Kernel Embeddings and Gaussian Processes: Applications in Causal Data Fusion and Statistical Downscaling

Dino Sejdinovic

Department of Statistics
University of Oxford

MARS 2022

Motivating Example 1

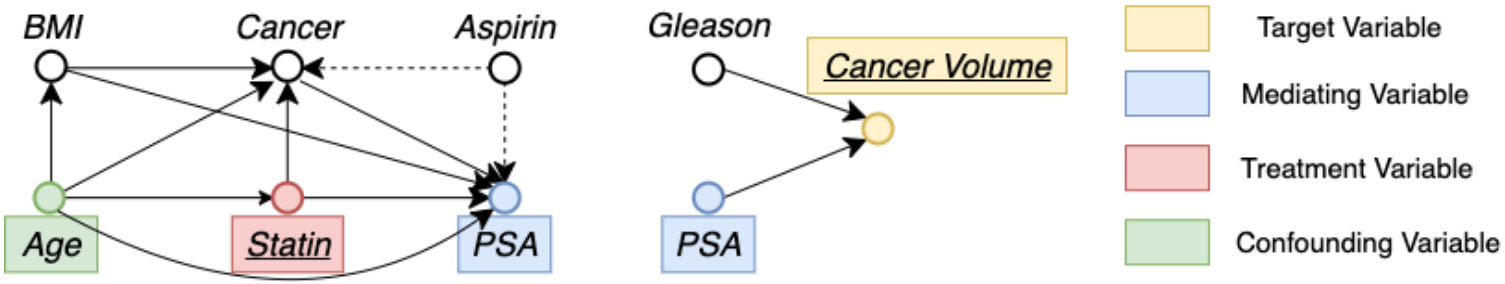
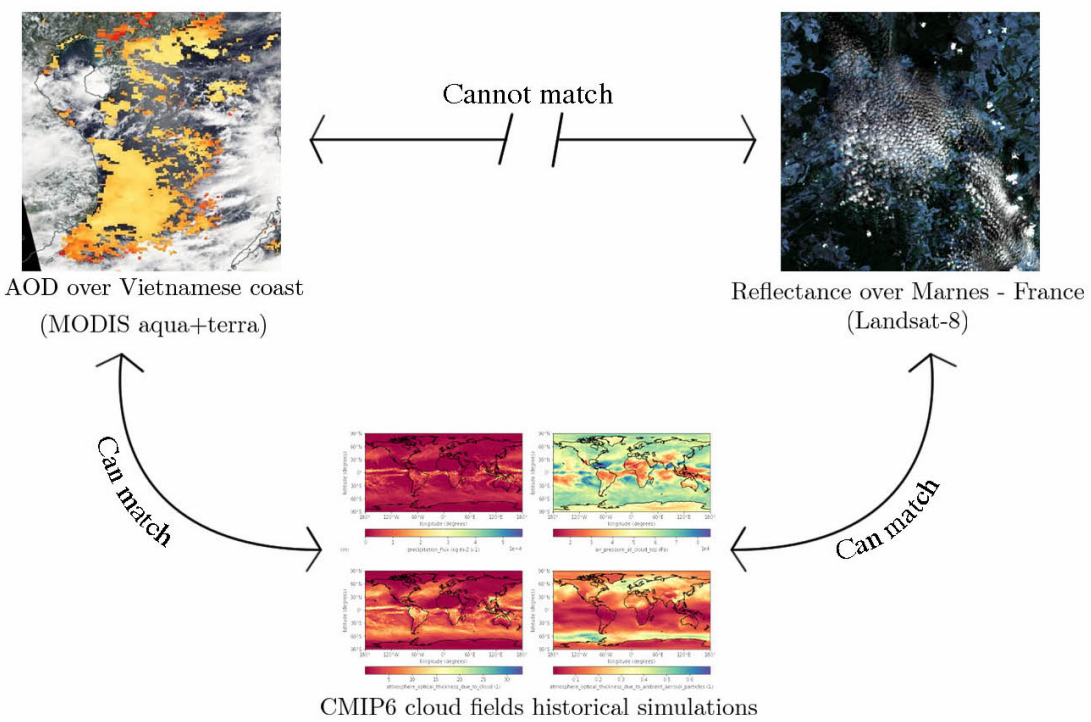
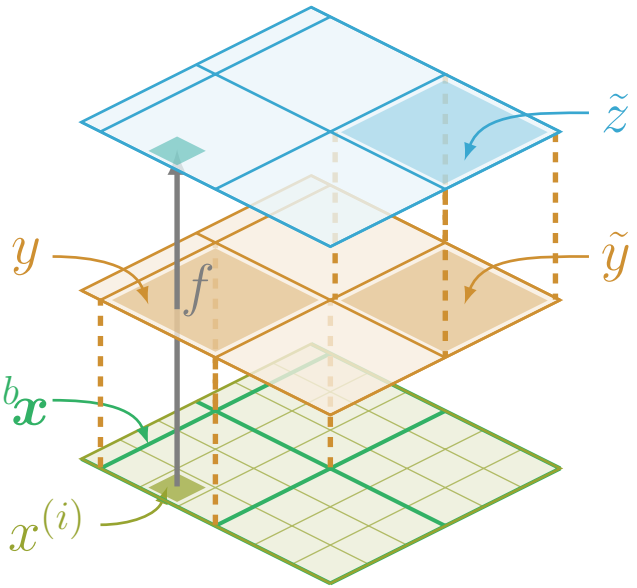


Figure: Causal Graphs corresponding to data collected in two separate medical studies. Left: Data describing the causal relationships between statin level and Prostate Specific Antigen (PSA). Right: Data from a prostate cancer study for patients about to receive a radical prostatectomy.

- Goal: estimate and optimize $\mathbb{E}[\text{Cancer Volume}|do(\text{Statin})]$, i.e. find a statin dosage such that *intervening* on statin would minimize the expected cancer volume.
- Need data fusion and uncertainty quantification.

Motivating Example 2





- Goal: downscale/disaggegate variables measured at a coarse resolution using potentially unmatched high resolution remote sensing data.
- Need data fusion and uncertainty quantification.

Outline

Outline

Kernels and Reproducing Kernel Hilbert Spaces

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation $\varphi : \mathcal{X} \to \mathcal{H}$.

Kernels and Reproducing Kernel Hilbert Spaces

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation $\varphi : \mathcal{X} \to \mathcal{H}$.
- Feature map φ and feature space $\mathcal{H}$ are not unique, but the inner product structure (kernel) is.

Kernels and Reproducing Kernel Hilbert Spaces

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation $\varphi : \mathcal{X} \to \mathcal{H}$.
- Feature map φ and feature space $\mathcal{H}$ are not unique, but the inner product structure (kernel) is.
- Kernel function is 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 $x, x' \in \mathcal{X}$.

Kernels and Reproducing Kernel Hilbert Spaces

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation $\varphi : \mathcal{X} \to \mathcal{H}$.
- Feature map φ and feature space $\mathcal{H}$ are not unique, but the inner product structure (kernel) is.
- Kernel function is 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 $x, x' \in \mathcal{X}$.
- There exists a special (canonical) feature space $\mathcal{H}_k$, called reproducing kernel Hilbert space (RKHS), with canonical feature map $x \mapsto k(\cdot, x)$, where:
  1. $\forall x \in \mathcal{X}, \;\; k(\cdot, x) \in \mathcal{H}_k$, and
  2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \;\; \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$.

  Thus also $k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k}$.

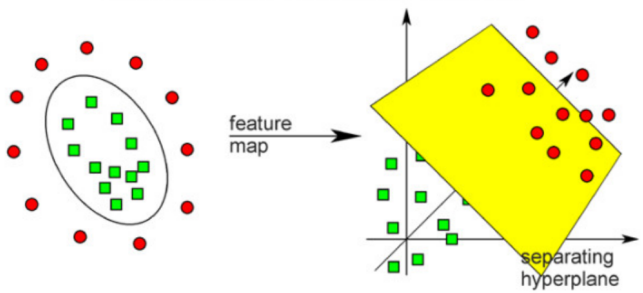
Kernels and Reproducing Kernel Hilbert Spaces

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation $\varphi : \mathcal{X} \to \mathcal{H}$.
- Feature map φ and feature space $\mathcal{H}$ are not unique, but the inner product structure (kernel) is.
- Kernel function is 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 $x, x' \in \mathcal{X}$.
- There exists a special (canonical) feature space $\mathcal{H}_k$, called reproducing kernel Hilbert space (RKHS), with canonical feature map $x \mapsto k(\cdot, x)$, where:
  1. $\forall x \in \mathcal{X}, \;\; k(\cdot, x) \in \mathcal{H}_k$, and
  2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \;\; \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$.

  Thus also $k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k}$.
- Moore-Aronszajn Theorem: *every positive semidefinite* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel of a *unique* RKHS $\mathcal{H}_k$.

Kernel Trick and Kernel Mean Trick

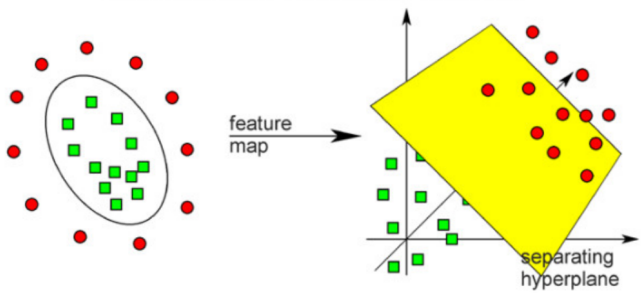
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions,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



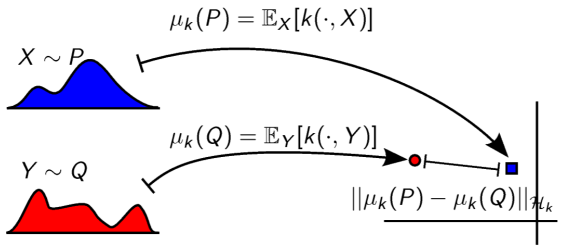
[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

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
- **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 $P \mapsto [\mathbb{E}\phi_1(X), \ldots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y)$
  inner products easy to estimate
  - nonparametric two-sample, independence, conditional independence, interaction testing, learning on distributions



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

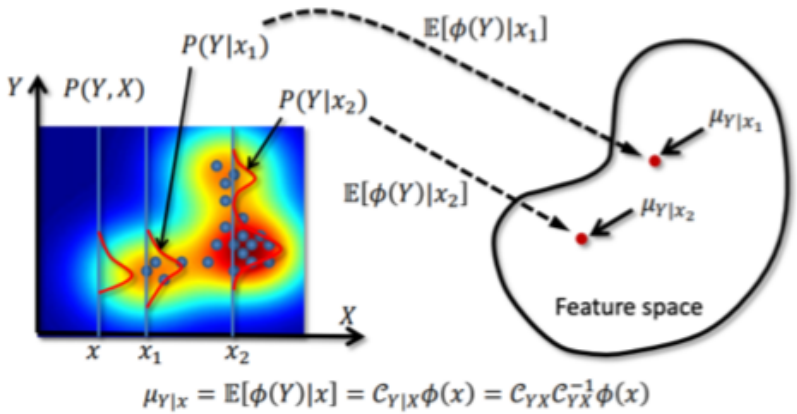


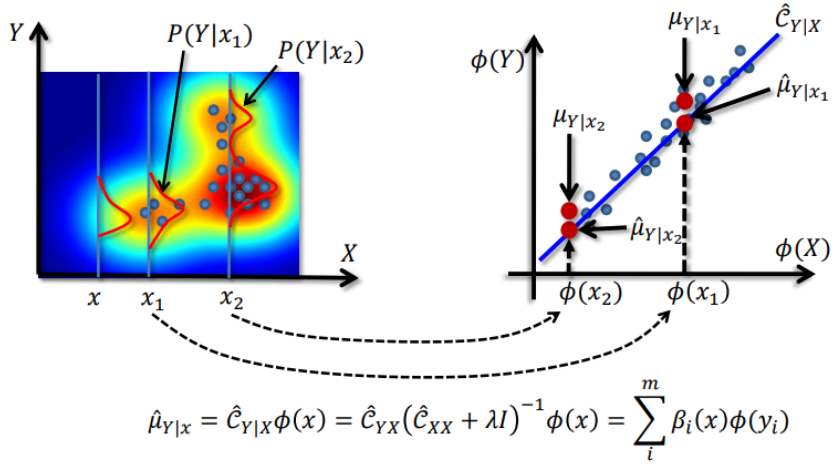
[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS et al, 2013; Muandet et al, 2012; Szabo et al, 2015]

Conditional Mean Embeddings

Consider a joint distribution P_{XY} over the random variables (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$. The conditional mean embedding (CME) of $Y|X = x$ is defined as:

$$\mu_{Y|X=x} := \mathbb{E}_{Y|X=x}[k_y(\cdot, Y)] = \int_{\mathcal{Y}} k_y(\cdot, y) dP(y|x) \in \mathcal{H}_{k_y}$$





To model conditional embeddings as functions of x, we associate them with a conditional mean operator (CMO) $\mathcal{C}_{Y|X} : \mathcal{H}_{k_x} \to \mathcal{H}_{k_y}$, which satisfies

$$\mu_{Y|X=x} = \mathcal{C}_{Y|X} k_x(\cdot, x).$$

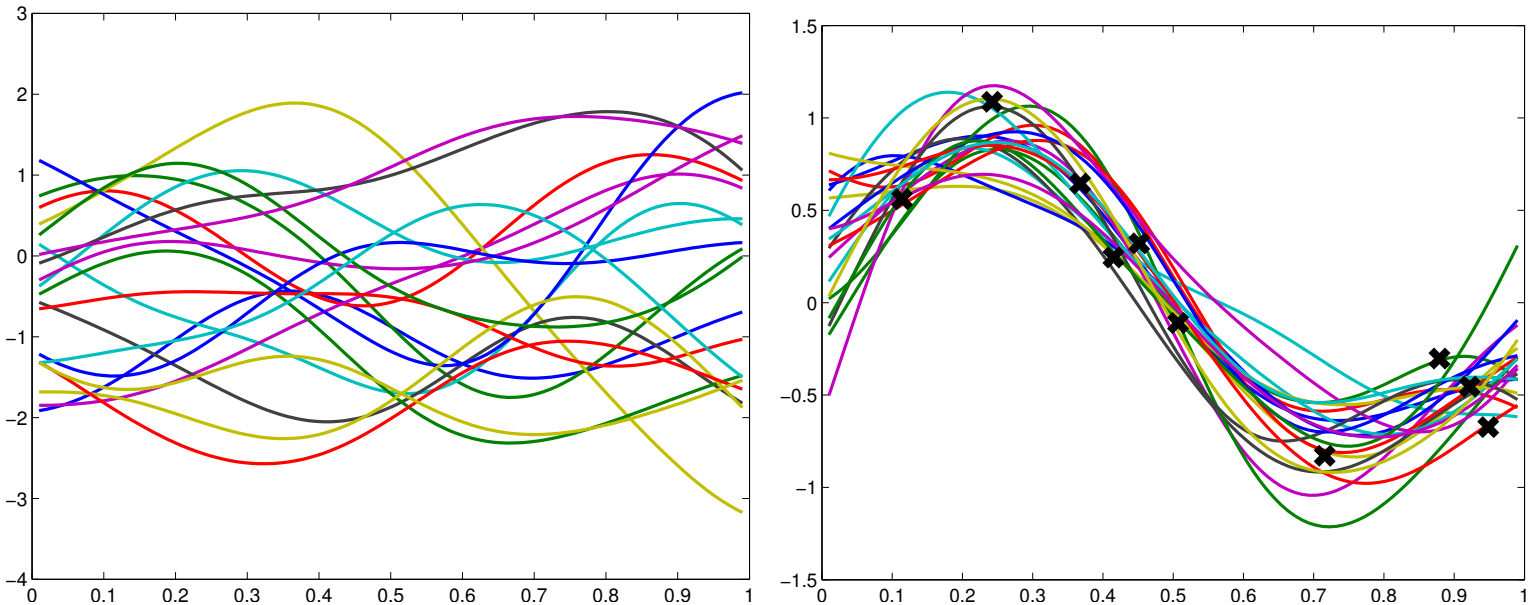
This is essentially feature-to-feature (RKHS-valued) ridge regression.

Gaussian Processes

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

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

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



GP Priors on RKHSs

Can we formulate a GP model for RKHS embeddings [Flaxman et al, 2016]? details

Since sample paths of a GP with kernel k lie outside RKHS $\mathcal{H}_k$ with probability 1 Kallianpur's 0-1 law, [Kallianpur, 1970; Wahba, 1990], we cannot use kernel k.

A smoother kernel, however, can be used, e.g.

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

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

For some simple cases, kernel r is analytically tractable, 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.

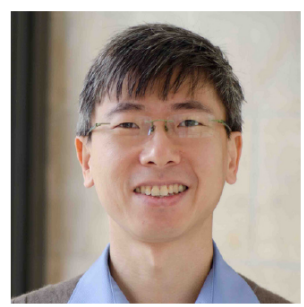
Outline

BayesIMP: Uncertainty Quantification for Causal Data Fusion

Siu Lun Chau* Jean-Francois Ton* Javier Gonzalez
Yee Whye Teh Dino Sejdinovic

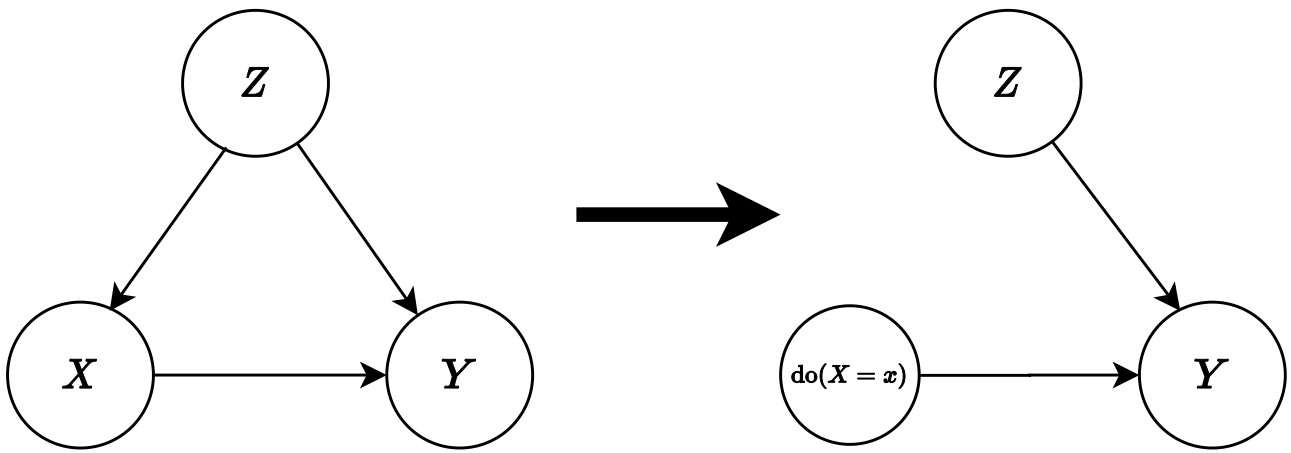
Advances in Neural Information System Processing, 2021



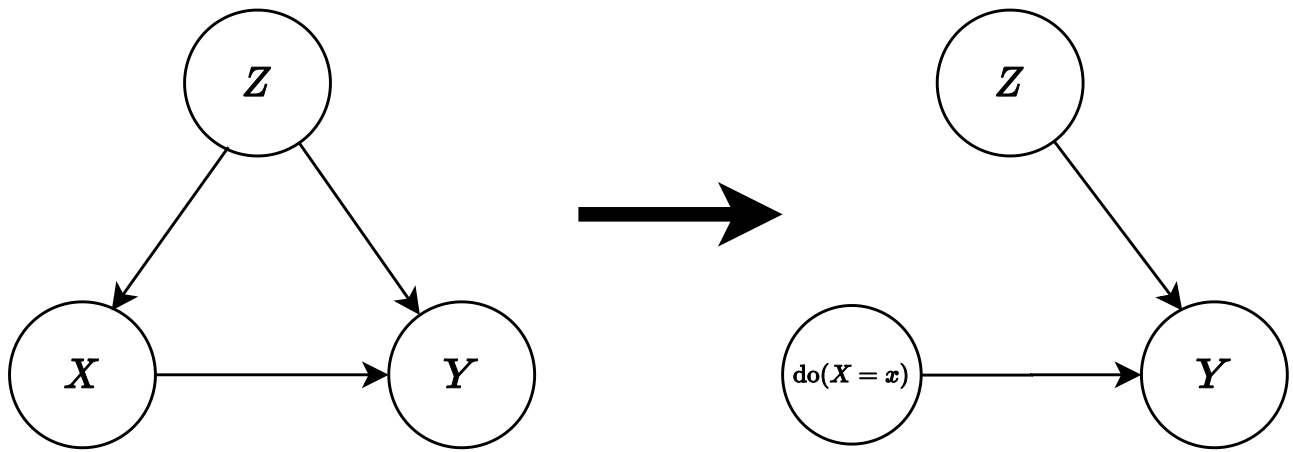
Interventional distribution

- We are interested in the effect that *intervening* on a treatment variable X (independently of all else) has on the response variable Y.
- However in most case we have a confounder Z which depends on both X and Y. Hence, we are not interested in the distribution from which we observe data, but in the distribution corresponding to a different graph where dependency between X and Z has been removed, but we have not affected the conditional distribution of Y given X, Z.



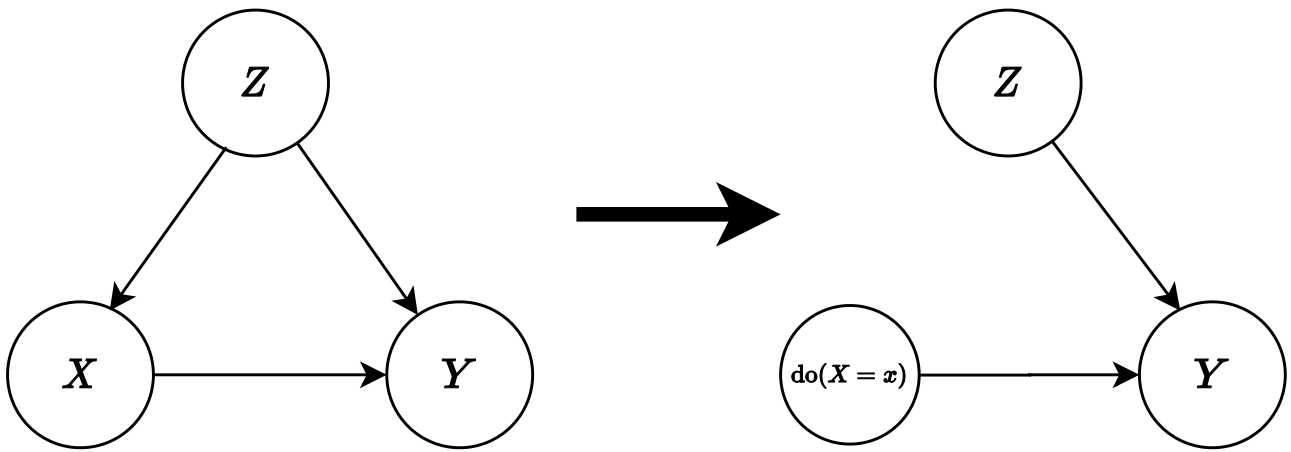
Causal Inference and do-Calculus

- We are interested in $p(Y|do(X) = x)$: a distribution of Y following an intervention on X whose value is set to x.
- How to estimate such *interventional distributions* from observational data?



Causal Inference and do-Calculus

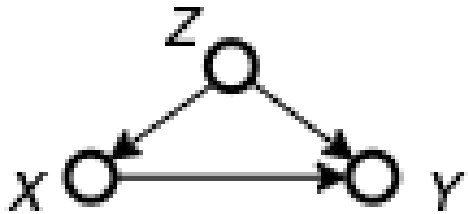
- We are interested in $p(Y|do(X) = x)$: a distribution of Y following an intervention on X whose value is set to x.
- How to estimate such *interventional distributions* from observational data?
- **do-calculus**



Backdoor and Frontdoor Adjustment

Backdoor Adjustment Formulae:

$$p(Y|do(X) = x) = \int_{\mathcal{Z}} p(Y|x, z)\, dP(z)$$



Frontdoor Adjustment Formulae:

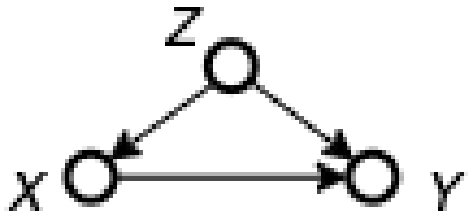
$$p(Y|do(X) = x) = \int_{\mathcal{Z}} \int_{\mathcal{X}} p(Y|X', Z) dP(Z|x) dP(X')$$

- **key idea**: obtain a model for the *do*-density by appropriately combining observational conditional densities, according to a given (known) DAG

Backdoor and Frontdoor Adjustment

Backdoor Adjustment Formulae:

$$p(Y|do(X)=x) = \int_{\mathcal{Z}} p(Y|x,z)\, dP(z)$$



Frontdoor Adjustment Formulae:

$$p(Y|do(X)=x) = \int_{\mathcal{Z}} \int_{\mathcal{X}} p(Y|X',Z) dP(Z|x) dP(X')$$

- **key idea**: obtain a model for the *do*-density by appropriately combining observational conditional densities, according to a given (known) DAG

Can we represent interventional distributions $p(Y|do(X)=x)$ in RKHSs?

YES: IME (Interventional Mean Embeddings (IME) by Singh et al. (2020)

Motivating Example

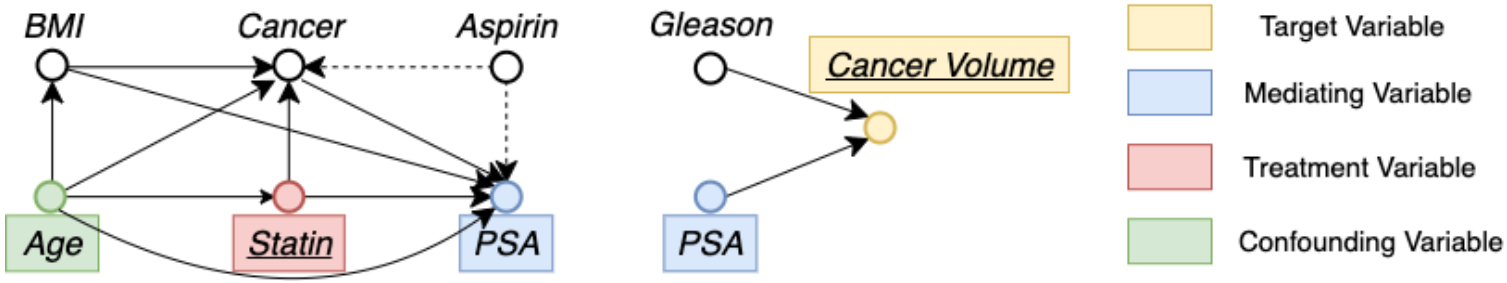
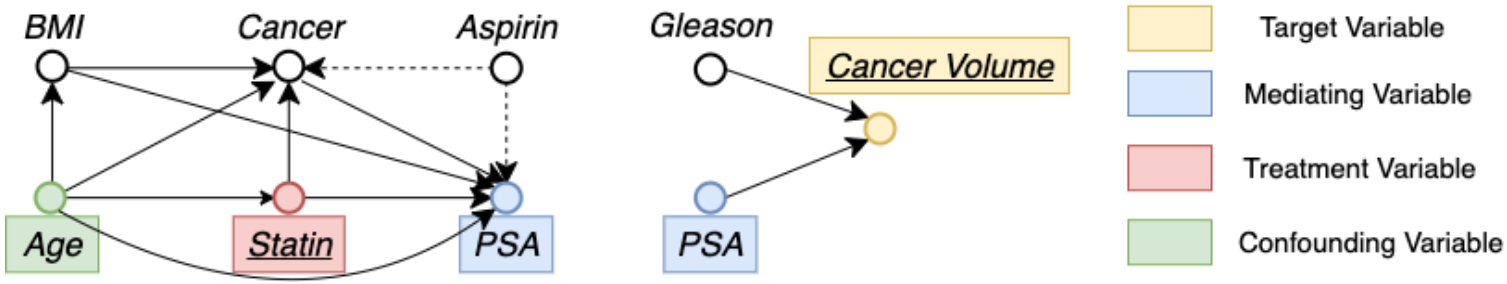


Figure: Causal Graphs corresponding to data collected in two separate medical studies. Left: Data describing the causal relationships between statin level and Prostate Specific Antigen (PSA). Right: Data from a prostate cancer study for patients about to receive a radical prostatectomy.

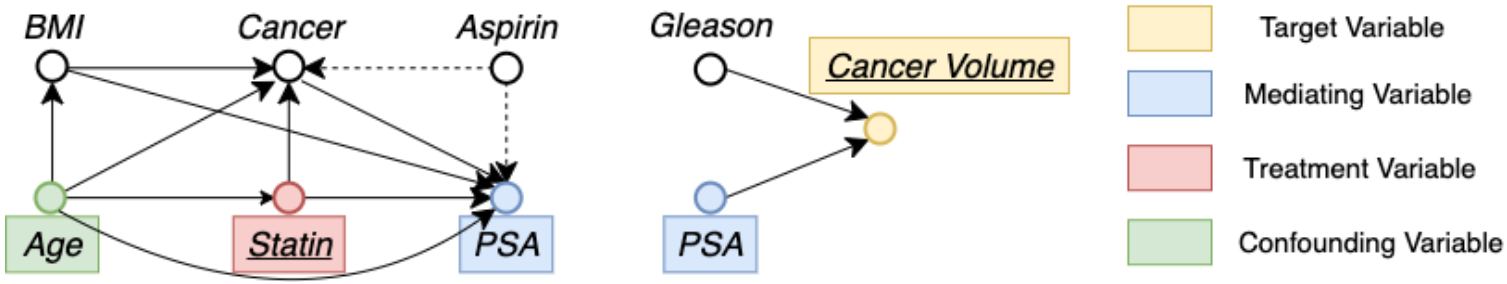
- Goal is to estimate $\mathbb{E}[\text{Cancer Volume}|do(\text{Statin})]$ while also quantifying **uncertainty** arising from *both datasets*.
- Principled uncertainty quantification would allow *Causal Bayesian Optimisation* (Aglietti et al, 2020): find a statin dosage such that *intervening* on statin would minimize the expected cancer volume.

Challenges



1. **Unmatched data**.
   Observed Cancer volume is not paired with Statin in observations. We thus need to perform *Causal Data Fusion* via the mediating variable PSA.
2. **Uncertainty quantification**.
   Datasets come from different studies and might have different quantity and/or quality.

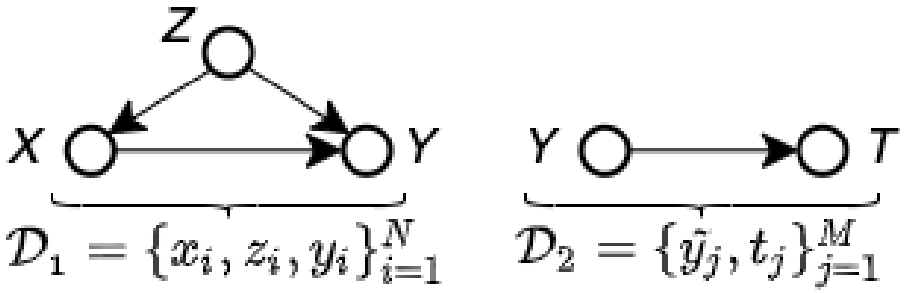
Challenges



1. **Unmatched data**.
   Observed Cancer volume is not paired with Statin in observations. We thus need to perform *Causal Data Fusion* via the mediating variable PSA.
2. **Uncertainty quantification**.
   Datasets come from different studies and might have different quantity and/or quality.

We propose a method termed *Bayesian Interventional Mean Process* (BayesIMP) to model average treatment effects which combines ideas from Gaussian Processes (GPs), conditional mean embeddings (CMEs) and do-calculus.

Causal Data Fusion Problem



- We are given two **causal graphs** and two corresponding **datasets** $\mathcal{D}_1, \mathcal{D}_2$.
  - X: treatment, Y: mediator, Z: confounder, T: response
- The goal is to infer $\mathbb{E}[T|do(X) = x]$ from these two datasets.
- We make the following assumptions:
  - **A1** Treatment only affects the target through the mediating variable, i.e $T \perp\!\!\!\perp do(X)|Y$
  - **A2** Function f given by $f(y) = \mathbb{E}[T|Y = y]$ belongs to an RKHS $\mathcal{H}_{k_y}$.

Causal Data Fusion Problem

- Using these assumptions and standard RKHS properties, we have:

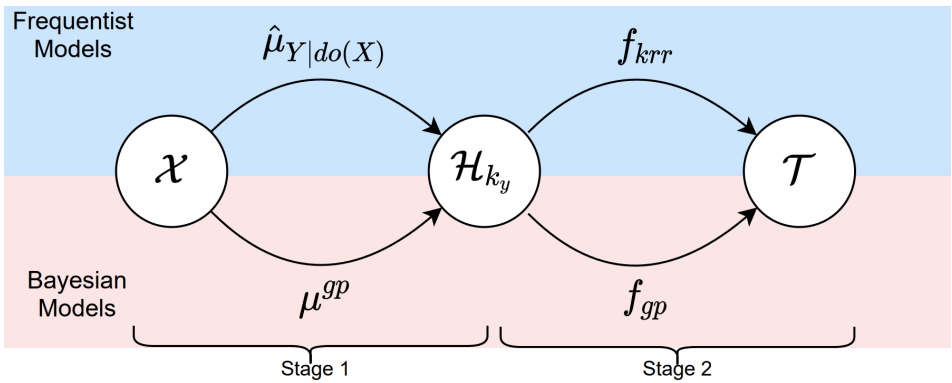
$$
\begin{aligned}
\mathbb{E}[T|do(X)=x] &= \int_{\mathcal{Y}} \underbrace{\mathbb{E}[T|do(x),y]}_{=\mathbb{E}[T|y],\ \text{since}\ T \perp\!\!\!\perp do(X)|Y} dP(y|do(x)) \\
&= \int_{\mathcal{Y}} f(y) dP(y|do(x)) \\
&= \int_{\mathcal{Y}} \langle f, k_y(\cdot,y)\rangle_{\mathcal{H}_{k_y}} dP(y|do(x)) \\
&= \langle f, \int_{\mathcal{Y}} k_y(\cdot,y) dP(y|do(x))\rangle_{\mathcal{H}_{k_y}} \\
&= \langle f, \underbrace{\mu_{Y|do(X)=x}}_{\text{IME}}\rangle_{\mathcal{H}_{k_y}}.
\end{aligned}
$$

Proposed Methods Summary

Function of interest is

$$g(x) = \mathbb{E}[T|do(X) = x] = \langle f, \mu_{Y|do(X)=x} \rangle_{\mathcal{H}_{k_y}}.$$

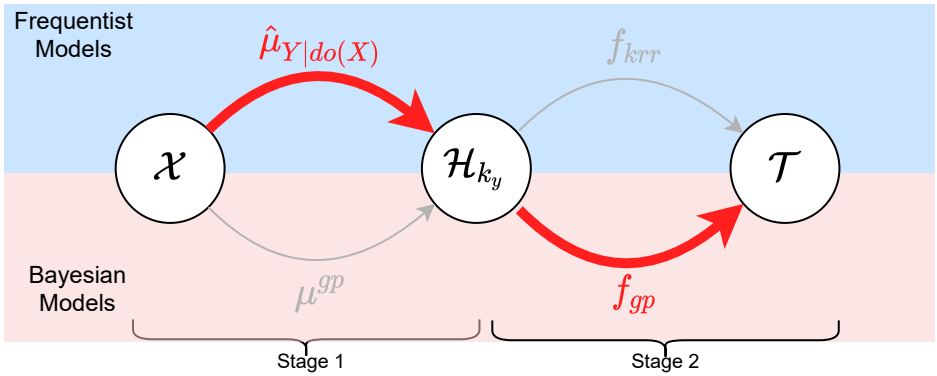
Both f and $\mu_{Y|do(X)=x}$ need to be estimated.



METHODS	$\mu_{Y\|do(X)}$	f
IME	KRR	KRR
IMP*	KRR	GP
BayesIME*	GP	KRR
BayesIMP*	GP	GP

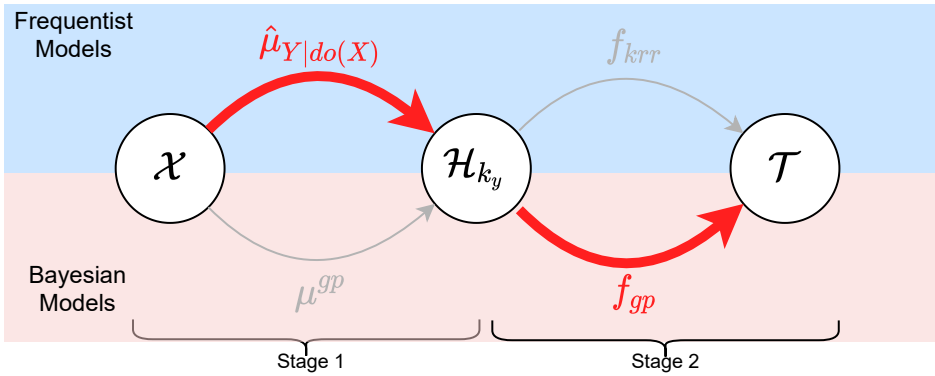
- Estimate $\mu_{Y|do(X)}$ from $\mathcal{D}_1 = \{x_i, z_i, y_i\}_{i=1}^N$
- Estimate f from $\mathcal{D}_2 = \{\tilde{y}_j, t_j\}_{j=1}^M$

Interventional Mean Process (IMP)



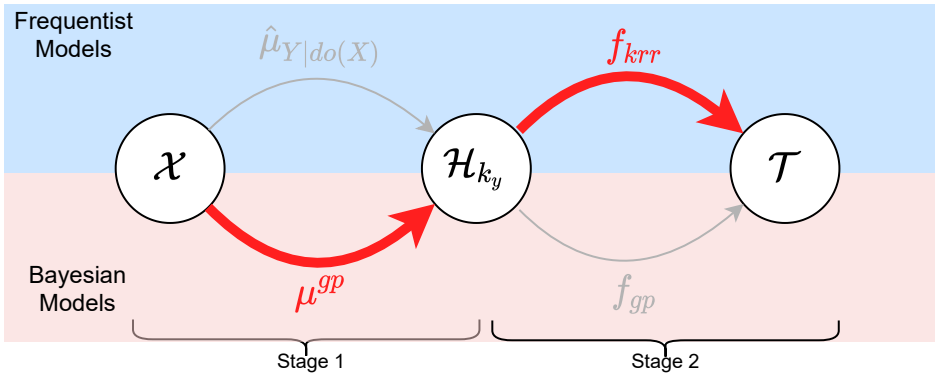
- IMP models f as a GP *with sample paths in* $\mathcal{H}_{k_y}$, and uses a vector-valued KRR estimate of IME $\hat{\mu}_{Y|do(X)}$.
- Then $g(x) = \langle f, \hat{\mu}_{Y|do(X)=x} \rangle_{\mathcal{H}_{k_y}}$ is by linearity also a GP and we can compute its mean and covariance directly from the posterior mean and covariance of f.

Interventional Mean Process (IMP)



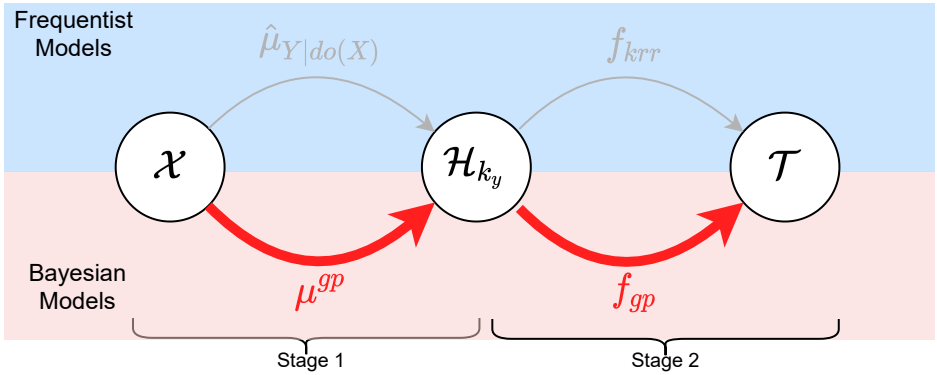
- IMP models f as a GP *with sample paths in* $\mathcal{H}_{k_y}$, and uses a vector-valued KRR estimate of IME $\hat{\mu}_{Y|do(X)}$.
- Then $g(x) = \langle f, \hat{\mu}_{Y|do(X)=x} \rangle_{\mathcal{H}_{k_y}}$ is by linearity also a GP and we can compute its mean and covariance directly from the posterior mean and covariance of f.
- But only uncertainty in $\mathcal{D}_2$ is taken into account!

Bayesian Interventional Mean Embedding (BayesIME)



- **Idea**: replace vv-KRR in CMEs with vv-GPs, corresponding to a *Bayesian model for CMEs*, and obtain the corresponding *Bayesian model for IMEs* via do-calculus.
- **Challenge**: need to ensure that GP posterior draws of $h(x, y) = \mu_{Y|X=x}(y)$ give a.s. an RKHS function of y, $\forall x \in \mathcal{X}$.
- We show that it is sufficient to take prior $h \sim \mathcal{GP}(0, k_x \otimes r_y)$ for a kernel r_y *nuclearly dominant* over k_y.
- Now if $\hat{f}$ comes from KRR, $g(x) = \langle \hat{f}, \mu_{Y|do(X)=x} \rangle_{\mathcal{H}_{k_y}}$ is again a GP by linearity.

Bayesian Interventional Mean Process (BayesIMP)



$$g(x) = \mathbb{E}[T|do(X) = x] = \langle f, \mu_{Y|do(X)=x} \rangle_{\mathcal{H}_{k_y}}.$$

- Finally, place GPs on both f and $\mu_{Y|do(X)}$ to quantify uncertainties in both datasets.
- Problem: inner product of Gaussians are not Gaussians so need to resort to moment matching to obtain a GP model for g to be used for e.g. Bayesian optimisation.

Ablation Study: Result

Simple causal graph: $X \to Y$ and $Y \to T$

GOAL: $\mathbb{E}[T|do(X)]$

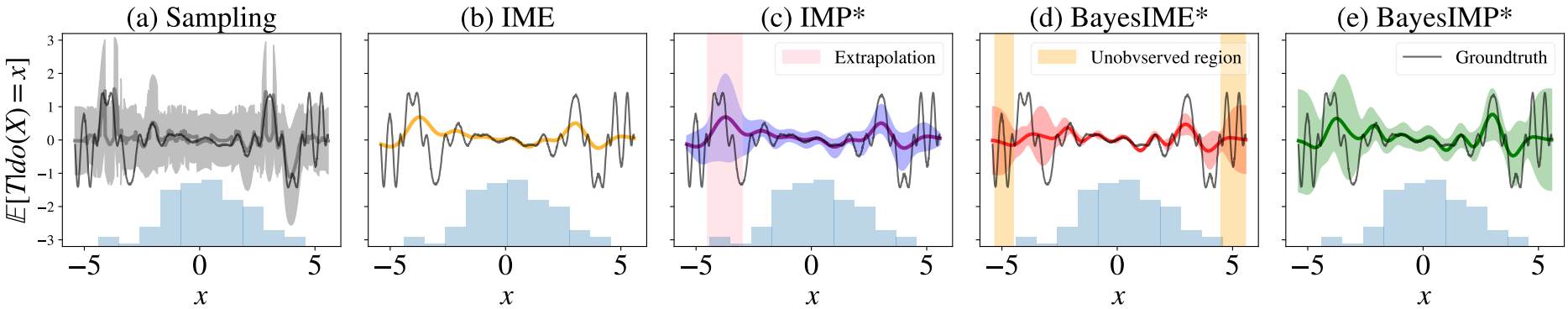


Figure: Ablation studies of various methods in estimating uncertainties for an illustrative experiment. IME does not come with uncertainty estimates. We see IMP and BayesIME covering different regions of uncertainty while BayesIMP takes the best of both worlds.

Ablation Study: Calibration

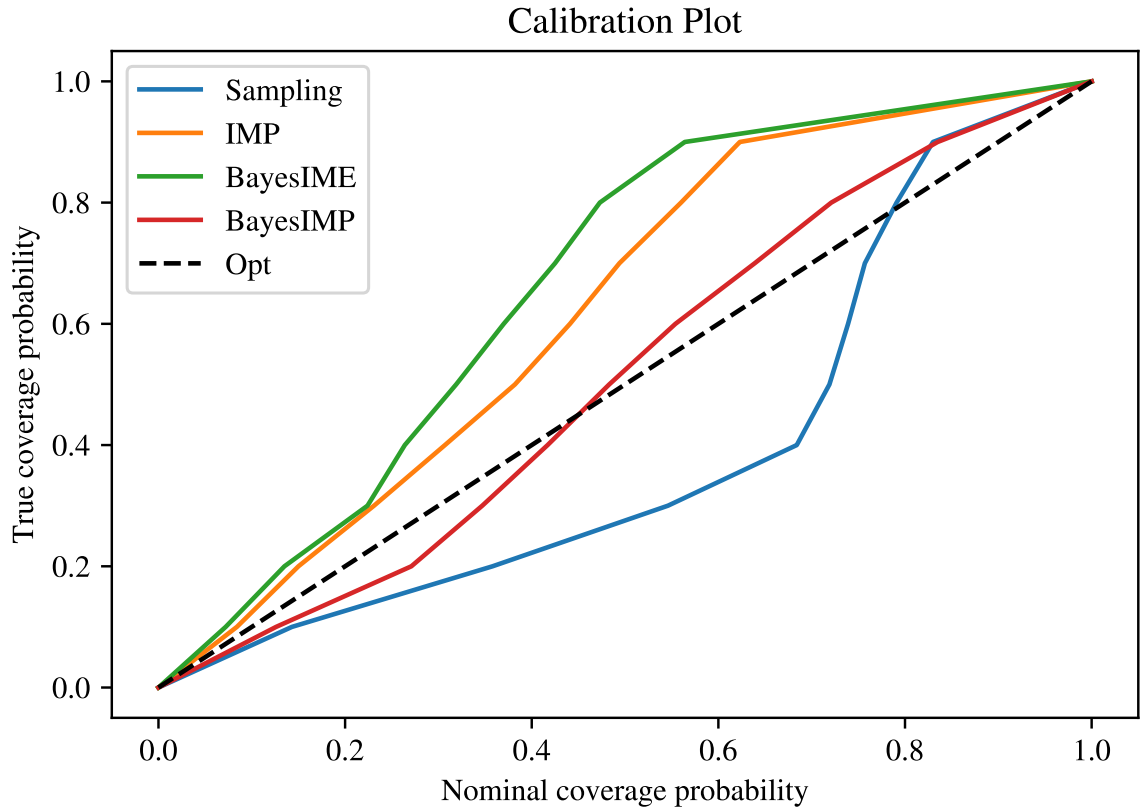


Figure: Calibration plots of Sampling method as well as our 3 proposed methods. We clearly see that BayesIMP is the best calibrated method amongst all other methods.

Causal Bayesian Optimisation (CBO)

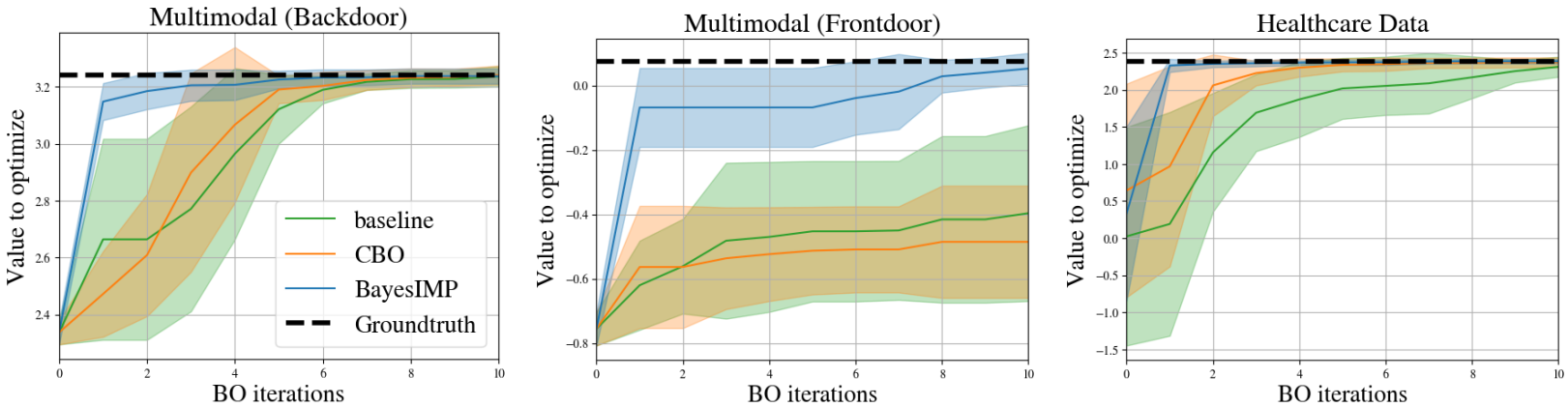


Figure: (Left) Backdoor adjustment and **multimodal** mediator Y, Frontdoor adjustment and **multimodal** mediator Y, (Right) Healthcare example, optimizing $\mathbb{E}[CancerVolume|do(Statin)]$. CBO is the sampling-based approach from Aglietti et al (2020).

Summary

- BayesIMP: a Bayesian method to estimate average treatment effect from unmatched observational data
  - GP model for representing interventional distributions in RKHSs
  - Can capture uncertainty arising from multiple datasets and combine them effectively
  - Leads to faster causal Bayesian optimisation
- Future directions:
  - Assumes full knowledge of the underlying causal graphs – can this be relaxed?

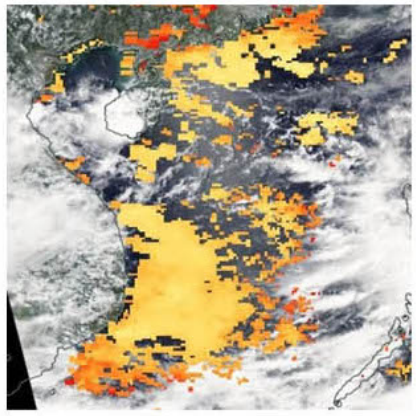
Outline

Deconditional Downscaling with Gaussian Processes

Siu Lun Chau* Shahine Bouabid* Dino Sejdinovic

Advances in Neural Information System Processing, 2021

Motivation



AOD over Vietnamese coast

(MODIS aqua+terra)

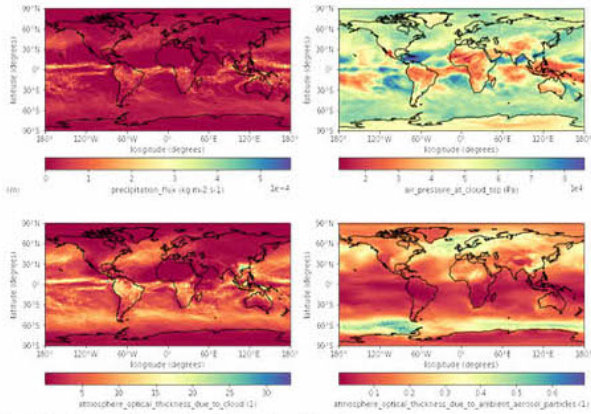
Cannot match

Reflectance over Marnes - France

(Landsat-8)

Can match

Can match



CMIP6 cloud fields historical simulations

Problem Setup

Data

- We have a dataset of N bags of high-resolution (HR) covariates ${}^b\boldsymbol{x}_j := \{x_j^{(1)}, \ldots, x_j^{(n_j)}\}$ each paired with a mediating low-resolution (LR) variable y_j

$$\mathcal{D}_1 = \{{}^b\boldsymbol{x}_j, y_j\}_{j=1}^N.$$

- We have a separate dataset of M mediating LR variables $\tilde{y}_j$ paired with a LR response of interest $\tilde{z}_j$.

$$\mathcal{D}_2 = \{\tilde{y}_j, \tilde{z}_j\}_{j=1}^M.$$

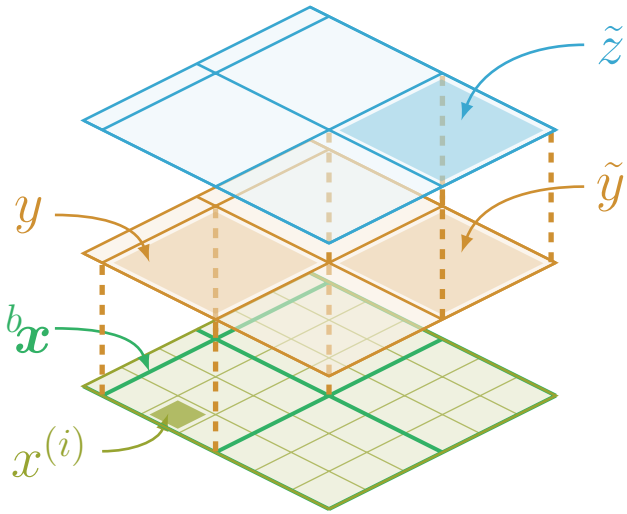


Figure: Illustration of HR and LR observations – indirect pairing

Problem Setup

Objective

- Downscale response z to the HR granularity level of $x_j^{(i)}$ covariates
  i.e. find a function $f : \mathcal{X} \to \mathbb{R}$ which maps between HR covariates and HR responses.

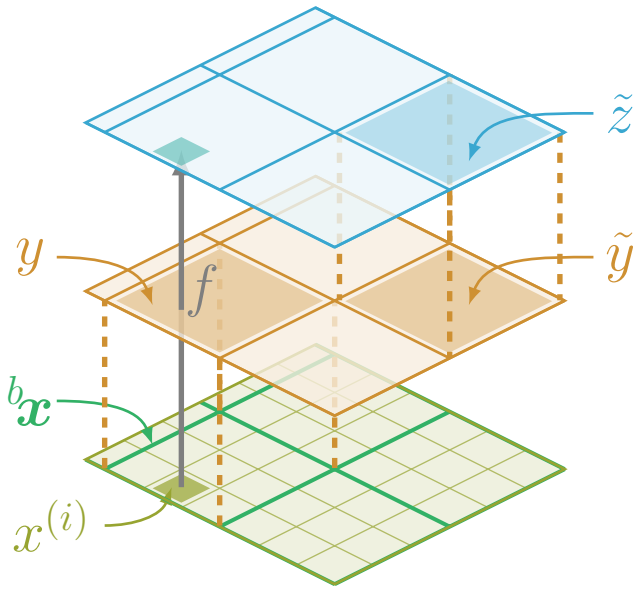


Figure: We wish to learn a map from HR covariates to an HR estimate of the response

Deconditional Formulation

Observation Model

- We assume that the HR responses $f(x)$ aggregate into the LR response $\tilde{z}_j$ as

$$\tilde{z}_j = \mathbb{E}_X[f(X)|Y = \tilde{y}_j] + \varepsilon_j$$

with noise $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$.

This is similar to the *deconditioning* problem studied by Hsu & Ramos (2019):

- Given an RKHS function $g : \mathcal{Y} \to \mathbb{R}$, infer an RKHS function $f : \mathcal{X} \to \mathbb{R}$ such that

$$g(y) = \mathbb{E}_X[f(X)|Y = y].$$

f is called the *deconditional mean of* g w.r.t. $\mathbb{P}_{X|Y}$.

Hsu & Ramos (2019) develop a deconditioning procedure based on estimating so called deconditional mean operators and complex chained inference derivations.

Bayesian formulation for f and g

- By placing a GP prior on $f \sim \mathcal{GP}(m, k)$, we can represent the LR field of responses as

$$g(y) = \mathbb{E}_X[f(X)|Y = y] = \int_{\mathcal{X}} f(x) \mathrm{d}\mathbb{P}_{X|Y=y}(x) \sim \mathcal{GP}(\nu, q).$$

  By linearity of expectation, g is also a GP where

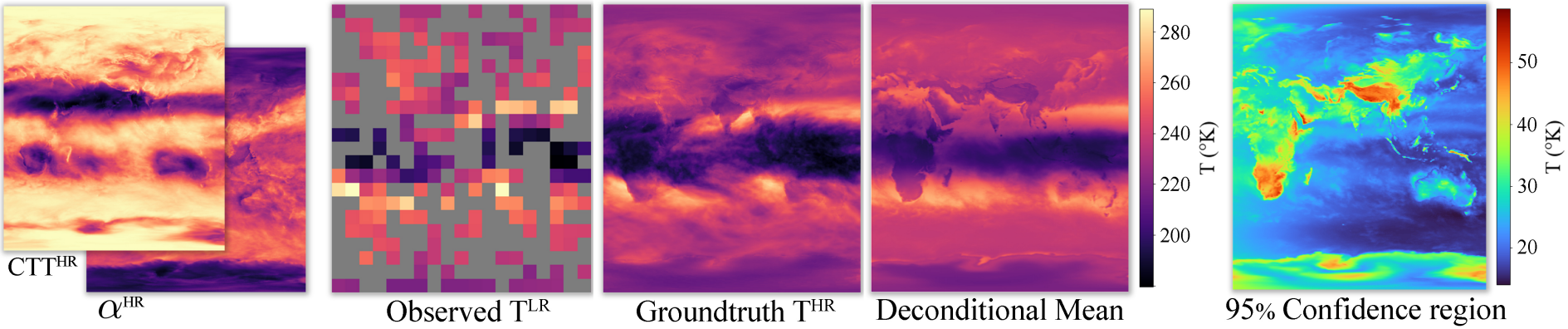
$$\nu(y) = \mathbb{E}_X[m(X)|Y = y]$$

$$q(y, y') = \mathbb{E}_{X,X'}[k(X, X')|Y = y, Y' = y'] = \langle \mu_{X|Y=y}, \mu_{X|Y=y'} \rangle$$

- Estimation of ν and q via conditional mean embeddings using $\mathcal{D}_1$.
- By joint normality between LR and HR fields, recover a posterior for HR field f based on $\mathcal{D}_2$.

Posterior mean has a form essentially identical to the estimator by Hsu & Ramos (2019).

Mediated Downscaling of Atmospheric Temperature



Model	RMSE ↓	MAE ↓	Corr. ↑	SSIM ↑
Kriging	8.02±0.28	5.55±0.17	0.831±0.012	**0.212**±0.011
VBAgg	8.25±0.15	5.82±0.11	0.821±0.006	0.182±0.004
Our method	**7.40**±0.25	**5.34**±0.22	**0.848**±0.011	**0.212**±0.013

Table: Downscaling similarity scores of posterior mean against HR groundtruth; reports 1 s.d. VBAgg approach from Law et al (2018) also operates on aggregate likelihoods but cannot handle unmatched data and thus requires to first estimate LR response for each bag of HR covariates. It can be thought of as a special case of the proposed method where mediating LR covariate is simply one-hot encoding of the bag.

Summary

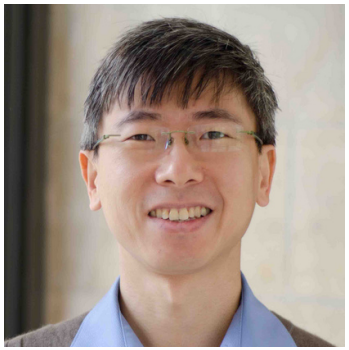
- A scalable Bayesian solution to the mediated statistical downscaling problem, which handles unmatched multi-resolution data.
- Combines Gaussian Processes with the framework of deconditioning using RKHSs and recovers previous approaches as its special cases.
- Future challenges: what if the mediating variable undergoes covariate shift between the two datasets?

References

- Siu Lun Chau, Shahine Bouabid, and DS, **Deconditional Downscaling with Gaussian Processes**, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Siu Lun Chau, Jean-Francois Ton, Javier Gonzalez, Yee Whye Teh, and DS, **BayesIMP: Uncertainty Quantification for Causal Data Fusion**, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.







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
- 0/1 laws: GP sample paths with (infinite-dimensional) covariance kernel k almost surely fall outside of $\mathcal{H}_k$.
  - The space of sample paths can be thought of as an "outer shell" of $\mathcal{H}_k$.
- Worst-case in RKHS $\Leftrightarrow$ Average-case in GPs

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

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

A Bayesian model of RKHS embeddings

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

Bayesian Learning of Kernel Embeddings
S. Flaxman, DS, J. P. Cunningham, and S. Filippi in **Uncertainty in Artificial Intelligence (UAI)**, 2016