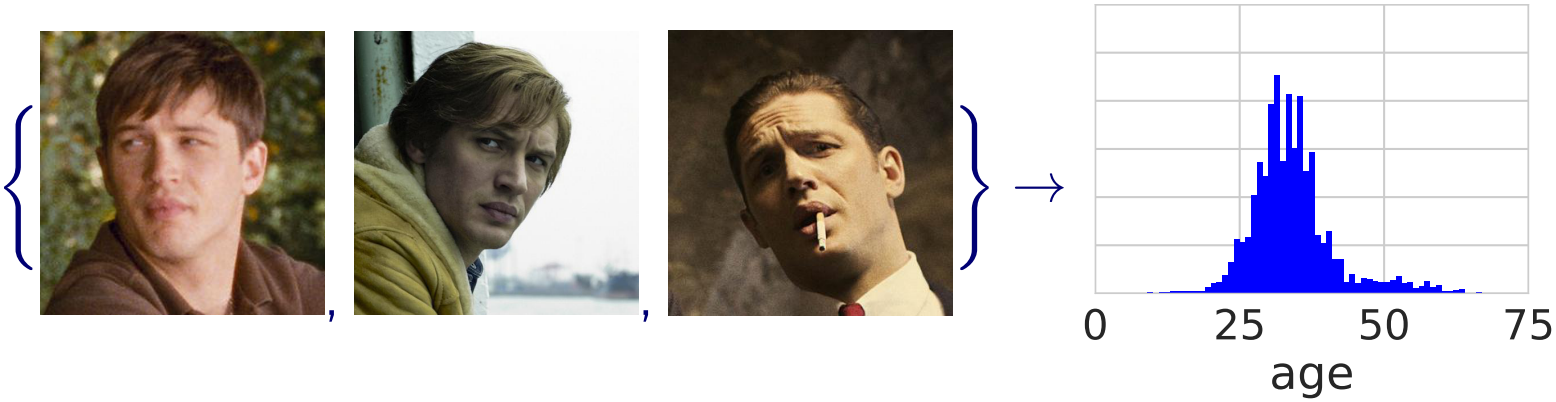
$$y_i \mid \{x_i^j\}, \beta \sim \mathcal{N}(\xi_i^\beta, \nu_i^\beta)$$

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

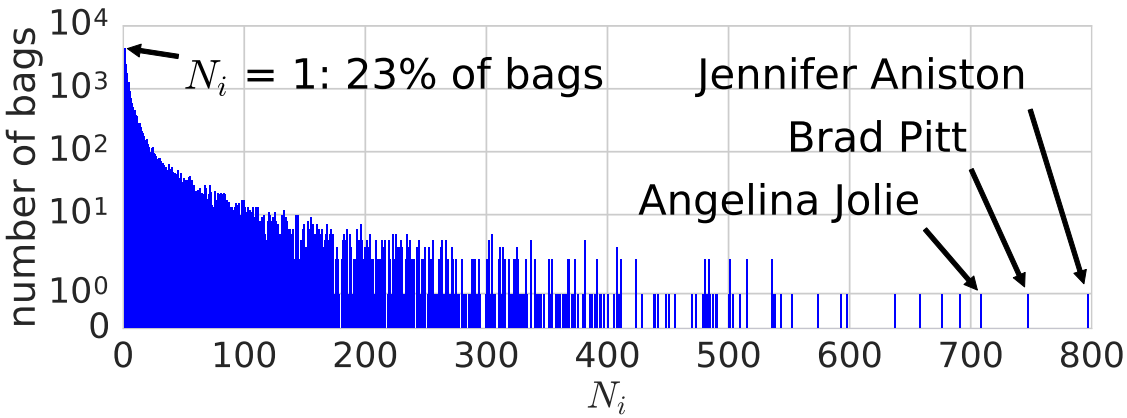
$$\nu_i^\beta = \beta^\top \left(R_{\mathbf{z}\mathbf{z}} - R_{\mathbf{z}\mathbf{x}_i} \left(R_{\mathbf{x}_i\mathbf{x}_i} + \frac{\Sigma_i}{N_i} \right)^{-1} R_{\mathbf{x}_i\mathbf{z}}^\top \right) \beta + \sigma^2.$$

- 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!

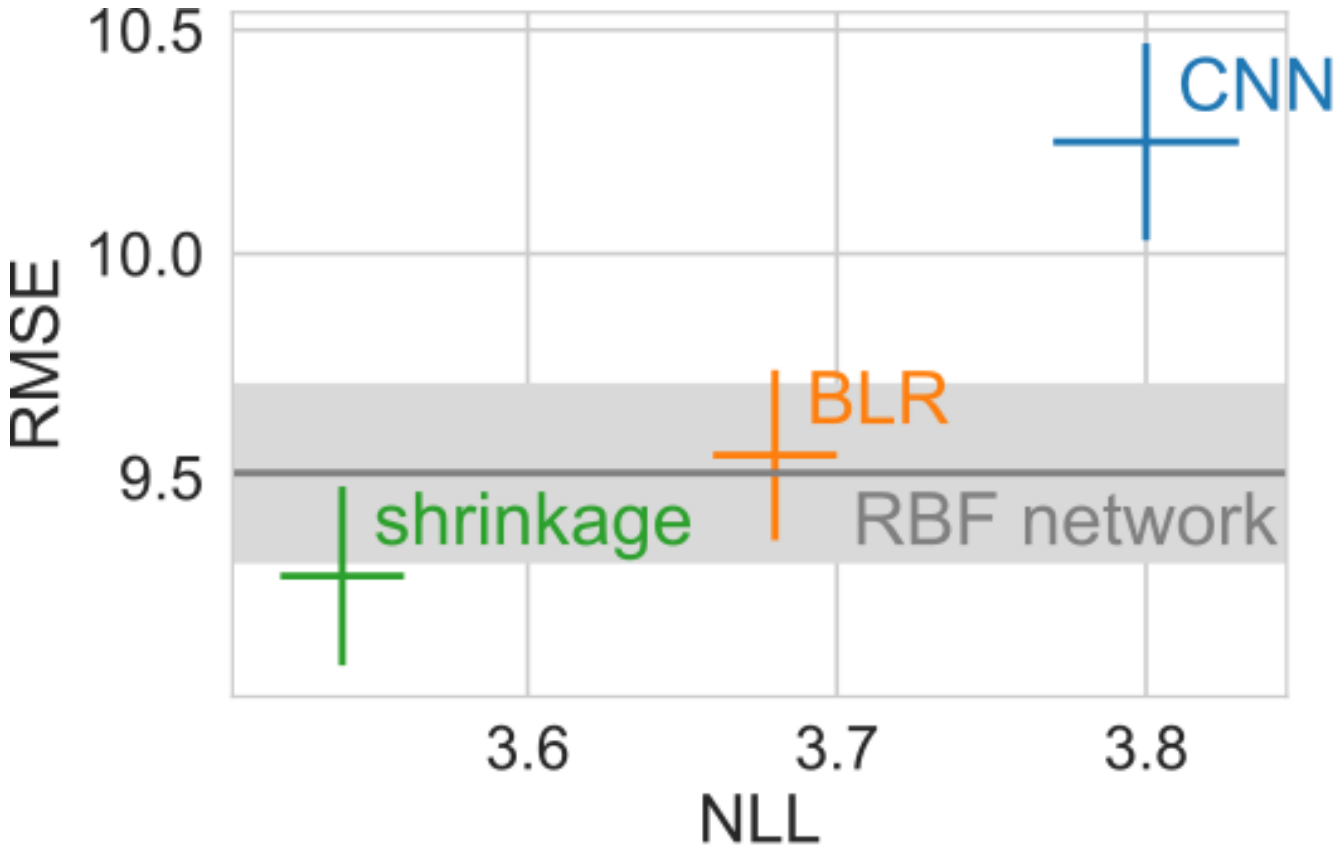


Figure: Results across 10 data splits (means and standard deviations). RBF net is tuned for RMSE, other methods for NLL. CNN takes the mean of the predictive distributions of [Rothe, 2016] for each point in the bag.

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

Outline

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.

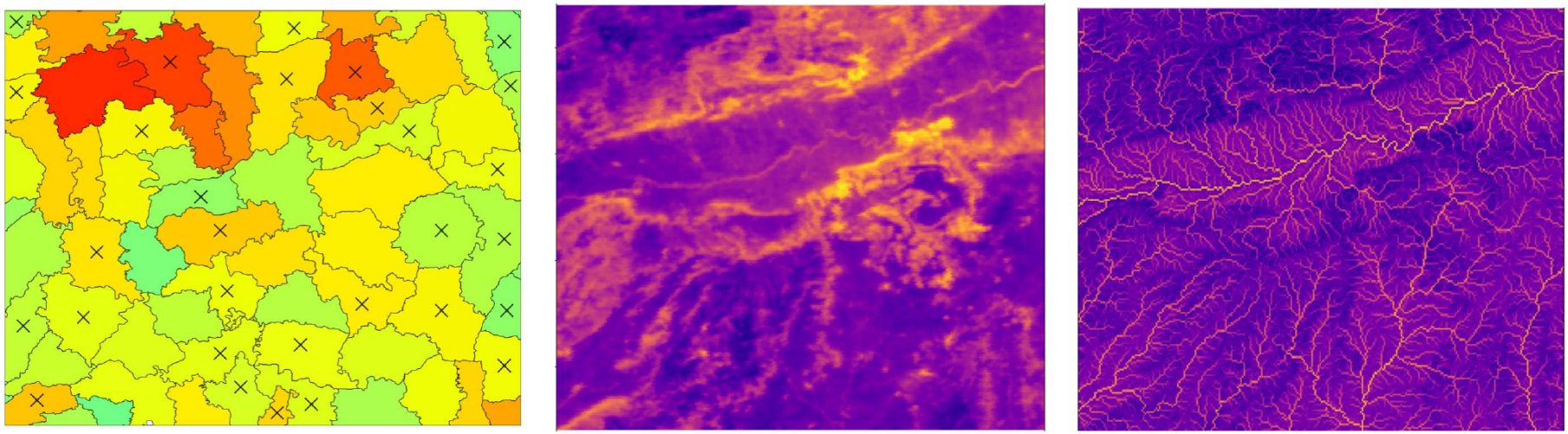
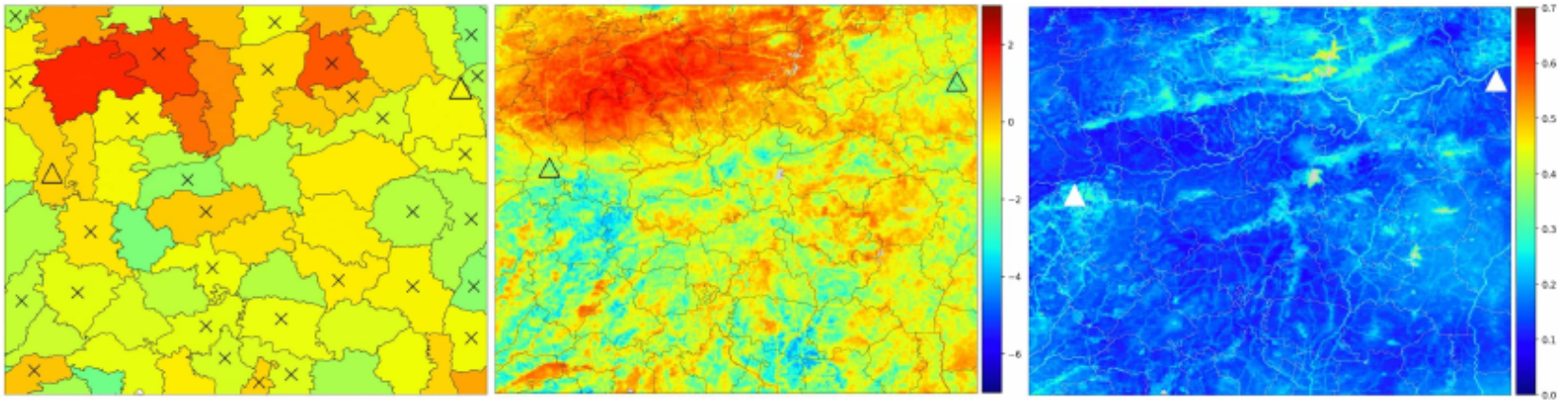


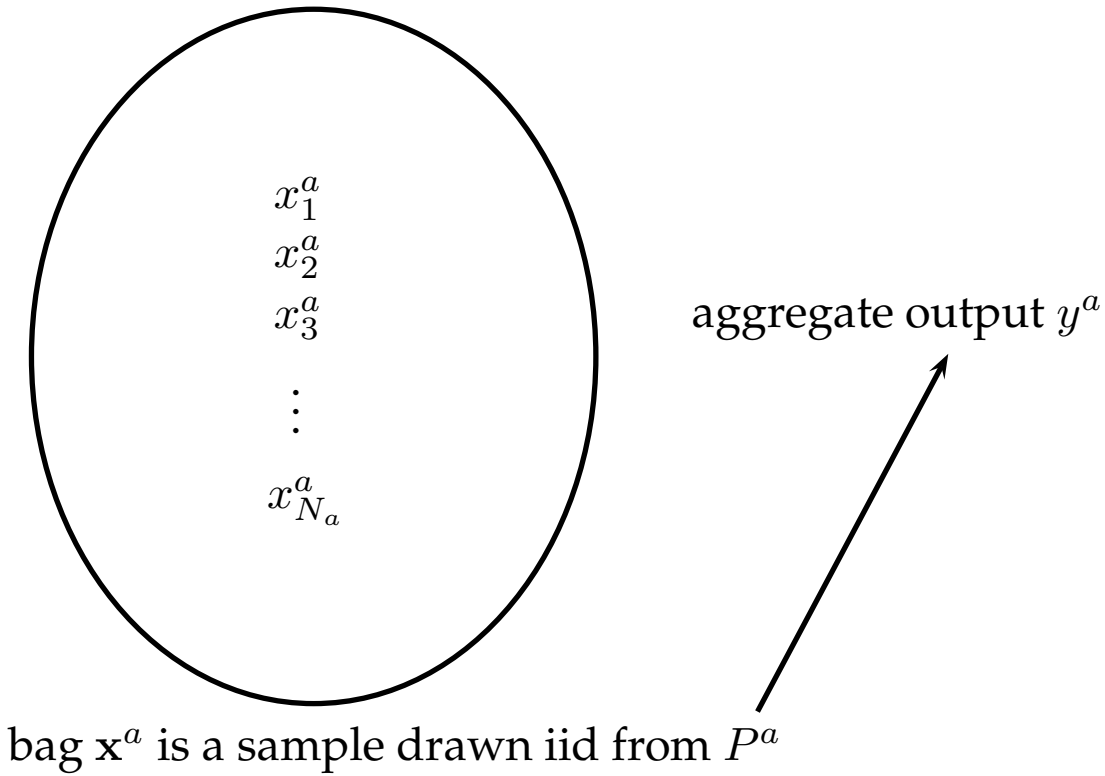
Figure: **left**: Malaria incidences reported per administrative unit; **centre**: land surface temperature at night; **centre**: topographic wetness index

Disaggregating Aggregate Outputs



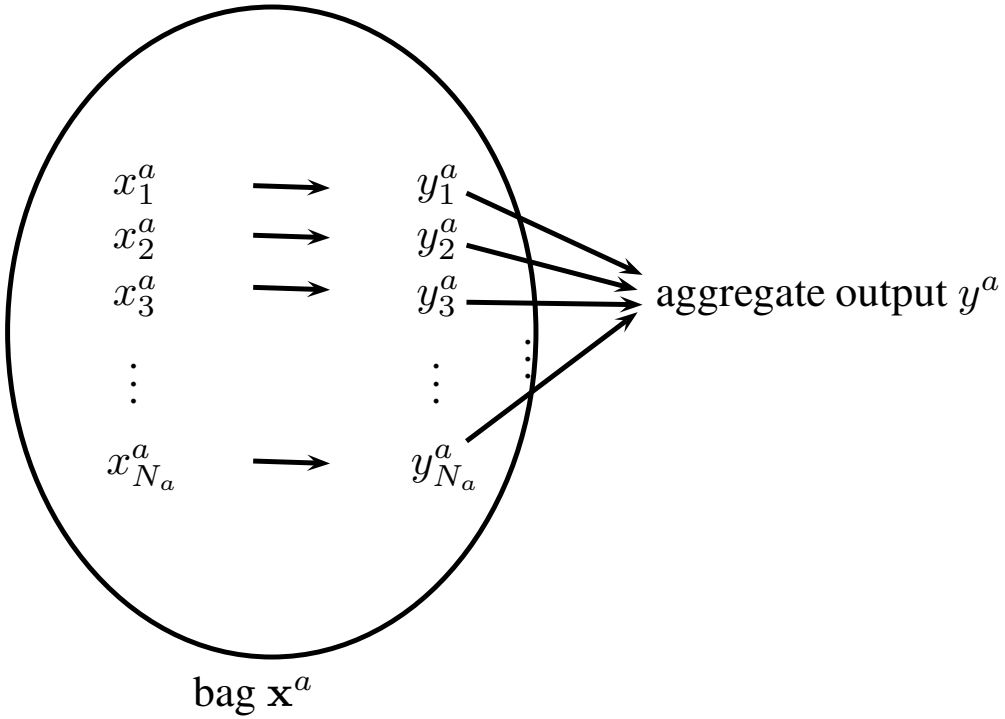
Variational Learning on Aggregate Outputs with Gaussian Processes
H. C. L. Law, DS, E. Cameron, T. C. D. Lucas, S. Flaxman, K. Battle, and K. Fukumizu
to appear in **NeurIPS 2018**
https://arxiv.org/abs/1805.08463

Distribution regression: train on bags, predict on bags



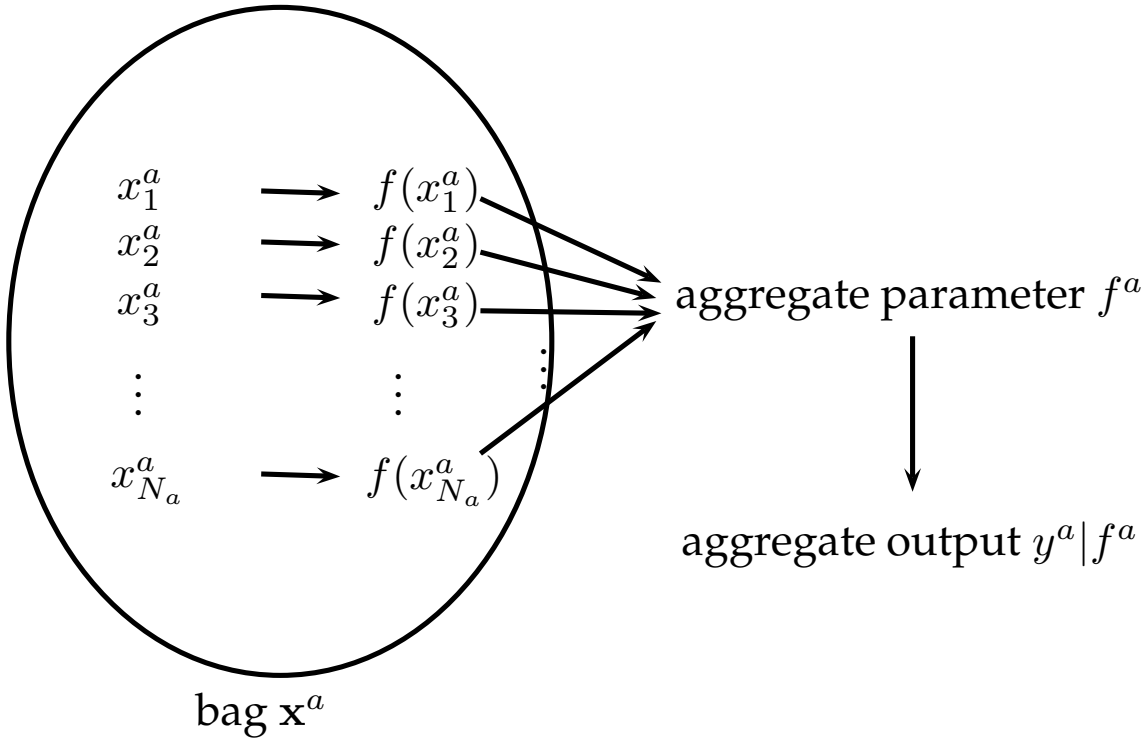
- Individual labels need not exist - the label is a function of the whole population.

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*:

$$y^a|\mathbf{x}^a \sim p(y|\eta^a), \qquad \eta^a = \sum_{i=1}^{N_a} p_i^a \eta(x_i^a), \tag{3}$$

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

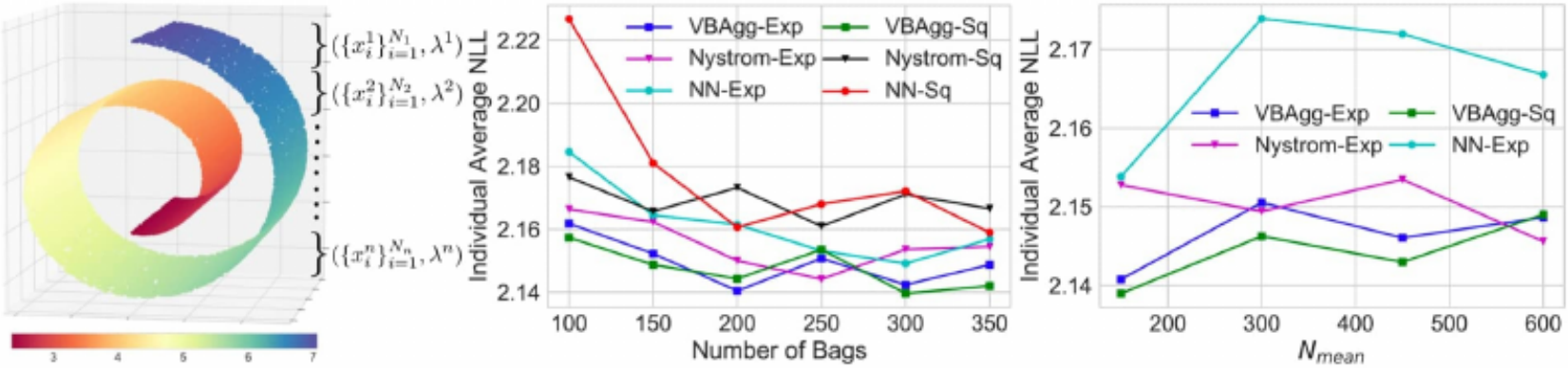


Figure 1: **Left**: Random samples on the Swiss roll manifold. **Middle, Right**: Individual Average NLL on train set for varying number of training bags n and increasing N_{mean}, over 5 repetitions. Constant prediction within bag gives a NLL of 2.22. bag-pixel model gives NLL above 2.4 for the varying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

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

Results

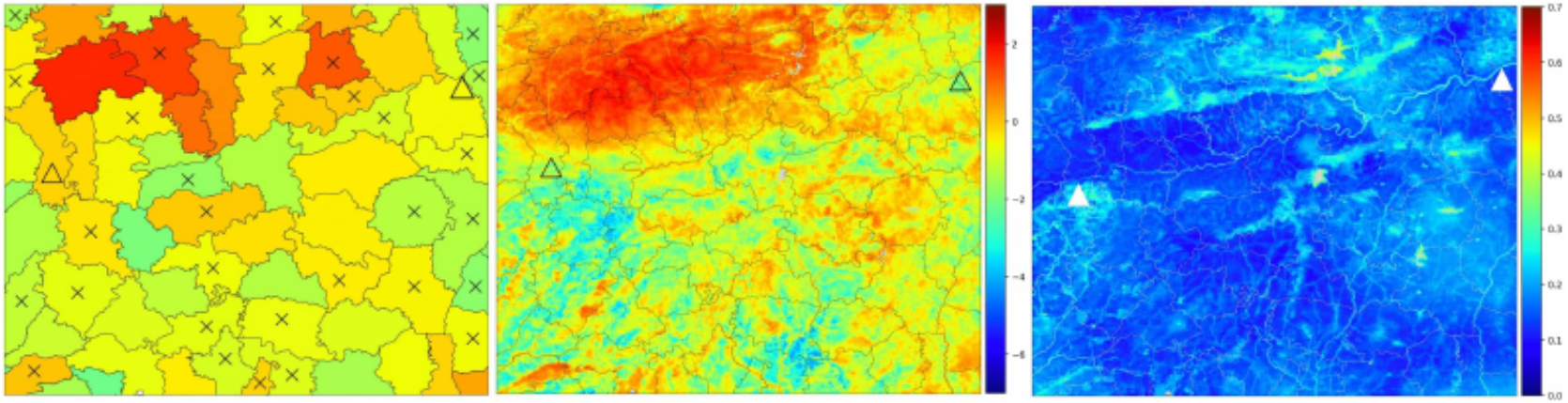


Figure 2: Triangle denotes approximate start and end of river location, crosses denotes non-train set bags. Malaria incidence rate λ_i^a is per 1000 people. **Left, Middle**: $\log(\hat{\lambda}_i^a)$, with constant model (Left), and VBAgg-Obj-Sq (tuned on $\mathcal{L}_1^s$) (Middle). **Right**: Standard 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 (NeurIPS)*, 2018, to appear. *ArXiv e-prints:1805.08463*, 2018.