Inference with Kernel Embeddings

Dino Sejdinovic

Department of Statistics
University of Oxford

RSS Conference, Manchester, 07/09/2016

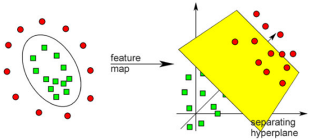
Outline

Outline

1. Preliminaries on Kernel Embeddings
2. Using Kernel MMD as a criterion in ABC
3. Bayesian Learning of Embeddings

Kernel Trick and Kernel Mean Trick

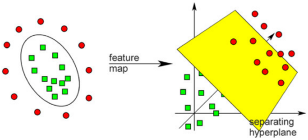
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\varphi_1(x), \ldots, \varphi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



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

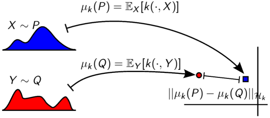
Kernel Trick and Kernel Mean Trick

- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\varphi_1(x), \ldots, \varphi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



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

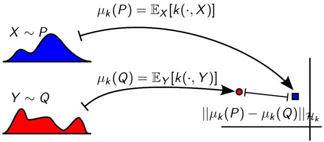
- **RKHS embedding**: implicit feature mean
  [Smola et al, 2007; Sriperumbudur et al, 2010]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\varphi_1(X), \ldots, \mathbb{E}\varphi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y)$
  inner products easy to estimate
  - nonparametric two-sample, independence, conditional independence, interaction testing, learning on distributions

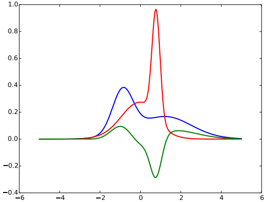


[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS et al, 2013; Muandet et al, 2012; Szabo et al, 2015]

Maximum Mean Discrepancy

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

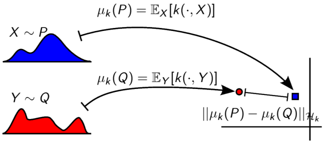


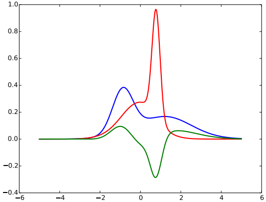


$$\text{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}f(X) - \mathbb{E}f(Y)|$$

Maximum Mean Discrepancy

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





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

- **Characteristic** kernels: $\mathrm{MMD}_k(P,Q) = 0$ iff $P = Q$.
  - Gaussian RBF $\exp(-\frac{1}{2\sigma^2}\|x - x'\|_2^2)$, Matérn family, inverse multiquadrics.
- For characteristic kernels on LCH $\mathcal{X}$, MMD metrizes weak* topology on probability measures [Sriperumbudur,2010],

$$\mathrm{MMD}_k(P_n, P) \to 0 \Leftrightarrow P_n \rightsquigarrow P.$$

Some uses of MMD

within-sample average similarity

–

between-sample average similarity

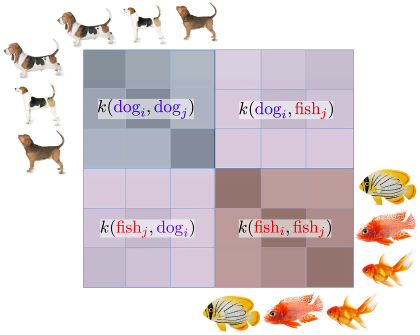


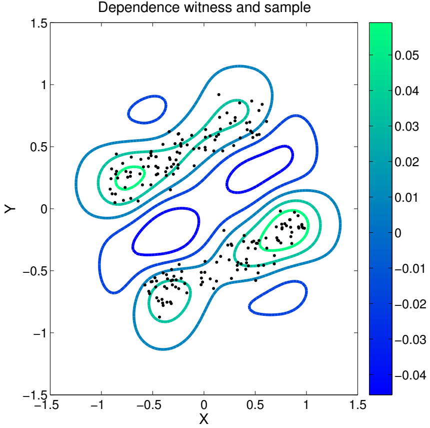
Figure by Arthur Gretton

MMD has been applied to:

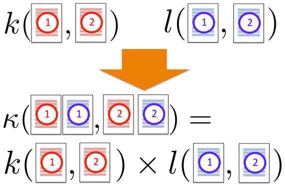
- independence tests [Gretton et al, 2009]
- two-sample tests [Gretton et al, 2012]
- training generative neural networks for image data [Dziugaite, Roy and Ghahramani, 2015]
- traversal of manifolds learned by convolutional nets [Gardner et al, 2015]
- similarity measure between observed and simulated data in ABC [Park, Jitkrittum and DS, 2015]

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

Kernel dependence measures



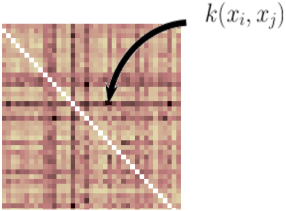
- $HSIC^2(X, Y; \kappa) = \|\mu_\kappa(P_{XY}) - \mu_\kappa(P_X P_Y)\|^2_{\mathcal{H}_\kappa}$
- dependence witness is a smooth function in the RKHS $\mathcal{H}_\kappa$ of functions on $\mathcal{X} \times \mathcal{Y}$

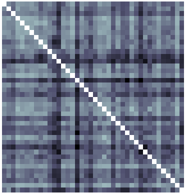


- Independence testing framework that generalises Distance Covariance (dCov): HSIC with Brownian motion covariance kernels

[Szekely et al, 2007; DS et al, 2013]

Kernel dependence measures (2)

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

$\ell($ 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... $) \to \mathbf{L} =$ 

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]

Outline

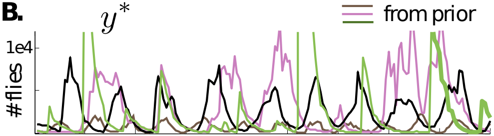
K2-ABC: Approximate Bayesian Computation with Kernel Embeddings.
AISTATS 2016
Mijung Park, Wittawat Jitkrittum, and DS.
http://arxiv.org/abs/1502.02558
Code: https://github.com/wittawatj/k2abc

Motivating example: ABC for modelling ecological dynamics

- Given: a time series $\mathbf{Y} = (Y_1, \ldots, Y_T)$ of population sizes of a blowfly.
- Model: A dynamical system for blowfly population (a discretised ODE) [Nicholson, 1954; Gurney et al, 1980; Wood, 2010; Meeds & Welling, 2014]

$$Y_{t+1} = P Y_{t-\tau} \exp\left(-\frac{Y_{t-\tau}}{Y_0}\right) e_t + Y_t \exp(-\delta \epsilon_t),$$

where $e_t \sim \text{Gamma}\left(\frac{1}{\sigma_P^2}, \sigma_P^2\right)$, $\epsilon_t \sim \text{Gamma}\left(\frac{1}{\sigma_d^2}, \sigma_d^2\right)$.
Parameter vector: $\theta = \{P, Y_0, \sigma_d, \sigma_p, \tau, \delta\}$.



- Goal: For a prior $p(\theta)$, sample from $p(\theta|\mathbf{Y})$.
  - Cannot evaluate $p(\mathbf{Y}|\theta)$. But, can sample from $p(\cdot|\theta)$.
  - For $\mathbf{X} = (X_1, \ldots, X_T) \sim p(\cdot|\theta)$, how to measure distance $\rho(\mathbf{X}, \mathbf{Y})$?

Data Similarity via Summary Statistics

- Distance ρ is typically defined via summary statistics

$$\rho(\mathbf{X}, \mathbf{Y}) = \|s(\mathbf{X}) - s(\mathbf{Y})\|_2.$$

- How to select the summary statistics $s(\cdot)$? Unless $s(\cdot)$ is sufficient, targets the incorrect (partial) posterior $p(\theta|s(\mathbf{Y}))$ rather than $p(\theta|\mathbf{Y})$.
- Hard to quantify additional bias.
  - Adding more summary statistics decreases "information loss": $p(\theta|s(\mathbf{Y})) \approx p(\theta|\mathbf{Y})$
  - ρ computed on a higher dimensional space - without appropriate calibration of distances therein, leads to a higher rejection rate so need to increase ϵ: $p_\epsilon(\theta|s(\mathbf{Y})) \not\approx p(\theta|s(\mathbf{Y}))$

Data Similarity via Summary Statistics

- Distance ρ is typically defined via summary statistics

$$\rho(\mathbf{X}, \mathbf{Y}) = \|s(\mathbf{X}) - s(\mathbf{Y})\|_2.$$

- How to select the summary statistics $s(\cdot)$? Unless $s(\cdot)$ is sufficient, targets the incorrect (partial) posterior $p(\theta|s(\mathbf{Y}))$ rather than $p(\theta|\mathbf{Y})$.
- Hard to quantify additional bias.
  - Adding more summary statistics decreases "information loss": $p(\theta|s(\mathbf{Y})) \approx p(\theta|\mathbf{Y})$
  - ρ computed on a higher dimensional space - without appropriate calibration of distances therein, leads to a higher rejection rate so need to increase ϵ: $p_\epsilon(\theta|s(\mathbf{Y})) \not\approx p(\theta|s(\mathbf{Y}))$
- Contribution: Use a nonparametric distance (MMD) between the empirical measures of datasets $\mathbf{X}$ and $\mathbf{Y}$).
  - No need to design $s(\cdot)$.
  - Rejection rate does not blow up since MMD penalises the higher order moments via Mercer expansion.

Embeddings via Mercer Expansion

Mercer Expansion

For a compact metric space $\mathcal{X}$, and a continous kernel k,

$$k(x,y) = \sum_{r=1}^{\infty} \lambda_r e_r(x) e_r(y),$$

with $\{\lambda_r, e_r\}_{r\geq 1}$ eigenvalue, eigenfunction pairs of $f \mapsto \int f(x)k(\cdot,x)dP(x)$ on $L_2(P)$, with $\lambda_r \to 0$, as $r \to \infty$. e_r are typically functions of increasing "complexity", i.e., Hermite polynomials of increasing degree.

$$\mathcal{H}_k \ni k(\cdot,x) \quad \leftrightarrow \quad \left\{\sqrt{\lambda_r} e_r(x)\right\} \in \ell_2$$

$$\mathcal{H}_k \ni \mu_k(P) \quad \leftrightarrow \quad \left\{\sqrt{\lambda_r} \mathbb{E} e_r(X)\right\} \in \ell_2$$

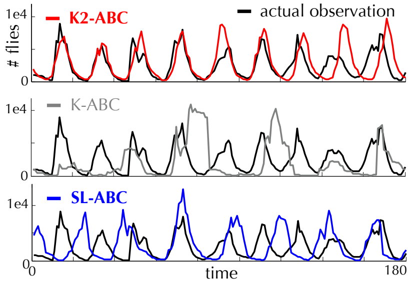
$$\left\|\mu_k(\hat{P}) - \mu_k(\hat{Q})\right\|_{\mathcal{H}_k}^2 = \sum_{r=1}^{\infty} \lambda_r \left(\frac{1}{n_x} \sum_{t=1}^{n_x} e_r(X_t) - \frac{1}{n_y} \sum_{t=1}^{n_y} e_r(Y_t) \right)^2$$

Blow Fly Population Modelling

Number of blow flies over time

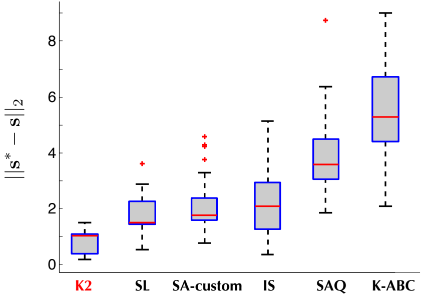
$$Y_{t+1} = PY_{t-\tau} \exp\left(-\frac{Y_{t-\tau}}{Y_0}\right) e_t + Y_t \exp(-\delta \epsilon_t)$$

- $e_t \sim$ Gam $\left(\frac{1}{\sigma_P^2}, \sigma_P^2\right)$ and $\epsilon_t \sim$ Gam $\left(\frac{1}{\sigma_d^2}, \sigma_d^2\right)$.
- Want $\theta := \{P, Y_0, \sigma_d, \sigma_p, \tau, \delta\}$.

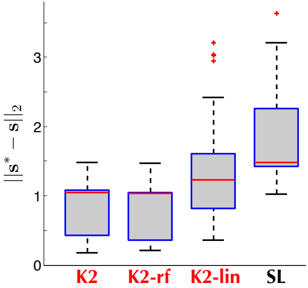


- Simulated trajectories with inferred posterior mean of θ
  - Observed sample of size 180.
  - Other methods use handcrafted 10-dimensional summary statistics $s(\cdot)$ from [Meeds & Welling, 2014]: quantiles of marginals, first-order differences, maximal peaks, etc.

Blowfly dataset



- Let $\tilde{\theta}$ be the posterior mean.
- Simulate $\mathbf{X} \sim p(\cdot|\tilde{\theta})$.
- $\mathbf{s} = s(\mathbf{X})$ and $\mathbf{s}^* = s(\mathbf{Y})$.
- Improved mean squared error on $\mathbf{s}$, even though SL-ABC, SA-custom explicitly operate on $\mathbf{s}$ while K2-ABC does not.



- Computation of $\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y})$ costs $O(n^2)$.
- Linear-time unbiased estimators of MMD^2 or random feature expansions reduce the cost to $O(n)$.

Summary: K2-ABC

- A dissimilarity criterion for ABC based on MMD between empirical distributions of observed and simulated data
- No “information loss” due to insufficient statistics.
- Simple and effective when parameters model marginal distribution of observations (variants for conditional distributions readily available).

Outline

Right... But how do you choose your kernel?

- Frequentists cross-validate, Bayesians optimize marginal likelihood...
- But with kernel embeddings, neither is typically available (e.g. hypothesis testing or ABC).
- Median heuristic: bandwidth parameter $\theta = \text{median}(\|x_i - x_j\|_2)$ for e.g. Gaussian kernel $k(x, x') = \exp(-\frac{\|x-x'\|^2}{2\theta^2})$

Bayesian Learning of Kernel Embeddings.
UAI 2016.
Seth Flaxman, DS, John Cunningham, and Sarah Filippi.
http://arxiv.org/abs/1603.02160

Bayesian Model for Embeddings

- In MMD and HSIC, we estimate embedding $\mu = \int k(\cdot, x)\mathsf{P}(dx)$ with its empirical mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i)$.
- Empirical mean over an infinite-dimensional case? Due to Stein's phenomenon, shrinkage estimators are better behaved [Muandet et al, 2013] and are reported to improve performance in kernel PCA and in testing power [Ramdas & Wehbe, 2015].
- Can we formulate a Bayesian inference procedure for kernel embeddings?
- Two challenges:
  - How to construct a valid prior over the RKHS?
  - What is the likelihood of our observations given the kernel embedding?

Priors on RKHS

A classical result, Kallianpur's 0-1 law, [Kallianpur, 1970; Wahba, 1990]: sample paths of a GP with kernel k lie outside RKHS $\mathcal{H}_k$ with probability 1.

However, one can use a prior $f \sim \mathcal{GP}(0, r)$ with

$$r(x, x') = \int k(x, u) k(u, x') \nu(du)$$

for any finite measure ν in which case $f \in \mathcal{H}_k$ with probability 1: nuclear dominance theory established by [Lukic and Beder, 2001; Pillai et al, 2007].

Kernel r analytically available, e.g. for a Gaussian kernel $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\theta^2}\right)$ and $\nu(du) \propto \exp\left(-\frac{\|u\|^2}{2\eta^2}\right) du$:

$$r(x, x') \propto \exp\left(-\frac{\|x - x'\|^2}{4\theta^2} - \frac{\|(x + x')/2\|^2}{4\theta^2 + \eta^2}\right).$$

- Has a nonstationary component, but similar to another (smoother) Gaussian kernel with bandwidth $\theta\sqrt{2}$ when η is large.

Likelihood

We need a likelihood linking the kernel mean embedding μ to the observations $\{x_i\}_{i=1}^n$ Consider evaluating $\widehat{\mu}$ induced by $\{x_i\}_{i=1}^n$ at some $x \in \mathcal{X}$ - we link $\widehat{\mu}(x)$ to $\mu(x)$ using a Gaussian distribution with variance τ^2/n:

$$p(\widehat{\mu}(x)|\mu(x)) = \mathcal{N}(\widehat{\mu}(x); \mu(x), \tau^2/n), \quad x \in \mathcal{X}.$$

Motivation by the Central Limit Theorem:

$$\sqrt{n}(\widehat{\mu}(x) - \mu(x)) \xrightarrow{D} \mathcal{N}(0, \mathsf{var}_{X\sim \mathsf{P}}[k(X,x)]).$$

A heteroscedastic noise model is certainly more appropriate, but let's keep this (obviously wrong) model for now.

Posterior of the embedding

Standard conjugacy results give:

$$\mu(\mathbf{x}) \mid \widehat{\mu}(\mathbf{x}) \sim \mathcal{N}(R(R + (\tau^2/n)I_n)^{-1}\widehat{\mu}(\mathbf{x}), R - R(R + (\tau^2/n)I_n)^{-1}R),$$

where R is the $n \times n$ matrix such that its (i, j)-th element is $r(x_i, x_j)$.

- Recovers the frequentist shrinkage estimator of [Muandet et al, 2013] as the posterior mean (with R instead of K).
- Allows to account for uncertainty in kernel embeddings in the inference procedures.

Learning hyperparameters

Kernel $k = k_\theta$ typically has hyperparameters θ, e.g., bandwidth of the Gaussian (SE) kernel.

Idea: Integrate out the kernel mean embedding μ_θ and consider the probability of our observations $\{x_i\}_{i=1}^n$ given the hyperparameters θ.

Fix a set of points $z_1, \dots, z_m$ in $\mathcal{X} \subset \mathbb{R}^D$, with $m \geq D$.

$$\widehat{\mu_\theta}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} \phi_{\mathbf{z}}(X_i) | \mu_\theta \sim \mathcal{N}\left(\mu_\theta(\mathbf{z}), \frac{\tau^2}{n} I_m\right),$$

with the mapping $\phi_{\mathbf{z}} : \mathbb{R}^D \mapsto \mathbb{R}^m$, given by

$$\phi_{\mathbf{z}}(x) := [k_\theta(x, z_1), \dots, k_\theta(x, z_m)] \in \mathbb{R}^m.$$

How good this model is depends on how far $\phi_{\mathbf{z}}(X_i)|\mu_\theta$ is from $\mathcal{N}\left(\mu_\theta(\mathbf{z}), \tau^2 I_m\right)$.

Marginal (pseudo)likelihood

Assume

$$\phi_{\mathbf{z}}(X_i)|\mu_\theta \sim \mathcal{N}\left(\mu_\theta(\mathbf{z}), \tau^2 I_m\right).$$

and apply change of variable to the mapping $x \mapsto \phi_{\mathbf{z}}(x)$, $\phi_{\mathbf{z}} : \mathbb{R}^D \mapsto \mathbb{R}^m$: what model does this imply on the original space?

$$p(x_1, \ldots, x_n|\theta) = \int p(x_1, \ldots, x_n|\mu_\theta, \theta) p(\mu_\theta|\theta) d\mu_\theta$$

$$= \int \mathcal{N}\left(\phi_{\mathbf{z}}(\mathbf{x}); \left[\mu_\theta(\mathbf{z})^\top \cdots \mu_\theta(\mathbf{z})^\top\right]^\top, \tau^2 I_{mn}\right) \left[\prod_{i=1}^n \gamma_\theta(x_i)\right] p(\mu_\theta|\theta) d\mu_\theta$$

$$= \mathcal{N}\left(\phi_{\mathbf{z}}(\mathbf{x}); \mathbf{0}, \mathbf{1}_n \mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}\right) \prod_{i=1}^n \gamma_\theta(x_i).$$

- Jacobian term: $\gamma_\theta(x) = \left(\det\left[\sum_{l=1}^m \frac{\partial k_\theta(x,z_l)}{\partial x^{(i)}} \frac{\partial k_\theta(x,z_l)}{\partial x^{(j)}}\right]_{ij}\right)^{1/2}$.
- Computational complexity: using Kronecker structure $\mathcal{O}(m^3 + mn)$ for the Gaussian log-likelihood and $\mathcal{O}(nD^3 + nmD^2)$ for the Jacobian term (Gaussian kernel).

Marginal (pseudo)likelihood for a challenging two-sample test

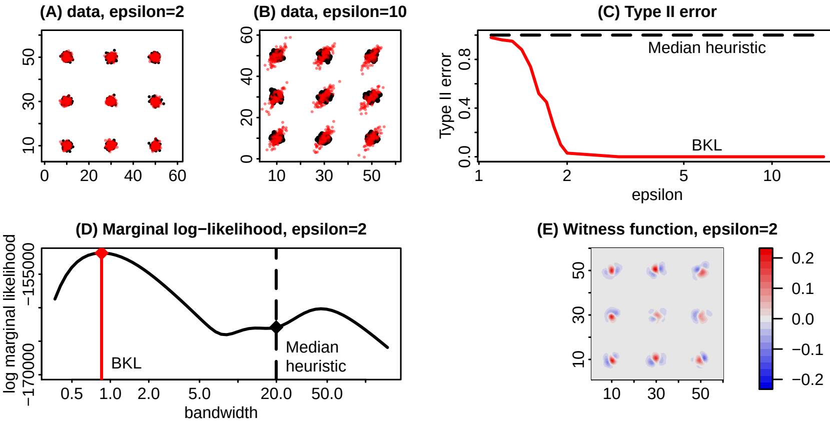
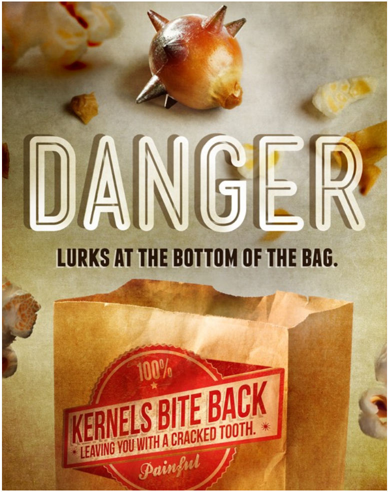


Figure : Comparing samples from a grid of isotropic Gaussians (black dots) to samples from a grid of non-isotropic Gaussians (red dots) with a ratio ϵ of largest to smallest covariance eigenvalues. BKL marginal log-likelihood is maximised for a lengthscale of 0.85 whereas the median heuristic suggests a value of 20.

Summary

- A simple Bayesian model on kernel embeddings recovers shrinkage estimators.
- Marginal (pseudo)likelihood of observations given the kernel hyperparameters allows optimization or sampling of hyperparameters as well.
- Can discover multiscale properties in the data – where there is a mismatch between the global scale of the distribution and the scale at which differences or dependencies are present.
- Potentially a drop-in replacement for median heuristic in unsupervised settings?


DANGER
LURKS AT THE BOTTOM OF THE BAG.
100%
KERNELS BITE BACK
LEAVING YOU WITH A CRACKED TOOTH.
Painful