Abstract

While causal models are becoming one of the mainstays of machine learning, the problem of uncertainty quantification in causal inference remains challenging. In this paper, we study the causal data fusion problem, where datasets pertaining to multiple causal graphs are combined to estimate the average treatment effect of a target variable. As data arises from multiple sources and can vary in quality and quantity, principled uncertainty quantification becomes essential. To that end, we introduce Bayesian Interventional Mean Processes, a framework which combines ideas from probabilistic integration and kernel mean embeddings to represent interventional distributions in the reproducing kernel Hilbert space, while taking into account the uncertainty within each causal graph. To demonstrate the utility of our uncertainty estimation, we apply our method to the Causal Bayesian Optimisation task and show improvements over state-of-the-art methods.

1 Introduction

Causal inference has seen a significant surge of research interest in areas such as healthcare [1], ecology [2], and optimisation [3]. However, data fusion, the problem of merging information from multiple data sources, has received limited attention in the context of causal modelling, yet presents significant potential benefits for practical situations [4, 5]. In this work, we consider a causal data fusion problem where two causal graphs are combined for the purposes of inference of a target variable (see Fig.1). In particular, our goal is to quantify the uncertainty under such a setup and determine the level of confidence in our treatment effect estimates.

Let us consider the motivating example in Fig.1, where a medical practitioner is investigating how prostate cancer volume is affected by a statin drug dosage. We consider the case where the doctor

---

* *Denotes equal contribution with alphabetical ordering
only has access to two separate medical studies describing the quantities of interest. On one hand we have observational data, from one medical study $D_1$ [1], describing the causal relationship between statin level and prostate specific antigen (PSA), and on the other hand we have observational data, from a second study $D_2$ [7], that looked into the link between PSA level and prostate cancer volume. The goal is to model the interventional effect between our target variable (cancer volume) and the treatment variable (statin). This problem setting is different from the standard observational scenario as it comes with the following challenges:

- **Unmatched data**: Our goal is to estimate $\mathbb{E}[\text{cancer volume}|do(\text{statin})]$ but the observed cancer volume is not paired with statin dosage. Instead, they are related via a mediating variable PSA.

- **Uncertainty quantification**: The two studies may be of different data quantity/quality. Furthermore, a covariate shift in the mediating variable, i.e. a difference between its distributions in two datasets, may cause inaccurate extrapolation. Hence, we need to account for uncertainty in both datasets.

Formally, let $X$ be the treatment (Statin), $Y$ be the mediating variable (PSA) and $T$ our target (cancer volume), and our aim is to estimate $\mathbb{E}[T|do(X)]$. The problem of unmatched data in a similar context has been previously considered by [5] using a two-staged regression approach ($X \rightarrow Y$ and $Y \rightarrow T$). However, uncertainty quantification, despite being essential if our estimates of interventional effects will guide decision-making, has not been previously explored. In particular, it is crucial to quantify the uncertainty in both stages as this takes into account the lack of data in specific parts of the space. Given that we are using different datasets for each stage, there are also two sources of epistemic uncertainties (due to lack of data) as well as two sources of aleatoric uncertainties (due to inherent randomness in $Y$ and $T$) [8]. It is thus natural to consider regression models based on Gaussian Processes (GP) [9], as they are able to model both types of uncertainties. However, as GPs, or any other standard regression models, are designed to model conditional expectations only and will fail to capture the underlying distributions of interest (e.g. if there is multimodality in $Y$ as discussed in [10]). This is undesirable since, as we will see, interventional effect estimation requires accurate estimates of distributions. While one could in principle resort to density estimation methods, this becomes challenging since we typically deal with a number of conditional/interventional densities.

In this paper, we introduce the framework of *Bayesian Interventional Mean Processes* (BAYESIMP) to circumvent the challenges in the causal data fusion setting described above. BAYESIMP considers kernel mean embeddings [11] for representing distributions in a reproducing kernel Hilbert space (RKHS), in which the whole arsenal of kernel methods can be extended to probabilistic inference (e.g. kernel Bayes rule [12], hypothesis testing [13], distribution regression [14]). Specifically, BAYESIMP uses kernel mean embeddings to represent the interventional distributions and to analytically marginalise out $Y$, hence accounting for aleatoric uncertainties. Further, BAYESIMP uses GPs to estimate the required kernel mean embeddings from data in a Bayesian manner, which allows to quantify the epistemic uncertainties when representing the interventional distributions. To illustrate the quality of our uncertainty estimates, we apply BAYESIMP to Causal Bayesian Optimisation [15], an efficient heuristic to optimise objective functions of the form $x^* = \arg\min_{x \in X} \mathbb{E}[T|do(X) = x]$. Our contributions are summarised below:

1. We propose a novel *Bayesian Learning of Conditional Mean Embedding* (BAYESCME) that allows us to estimate conditional mean embeddings in a Bayesian framework.

2. Using BAYESCME, we propose a novel *Bayesian Interventional Mean Process* (BAYESIMP) that allows us to model interventional effect across causal graphs without explicit density estimation, while obtaining uncertainty estimates for $\mathbb{E}[T|do(X) = x]$.

3. We apply BAYESIMP to Causal Bayesian Optimisation, a problem introduced in [15] and show significant improvements over existing state-of-the-art methods.

Note that [16] also considered a causal fusion problem but with a different objective. They focused on extrapolating experimental findings across treatment domains, i.e. inferring $\mathbb{E}[Y|do(S)]$ when only data from $p(Y|do(S))$ is observed, where $S$ is some other treatment variable. In contrast, we focus on modelling combined causal graphs, with a strong emphasis on uncertainty quantification. While [17] considered mapping interventional distributions in the RKHS to model quantities such as $\mathbb{E}[T|do(X)]$, they only considered a frequentist approach, which does not account for epistemic uncertainties.
Notations. We denote \( X, Y, Z \) as random variables taking values in the non-empty sets \( \mathcal{X}, \mathcal{Y} \) and \( \mathcal{Z} \) respectively. Let \( k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) be positive definite kernels on \( \mathcal{X} \) with an associated RKHS \( \mathcal{H}_{k_x} \). The corresponding canonical feature map \( k_x(x, \cdot) \) is denoted as \( \phi_x(x) \). Analogously for \( Y \) and \( Z \).

In the simplest setting, we observe i.i.d samples \( \mathcal{D}_1 = \{ x_i, y_i, z_i \}_{i=1}^N \) from joint distribution \( \mathbb{P}_{XYZ} \) which we concatenate into vectors \( x := [x_1, ..., x_N]^\top \). Similarly for \( y \) and \( z \). For this work, \( X \) is referred as the treatment variable, \( Y \) as the mediating variable and \( Z \) as adjustment variables accounting for confounding effects. With an abuse of notation, features matrices are defined by stacking feature maps along the columns, i.e \( \Phi_x := [\phi_x(x_1), ..., \phi_x(x_N)] \). We denote the Gram matrix as \( K_{xx} := \Phi_x \Phi_x^\top \) and the vector of evaluations \( k_{xx} \) as \( [k_x(x, x_1), ..., k_x(x, x_N)] \). We define \( \Phi_y, \Phi_z \) analogously for \( y \) and \( z \).

Lastly, we denote \( T = f(Y) + \epsilon \) as our target variable, which is modelled as some noisy evaluation of a function \( f : \mathcal{Y} \to \mathcal{T} \) on \( Y \) while \( \epsilon \) being some random noise. For our problem setup we observe a second dataset of i.i.d realisations \( \mathcal{D}_2 = \{ \tilde{y}_j, t_j \}_{j=1}^M \) from the joint \( \mathbb{P}_{YT} \) independent of \( \mathcal{D}_1 \). Again, we define \( \tilde{y} := [\tilde{y}_1, ..., \tilde{y}_M]^\top \) and \( t := [t_1, ..., t_M]^\top \) just like for \( \mathcal{D}_1 \). See Fig.2 for illustration.

2 Background

Representing interventional distributions in an RKHS\(^1\) has been explored in different contexts [18, 17, 19]. In particular, when the treatment is continuous, [17] introduced the Intervenional Mean Embeddings (IMEs) to model densities in an RKHS by utilising smoothness across treatments. Given that IME is an important building block to our contribution, we give it a detailed review by first introducing the key concepts of do-calculus [20] and conditional mean embeddings [21].

2.1 Interventional distribution and do-calculus

In this work, we consider the structural causal model [20] (SCM) framework, where a causal directed acyclic graph (DAG) \( \mathcal{G} \) is given and encodes knowledge of existing causal mechanisms amongst the variables in terms of conditional independencies. Given random variables \( X \) and \( Y \), a central question in interventional inference [20] is to estimate the distribution \( p(Y|do(X) = x) \), where \( \{do(X) = x\} \) represents an intervention on \( X \) whose value is set to \( x \). Note that this quantity is not directly observed given that we are usually only given observational data, i.e., data sampled from the conditional \( p(Y|X) \) but not from the interventional density \( p(Y|do(X)) \). However, Pearl [20] developed do-calculus which allows us to estimate interventional distributions from purely observational distributions under the identifiability assumption. Here we present the backdoor and front-door adjustments, which are the fundamental components of DAG based causal inference.

The backdoor adjustment is applicable when there are observed confounding variables \( Z \) between the cause \( X \) and the effect \( Y \) (see Fig. 3 (Top)). In order to correct for this confounding bias we can use the following equation, adjusting for \( Z \) as \( p(Y|do(X) = x) = \int_Z p(Y|X = x, z)p(z)dz \).

The front-door adjustment applies to cases when confounders are unobserved (see Fig. 3 (Bottom)). Given a set of front-door adjustment variables \( Z \), we can again correct the estimate for the causal effect from \( X \) to \( Y \) with \( p(Y|do(X) = x) = \int_Z \int_{X'} p(Y|x', z)p(z|X = x)p(x')dx'dz \).

We rewrite the above formulae in a more general form as we show below. For the remainder of the paper we will opt for this notation:

\[
p(Y|do(X) = x) = \mathbb{E}_{\Omega_x}[p(Y|\Omega_x)] = \int p(Y|\Omega_x)p(\Omega_x)d\Omega_x \tag{1}
\]

For backdoor we have \( \Omega_x = \{X = x, Z\} \) and \( p(\Omega_x) = \delta_x p(Z) \) where \( \delta_x \) is the Dirac measure at \( X = x \). For front-door, \( \Omega_x = \{X', Z\} \) and \( p(\Omega_x) = p(X')p(Z|X = x) \).

---

\(^1\)We refer the reader to a detailed review of RKHS methods provided in the Appendix of [5]
2.2 Conditional Mean Embeddings

Kernel mean embeddings of distributions provide a powerful framework for representing probability distributions [11, 21] in an RKHS. In particular, we work with conditional mean embeddings (CMEs) in this paper. Given random variables $X, Y$ with joint distribution $P_{XY}$, the conditional mean embedding with respect to the conditional density $p(Y|X = x)$, is defined as:

$$\mu_{Y|X=x} := E_{Y|X=x}[\phi_y(Y)] = \int_Y \phi_y(y)p(y|X = x)dy$$

CMEs allow us to represent the distribution $p(Y|X = x)$ as an element $\mu_{Y|X=x}$ in the RKHS $H_{k_y}$ without having to model the densities explicitly. Following [21], CMEs can be associated with a Hilbert-Schmidt operator $C_{Y|X} : H_{k_y} \rightarrow H_{k_y}$, known as the conditional mean embedding operator, which satisfies $\mu_{Y|X=x} = C_{Y|X}\phi_x(x)$ where $C_{Y|X} := C_{YX}C_{XX}^{-1}$ with $C_{YX} := E_{Y,X}[\phi_y(Y) \otimes \phi_x(X)]$ and $C_{XX} := E_{X,X}[\phi_x(X) \otimes \phi_x(X)]$ being the covariance operators. As a result, the finite sample estimator of $C_{Y|X}$ based on the dataset $\{x, y\}$ can be written as:

$$\hat{C}_{Y|X} = \Phi_Y(K_{xx} + \lambda I)^{-1}\Phi_X^T$$

where $\lambda > 0$ is a regularization parameter. Note that from Eq.3, [22] showed that the CME can be interpreted as a vector-valued kernel ridge regressor (V-KRR) i.e. $\phi_x(x)$ is regressed to an element in $H_{k_y}$. This is crucial as CMEs allow us to turn the integration, in Eq.2, into a regression task and hence remove the need for explicit density estimation. This insight is important as it allows us to derive analytic forms for our algorithms. Furthermore, the regression formalism of CMEs motivated us to derive a Bayesian version of CME using vector-valued Gaussian Processes (V-GP), see Sec.3.

2.3 Interventional Mean Embeddings

Interventional Mean Embeddings (IME) [17] combine the above ideas to represent interventional distributions in RKHSs. We derive the front-door adjustment embedding here but the backdoor adjustment follows analogously. Denote $\mu_{Y|do(X) = x}$ as the IME corresponding to the interventional distribution $p(Y|do(X) = x)$, which can be written as:

$$\mu_{Y|do(X) = x} := \int_Y \int_Z \int_X \int_Z \int_X \phi_y(y)p(y|x', z)dydzdx$$

using the front-door formula with adjustment variable $Z$, and rearranging the integrals. By definition of CME, $\int \phi_y(y)p(y|x', z)dy = C_{Y|X,Z}(\phi_x(x') \otimes \phi_z(z))$ and linearity of integration, we have

$$= C_{Y|X,Z}(\phi_x(x')p(x')dx' \otimes \phi_z(z)p(z|x)dz) = C_{Y|X,Z}(\mu_x \otimes \mu_Z|X=x)$$

Using notations from Sec.2.1, embedding interventional distributions into an RKHS is as follows.

**Proposition 1.** Given an identifiable do-density of the form $p(Y|do(X) = x) = \mathbb{E}_{\Omega_x}[p(Y|\Omega_x)]$, the general form of the empirical interventional mean embedding is given by,

$$\hat{\mu}_{Y|do(X) = x} = \Phi_Y(K_{\Omega_x} + \lambda I)^{-1}\Phi_{\Omega_x}(x)^T$$

where $K_{\Omega_x} = K_{XX} \otimes K_{ZZ}$ and $\Phi_{\Omega_x}(x)$ is derived depending on $p(\Omega_x)$. In particular, for backdoor adjustments, $\Phi_{\Omega_x}^{(bd)}(x) = \Phi_X^T k_X(x, \cdot) \otimes \Phi_Z^T \hat{\mu}_z$ and for front-door $\Phi_{\Omega_x}^{(fd)}(x) = \Phi_X^T \mu_X \otimes \Phi_Z^T \hat{\mu}_Z|X=x$.

3 Our Proposed Method

**Two-staged Causal Learning.** Given two independent datasets $D_1 = \{(x_i, z_i, y_i)\}_{i=1}^N$ and $D_2 = \{(y_j, t_j)\}_{j=1}^M$, our goal is to model the average treatment effect in $T$ when intervening on variable $X$, i.e. model $g(x) = \mathbb{E}[T|do(X) = x]$. Note that the target variable $T$ and the treatment variable $X$ are never jointly observed. Rather, they are linked via a mediating variable $Y$ observed in both datasets.
We can thus express the average treatment effect as:

\[ g(x) = \mathbb{E}[T|do(X) = x] = \int_y \mathbb{E}[T|do(X) = x, Y = y] \ p(y|do(X) = x) \ dy \]

\[ = \int_y f(y)p(y|do(X) = x) \ dy = \langle f, \mu_{Y|do(X) = x} \rangle_{\mathcal{H}_k}. \]  

The final expression decomposes the problem of estimating \( g \) into that of estimating the IME \( \mu_{Y|do(X)} \) (which can be done using \( D_1 \)) and that of estimating the integrand \( f : \mathcal{Y} \to \mathcal{T} \) (which can be done using \( D_2 \)). Each of these two components can either be estimated using a GP or KRR approach (See Table 1). Furthermore, the reformulation as an RKHS inner product is crucial, as it circumvents the need for density estimation as well as the need for subsequent integration in Eq.6. Rather, the main parts of the task can now be viewed as two instances of regression (recall that mean embeddings can be viewed as vector-valued regression).

To model \( g \) and quantify its uncertainty, we propose 3 GP-based approaches. While the first 2 methods, Interventional Mean Process (IMP) and Bayesian Interventional Mean Embedding (BAYISEME) are novel derivations that allow us to quantify uncertainty from either one of the datasets, we treat them as intermediate yet necessary steps to derive our main algorithm, Bayesian Interventional Mean Process (BAYESIMP), which allows us to quantify uncertainty from both sources in a principled way. For a summary of the methods, see Fig.4 and Table 1. All derivations are included in the appendix.

**Interventional Mean Process**: Firstly, we train \( f \) as a GP using \( D_2 \) and model \( \mu_{Y|do(X) = x} \) as V-KRR using \( D_1 \). By drawing parallels to Bayesian quadrature [24] and conditional mean process introduced in [25], the integral of interest \( g(x) = \int y f(y)p(y|do(X) = x) \ dy \) will be a GP indexed by the treatment variable \( X \). We can then use the empirical embedding \( \mu_{Y|do(X)} \) learnt in \( D_1 \) to obtain an analytic mean and covariance of \( g \).

**Bayesian Interventional Mean Embedding**: Next, to account for the uncertainty from \( D_1 \), we model \( f \) as a KRR and \( \mu_{Y|do(X) = x} \) using a V-GP. We introduce our novel Bayesian Learning of Conditional Mean Embeddings (BAYESEME), which uses a nuclear dominant kernel [26] construction, similar to [27], to ensure that the inner product \( \langle f, \mu_{Y|do(X) = x} \rangle \) is well-defined. As the embedding is a GP, the resulting inner product is also a GP and hence takes into account the uncertainty in \( D_1 \). (See Prop. 4).

**Bayesian Interventional Mean Process**: Lastly, in order to account for uncertainties coming from both \( D_1 \) and \( D_2 \), we combine ideas from the above IMP and BAYESEME. We place GPs on both \( f \) and \( \mu_{Y|do(X)} \) and use their inner product to model \( \mathbb{E}[T|do(X)] \). Interestingly, the resulting uncertainty can be interpreted as the sum of uncertainties coming from IMP and BAYESEME with an additional interaction term (See Prop.5).

### 3.1 Interventional Mean Process

Firstly, we consider the case where \( f \) is modelled using a GP and \( \mu_{Y|do(X) = x} \) using a V-KRR. This allows us to take into account the uncertainty from \( D_2 \) by modelling the relationship between \( Y \) and

![Figure 4: Two-staged causal learning problem](image)

<table>
<thead>
<tr>
<th>METHODS</th>
<th>Stage 1</th>
<th>Stage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>IME [17]</td>
<td>KRR</td>
<td>KRR</td>
</tr>
<tr>
<td>IMP (Ours)</td>
<td>KRR</td>
<td>GP</td>
</tr>
<tr>
<td>BAYESEME (Ours)</td>
<td>GP</td>
<td>KRR</td>
</tr>
<tr>
<td>BAYESIMP (Ours)</td>
<td>GP</td>
<td>GP</td>
</tr>
</tbody>
</table>

Table 1: Summary of our proposed methods

---

We note that this two-staged setup resembles Instrumental Variable [23] (IV) regression. However, our general setup allows the IV and the treatment to be confounded, which is not the case in standard IV regression.
where we cannot define a GP well defined, we need to ensure $\mu$ to account for the uncertainty in version of CME, which then lead to $B$. In addition, the following marginal likelihood can be used for hyperparameter optimisation, mean and covariance:  

\[ \text{Proposition 3} \]

Let $\phi(x, y)$ be a GP that models $\mu_{\|Y\|}(x, y)$, and $\mu_{\|Y\|}(x, y)$ to be well defined, we need to ensure $\mu_{\|Y\|}(x, y)$ is also restricted to $H_{k_x}$ for any fixed $x$. Consequently, we cannot define a $\mathcal{GP}(0, K_x \otimes K_y)$ prior on $\mu_{\|Y\|}$ as usual, as draws from such prior will almost surely fall outside $H_{k_x} \otimes H_{k_y}$ [26]. Instead we define a prior over $\|Y\|$ that is a nuclear dominant kernel [26] over $k_{\|Y\|}$, which ensures that samples paths of $\mu_{\|Y\|}$ live in $H_{k_x} \otimes H_{k_y}$ almost surely. In particular, we follow a similar construction as in [27] and model $r_y$ as $r_y(y_i, y_j) = \int k_y(y_i, u)k_y(u, y_j)\nu(du)$ where $\nu$ is some finite measure on $Y$. Hence we can now setup a vector-valued regression in $H_{k_y}$ as follows:

\[ \phi_y(y_i) = \mu_{gp}(x_i, \cdot) + \lambda \epsilon_i \]

where $\epsilon_i \sim \mathcal{GP}(0, r)$ are independent noise functions. By taking the inner product with $\phi_y(y')$ on both sides, we then obtain $k_y(y_i, y') = \mu_{gp}(x_i, y') + \lambda \epsilon_i(y')$. Hence, we can treat $k(y_i, y_j)$ as noisy evaluations of $\mu_{gp}(x_i, y_j)$ and obtain the following posterior mean and covariance for $\mu_{gp}$.

\[ \text{Proposition 3 (BAYESCME).} \]

The posterior GP of $\mu_{gp}$ given observations $\{x, y\}$ has the following mean and covariance:

\[ m_{\mu}(x, y) = k_x(x, y) = k_{xx}(K_{xx} + \lambda I)^{-1}K_{yy}R_{yy}^{-1}r_{yy} \]

\[ \kappa_{\mu}(x, y, x', y') = k_{xx}(x, x') - k_{xx}(x, y)k_{yy}(y, y')k_{yy}(y', x') \]

In addition, the following marginal likelihood can be used for hyperparameter optimisation,

\[ -\frac{N}{2} \left( \log |K_{xx} + \lambda I| + \log |R| \right) - \frac{1}{2} \text{Tr} \left( (K_{xx} + \lambda I)^{-1}K_{yy}R_{yy}^{-1}K_{yy} \right) \]
Note that in practice we fix the lengthscale of \( k_y \) and \( r_y \) when optimising the above likelihood. This is to avoid trivial solutions for the vector-valued regression problem as discussed in [10]. The Bayesian version of the IME is derived analogously and we refer the reader to appendix due to limited space.

Finally, with V-GPs on embeddings defined, we can model \( g(x) \) as \( \langle f, \mu_{gp}(x,) \rangle_{\mathcal{H}_y} \), which due to the linearity of the inner product, is itself a GP. Here, we first considered the case where \( f \) is a KRR learnt from \( \mathcal{D}_2 \) and call the model BAYSIME.

**Proposition 4 (BAYSIME).** Given dataset \( D_1 = \{(x_i, y_i, z_i)\}_{i=1}^N \) and \( D_2 = \{(\tilde{y}_j, t_j)\}_{j=1}^M \), if \( f \) is a KRR learnt from \( \mathcal{D}_2 \) and \( \mu_{\mathcal{Y}|do(X)} \) modelled as a V-GP using \( \mathcal{D}_1 \), then \( g = \langle f, \mu_{\mathcal{Y}|do(X)} \rangle \sim \mathcal{GP}(m_2, \kappa_2) \) where,

\[
\begin{align*}
m_2(x) &= \Phi_{\Omega_x}(x)^T (K_{\Omega_x} + \lambda I)^{-1} K_{yy} R_{yy}^{-1} y \quad \text{(14)} \\
\kappa_2(x, x') &= B \Phi_{\Omega_x}(x)^T \Phi_{\Omega_x}(x) - C \Phi_{\Omega_x}(x)^T (K_{\Omega_x} + \lambda I)^{-1} \Phi_{\Omega_x}(x') \quad \text{(15)}
\end{align*}
\]

where \( A = (K_{xx} + \lambda I)^{-1} t \), \( B = A^T R_{yy} A \) and \( C = A^T R_{yy} R_{yy}^T R_{yy} A \).

**Summary:** Constants \( B \) and \( C \) in \( \kappa_2 \) can be interpreted as different estimation of \( \| f \|_{\mathcal{H}_y} \), i.e the RKHS norm of \( f \). As a result, for problems that are “harder” to learn in \( \mathcal{D}_2 \), i.e. corresponding to larger magnitude of \( \| f \|_{\mathcal{H}_y} \), will result into larger values of \( B \) and \( C \). Therefore the covariance \( \kappa_2 \) can be interpreted as uncertainty in \( \mathcal{D}_1 \) scaled by the difficulty of the problem to learn in \( \mathcal{D}_2 \).

### 3.3 Bayesian Interventional Mean Process

To incorporate both uncertainties in \( \mathcal{D}_1 \) and \( \mathcal{D}_2 \), we combine ideas from IMP and BAYSIME to estimate \( g = \langle f, \mu_{\mathcal{Y}|do(X)} \rangle \) by placing GPs on both \( f \) and \( \mu_{\mathcal{Y}|do(X)} \). Again as before, a nuclear dominant kernel \( r_y \) was used to ensure the GP \( f \) is supported on \( \mathcal{H}_y \). For ease of computation, we consider a finite dimensional approximation of the GPs \( f \) and \( \mu_{\mathcal{Y}|do(X)} \) and estimate \( g \) as the RKHS inner product between them. In the following we collate \( y \) and \( \hat{y} \) into a single set of points \( \tilde{y} \), which can be seen as landmark points for the finite approximation [29]. We justify this in the Appendix.

**Proposition 5 (BAYSIMP).** Let \( f \) and \( \mu_{\mathcal{Y}|do(X)} \) be GPs learnt as above. Denote \( \hat{f} \) and \( \hat{\mu}_{\mathcal{Y}|do(X)} \) as the finite dimensional approximation of \( f \) and \( \mu_{\mathcal{Y}|do(X)} \) respectively. Then \( \hat{g} = \langle \hat{f}, \hat{\mu}_{\mathcal{Y}|do(X)} \rangle \) has the following mean and covariance:

\[
\begin{align*}
m_3(x) &= E_x K_{xx} K_{yy}^{-1} \tilde{R}_{yy} (R_{yy} + \lambda I)^{-1} t \quad \text{(16)} \\
\kappa_3(x, x') &= E_x \Theta^{(a)}_{\Omega_x} \tilde{R}_{yy} \Theta_{\Omega_x}^T + \Theta^{(b)}_{\Omega_x} G_{xx'} + \Theta^{(a)}_{\Omega_y} F_{xx'} - \Theta^{(b)}_{\Omega_y} G_{xx'} \quad \text{(17)}
\end{align*}
\]

where \( E_x = \Phi_{\Omega_x}(x)^T (K_{\Omega_x} + \lambda I)^{-1} \), \( F_{xx'} = \Phi_{\Omega_x}(x)^T \Phi_{\Omega_x}(x') \), \( G_{xx'} = \Phi_{\Omega_x}(x)^T (K_{\Omega_x} + \lambda I)^{-1} \Phi_{\Omega_x}(x') \), and \( \Theta_{\Omega_x}^{(a)} = \Theta_{\Omega_x}^{(b)} \) are the uncertainy from \( \mathcal{D}_1 \). \( \Theta_{\Omega_y}^{(a)} = \Theta_{\Omega_y}^{(b)} \) are the uncertainty from \( \mathcal{D}_2 \). \( \Theta_{\Omega_y}^{(a)} = \Theta_{\Omega_y}^{(b)} \) are the uncertainty from interaction.

\( \Phi_{\Omega_y}^{(a)} = tr(K_{yy}^{-1} \tilde{R}_{yy} K_{yy}^{-1} \tilde{R}_{yy}), \Theta_{\Omega_y}^{(b)} = tr(R_{yy}^{-1} R_{yy} K_{yy}^{-1}) \) and \( \Theta_4 = K_{yy}^{-1} \tilde{R}_{yy} (R_{yy} + \lambda I)^{-1} t \). \( \tilde{R}_{yy} \) is the posterior covariance of \( f \) evaluated at \( \tilde{y} \).

**Summary:** While the first two terms in \( \kappa_3 \) resemble the uncertainty estimates from IMP and BAYSIME, the last term acts as an extra interaction between the two uncertainties from \( \mathcal{D}_1 \) and \( \mathcal{D}_2 \). We note that unlike IMP and BAYSIME, \( \hat{g} \) from Prop.5 is not a GP as inner products between Gaussian vectors are not Gaussian. Nonetheless, the mean and covariance can be estimated.

### 4 Experiments

In this section, we first present an ablation studies on how our methods would perform under settings where we have missing data parts at different regions of the two datasets. We then demonstrate BAYSIMP’s proficiency in the Causal Bayesian Optimisation setting.

In particular, we compare our methods against the sampling approach considered in [15]. [15] start by modelling \( f : Y \to T \) as GP and estimate the density \( p(Y|do(X)) \) using a GP along with do-calculus. Then given a treatment \( x \), we obtain \( L \) samples of \( y_l \) and \( R \) samples of \( f_t \) from their posterior GPs.
The empirical mean and standard deviation of the samples \( \{ f_i(y_i) \}_{i=1,r=1}^{L,R} \) can now be taken to estimate \( \mathbb{E}[T|do(X) = x] \) as well as the correspondingly uncertainty. We emphasize that this point estimation requires repeated sampling and is thus inefficient compared to our approaches, where we explicitly model the uncertainty as covariance function.

Ablation study. In order to get a better intuition into our methods, we will start off with a preliminary Ablation study. The empirical mean and standard deviation of the samples \( \{ f_i(y_i) \}_{i=1,r=1}^{L,R} \) can now be taken to estimate \( \mathbb{E}[T|do(X) = x] \) as well as the correspondingly uncertainty. We emphasize that this point estimation requires repeated sampling and is thus inefficient compared to our approaches, where we explicitly model the uncertainty as covariance function.

BayesIMP for Bayesian Optimisation (BO). We now demonstrate, on both synthetic and real-world data, the usefulness of the uncertainty estimates obtained using our methods in BO tasks. Our goal is to utilise the uncertainty estimates to direct the search for the optimal value of \( \mathbb{E}[T|do(X) = x] \) by querying as few values of the treatment variable \( X \) as possible, i.e. we want to optimize for \( x^* = \arg \min_{x \in X} \mathbb{E}[T|do(X) = x] \). For the first synthetic experiment (see Fig.6 (Top)), we will use the following two datasets: \( D_1 = \{ x_i, u_i, z_i, y_i \}_{i=1}^N \) and \( D_2 = \{ y_j, t_j \}_{j=1}^M \). Note that BayesIMP from Prop.5 is not a GP as inner products between Gaussian vectors are not Gaussian. Nonetheless, with the mean and covariance estimated, we will use moment matching to construct a GP out of BayesIMP for posterior inference. At the start, we are given \( D_1 \) and \( D_2 \), where these observations are used to construct a GP prior for the interventional effect of \( X \) on \( T \), i.e \( \mathbb{E}[T|do(X) = x] \), to provide a “warm” start for the BO.

Again we compare BayesIMP with the sampling-based estimation of \( \mathbb{E}[T|do(X)] \) and its uncertainty, which is exactly the formulation used in the Causal Bayesian Optimisation algorithm (CBO) [15]. In order to demonstrate how BayesIMP performs in the multimodal setting, we will be considering the case where we have the following distribution on \( Y \) i.e. \( p(y|u,z) = \pi p_1(y|u,z) + (1 - \pi)p_2(y|u,z) \) where \( Y \) is a mixture and \( \pi \in [0,1] \). These scenarios might arise when there is an unobserved binary variable which induces a switching between two regimes on how \( Y \) depends on \( (U,Z) \). In this case, the GP model of [15] would only capture the conditional expecta-

Figure 5: Ablation studies of various methods in estimating uncertainties for an illustrative experiment. * indicates our methods. \( N = M = 100 \) data points are used. Uncertainty from sampling gives a uniform estimate of uncertainty and 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.

From Fig.5(a), we see that the uncertainty for sampling is rather uniform across the ranges of \( x \) despite the fact we have more data around \( x = 0 \). This is contrary to our methods, which show a reduction of uncertainty at high \( x \) density regions. In particular, \( x = -5 \) corresponds to an extrapolation of data, where \( x \) gets mapped to a region of \( y \) where there is no data in \( D_2 \). This fact is nicely captured by the spike of credible interval in Fig.5(c) since IMP utilises uncertainty from \( D_2 \) directly. Nonetheless, IMP failed to capture the uncertainty stemming from \( D_1 \), as seen from the fact that the credible interval did not increase as we have less data in the region \( |x| > 5 \). In contrast, BayesIME (Fig.5(d)) gives higher uncertainty around low \( x \) density regions but failed to capture the extrapolation phenomenon. Finally, BayesIMP Fig.5(e) seems to inherit desirable characteristics from both IMP and BayesIME, due to taking into account uncertainties from both \( D_1, D_2 \). Hence, in the our experiments, we focus on BayesIMP and refer the reader to the appendix for the remaining methods.

BayesIMP for Bayesian Optimisation (BO). We now demonstrate, on both synthetic and real-world data, the usefulness of the uncertainty estimates obtained using our methods in BO tasks. Our goal is to utilise the uncertainty estimates to direct the search for the optimal value of \( \mathbb{E}[T|do(X) = x] \) by querying as few values of the treatment variable \( X \) as possible, i.e. we want to optimize for \( x^* = \arg \min_{x \in X} \mathbb{E}[T|do(X) = x] \). For the first synthetic experiment (see Fig.6 (Top)), we will use the following two datasets: \( D_1 = \{ x_i, u_i, z_i, y_i \}_{i=1}^N \) and \( D_2 = \{ y_j, t_j \}_{j=1}^M \). Note that BayesIMP from Prop.5 is not a GP as inner products between Gaussian vectors are not Gaussian. Nonetheless, with the mean and covariance estimated, we will use moment matching to construct a GP out of BayesIMP for posterior inference. At the start, we are given \( D_1 \) and \( D_2 \), where these observations are used to construct a GP prior for the interventional effect of \( X \) on \( T \), i.e \( \mathbb{E}[T|do(X) = x] \), to provide a “warm” start for the BO.

Figure 6: Illustration of synthetic data experiments.
tion of $Y$ with an inflated variance, leading to slower convergence and higher variance in the estimates of the prior as we will show in our experiments. Throughout our experiments, similarly to [15], we will be using the expected improvement (EI) acquisition function to select the next point to query.

**Synthetic data experiments.** We compare BayesIMP to CBO as well as to a simple GP with no learnt prior as baseline. We will be using $N = 100$ datapoints for $D_1$ and $M = 50$ datapoints $D_2$. We ran each method 10 times and plot the resulting standard deviation for each iteration in the figures below. The data generation and details were added in the Appendix. We see from the Fig.7 that

BayesIMP is able to find the maxima much faster and with smaller standard deviations, than the current state-of-the-art method, CBO, using both front-door and backdoor adjustments (Fig.7(Right, Middle)). Given that our method uses more flexible representations of conditional distributions, we are able to circumvent the multimodality problem in $Y$. In addition, we also consider the unimodal version, i.e. $\pi = 0$ (see right Fig.7). We see that the performance of CBO improves in the unimodal setting, however BayesIMP still converges faster than CBO even in this scenario.

Next, we consider a harder causal graph (see Fig.6 (Bottom)), previously considered in [15]. We again introduce multimodality in the $Y$ variable in order to explore the case of more challenging conditional densities. We see from Fig.8 (Left, Middle), that BayesIMP again converges much faster to the true optima than CBO [15] and the standard GP prior baseline. We note that the fast convergence of BayesIMP throughout our experiments is not due to simplicity of the underlying BO problems. Indeed, the BO with a standard GP prior requires significantly more iterations. It is rather the availability of the observational data, allowing us to construct a more appropriate prior, which leads to a “warm” start of the BO procedure.

**Healthcare experiments.** We conclude with a healthcare dataset corresponding to our motivating medical example in Fig.1. The causal mechanism graph, also considered in the CBO paper [15], studies the effect of certain drugs (Aspirin/Statin) on Prostate-Specific Antigen (PSA) levels [6]. In our case, we modify statin to be continuous, in order to optimize for the correct drug dosage. However, in contrast to [15], we consider a second experimental dataset, arising from a different medical study, which looks into the connection between PSA levels and cancer volume amount in patients [7]. Similar to the original CBO paper [15], given that interventional data is hard to obtain, we construct data generators based on the true data collected in [7]. This is done by firstly fitting a GP on the data and then sampling from the posterior (see Appendix for more details). Hence this is the
perfect testbed for our model where we are interested in $\mathbb{E}[\text{Cancer Volume}|do(\text{Statin})]$. We see from Fig.8 (Right) that BAYESIMP again converges to the true optima faster than CBO hence allowing us to find the fastest ways of optimizing cancer volume by requesting much less interventional data. This could be critical as interventional data in real-life situations can be very expensive to obtain.

5 Discussion and Conclusion

In this paper we propose BAYESIMP for quantifying uncertainty in the setting of causal data fusion. In particular, our proposed method BAYESIMP allows us to represent interventional densities in the RKHS without explicit density estimation, while still accounting for epistemic and aleatoric uncertainties. We demonstrated the quality of the uncertainty estimates in a variety of Bayesian optimization experiments, in both synthetic and real-world healthcare datasets, and achieve significant improvement over current SOTA in terms of convergence speed. However, we emphasize that BAYESIMP is not designed to replace CBO but rather an alternative model for interventional effects.

In the future, we would like to improve BAYESIMP over several limitations. As in [15], we assumed full knowledge of the underlying causal graph, which might be limiting in practice. Furthermore, as the current formulation of BAYESIMP only allows combination of two causal graphs, we hope to generalise the algorithm into arbitrary number of graphs in the future. Causal graphs with recurrent structure will be an interesting direction to explore. Lastly, we would also like to include a cost function as in [15] to constrain the search space to the most sensible solutions.

6 Acknowledgements

The authors would like to thank Bobby He, Robert Hu, Kaspar Martens, Jake Fawkes and Joost van Amersfoort for helpful comments. SLC and JFT are supported by the EPSRC and MRC through the OxWaSP CDT programme EP/L016710/1. YWT and DS are supported in part by Tencent AI Lab and DS is supported in part by the Alan Turing Institute (EP/N510129/1). YWT’s research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section Discussion
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See below
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Appendix
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Supp material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Experiment section
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Experiment section and Appendix
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Societal Impact

In this paper we propose a general framework for embedding interventional distributions into a RKHS while accounting for uncertainty in causal datasets. We believe that this is a crucial problem that has not gotten much attention yet, but is nonetheless important for the future of causal inference research. In particular, given that our proposed method allows us to combine datasets from different studies, we envision that this could potentially be used in a variety of scientific areas such as healthcare, drug discovery etc. Finally, the only potential negative impact, would be, when using biased data. Our method relies on knowing the correct causal graph and hence could be misinterpreted when this is not the case.