From LATE to ATE: A Bayesian Approach[*]

Isaac M. Opper[†]

June 18, 2024

Abstract

We develop a Bayesian model that produces a posterior distribution of the marginal treatment effect (MTE) function. The method provides researchers with a principled way to extrapolate from the observed moments using flexible assumptions, thereby allowing researchers to generate plausible ranges of important and potentially policy-relevant quantities of interest. We then use the model to propose a natural decomposition of the posterior variance into "statistical uncertainty," i.e., variance that stems from the imprecise estimation of the observed moments, and "extrapolation uncertainty," i.e., variance that stems from uncertainty in how to extrapolate away from the observed moments. We conclude by showing that under our preferred priors, even in an experiment as large as the Oregon Health Insurance Experiment, the main source of uncertainty in the ATE comes from uncertainty in the true values of the observed moments.

Keywords: marginal treatment effects; Bayesian models; instrumental variables; compliers; Gaussian process; RCTs

JEL Classification: C11; C26

[*] I am very thankful to Aarya Suryavanshi, who took the lead in developing the R package, and to Joshua Eagan who helped QA the code. An R package to implement the method and replicate the empirical results is available at https://github.com/isaacopper/BayesianMTEs. I'm also thankful to Stephanie Rennane and Jessie Cox, who both provided valuable feedback on an early draft, as well as Lisa Jaycox for all her help, support, and patience during my particularly challenging – and not particularly on-time – RIR. Finally, the comments I have received from two anonymous referees, the AE, and editor truly transformed the paper and I am grateful for their careful read and thoughtful comments. Funding for this research was provided by gifts from RAND supporters and income from operations.

[†] RAND Corporation. Email: iopper@rand.org.

I Introduction

One of the most commonly used econometric techniques is the instrumental variables (IV) design. While it is an incredibly powerful technique, a downside of the IV design is that the resulting treatment effects are only valid locally, resulting in what is commonly referred to as the local average treatment effect (LATE) (Imbens and Angrist, 1994). In an RCT with imperfect compliance, for example, that means that the resulting estimates represent the average treatment effect on the set of individuals whose treatment status is impacted by instrument, i.e., who enroll in the treatment if and only if they are assigned to it. While the LATE is often an important parameter of interest, it is rarely the only parameter of interest (Heckman and Vytlacil, 2001). In the RCT setting, for example, it is also valuable to know what the effect of the treatment was on individuals who enrolled regardless of their treatment assignment or to know the average effect on the entire population.

Unfortunately, these other parameters can generally be estimated only by extrapolating away from the observed moments. This necessity has led to a burgeoning line of research that aims to better understand how this extrapolation occurs and what assumptions about this extrapolation are needed to estimate various parameters of interest (e.g., Brinch et al. (2017); Kline and Walters (2019); Kowalski (2023a)). While such research has been incredibly valuable, approaches that rely on such parametric assumptions suffer from two related shortcomings. First, it is not immediately clear how much the resulting estimates and, especially, uncertainty in the resulting estimates stem from the observed data and how much depend on the extrapolation assumption. Second, we may be hesitant to make the necessary parametric assumptions, but comfortable with more flexible assumptions; for example, we may be uncomfortable assuming that the marginal treatment effects are linear, but willing to assume that they are (in some sense) “close to linear.”

Motivated by these shortcomings, we develop a Bayesian model in this paper (Meager (2019, 2022)). Like many existing approaches, we start with a generalized Roy model in which individuals have different propensities to self-select into treatment. Key to this model are two functions: one that illustrates how individuals’ untreated outcome varies with their implied cost of enrolling in the treatment and the other which illustrates how the treatment’s effect on individuals’ outcomes varies with their implied cost of enrolling. The second of these functions is usually referred to as the marginal treatment effect (MTE) function and from which one can construct most estimands of interest (Heckman and Vytlacil, 2007a,b). Rather than viewing these functions as fixed, however, we assume that the functions themselves are generated probabilistically, specifically via a Gaussian process.

We then show that the generalized Roy model, the Gaussian process, and the observed moments can be combined in a straightforward way to output a posterior distribution of the MTE function, and therefore of most estimands of interest. The model can be thought of as nesting the parametric approaches, in that by appropriately restricting the set of potential functions we can arrive at equivalent estimates. It also provides an alternative way to estimate bounds on the estimands of interest (e.g., Manski (1990); Balke and Pearl (1997); Bhattacharya et al. (2008); Mogstad et al. (2018); Kowalski (2023b)). Instead of starting with sharp bounds on the set of permissible functions, we can, for example, capture the belief – prevalent in data science and machine learning – that smooth functions are more likely than functions which oscillate wildly and from there compute the range of plausible treatment effects.

Furthermore, the Bayesian model is particularly useful in that allows us to easily and explicitly quantify how uncertain the estimates are for any parameter of interest – even those that require extrapolation such as the average treatment effect (ATE), always taker average treatment effect (ATATE), and never taker average treatment effect (NTATE) – in a way that accounts for both uncertainty in the unobserved moments and uncertainty in how one should extrapolate away from the observed moments. The posterior variance captures both sources of uncertainty and does not require distinguishing between estimands that can be estimated using the observed moments and those that require extrapolation. To better understand where the uncertainty stems from, we show the model permits a natural decomposition of the posterior variance into variance due to "statistical uncertainty," i.e., variance that stems from the imprecise estimation of the observed moments, and variance due to "extrapolation uncertainty," i.e., variance that stems from uncertainty in how to extrapolate away from the observed moments.

We then turn to questions of implementation. After illustrating how the hyperparameters of the model govern the distribution of potential MTE functions and specifying our preferred hyperprior, i.e., prior distribution over these hyperparameters, we use data from the Oregon Health Insurance Experiment (OHIE) to explore how the model works in practice. We first show that our specification of the hyperprior is important, since there are a wide range of hyperparameters consistent with the observed data (at least in cases like the OHIE where there is a single binary instrument). In contrast, the observed moments provide important information about the treatment effects; for example, the posterior variance of the ATE is 93% smaller than the prior variance of the ATE. To better understand where the remaining uncertainty stems from, we use the proposed decomposition of overall uncertainty into extrapolation and statistical uncertainty. In doing so, we show that under our preferred hyperprior the extrapolation uncertainty is much less important than the statistical uncertainty, even in an RCT as large as the OHIE and one in which a large fraction of the population are never-takers.

II Defining the Bayesian Hierarchical Model

We start by describing the two building blocks of our approach: the generalized Roy model and Gaussian processes. Both have been studied extensively and the purpose of Section II.A and II.B is to ensure the reader starts with the necessary background and to clarify our notation, rather than to introduce any new ideas. For the interested reader, see Heckman (2010) (among others) for more discussion about the generalized Roy model and Rasmussen and Williams (2006) for more details about Gaussian processes. We then discuss how the two building blocks can be naturally combined into a Bayesian hierarchical model and how the resulting Bayesian hierarchical model relates to other commonly used models.

II.A Generalized Roy Model

We consider the effect of a binary treatment on a single outcome. We assume that each individual is defined by three latent variables: their outcome if they are not treated, the effect that the treatment has on their outcome, and their implied cost of enrolling in the treatment; we denote these as μ_i, τ_i, η_i, respectively. In other words, we use μ_i to denote individual i's outcome in the absence of treatment and τ_i to denote the causal effect of the treatment on individual i's outcome; clearly $\mu_i + \tau_i$ is then their outcome if they are treated.

The researcher does not observe these three latent variables and instead observes each individuals' outcome, treatment status, and an instrument; we denote these as,

Y_i, T_i, and Z_i, respectively. For simplicity, we focus here on the case where there are no additional covariates, although the model can be extended to include these; see Section C in the Appendix for a discussion of how to do so. Given T_i, we can write the observed outcome as a function of the latent variables without further assumptions as follows: $Y_i = \mu_i + \tau_i T_i$. The restrictions to the model appear in how we relate the latent variables to the treatment status. To do so, we assume that we can relate η_i and Z_i to treatment status via a threshold-crossing representation, i.e.,:

$$T_i = \mathbf{1}\big(\nu(Z_i) \geq \eta_i\big) \tag{1}$$

for some (unknown) function of the instrument $\nu(Z_i)$ As is common, we will assume that η_i is continuously distributed, which means without loss of generality we can normalize this distribution to be uniform between zero and one. Note that individuals with higher η_i are less likely to enroll in the treatment, i.e., have a higher implied cost of enrolling in the treatment. In doing so, it then follows that $\nu(Z_i) = \mathbb{E}[T_i|Z_i]$, which hints at how we can estimate ν. Finally, we will assume that $(\mu_i, \tau_i, \eta_i) \perp\!\!\!\perp Z_i$. This assumption captures the idea that Z_i is a valid instrument, in that it only affects the outcomes by affecting the likelihood that an individual is treated.[1]

We then define the following two conditional moments, which we assume exist and serve as the objects of interest:

$$\tau(\eta) = \mathbb{E}[\tau_i|\eta_i = \eta] \qquad \text{and} \qquad \mu(\eta) = \mathbb{E}[\mu_i|\eta_i = \eta] \tag{2}$$

The first function $\tau(\eta)$, in particular, is the marginal treatment effect (MTE) as defined in Heckman and Vytlacil (1999, 2005) and others. Again, implicit in these definitions is the IV assumption, that once we condition on η, we do not need to condition on Z. More specifically, the assumption is that $\mathbb{E}[\tau_i|\eta_i = \eta] = \mathbb{E}[\tau_i|\eta_i = \eta, Z_i = Z]$ for all Z, with a similar expression for μ_i.

In addition, we define two additional conditional moments as follows:

$$y_0(\eta) = \mathbb{E}[\mu_i|\eta_i > \eta] \qquad \text{and} \qquad y_1(\eta) = \mathbb{E}[\mu_i + \tau_i|\eta_i \leq \eta] \tag{3}$$

[1] We do not explicitly assume that Z_i is related to T_i, or in our model that $\nu(Z_i)$ varies with Z_i, which is often included as the second condition that Z_i is a valid instrument. This is because the proposed method will be valid even in this case, it is just uninteresting since the resulting lack of variation makes the posterior generally uninformative.

Defining four sets of conditional moments is redundant, in that the functions defined in Equation (2) would imply the values of the functions in Equation (3) and vice versa. Most useful for our purposes, we can relate y_0 and y_1 to τ as follows:

$$\mathbb{E}[\tau(\eta)|\eta \in (\underline{\eta}, \overline{\eta})] = \frac{1}{\overline{\eta} - \underline{\eta}} \cdot \begin{bmatrix} \overline{\eta} \\ -\underline{\eta} \\ 1 - \overline{\eta} \\ -(1 - \underline{\eta}) \end{bmatrix}' \cdot \begin{bmatrix} y_1(\overline{\eta}) \\ y_1(\underline{\eta}) \\ y_0(\overline{\eta}) \\ y_0(\underline{\eta}) \end{bmatrix} \tag{4}$$

Note that while $y_0(\eta)$ conditions on η_i being larger than η, $y_1(\eta)$ conditions on η_i being smaller than η. The reasons for this peculiar conditioning is that it makes y_0 and y_1 more directly reflect the values we actually observe in the data, which – as we will explain below – will be quite useful. In particular, we observed estimates $\hat{y}_k(\eta)$ of the true moments $y_k(\eta)$ for every $\eta \in Range(\nu)$.

II.B Gaussian Process Prior

While there are multiple ways to define a Gaussian process, for our purposes the most useful definition is taken from Rasmussen and Williams (2006):

Definition (Rasmussen and Williams, 2006) 1. *A* Gaussian process *(GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

To highlight how a GP is useful in our context, let's restrict our attention to the MTE function $\tau(\eta)$ and denote $\mathcal{T}$ as the space of potential MTE functions. It is intuitive to imagine placing a prior distribution over the various functions $\tau \in \mathcal{T}$, which governs how likely any one function is to be drawn at random from $\mathcal{T}$. Any function τ is simply defined by its value at every point η in its domain, and so an equivalent formulation is to consider each point $\tau(\eta)$ as its own random variable; one draw of $\tau \in \mathcal{T}$ is therefore equivalent to one draw from an infinite number of (potentially correlated) random variables $\tau(\eta)$. A consequence of this is that defining how the random variables $\tau(\eta)$ themselves co-vary is an alternative way of defining how likely it is to draw a particular function τ, i.e., to define our prior distribution of functions.

The GP is a common way to specify the covariance and hence the prior distribution over functions. We can define the GP prior by the mean and covariance function.

For example, consider some GP that models the relationship between x_i and y_i and then let $m(x_i)$ and $k\big(x_i, x_i'\big)$ be the mean and covariance functions, respectively. One of the main advantages of a GP is that with a GP it is quite easy to transition from the prior distribution over functions, defined implicitly by $m(x_i)$ and $k\big(x_i, x_i'\big)$, to the posterior distribution over functions after one conditions on a set of observations. For example, suppose we observe a single observation (y_i, x_i) and assume for now that there is no additional error term so $y_i = f(x_i)$ for some function f. Then from the definition of a GP, it clearly follows that:

$$f(x')\big|y_i, x_i = N\Big(\mu, \sigma^2\Big) \quad \text{with}$$
$$\mu = m(x') + \frac{k(x_i, x')}{k(x_i, x_i)}\big(f(x_i) - m(x_i)\big) \quad \text{and} \quad \sigma^2 = k(x', x') - \frac{k(x_i, x')^2}{k(x_i, x_i)}$$

for any x'. We can write a similar expression if we observe multiple observations and/or want to generate posterior predictions at multiple points on the domain of f.

II.C Bayesian Hierarchical Model

We now combine the two building blocks to specify the full Bayesian hierarchical model. To do so, we add to the Generalized Roy model the assumption that the functions $\tau(\eta)$ and $\mu(\eta)$ themselves follow a Gaussian process. Specifically, we have that:

$$\text{Gaussian process for } \mu: \qquad \mu(\eta)|\theta_\mu \sim \mathcal{GP}\big(0, k_\mu(\eta, \eta'|\theta_\mu)\big) \tag{5}$$
$$\text{Gaussian process for } \tau: \qquad \tau(\eta)|\theta_\tau \sim \mathcal{GP}\big(0, k_\tau(\eta, \eta'|\theta_\tau)\big) \tag{6}$$

for known covariance functions (or kernels) $k_\mu(\eta, \eta'|\theta_\mu)$ and $k_\tau(\eta, \eta'|\theta_\tau)$ with hyperparameters θ_μ and θ_τ.[2]

In the next sections, we discuss the choice of the covariance functions and hy-

[2] We specify that the mean of both Gaussian processes is zero to reflect the fact that – absent the data – we usually do not want to encode a strong prior about the expected sign of the functions. As discussed in Chapter 2 of Rasmussen and Williams (2006), this specification is actually quite flexible. If, for example, we prefer that the Gaussian process for $\mu(\eta)$ is centered at a mean of $\overline{\mu}(\eta)$ for some known function, then we can just add that value to the posterior derived in the next section. If, instead, we want to specify that the Gaussian process for $\mu(\eta)$ is centered at a mean of $\overline{\mu}(\eta)$ for some function that depends on a set of unknown parameters, we can generally reformulate that as being a Gaussian process centered at zero by adjusting the covariance term.

perparameters, but for now will take those as given and only make the assumption that the covariance functions imply that the process is sample-continuous, i.e., that every realization of $\mu(\eta)$ and $\tau(\eta)$ results in continuous functions. Rather than make the weakest assumption possible, we opt for a more readily interpretable sufficient condition regarding the covariance functions, stated below:

Assumption 1. *The covariance functions $k_\mu(\eta, \eta'|\theta_\mu)$ and $k_\tau(\eta, \eta'|\theta_\tau)$ are both Lipschitz continuous functions.*

We further assume that the outcomes also contain a normally distributed error term. Specifically, defining the error term $\epsilon_i = Y_i - \mu(\eta) - \tau(\eta)T_i$ and letting ϵ be the vector of all individuals' error terms, we assume that $\epsilon \sim N(0, \Sigma)$ for some positive semi-definite matrix Σ. It is common to assume that the error term is distributed i.i.d., in which case $\Sigma = \sigma_\epsilon^2 \mathbf{I}$ where $\mathbf{I}$ is the identity matrix; however, we keep this more general form to highlight how the method can be used in cases where the errors are not all independent, as would be the case in cluster randomized trials, for example. Finally, we note that the assumption that the error terms are distributed normally can be relaxed if one prefers take the asymptotic perspective and apply the central limit theorem to infer that the observed moments are approximately normal.

III Implications of the Bayesian Hierarchical Model

Having fully specified the model in the above section, we now turn to an analysis of the general model. We first discuss how it can be used to create posteriors of the MTE function and other estimands of interest, and then highlight how the model allows one to decompose the uncertainty in the resulting estimates into uncertainty that is due to imprecise estimation of the observed moments and uncertainty that is due to the required extrapolation away from these observed moments.

III.A Bayesian Posterior of the MTE

To generate posterior predictions we have to grapple with the fact that we do not directly observe $\tau(\eta)$ or $\mu(\eta)$ at any point η. Instead, we observe the functions $y_0(\eta)$ and $y_1(\eta)$, and only do so at points η in the image of $\nu(Z_i)$. Luckily, if $\tau(\eta)$ and $\mu(\eta)$ are both GPs, then together $y_0(\eta)$ and $y_1(\eta)$ form one large Gaussian process.

Formally, define $y_1(\eta)$ and $y_0(\eta)$ as in Equation (3) and let:

$$\tilde{y}(t,\eta) = ty_1(\eta) + (1-t)y_0(\eta) \tag{7}$$

for $t \in \{0,1\}$. We then have the following remark, which we prove in Appendix A.

Remark 1. *Under the the Bayesian hierarchical model defined in Section II and Assumption 1, $\tilde{y}(t,\eta)$ as defined in Equation (7) also follows a mean-zero Gaussian process with a covariance function – denoted $k_{\tilde{y}}$ – that depends on $k_\mu(\eta,\eta'|\theta_\mu)$ and $k_\tau(\eta,\eta'|\theta_\tau)$. In particular, we have that:*

$$k_{\tilde{y}}\big((t,\eta),(t',\eta')|\theta_\mu,\theta_\tau\big) = \begin{cases} \mathbb{E}\big[k_\mu(\tilde{\eta},\tilde{\eta}'|\theta_\mu) + k_\tau(\tilde{\eta},\hat{\eta}|\theta_\tau)|\tilde{\eta} \leq \eta, \tilde{\eta}' \leq \eta'\big] \ \textit{if}\ t = t' = 1 \\ \mathbb{E}\big[k_\mu(\tilde{\eta},\tilde{\eta}'|\theta_\mu)|\tilde{\eta} > \eta, \tilde{\eta}' > \eta'\big] \ \textit{if}\ t = t' = 0 \\ \mathbb{E}\big[k_\mu(\tilde{\eta},\tilde{\eta}'|\theta_\mu)|\tilde{\eta} > \eta, \tilde{\eta}' \leq \eta'\big] \ \textit{if}\ t = 0 \neq t' \\ \mathbb{E}\big[k_\mu(\tilde{\eta},\tilde{\eta}'|\theta_\mu)|\tilde{\eta} \leq \eta, \tilde{\eta}' > \eta'\big] \ \textit{if}\ t = 1 \neq t' \end{cases} \tag{8}$$

As mentioned, the fact that $\tilde{y}(t,\eta)$ is generated via a Gaussian process is helpful because that corresponds directly to what is observed by the researcher. To more fully highlight why this is helpful, denote Y^{obs} to be the vector of observed outcomes and consider a finite vector $\tilde{Y}'$ of potentially observed conditional moments. For example, we could have $\tilde{Y}$ be:

$$\tilde{Y}' = [y_0(0), y_0(\Delta), ..., y_0(1-\Delta), y_0(1), y_1(0), y_1(\Delta), ..., y_1(1-\Delta), y_1(1)] \tag{9}$$

for some arbitrarily small step size Δ. From Remark 1, we get that:

$$\tilde{Y} \sim N\big(0, K_{\tilde{Y}}\big) \tag{10}$$

where $K_{\tilde{Y}}$ is a matrix that can be inferred by the covariance functions $k_\mu(\eta,\eta'|\theta_\mu)$ and $k_\tau(\eta,\eta'|\theta_\tau)$. In particular, if the i^{th} row of $\tilde{Y}$ is $y_t(\eta)$ and the j^{th} row of $\tilde{Y}$ is $y_{t'}(\eta')$, the the $(i,j)^{th}$ element of $K_{\tilde{Y}}$ is $k_{\tilde{y}}\big((t,\eta),(t',\eta')\big)$. We also get that

$$Y^{obs} \sim N\big(0, K_{Y^{obs}} + \Sigma\big) \tag{11}$$

where $K_{Y^{obs}}$ is defined similarly to $K_{\tilde{Y}}$ and Σ is the variance of the error term (as defined in Section II). Finally, define $K_{\tilde{Y},Y^{obs}}$ such that if i^{th} row of $\tilde{Y}$ is $y_t(\eta)$ then the $(i,j)^{th}$ element of $K_{\tilde{Y},Y^{obs}}$ is $k_{\tilde{y}}\big((t,\eta),(T_j,\nu(Z_j))\big)$.

Given these matrices, we can derive the Bayesian posterior of $\tilde{Y}$ as stated below:

Remark 2. *The Bayesian posterior of $\tilde{Y}$ given the observed data:*

$$\tilde{Y}|Y^{obs} \sim N\big(\mu_{\tilde{Y}}, \Sigma_{\tilde{Y}}\big) \tag{12}$$

where

$$\mu_{\tilde{Y}} = K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}} + \Sigma\big)^{-1}Y^{obs} \tag{13}$$

$$\Sigma_{\tilde{Y}} = K_{\tilde{Y}} - K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}} + \Sigma\big)^{-1}K'_{\tilde{Y},Y^{obs}} \tag{14}$$

One nice thing about this expression is that posterior described above gives a closed-form solution to quantify how much is learned from the data (as opposed to reflecting the prior). In particular, one way to summarize this is through $Var(\tilde{Y}) - Var(\tilde{Y}|Y^{obs})$, i.e., how much posterior variance shrinks relative to the prior variance, which from above can be expressed as $K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}} + \Sigma\big)^{-1}K'_{\tilde{Y},Y^{obs}}$.

Finally, note that although the expressions above give the posterior of $\tilde{Y}$ rather than of the treatment effects themselves, the expression can be combined with Equation (4) to generate Bayesian posteriors of nearly any target parameter of interest, e.g., the average treatment effect, the average treatment on the treated, or average treatment on the controls. Furthermore, since the marginal treatment effect function $\tau(\eta)$ is a continuous function, we can also use the expressions above to generate a Bayesian posterior of the full marginal treatment effect function.[3]

III.B Understanding the Sources of Uncertainty

Broadly speaking, there are two sources of uncertainty in the estimates: statistical uncertainty, which stems from the fact that the true values of the observed moments are unknown due to the finite sample size and extrapolation uncertainty, which stems from the fact that we may want to extrapolate away from these moments to generate estimates of various other average effects such as ATE, Always Taker ATE

[3] The continuity of $\tau(\eta)$ is a consequence of Assumption 1 and ensures that we can approximate the function with a finite grid, where each point represents the average treatment effect within a small window.

(ATATE), and Never Taker ATE (NTATE). In fact, much of the initial motivation for Bayesian approach developed here was to generate a continuous measure of extrapolation uncertainty, rather than have it be zero for the the local average treatment effect (i.e., LATE) and infinite for all other estimands of interest (e.g., ATE, ATATE, and NTATE). We now formally define "statistical uncertainty" and "extrapolation uncertainty" to allow for researchers to understand which (if any) source dominates the overall uncertainty measures.

From the above intuition, it follows that extrapolation uncertainty can be thought of as the variance of the posteriors in the (hypothetical) case where the observed moments are known with certainty. Note that in the model, this corresponds precisely to the case in which Σ vanishes, which we can use along with Equation (14) to compute the extrapolation uncertainty. We can similarly define the statistical uncertain as the variance of the posteriors that is due to the fact that the observed moments are measured with noise.[4] Formally, we define these as follows:

$$\Sigma_{extrap} = K_{\tilde{Y}} - K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}}\big)^{-1}K'_{\tilde{Y},Y^{obs}} \tag{15}$$

$$\Sigma_{stat} = K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}}^{-1} - (K_{Y^{obs}} + \Sigma)^{-1}\big)K'_{\tilde{Y},Y^{obs}} \tag{16}$$

To more better motivate these expressions, we document a number of properties about the uncertainty measures in the following remark.

Remark 3. *Let $\tilde{Y}$ to be a set of outcomes we want to estimate.*[5] *Define $\Sigma_{\tilde{Y}}$ as in Equation (14) and Σ_{extrap} and Σ_{stat} as in Equations (15)-(16). Then we have the following:*

- *Both Σ_{extrap} and Σ_{stat} are positive semi-definite matrices.*
- *If $\Sigma = 0$, then $\Sigma_{stat} = 0$.*
- *Let $\mathcal{Y}\big(Y^{obs}\big)$ be the set of possible values of $\tilde{Y}$ under the model specified in Section II.C and the constraint that $\tilde{y}_k(\eta) = \hat{y}_k(\eta)$ for all $\hat{y}_k(\eta) \in Y^{obs}$. Then $\Sigma_{extrap} = 0$ if there exists a single $\hat{y} \in \mathcal{Y}\big(Y^{obs}\big)$.*

[4] In a previous version of the paper, we defined statistical uncertainty in the way we currently define frequentist uncertainty (as defined below) and used Σ_{stat} to denote this measure. We apologize for the major change in terminology and notation, but feel this version better captures the concepts.

[5] We acknowledge a slight abuse of notation; here we use $\tilde{Y}$ and Y^{obs} to refer to sets of outcomes, while in Equation (2) we use them to refer to vectors of outcomes.

- *Extrapolation and statistical uncertainty combine to equal the overall uncertainty, e.g.,* $\Sigma_{\tilde{Y}} = \Sigma_{extrap} + \Sigma_{stat}$

The first property – that the measures are all positive semi-definite matrices – simply highlights that they can be considered valid measures of uncertainty. The next helps justify the fact that we refer to Σ_{stat} as the "statistical uncertainty" by noting that it, as well as the frequentist measure of uncertainty defined below, are zero if the observed moments are known with certainty. Similarly, the third helps justify why we refer to Σ_{extrap} as the "extrapolation uncertainty," by highlighting that it stems from the the lack of certainty in how we should extrapolate away from the observed moments.

The third statement says, roughly, that the extrapolation uncertainty is zero if the treatment effects are point identified in a classical sense that there is only a single value of the parameter that is consistent with the observed moments. There is some subtlety in this, however, since identification is generally an asymptotic statement, e.g., with enough data we would be able to estimate the underlying functions with near perfect precision, while Σ_{extrap} is defined for any sample size. Here, the problem with a finite sample size does not stem from the fact that the outcomes are measured with noise, but from the fact that only a finite number of points in the domain of y_0 and y_1 are observed. For example, under an assumed sampling scheme where individuals are sampled randomly with $Z_i \sim U(0,1)$ and $\nu(Z_i) = Z_i$, the functions y_0 and y_1 would be non-parametrically identified, but still potentially require substantial extrapolation if the sample size is only 10 individuals. Thus, the statement here is a finite sample statement stating that there the extrapolation uncertainty is zero if there are enough unique points in the domain of the observed sample to fully pin-down the parameter of interest.

Finally, the fourth statement then highlights that the statistical uncertainty and extrapolation uncertainty serve as a true decomposition of the overall uncertainty, in that the sum to the two sources of uncertainty equal to the overall uncertainty. The formal proofs of all the statements in Remark 3 are in Appendix A.

We also find is useful to define the "frequentist uncertainty," which we define as the variance of the maximum a posteriori (MAP) estimates, i.e., of $\mu_{\tilde{Y}}$, that is due to uncertainty in the observed moments stemming from the error term. This corresponds to the conventional standard error estimates under a frequentist approach where, for example, the formula for $\mu_{\tilde{Y}}$ is motivated as a regularized basis expansion rather than via a Bayesian model (e.g., see Chapter 5 in Hastie et al. (2009)). Using Equations 13 and 14 along with the fact that $Var(Y^{obs}|\mu,\tau) = \Sigma$, it is easy to derive the fact that this corresponds to:

$$\Sigma_{freq} = K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}} + \Sigma\big)^{-1}\Sigma\big(K_{Y^{obs}} + \Sigma\big)^{-1}K'_{\tilde{Y},Y^{obs}} \tag{17}$$

With this definition, we can compare the statistical uncertainty in this model to the traditional frequentist measures uncertainty, i.e., traditional standard errors. Interestingly, we show that our measure of statistical uncertainty is strictly larger than the traditional frequentist measure of uncertainty. The formal statement is documented in the following remark:

Remark 4. *Define Σ_{stat} as in Equation (16) and Σ_{freq} as in Equation (17). Then if $\Sigma \neq 0$, we have that $\Sigma_{stat} - \Sigma_{freq}$ is a positive definite matrix.*

Note that, along with the statements in Remark 3, this result implies that the standard deviation of the Bayesian posterior is strictly larger than traditional frequentist standard errors for any estimand of interest, i.e., any measure of $\mathbb{E}[\tau(\eta)|\eta \in (\underline{\eta}, \overline{\eta})]$, including those ones such as the LATE that require no extrapolation.

III.C Discussion of the Model and Relationship to Alternative Approaches

To better understand the model's assumptions and construction, it is helpful to use an alternative formulation of the Gaussian process. In particular, it is possible to think of the covariance functions as implying a mapping of η to a larger feature space and a set of priors on the coefficients in that feature space. For example, consider the basis expansion that maps η to $\phi(\eta)' = [1, \eta, \eta^2, \eta^3]$ and a Bayesian linear regression specified as $y = \phi(\eta)'\beta + \epsilon$ under a prior $\beta \sim N(0, \Sigma_\beta)$ and a normally distributed and i.i.d. error term. While formulated quite differently, this is equivalent to a Gaussian process with a covariance function specified as: $k(\eta, \eta') = \phi(\eta)'\Sigma_\beta\phi(\eta')$.

This formulation suggests that the model described in Section II.C nests other models which extrapolate away from the observed moments via linear (Kowalski (2023a)), polynomial (Brinch et al. (2017)), or structural (Kline and Walters (2019)) assumptions. We can formalize this understanding in the following remark:

Remark 5. *Suppose that Z_i is a binary instrument and that ϵ_i is independent across individuals. Then:*

- *Let $k_\tau(\eta, \eta'|\theta) = k_\mu(\eta, \eta'|\theta) = 1 + \eta\eta'$. Then the resulting Bayesian marginal treatment effect curve converges to the same marginal treatment effect function as a linear extrapolation of the observed moments, e.g., Kowalski (2023a) and Brinch et al. (2017).*
- *Let $k_\tau(\eta, \eta'|\theta) = k_\mu(\eta, \eta'|\theta) = 1 + \Phi^{-1}(\eta)\Phi^{-1}(\eta')$, where Φ^{-1} is the inverse of a standard normal CDF. Then the resulting Bayesian marginal treatment effect curve converges to the same marginal treatment effect function as a traditional Heckit, e.g., Kline and Walters (2019).*

While these assumptions are at times palatable, there are other times when the researcher may be uncomfortable making the assumption that, for example, the MTE function is necessarily linear, but still wants to say something about policy-relevant parameters beyond the LATE. By using a Bayesian framework, the model allows researchers to relax the assumptions imposed on the MTE and $\mu(\eta)$ functions and still generate the posterior distributions of important parameters such as the ATE. We should highlight explicitly, however, that there is "no Bayesian free lunch" (Poirier (1998)) and while the model does allow researchers to relax the assumptions imposed set of potential functions, it still requires the researchers to make an assumption (implicit in the choice of the covariance specification) about which functions are more or less likely.

This formulation also helps illustrate how the choice of covariance function implicitly places priors on how the relative likelihood of functions in $\mathcal{T}$ is related to how smooth they are in η. From Mercer's theorem it is always possible to associate the covariance function with a (possibly infinite dimensional) feature space $\phi(\eta)$ and prior on the error terms. In particular, while we will not cover the details here, the squared exponential covariance function that we use for our empirical analysis implies an infinite dimensional feature space $\phi(\eta)$ that consists of mappings of the form $\phi_n(\eta) = -exp(-\alpha\eta^2)H_n(\beta\eta)$ where H_n is n^{th} order Hermite polynomial and α and β are constants that depend on the hyperparameters of the squared exponential covariance function. Crucially, the variance of the priors for $\phi_n(\eta)$ are decreasing in n, which implies that the impact of the higher order (and therefore more oscillating) terms on the resulting function are minimized, leading to a smooth function.

It is also worth comparing this approach to approaches that aim to bound the treatment effects by restricting the set of potential functions, e.g., Mogstad et al. (2018). In many ways, the two approaches are quite similar: both respond to the limited number of observed moments by restricting the set of potential functions, while not (necessarily) restricting the set of potential functions enough to guarantee that a single treatment effect is estimated. The key difference is that, in contrast to traditional bounding approaches, the Bayesian approach developed in this paper also places a probability measure on the set of restricted functions. Of course, one could also ignore the probability measure and calculate the minimum and maximum of the posterior, therefore using the Bayesian approach as an alternative way to bound the treatment effects.

The additional structure provided by not only restricting the set of potential functions but also specifying a probability measure on this set has a few key advantages.[6] First, there are many instances – like the one described in the empirical example below – in which the set of potential functions alone does not narrow down the set of possible treatment effects, but in which the 95% credible interval covers a small fraction of these potential values. Second, and relatedly, the method outlined in this paper provides a straightforward way to accommodate restrictions such as the fact that the functions are "nearly linear" or that "smooth functions are more likely than those which oscillate wildly." One could certainly formalize these restrictions and bound the treatment estimates, but the approach outlined here allows one to leverage the large literature on the kernel functions of Gaussian processes to quickly and easily incorporate these assumptions. Finally, while it is certainly possible to conduct hypothesis tests and therefore to construct confidence intervals using the bounding approach – see the NBER version of Mogstad et al. (2018) for a discussion of this in the context of MTE functions – the Bayesian approach naturally accounts for uncertainty in both the true values of the observed moments and (potentially) uncertainty in how to extrapolate away from these moments. Furthermore, the decomposition described in this paper provide a way to explicitly quantify the relative importance of these two sources of uncertainty. It is worth emphasizing, however, that by specifying that the prior takes the form of a Guassian process the approach in this paper is not a

[6] It goes without saying that there has been a substantial and ongoing debate between frequentist and Bayesian approaches. Our goal in this paragraph is not to contribute to the philosophical debate about which approach is preferred and instead to highlight a few practical advantages of the Bayesian approach in this specific context.

strict generalization of bounding approaches and there are certain restrictions – most notably that the functions $\tau(\eta)$ and $\mu(\eta)$ are monotonic – which are quite challenging to implement with a Guassian process prior.[7]

IV Implementation and an Empirical Example

In the above sections, we outlined the Bayesian hierarchical model and discussed both how it can be used to generate posterior estimates of the marginal treatment effects (and other objects of interest) as well as how the model can allow for a better understanding of the sources of uncertainty in the resulting estimates. In doing so, we have been agnostic about the choice of the prior.

Of course, this choice is quite important when implementing the method. We therefore now discuss a commonly used covariance function, the hyperparameters associated with it, and the hyperprior, i.e., the imposed prior distribution over those hyperparameters. We also discuss another important implementation decision: whether to integrate over or estimate the hyperparameters. We then illustrate how the method can work in practice by focusing on a specific empirical example, the Oregon Health Insurance Experiment.

For even more implementation details, we provide a sketch of the overall algorithm in Appendix B. Furthermore, an R package that implements the method and which contains even more implementation details can be found at:
https://github.com/isaacopper/BayesianMTEs.

IV.A Covariance Function and Hyperpriors

In our empirical analysis below we will use the squared exponential to define the covariance, which is a common choice when modeling Gaussian processes and which – as discussed briefly in Section III.C – captures the notion that smooth functions are more likely than those which oscillate wildly. Specifically, with a squared exponential

[7] Another commonly used bound is to set limits on the values of the potential outcomes. The approach outlined in this paper could likely accommodate these bounds by relating the linear model specified here to the outcomes via a flexible link function, i.e., to use this specification as a portion of a generalized linear model. However, doing so complicates the relationship between the $\mu(\eta)$ and $\tau(\eta)$ functions and the observed moments $y_0(\eta)$, and $y_1(\eta)$. Generalizing the method here to allow for a flexible link function seems like an important avenue for future research.

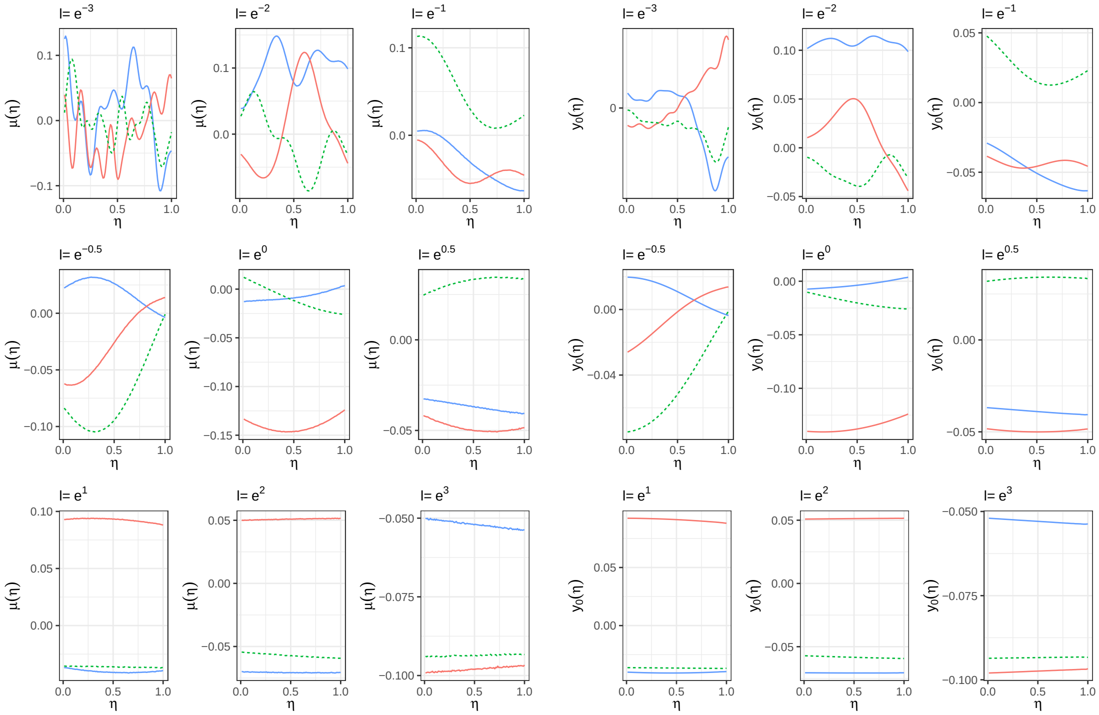
covariance function we get that:

$$k_\mu(\eta, \eta'|\theta_\mu) = \sigma^2_\mu exp\Bigg(- \frac{(\eta - \eta')^2}{2l^2_\mu}\Bigg) \quad \text{and} \quad k_\tau(\eta, \eta'|\theta_\tau) = \sigma^2_\tau exp\Bigg(- \frac{(\eta - \eta')^2}{2l^2_\tau}\Bigg) \tag{18}$$

Each covariate function has two hyperparamters: σ^2 which controls the amplitude and l which is referred to as the lengthscale.

The amplitude (σ^2) is simply a scale factor and is present in front of nearly every covariance function. The lengthscale (l) is more unique to the squared exponential function form and, roughly speaking, determines how much the function oscillates. As a visual example, consider Panel (a) of Figure 1. In it, each panel illustrates three random $\mu(\eta)$ functions generated by a Gaussian process with different lengthscales specified in the title. As can be seen, with lengthscales less than e^{-1} the random functions tend to oscillate widely, while with lengthscales greater than e^1 the random functions are all virtually flat. It is worth noting, however, that the interpretation of the lengthscale is bit different if one instead considers the functions $y_0(\eta)$ and $y_1(\eta)$ instead of the functions $\mu(\eta)$ and $\tau(\eta)$. In Panel (b) of Figure 1, we three random functions of $y_0(\eta)$ under the different lengthscales as before. Since $y_0(\eta)$ corresponds to the conditional average, we find that the resulting functions are smoother than the three random $\mu(\eta)$ functions, at least for relatively small lengthscales.

Figure 1: Random Functions with Different Lengthscales



(a) Three random $\mu(\eta)$ functions

(b) Three random $y_0(\eta)$ functions

Note: These figure shows three random functions sampled from a Gaussian process with varying lengthscales and the same output variance, equal to 0.05. Panel (a) shows the functions $\mu(\eta)$, while panel (b) shows the resulting $y_0(\eta)$ functions.

Of course, while above analysis suggests that the lengthscales are important parameters in the model, it does not answer the question of how they should be chosen. To define the two potential approaches, it will be useful to let θ be the vector of hyperparameters – e.g., in our model: $\theta = [l_\mu, l_\tau, \sigma_\mu, \sigma_\tau]'$. We then can use $p(\theta|Y, T, Z)$ to be the likelihood of θ conditional on the data. (We will discuss shortly how $p(\theta|Y, T, Z)$ can be calculated and what additional assumptions are needed to do so.)

There are then two potential approaches about how to handle the hyperparameters: to choose the hyperparameters that be best fit the data or to integrate over the hyperparameters that are consistent with the data. More precisely, the first, which we refer to as the "empirical Bayes approach" consists of choosing the vector of hyperparameters that maximizes the empirical likelihood, i.e., $\hat{\theta}^{EB} = \arg\max_\theta p(\theta|Y, T, Z)$. The second approach, which we refer to as "the full Bayes approach," samples from the distribution $p(\theta|Y,T,Z)$ estimating $\mu_{\tilde{Y}}$ and $\Sigma_{\tilde{Y}}$ for each sampled vector of hyperparamters, and then combining each estimate of these into the final estimate.

Clearly, both approaches require us to be able to compute $p(\theta|Y,T,Z)$. To do so, we note that from Bayes' law, we have that:

$$p(\theta|Y,T,Z) \propto p(Y|\theta,T,Z) \cdot p(\theta) \tag{19}$$

Furthermore, the values of $p(Y|\theta,T,Z)$ comes directly from Equation (11), so the only object we need to know is $p(\theta)$. This is not something we can determine from the data and corresponds to the hyperprior, i.e., the priors over the hyperparameters. Thus, regardless of whether we employ an empirical Bayes or full Bayes approach, to complete the Bayesian hierarchical model, we need to specify how likely the hyperparameters are to have been the ones used to generate the data.

A seemingly appealing option is to choose a very diffuse prior with roughly equal weight on a wide-range of hyperparameters. This reflects the general preference to "let the data speak" rather than imposing – even inadvertently – the result through our initial assumptions on the data generating process. It is worth noting, however, that choosing a diffuse prior is itself an initial assumption. Our view is that we generally do have a prior belief that, for example, a monotonic MTE function is more likely than one with multiple peaks and valleys. Our preferred approach is to therefore to choose a hyperprior that suggests lengthscales in the middle row of Figure 1 are more likely than the lengthscales in either the top or bottom row. This is particularly important in the case where there is a single binary instrument; as highlighted in Appendix E, in this case there are many hyperparameters that are consistent with the observed data and so a diffuse prior could lead to the model implying a very restricted set of potential MTE functions.

Specifically, in the empirical example we specify that the hyperpriors take the form of a log-normal distribution as follows:

$$log(l) \sim N\big(0, 0.5^2\big) \text{ and } log(\sigma) \sim N\big(.5log(\sigma_\epsilon), 1.5^2\big) \tag{20}$$

$$l_\mu = l_\tau \text{ and } \sigma_\mu = \sigma_\tau \tag{21}$$

where σ_ϵ is the standard error of the residuals. Note that in this specification, we assume that the hyperparameters for the function $k_\mu(\eta,\eta'|\theta_\mu)$ are the same as for the function $k_\tau(\eta, \eta'|\theta_\tau)$.

IV.B Empirical Example

We now illustrate the method using the Oregon Health Insurance Experiment (OHIE), in which participating individuals were randomly assigned to be eligible or ineligible to enroll in Medicaid. See the OHIE website for more detail about the OHIE and links to the public data, as well as Finkelstein et al. (2012); Taubman et al. (2014); Finkelstein et al. (2016) for other work on the OHIE. We also provide additional examples via simulations for both binary instruments and continuous instruments in Appendix E.

We chose to use the OHIE for a handful of reasons. First and foremost, the data is publicly available and so interested readers can easily explore how our subjective choices (e.g., hyperpriors) impact the results. Second, the OHIE is a particularly interesting context for us to study. Not only does it provide some of the most compelling evidence on an important public policy question, but it had high levels of non-compliance; many of those that were randomly given eligibility did not enroll in Medicaid and many of those that were not randomly given eligibility gained eligibility in another way and so ended up enrolling in Medicaid. Finally, by using the same data as a previous study that used a linear extrapolation, namely Kowalski (2023a), it is easy to compare the two approaches and understand the relative benefits of the two approaches. For that reason, we will focus on the same outcome as used in Kowalski (2023a), namely the likelihood that an individual will go the emergency room (ER).

We start by illustrating how the data informs the plausible values of the hyperparameters. To do so, we take a slightly different approach than the one discussed in Section B and instead of random sampling from the hyperposterior we calculate the marginal likelihood, i.e., $p(Y|\theta, Z, T)$, for all values of θ in a discrete grid of step-size 0.1 with $log(l) \in [-3, 3]$ and $log(\sigma) \in [-3, 3]$.[8] We also calculate the hyperprior, i.e. $p(\theta)$ for each of these values, using the hyperprior specified in Equation (21). To help visualize the results, we then turn the marginal likelihood into a function of the lengthscale by choosing the value for σ that maximizes the marginal likelihood for each value of l.[9] We also do the same for the hyperprior and then combine these to

[8] In other words, $log(l) = -3 + 0.1k$ for $0 \leq k \leq 60$ and $log(\sigma) = -3 + 0.1j$ for $0 \leq j \leq 60$.

[9] Formally, we plot $p(Y|l, Z, T)$ as a function of l where $p(Y|l, Z, T) = \max_{log(\sigma)\in[-3,1]} p(Y|\sigma, l, Y, Z, T)$.

produce a plot of the marginal distribution of the hyperposterior, i.e., the posterior as a function of l. We also do the same to plot the marginal likelihood, hyperprior, and hyperposterior as a function of σ.

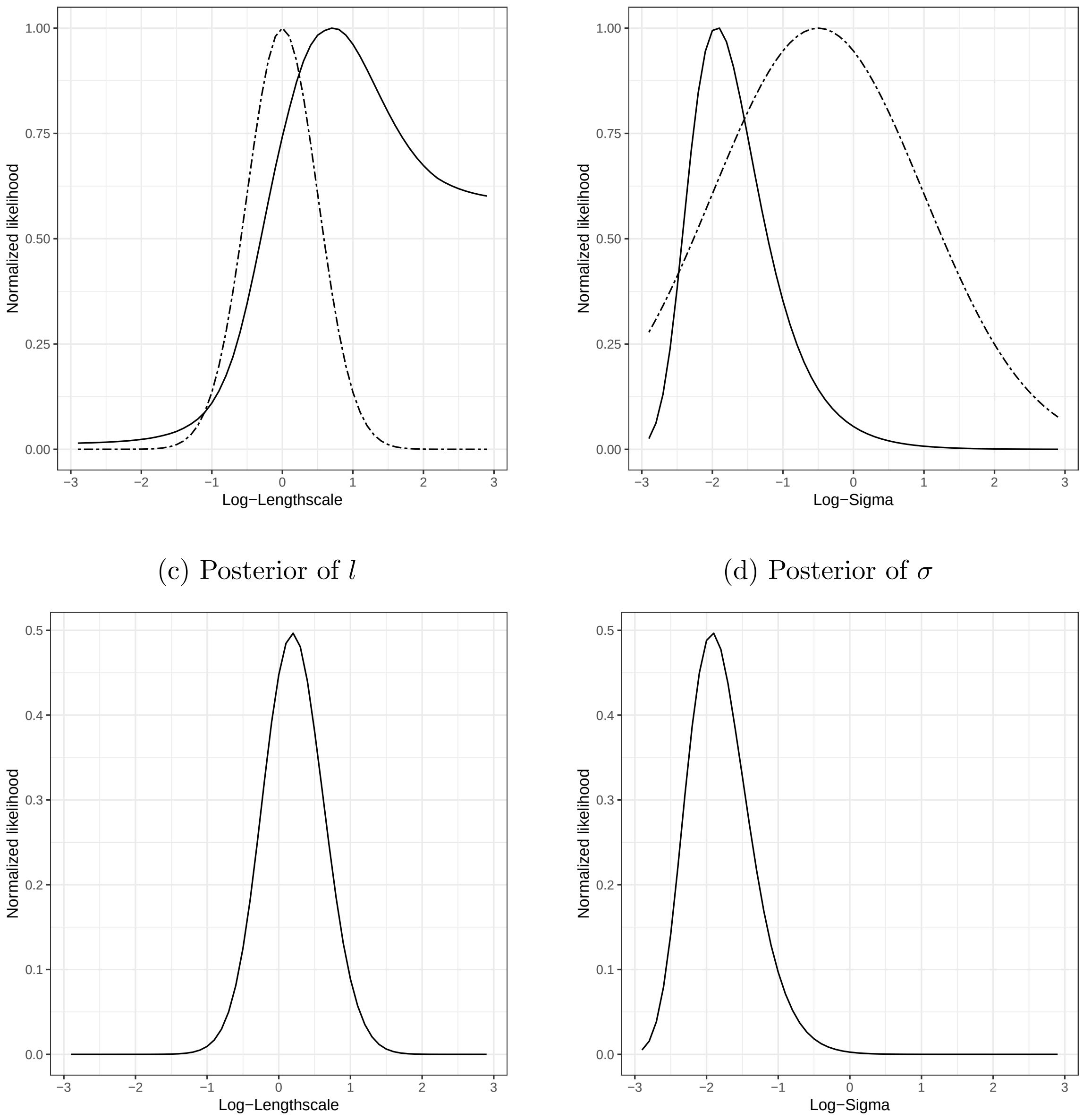
The results are shown in Figure 2. The top two subfigures show the marginal likelihood of the hyperparameters in solid lines and the hyperpriors in dashed lines and the bottom two subfigures show the hyperposterior; the left two subfigures show the results for the lengthscale (l) and the right two show the results for the scale (σ). The main result is that the marginal likelihood of the lengthscale (l) does not have a well-defined peak – as seen in Figure 2a – while the scale does – as seen in Figure 2b. As a consequence, the hyperposterior of l – shown in Figure 13b – is more similar to the hyperprior of l than of the marginal likelihood, while the hyperposterior of σ - shown in Figure 2d – is more similar to the marginal likelihood of σ than to the hyperprior. In other words, in cases like this one where there is a single binary instrument the hyperposterior of l is determined more by the researchers' choice of the hyperprior more than the data; however, as we discuss more in Appendix D.A, a more diffuse prior often results in a smaller credible interval for most of the estimands rather than a larger one.

We next turn our attention to the estimands of interest, namely the MTE function $\hat{\tau}(\eta)$ and various averages of the MTE function such as the overall average treatment effect (ATE), the always taker average treatment effect (ATATE), the never taker average treatment effect (NTATE), and the complier average treatment effect also known as the local average treatment effect (LATE). We start by initially using a single value of the hyperparameters: the maximum a posteriori (MAP) values, i.e., the value of θ that maximizes the $p(\theta|Y, Z, T)$.[10] This approach is often referred to as empirical Bayes approach and results in the following values for the hyperparameters: $log(l) = 0.11$ and $log(\sigma) = -1.97$ or $l = 1.11$ and $\sigma = 0.14$.

Holding these hyperparameters fixed, we then simulate ten random functions y_1 and y_0. These simulations are shown in Figure 3; Panel (a) shows ten random functions drawn without conditioning on any data, while panel (b) shows ten random functions drawn when conditioning on the observed four moments. While not a particularly subtle point, Figure 3 illustrates nicely that conditioning on the four observed

[10] Note that we choose the MAP value, rather than the value that maximizes the log marginal likelihood. These coincide under a very diffuse prior on the hyperparameters, however, for reasons we discuss above and in Appendix D.A we view it preferable to use an informed prior on the hyperparameters.

Figure 2: Bayesian Priors, Marginal Likelihood, and Posterior of the Hyperparameters



(a) Marginal Likelihood and Prior of l

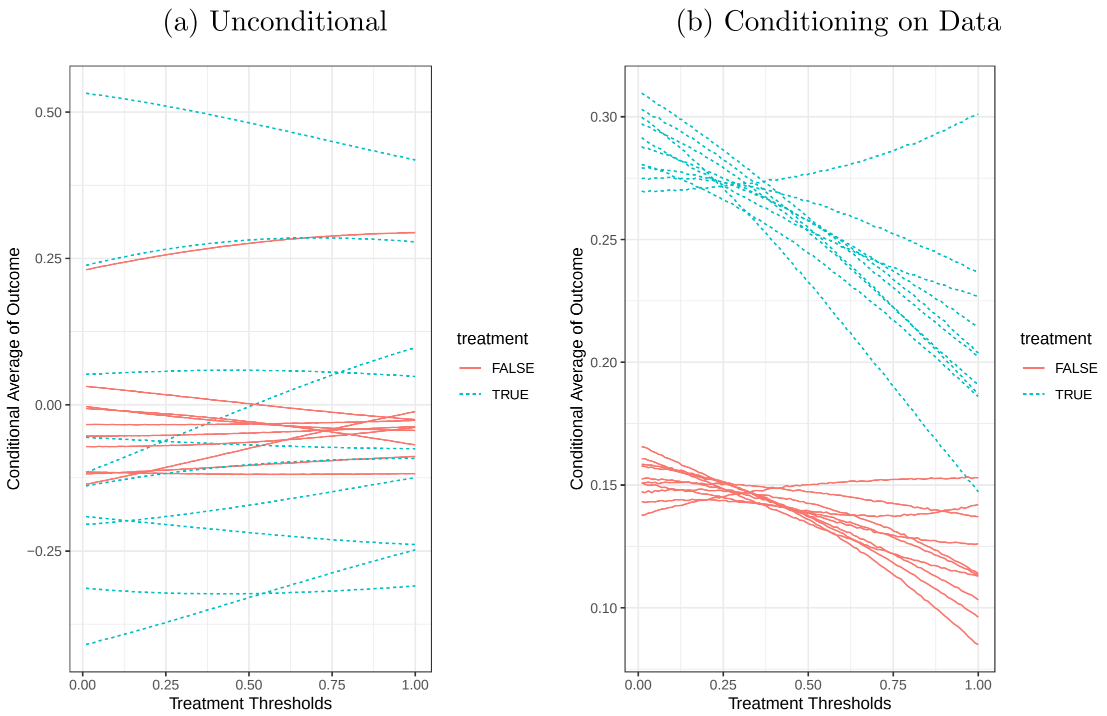
(b) Marginal Likelihood and Prior of σ

(c) Posterior of l

(d) Posterior of σ

Note: The top two subfigures show the marginal likelihood of the hyperparameters in solid lines and the hyperpriors in dashed lines, while the bottom two figures show the hyperposterior in solid lines. The left two figures show the results for the lengthscale parameter (l), while the right two figures show the result for the scale parameter (σ); see Section IV.A for a discussion of the two hyperparameters. To illustrate the results on a two-dimensional graph, we show plots as a function of one hyperparameter while optimizing over the second. For example, in the top-left figure we show the marginal likelihood as a function of l by plotting $p(Y|l, Z, T) = \max_\sigma p(Y|l, \sigma, Z, T)$.

Figure 3: Random Functions via Empirical Bayes


(a) Unconditional

(b) Conditioning on Data

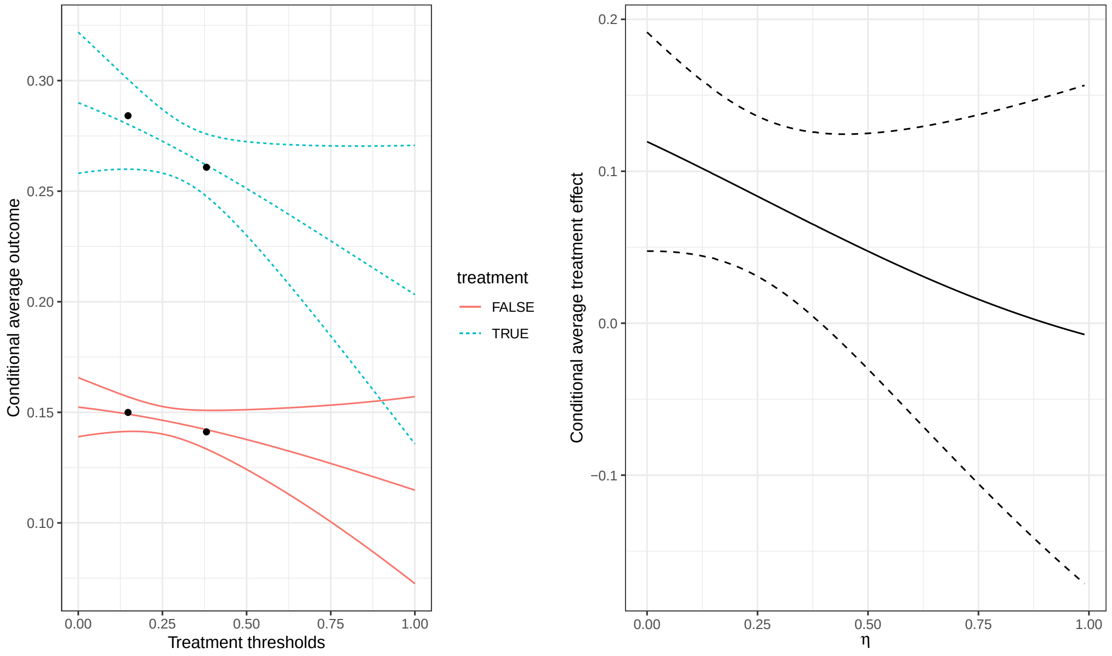
Note: This figure shows ten random functions sampled from a Gaussian process with estimated hyperparameters. In Panel (a), the functions are drawn unconditionally while in Panel (b) we condition on the four observed moments.

moments has a large impact on the plausible functions, even if it does not fully pin down the entirety of the functions.

Figure 4a uses the above equations to calculate the posterior distribution of y_1 and y_0 on a 101 point grid. The red solid lines illustrate the posterior mean of y_0 and the 95% credible interval, while the blue dashed lines show the posterior mean of y_1 and the 95% CI. The black dots indicate the four estimated moments. Note that the posterior means come close – but do not go directly through – the four black dots and there is still uncertainty in the posterior distributions of the y functions at the point where the four moments are observed. This is because the four moments are estimated with error rather than being observed directly and so the mean of the posterior does not correspond exactly to the observed mean.[11] However, as would

[11] This depends on the choice of the prior; if we include in the kernel an additional linear term with infinite prior variance, the posterior means will go directly through the observed moments. See Opper and Özek (2023), for example.

Figure 4: Posterior Mean and 95% CIs



(a) Conditional Moment Functions

(b) Marginal Treatment Effects

Note: Panel (a) shows the posterior mean and 95% credible interval of the function y_0 (in the red solid lines) and y_1 (in the blue dashed lines). The black dots represent the estimated moments observed in the data. Panel (b) shows the posterior distribution of the MTE function, i.e. $\tau(\eta)$, with the solid indicating the posterior mean and the dashed lines indicating the 95% credible interval.

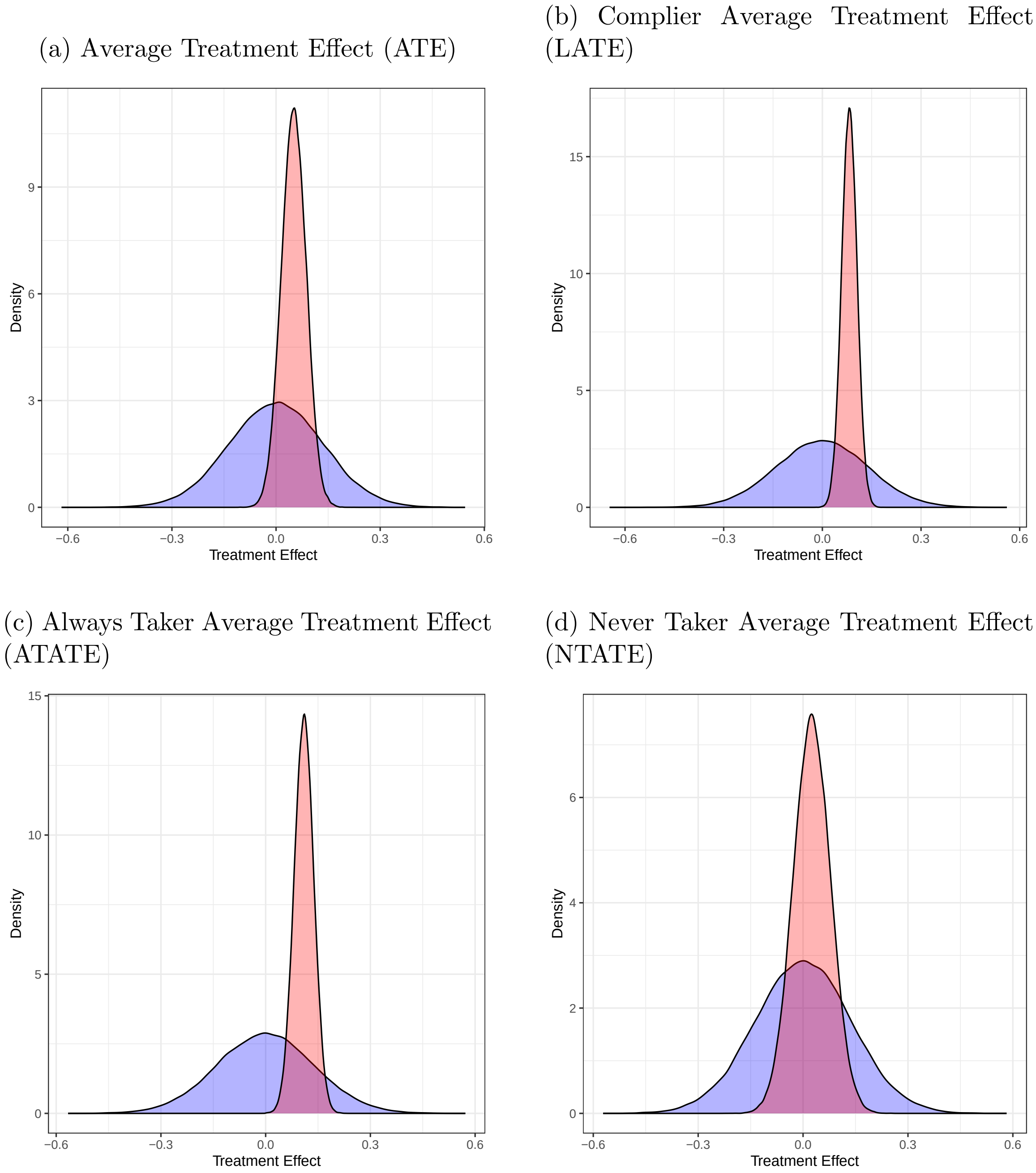
be expected, the posterior variance increases significantly away from the observed moments.

The posterior distributions of y_1 and y_0, while suggestive, are mainly useful because they allow us to construct a posterior distribution of the marginal treatment effect (MTE) function. In particular, as illustrated in Equation (4) the MTEs can be approximated via a linear combination of four points on the curves y_1 and y_0. The posterior of the MTE function is therefore also normally distributed and we can use Equation (4) to compute the posterior distribution of the MTE at the same 101 point grid used to estimate y_1 and y_0. This is illustrated in Figure 4b, which shows that the effect of Medicaid enrollment on ER visits is higher for those more likely to enroll (if given eligibility) than those less likely to enroll. This is consistent with Kowalski (2023a). As illustrated in the dashed lines, however, we cannot be particularly confident in these point estimates, with the 95% CIs spanning from an effect of approximately 0.15 to -0.15 among those who are least likely to enroll in treatment.

We can similarly use Equation (4) to compute various average effects, such as the overall average treatment effect (ATE), always taker average treatment effect (ATATE), never taker average treatment effect (NTATE), and the local average treatment effect (LATE). While there is a lot of uncertainty in MTEs, we can be much more confident in the implied average effects. This is true even though (with the exception of the LATE) the average effects require some extrapolation beyond the observed moments. After the RCT we have the most confidence in the LATE estimate – which requires no extrapolation beyond the observed measures – and the least confidence in the NTATE estimate – which requires the most extrapolation. This result can be seen in Table 1, which shows that the standard deviation of the LATE posterior is approximately half that of the NTATE posterior. In addition as suggested by Figure 4b, the mean posterior of the ATATE is greater than the LATE, which is in turn greater than the NTATE. This can also be seen in Figure 5, which shows the prior distributions of the four estimands discussed above in blue and the posterior distributions in red; as can be seen, the observed data significantly changes the posterior distribution of even the estimands that require large amounts of extrapolation.

An important caveat is that results in the first two columns of Table 1 take the hyperparameter values as fixed. An alternative is to incorporate uncertainty in the hyperparameters by integrating over the posterior distribution of hyperparameters. As discussed in Section B, we therefore repeat analysis for a range of hyperparameters

Figure 5: Bayesian Prior and Posterior Distributions of Four Estimands



(a) Average Treatment Effect (ATE)

(b) Complier Average Treatment Effect (LATE)

(c) Always Taker Average Treatment Effect (ATATE)

(d) Never Taker Average Treatment Effect (NTATE)

Note: The four subfigures show the prior distribution of the estimand of interest in blue and the posterior distribution in red.

that are consistent with the data using a simple accept/reject algorithm. In doing so, we accepted approximately 20% of the 10,000 proposed values of the hyperparameters; for each accepted value, we can use the approach outlined above to generate posterior distributions of the MTE function as well as the LATE, ATE, ATATE, and NTATE. The resulting posterior is therefore a mixture of normally distributed random variables, which means there is a closed form solution for the mean and variance of the posterior. The results of this are shown in the final two columns of Table 1. Intuitively, the posterior standard deviations are mostly unchanged for the averages which require the least extrapolation from the observed moments – in particular, the ATATE and LATE – while the posterior standard deviations of the NTATE increases by 40% when accounting for uncertainty in the hyperparameters.

Table 1: Posterior Means and Standard Deviations

	Empirical Bayes		Full Bayes	
	Mean	Std.	Mean	Std.
Avg. Treat Effect (ATE)	0.054	0.035	0.053	0.045
Always Taker Avg. Treat Effect (ATATE)	0.108	0.027	0.110	0.031
Local Avg. Treat Effect (LATE)	0.083	0.023	0.083	0.024
Never Taker Avg. Treat Effect (NTATE)	0.030	0.051	0.029	0.070

Note: This table shows the mean and standard deviation of the posterior distribution of the estimated effects. The first two columns estimate the value of the hyperparameters, while the last two columns integrates over the hyperposterior distribution. See Section B for more discussion of the differences between the two approaches.

We can use Section III.B to better understand the sources of uncertainty in the posterior estimates. The results are shown in Table 2. In the four columns, we show the standard deviation of the posterior of four average effects: ATE, ATATE, LATE, and NTATE. Even the NTATE, which requires the most extrapolation away from the observed moments the statistical uncertainty is significantly larger than the extrapolation uncertainty. From a practical perspective, it means that increasing the sample size and/or decreasing the variance of the residual will significantly reduce the uncertainty of the resulting average treatment effects, even the ones that are require substantial extrapolation. Note that the relative unimportance of extrapolation uncertainty in the resulting estimate of the NTATE, for example, stems in part from the fact that the NTATE involves averaging $\tau(\eta)$ over a relatively large range of η values, i.e., $NTATE = \mathbb{E}[\tau(\eta)|\eta \in (0.38, 1)]$ in the OHIE experiment. The extrapolation uncertainty is a higher fraction of the posterior variance of a single point on the MTE function, e.g., $\hat{\tau}(1)$ for example, than for the NTATE. This can be seen by comparing Figure 4 to Figure 6.

Table 2: Statistical vs. Extrapolation vs. Frequentist Uncertainty

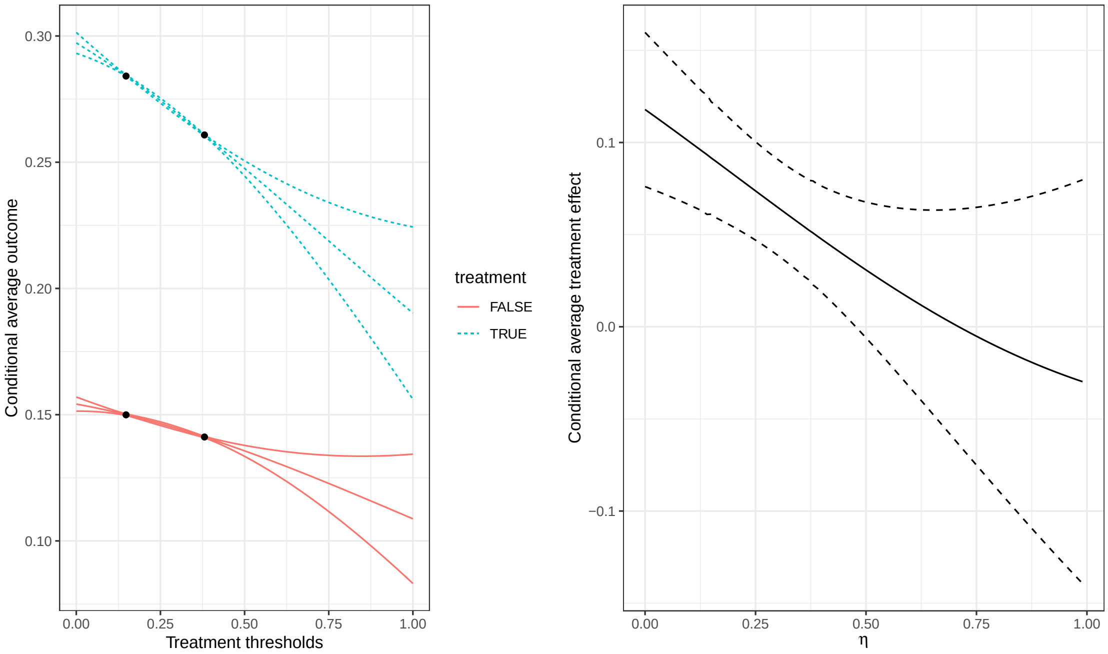
	Full Sample			
	Full	Stat.	Extrap.	Freq.
Avg. Treat Effect (ATE)	0.035	0.031	0.015	0.025
Always Taker ATE (ATATE)	0.027	0.025	0.008	0.021
Local ATE (LATE)	0.023	0.023	0.000	0.021
Never Taker ATE (NTATE)	0.051	0.044	0.025	0.033
	N = $1,000$			
	Full	Stat.	Extrap.	Freq.
Avg. Treat Effect (ATE)	0.066	0.064	0.016	0.036
Always Taker ATE (ATATE)	0.057	0.056	0.007	0.036
Local ATE (LATE)	0.058	0.058	0.000	0.037
Never Taker ATE (NTATE)	0.079	0.075	0.025	0.036

Note: This table shows the standard deviations of the posterior distributions. The definition of statistical uncertainty, abbreviated as "Stat." – extrapolation uncertainty – abbreviated as "Extrap." – and frequentist uncertainty – abbreviated as "Freq." are in Section III.B. To ease the comparison, we use the same set of hyperparamters for all eight estimates, which are the hyperparameters that maximize the hyper-posterior in the full sample.

It's worth highlighting that the OHIE was much larger than most RCTs, with a sample size of approximately $17,000$ (when restricting to the Portland sample, as we do). This, along with the fact that compliance rates were low, suggests that the relative importance of statistical uncertainty would be even larger for most other RCTs. To illustrate this, we consider the relative importance of the sources if the OHIE instead had a sample size of only $1,000$. As can be seen, the statistical uncertainty increases – as would be expected – but the extrapolation uncertainty does not change; hence, the importance of statistical uncertainty increases relative to the extrapolation uncertainty when the sample size decreases.[12] For more discussion of

[12] The minor changes in the extrapolation uncertainty between the full sample and the sample of

Figure 6: Posterior Mean and 95% CIs – Extrapolation Uncertainty Only



(a) Conditional Moment Functions

(b) Marginal Treatment Effects

Note: Here, we reproduce Figure 4, while ignoring uncertainty in the estimated moments. The Panel (a) shows the posterior mean and 95% credible interval of the function y_0 (in the red solid lines) and y_1 (in the blue dashed lines). The black dots represent the estimated moments observed in the data. Panel (b) shows the posterior distribution of the MTE function, i.e. $\tau(\eta)$, with the solid indicating the posterior mean and the dashed lines indicating the 95% credible interval.

how the uncertainty scales with the sample size, see Appendix D.B.

Finally, we can contrast the results of the proposed approach to two other commonly used approaches: a linear extrapolation (i.e., one that restricts the functions $\mu(\eta)$ and $\tau(\eta)$ to be linear functions of η) and a Heckit extrapolation (i.e., one that restricts the functions $\mu(\eta)$ and $\tau(\eta)$ to take the form $\alpha_k + \beta_k \Phi^{-1}(\eta)$). The resulting estimates of the ATE, ATATE, LATE, and NTATE are shown in Table 3. In the top panel, we see that for the full OHIE sample the estimates and the confidence/credible intervals are quite similar for for all three approaches. It is not the case, however, that the Bayesian MTE estimates will always be similar to those obtained via a linear extrapolation and the fact that they are in this case is dependent on the sample size.

This can be seen in the next two panels, where we assume that the sample is either one-tenth or ten times as large at the full OHIE sample.[13] When the sample size is one-tenth as large as the OHIE experiment the Bayesian MTE approach produces smaller credible intervals than the other two approaches. In this case, there is enough uncertainty in the true values of the observed moments that the Bayesian MTE approach places larger weight on the prior and therefore (mostly) does not attempt to distinguish between the ATATE, LATE, and NTATE. This produces similar posterior means for the three estimands, while also reducing the posterior variance the Bayesian MTE estimates. In contrast, when the sample size is ten times as large as the OHIE experiment the Bayesian MTE approach produces larger credible intervals than the other two approaches. In this case, the observed moments are known with near certainty. This means that the linear and Heckit models also treat the estimands that require substantial extrapolation as being estimated with near certainty, while the Bayesian MTE approach accounts for the fact that there is uncertainty in how one should extrapolate away from the observed moments. These results are also clear when plotting the estimated MTEs, which we do in Figure 13 in the Appendix.

V Conclusion

This paper developed a Bayesian model that generates posterior distributions of the marginal treatment effects (MTEs) and hence of various estimates of interest, e.g., the

$1,000$ in Table 2 are because the estimated cutoffs – and hence the extrapolation uncertainty – will vary slightly depending on the sample.

[13] We get similar results if we assume the sample is either one-third or three times as large as the full OHIE sample.

Table 3: Comparing Different Extrapolation Approaches

	Full Sample (N = 17,092)					
	Bayesian MTEs		Linear Extrap.		Heckit Extrap.	
	Mean	Std.	Mean	Std.	Mean	Std.
Avg. Treat Effect (ATE)	0.053	0.045	0.043	0.038	0.062	0.028
Always Taker ATE (ATATE)	0.110	0.031	0.097	0.034	0.087	0.041
Local ATE (LATE)	0.083	0.024	0.073	0.025	0.073	0.025
Never Taker ATE (NTATE)	0.029	0.070	0.020	0.068	0.053	0.040

	One-Tenth Sample					
	Bayesian MTEs		Linear Extrap.		Heckit Extrap.	
	Mean	Std.	Mean	Std.	Mean	Std.
Avg. Treat Effect (ATE)	0.104	0.069	0.161	0.114	0.144	0.086
Always Taker ATE (ATATE)	0.128	0.062	0.020	0.107	-0.002	0.127
Local ATE (LATE)	0.118	0.054	0.086	0.077	0.086	0.077
Never Taker ATE (NTATE)	0.092	0.100	0.227	0.176	0.205	0.126

	Ten Times Sample					
	Bayesian MTEs		Linear Extrap.		Heckit Extrap.	
	Mean	Std.	Mean	Std.	Mean	Std.
Avg. Treat Effect (ATE)	0.040	0.030	0.043	0.012	0.062	0.009
Always Taker ATE (ATATE)	0.106	0.021	0.097	0.011	0.087	0.013
Local ATE (LATE)	0.074	0.008	0.073	0.008	0.073	0.008
Never Taker ATE (NTATE)	0.011	0.049	0.020	0.018	0.053	0.013

Note: This table shows the mean and standard deviation of three different estimation approaches: the Bayesian MTE approach developed here, a linear extrapolation approach (e.g., Brinch et al. (2017) and Kowalski (2023a)), and a Heckit model (e.g., Kline and Walters (2019)). For the Bayesian MTEs, Mean corresponds to the mean of the posterior and Std. corresponds to the standard deviation of the posterior; for the other two approaches, Mean corresponds to the point estimate and Std. to the standard error of this estimate.

average treatment effect (ATE), always taker average treatment effect (AT ATE), or never taker average treatment effect (NT ATE). By providing a principled approach to extrapolate from observed estimands, this model can help researchers generate plausible ranges for important and potentially policy-relevant quantities of interest.

We conclude by noting that our focus here is on a simple model: one with a single binary treatment, a single instrument, and no covariates. One can also start thinking of ways to extend the model in Section II to capture more complex settings – e.g., those with multiple or continuous treatments, multiple instruments, an/or many covariates – in which a similar approach could be used. In these cases, the amount of extrapolation required for many estimands of interest is less apparent and so the decomposition of extrapolation vs. statistical uncertainty is likely to be particularly illuminating. One can also think about extending the model to apply to different research designs – e.g., designs that combine experimental and non-experimental variation and/or estimates from multiple studies (Meager (2019, 2022); Gechter and Meager (2022)) or fuzzy regression discontinuity designs (Opper and Özek (2023)). Thus, while we view this paper as a useful method in and of itself, we also hope that it helps illustrate that utility of adding a hierarchical Bayesian model on top of traditional econometric models and of the relative ease with which the Gaussian process Bayesian model and generalized Roy models can work together.

References

Balke, Alexander and Judea Pearl, "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 1997, *92* (439), 499–527.

Bhattacharya, Jay, Azeem M. Shaikh, and Edward Vytlacil, "Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization," *American Economic Review: Papers and Proceedings*, 2008, *98* (2), 351–356.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall, "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 2017, *125* (4), 985–1039.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group, "The Oregon Health Insurance Experiment: Evidence from the First Year," *Quarterly Journal of Economics*, 2012, *127* (3), 1057–1106.

_ , _ , Heidi Allen, Bill Wright, and Katherine Baicker, "Effect of Medicaid Coverage on ED Use - Further Evidence from Oregon's Experiment," *New England Journal of Medicine*, 2016, *375* (16), 1505–1507.

Gechter, Michael and Rachael Meager, "Combining Experimental and Observational Studies in Meta-Analysis: A Debiasing Approach," 2022.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, 2009.

Heckman, James J., "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 2010, *48* (2), 356–398.

_ and Edward J. Vytlacil, "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in J. J. Heckman and E. E. Leamer, eds., *Handbook of Econometrics*, Vol. 6, Elsevier, 2007, chapter 70, pp. 4779–4874.

— **and** — , “Econometric Evaluation of Social Programs, Part II: Using the Marginal Treat- ment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments,” in J. J. Heckman and E. E. Leamer, eds., *Handbook of Econometrics*, Vol. 6, Elsevier, 2007, chapter 71, pp. 4785–5143.

— **and Edward Vytlacil**, “Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects,” *Proceedings of the National Academy of Sciences*, 1999, *96* (8), 4730–4734.

— **and** — , “Policy-Relevant Treatment Effects,” *American Economic Review*, 2001, *91* (2), 107–11.

— **and** — , “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 2005, *73* (3), 669–738.

Imbens, Guido W. and Joshua D. Angrist, “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 1994, *62* (2), 467–475.

Kline, Patrick and Christopher R. Walters, “On Heckits, LATE, and Numerical Equivalence,” *Econometrica*, 2019, *87* (2).

Kowalski, Amanda, “Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform,” *Review of Economics and Statistics*, 2023, *105* (3), 646–664.

Kowalski, Amanda E, “Behaviour within a Clinical Trial and Implications for Mammography Guidelines,” *The Review of Economic Studies*, 2023, *90* (1), 432–462.

Manski, Charles F., “Nonparametric Bounds on Treatment Effects,” *American Economic Review*, 1990, *80* (2), 319–323.

Meager, Rachael, “Understanding the Average Impact of Microcredit Expansion: A Bayesian Hierarchical Analysis of Seven Randomized Experiments,” *American Economic Journal: Applied Economics*, 2019, *11* (1), 57–91.

— , “Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature,” *American Economic Review*, 2022, *112* (6), 1818–1847.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky, "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters," *Econometrica*, 2018, *86* (5), 1589–1619.

Opper, Isaac M. and Umut Özek, "A Global Regression Discontinuity Design: Theory and Application to Grade Retension Policies," 2023.

Poirier, Dale J., "Revising Beliefs in Nonidentified Models," *Econometric Theory*, 1998, *14*.

Rasmussen, Carl Edward and Chrisopher K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.

Taubman, Sarah, Heidi Allen, Bill Wright, Katherine Baicker, Amy Finkelstein, and the Oregon Health Study Group, "Medicaid Increases Emergency Depratment Use: Evidence from Oregon's Health Insurance Experiment," *Science*, 2014, *343* (6168), 263–268.

A Proofs

Remark 1. *Under the the Bayesian hierarchical model defined in Section II and Assumption 1, $\tilde{y}(t,\eta)$ as defined in Equation (7) also follows a mean-zero Gaussian process with a covariance function – denoted $k_{\tilde{y}}$ – that depends on $k_\mu(\eta,\eta'|\theta_\mu)$ and $k_\tau(\eta,\eta'|\theta_\tau)$. In particular, we have that:*

$$k_{\tilde{y}}\big((t,\eta),(t',\eta')|\theta_\mu,\theta_\tau\big) = \begin{cases} \mathbb{E}\big[k_\mu(\tilde{\eta},\tilde{\eta}'|\theta_\mu) + k_\tau(\tilde{\eta},\hat{\eta}|\theta_\tau)|\tilde{\eta} \le \eta, \tilde{\eta}' \le \eta'\big] & \text{if } t = t' = 1 \\ \mathbb{E}\big[k_\mu(\tilde{\eta},\tilde{\eta}'|\theta_\mu)|\tilde{\eta} > \eta, \tilde{\eta}' > \eta'\big] & \text{if } t = t' = 0 \\ \mathbb{E}\big[k_\mu(\tilde{\eta},\tilde{\eta}'|\theta_\mu)|\tilde{\eta} > \eta, \tilde{\eta}' \le \eta'\big] & \text{if } t = 0 \ne t' \\ \mathbb{E}\big[k_\mu(\tilde{\eta},\tilde{\eta}'|\theta_\mu)|\tilde{\eta} \le \eta, \tilde{\eta}' > \eta'\big] & \text{if } t = 1 \ne t' \end{cases} \tag{8}$$

Proof. We first show that under the model specified in Section II and Assumption 1, we can infer that the Guassian process for both τ and μ are continuous processes. To do so, we note that:

$$\begin{aligned} \mathbb{E}\Big[\big(\mu(\eta) - \mu(\eta')\big)^4\Big] =& 12 \cdot \big(k_\mu(\eta,\eta|\theta_\mu) - k_\mu(\eta,\eta'|\theta_\mu)\big)^2 \\ \le& C\big(\eta - \eta'\big)^2 \end{aligned}$$

where C is a constant. The first equality stems from the fact that $\mu(\eta) - \mu(\eta')$ is distributed normally and the inequality comes from the assumption that k_μ is Lipschitz continuous. From this expression, we can used the Kolmogorov continuity theorem to infer that μ is a continuous process. Note that the fact that $\mu(\eta)$ can be assumed to be continuous implies that $y_0(\eta) = \frac{1}{1-\eta}\int_\eta^1 \mu(\tilde{\eta})d\tilde{\eta}$ can be arbitrarily approximated by a finite sum of jointly Gaussian variables. In the same way, we can show that τ is a continuous process and that $y_1(\eta)$ can be arbitrarily approximated by a finite sum of jointly Gaussian variables.

We can then show that $\tilde{y}(t,\eta)$ is itself a Gaussian process. For this, consider any finite linear combination $\sum a_k\tilde{y}(t_k,\eta_k)$. As highlighted above, we have that each $\tilde{y}(t_k,\eta_k)$ can be arbitrarily approximated by a finite sum of jointly Gaussian variables and so $\sum a_k\tilde{y}(t_k,\eta_k)$ can be arbitrarily approximated by a finite sum of jointly Gaussian variables and is therefore Gaussian. Thus, $\tilde{y}$ is a Gaussian process.

We therefore need only show that the covariance of $\tilde{y}(t,\eta)$ accords to the definition above, which is mostly just algebra. For example, consider two points $\tilde{y}(1,\eta)$ and

$\tilde{y}(1,\eta')$. From the definition of $\tilde{y}$, we get that:

$$\begin{aligned}
Cov(\tilde{y}(1,\eta),\tilde{y}(1,\eta)) &= \left(\frac{1}{N}\lim_{N\to\infty}\sum_{i=0}^{N}\mu(\eta\frac{i}{N})+\tau(\eta\frac{i}{N})\right)\cdot\left(\frac{1}{N}\lim_{N\to\infty}\sum_{i=0}^{N}\mu(\eta'\frac{i}{N})+\tau(\eta'\frac{i}{N})\right)\\
&= \lim_{N\to\infty}\sum_{i=0}^{N}\Big(\mu(\eta\frac{i}{N})+\tau(\eta\frac{i}{N})\Big)\cdot\sum_{i=0}^{N}\Big(\mu(\eta'\frac{i}{N})+\tau(\eta'\frac{i}{N})\Big)\\
&= \lim_{N\to\infty}\sum_{i=0}^{N}\mu(\eta\frac{i}{N})\cdot\sum_{i=0}^{N}\mu(\eta'\frac{i}{N})+\lim_{N\to\infty}\sum_{i=0}^{N}\tau(\eta\frac{i}{N})\cdot\sum_{i=0}^{N}\tau(\eta'\frac{i}{N})\\
&= \mathbb{E}\big[k_\mu(\tilde{\eta},\tilde{\eta}'|\theta_\mu)|\tilde{\eta}\leq\eta,\tilde{\eta}'\leq\eta'\big]+\mathbb{E}\big[k_\tau(\tilde{\eta},\tilde{\eta}'|\theta_\mu)|\tilde{\eta}\leq\eta,\tilde{\eta}'\leq\eta'\big]
\end{aligned}$$

□

The only somewhat subtle point in here is the third equality, which uses the fact that $\tau(\eta)$ and $\mu(\eta)$ are assumed to be independent GPs. The covariance functions when $t \neq 1$ and/or $t' \neq 1$ are derived similarly.

Remark 2. *The Bayesian posterior of $\tilde{Y}$ given the observed data:*

$$\tilde{Y}|Y^{obs} \sim N\big(\mu_{\tilde{Y}}, \Sigma_{\tilde{Y}}\big) \tag{12}$$

where

$$\mu_{\tilde{Y}} = K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}}+\Sigma\big)^{-1}Y^{obs} \tag{13}$$

$$\Sigma_{\tilde{Y}} = K_{\tilde{Y}} - K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}}+\Sigma\big)^{-1}K'_{\tilde{Y},Y^{obs}} \tag{14}$$

Proof. This is a simple application of conditional multivariate normal distributions.

□

Remark 3. *Let $\tilde{Y}$ to be a set of outcomes we want to estimate.*[14] *Define $\Sigma_{\tilde{Y}}$ as in Equation (14) and Σ_{extrap} and Σ_{stat} as in Equations (15)-(16). Then we have the following:*

- *Both Σ_{extrap} and Σ_{stat} are positive semi-definite matrices.*
- *If $\Sigma = 0$, then $\Sigma_{stat} = 0$.*

[14] We acknowledge a slight abuse of notation; here we use $\tilde{Y}$ and Y^{obs} to refer to sets of outcomes, while in Equation (2) we use them to refer to vectors of outcomes.

- *Let $\mathcal{Y}\big(Y^{obs}\big)$ be the set of possible values of $\tilde{Y}$ under the model specified in Section II.C and the constraint that $\tilde{y}_k(\eta) = \hat{y}_k(\eta)$ for all $\hat{y}_k(\eta) \in Y^{obs}$. Then $\Sigma_{extrap} = 0$ if there exists a single $\hat{y} \in \mathcal{Y}\big(Y^{obs}\big)$.*

- *Extrapolation and statistical uncertainty combine to equal the overall uncertainty, e.g., $\Sigma_{\tilde{Y}} = \Sigma_{extrap} + \Sigma_{stat}$*

Proof. For the first statement, we get that Σ_{stat} is a positive semi-definite matrix using the fact that $K_{Y^{obs}}^{-1} - \big(K_{Y^{obs}} + \Sigma\big)^{-1}$ is a positive semi-definite matrix. This, in turn, stems from the fact that $A - B$ is positive semi-definite, then $B^{-1} - A^{-1}$ is positive semi-definite and $K_{Y^{obs}}^{-1} - \big(K_{Y^{obs}} + \Sigma\big)^{-1}$ is positive definite from the fact that $\big(K_{Y^{obs}} + \Sigma\big) - K_{Y^{obs}} = \Sigma$ is positive definite. The fact that Σ_{extrap} is positive semi-definite comes from the fact that it is the covariance matrix of a conditional multivariate normal.

The second statement is immediately clear from the definition of Σ_{stat} and the fourth statement can be shown using simple arithmetic.

We therefore turn our attention to the third statement. We first note that one reason why there could be a single $\hat{y} \in \mathcal{Y}\big(Y^{obs}\big)$ is if $\tilde{Y} = Y^{obs}$. We can think of this roughly as non-parametric identification in the classical sense, and from this assumption it follows that $K_{\tilde{Y}} = K_{Y^{obs}} = K_{\tilde{Y},Y^{obs}}$ and so $\Sigma_{extrap} = 0$. We can therefore turn our attention to the second reason why there could be a single $\hat{y} \in \mathcal{Y}\big(Y^{obs}\big)$, which corresponds roughly to the the case where the outcomes are identified parametrically. To write this in terms of our model, we can write our model – as discussed in Section III.C – by specifying that $\tau(\eta) = \phi_\tau(\eta)'\beta$ and $\mu(\eta) = \phi_\mu(\eta)'\alpha$ for some (unknown) parameters α and β and some (known) vectors ϕ_τ and ϕ_μ that are basis expansions of η, and a prior on α and β such that $k_\tau(\eta, \eta') = \phi(\eta)'\Sigma_\beta\phi(\eta')$ and $k_\mu(\eta, \eta') = \phi(\eta)'\Sigma_\alpha\phi(\eta')$. For notation, let Φ denote the stacked values of $\phi(\eta)$ that correspond to Y^{obs}, $\tilde{\Phi}$ denote the stacked values of $\phi(\eta)$ that correspond to $\tilde{Y}$, and Ω be the implied prior (e.g., a block diagonal with Σ_β and Σ_α as diagonals). Then, as discussed in Rasmussen and Williams (2006) among others, we can write:

$$K_{Y^{obs}} = \Psi'\Psi \tag{22}$$

$$K_{\tilde{Y}} = \tilde{\Psi}'\tilde{\Psi} \tag{23}$$

$$K_{\tilde{Y},Y^{obs}} = \tilde{\Psi}'\Psi \tag{24}$$

where $\Psi = \Omega^{1/2}\Phi$ and $\tilde{\Psi} = \Omega^{1/2}\tilde{\Phi}$. We can then write:

$$K_{\tilde{Y},Y^{obs}} K_{Y^{obs}}^{-1} K'_{\tilde{Y},Y^{obs}} = \big(\tilde{\Psi}'\Psi\big)\big(\Psi'\Psi\big)^{-1}\big(\tilde{\Psi}'\Psi\big)' \tag{25}$$

Next letting A^+ denote the Moore-Penrose inverse of A, we can re-write this as:

$$\begin{aligned} K_{\tilde{Y},Y^{obs}} K_{Y^{obs}}^{-1} K'_{\tilde{Y},Y^{obs}} &= \big(\tilde{\Psi}'\Psi\big)\big(\Psi^+\Psi^{+'}\big)\big(\tilde{\Psi}'\Psi\big)' && (26)\\ &= \tilde{\Psi}'\big(\Psi\Psi^+\big)\big(\Psi^{+'}\Psi'\big)\tilde{\Psi} && (27) \end{aligned}$$

The key step is to realize that if the outcomes are parametrically identified, then Ψ has full row rank. For example, if $\tau(\eta)$ and $\mu(\eta)$ are assumed to be polynomials of degree K, then Ψ having full row rank is equivalent to the assumption that $\nu(Z)$ takes at least $K+1$ values. This is important, because if Ψ has full row rank we get that $\Psi\Psi^+ = \Psi^+\Psi^{+'} = I$, where I is the identity matrix. Thus, τ being parametrically identified implies that:

$$K_{\tilde{Y},Y^{obs}} K_{Y^{obs}}^{-1} K'_{\tilde{Y},Y^{obs}} = \tilde{\Psi}'\tilde{\Psi} \tag{28}$$

Since $K_{\tilde{Y}}$ also equals $\tilde{\Psi}'\tilde{\Psi}$ – see Equation (23) – it immediately follows that $\Sigma_{extrap} = 0$. □

Remark 4. *Define Σ_{stat} as in Equation (16) and Σ_{freq} as in Equation (17). Then if $\Sigma \neq 0$, we have that $\Sigma_{stat} - \Sigma_{freq}$ is a positive definite matrix.*

Proof. Using the definitions of Σ_{stat} and Σ_{freq}, we can start by noting that:

$$\begin{aligned} &\Sigma_{stat} - \Sigma_{freq} = \\ &\quad K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}}^{-1} - (K_{Y^{obs}} + \Sigma)^{-1}\big)K'_{\tilde{Y},Y^{obs}} - K_{\tilde{Y},Y^{obs}}\big(K_{Y^{obs}} + \Sigma\big)^{-1}\Sigma\big(K_{Y^{obs}} + \Sigma\big)^{-1}K'_{\tilde{Y},Y^{obs}} = \\ &\quad K_{\tilde{Y},Y^{obs}}\left[\big(K_{Y^{obs}}^{-1} - (K_{Y^{obs}} + \Sigma)^{-1}\big) - \big(K_{Y^{obs}} + \Sigma\big)^{-1}\Sigma\big(K_{Y^{obs}} + \Sigma\big)^{-1}\right]K'_{\tilde{Y},Y^{obs}} \end{aligned}$$

Since $K_{\tilde{Y},Y^{obs}}$ has full column rank, it then follows that if:

$$\big(K_{Y^{obs}}^{-1} - (K_{Y^{obs}} + \Sigma)^{-1}\big) - \big(K_{Y^{obs}} + \Sigma\big)^{-1}\Sigma\big(K_{Y^{obs}} + \Sigma\big)^{-1}$$

is positive definite, then so is $\Sigma_{stat} - \Sigma_{freq}$. With some algebra, we get that:

$$\begin{aligned}\big(K_{Y^{obs}}^{-1} - (K_{Y^{obs}} + \Sigma)^{-1}\big) &= \Big[K_{Y^{obs}}^{-1}(K_{Y^{obs}} + \Sigma) - I\Big](K_{Y^{obs}} + \Sigma)^{-1}\\ &= \Big[I + K_{Y^{obs}}^{-1}\Sigma - I\Big](K_{Y^{obs}} + \Sigma)^{-1}\\ &= K_{Y^{obs}}^{-1}\Sigma(K_{Y^{obs}} + \Sigma)^{-1}\end{aligned}$$

where I corresponds to the identity matrix. We can therefore write:

$$\begin{aligned}&\big(K_{Y^{obs}}^{-1} - (K_{Y^{obs}} + \Sigma)^{-1}\big) - \big(K_{Y^{obs}} + \Sigma\big)^{-1}\Sigma\big(K_{Y^{obs}} + \Sigma\big)^{-1} =\\ &\qquad K_{Y^{obs}}^{-1}\Sigma(K_{Y^{obs}} + \Sigma)^{-1} - \big(K_{Y^{obs}} + \Sigma\big)^{-1}\Sigma(K_{Y^{obs}} + \Sigma)^{-1}\end{aligned}$$

We will then use the fact that if $B^{-1} - A^{-1}$ is positive definite, then $A - B$ is positive definite and so aim to show that:

$$\Big[\big(K_{Y^{obs}} + \Sigma\big)^{-1}\Sigma(K_{Y^{obs}} + \Sigma)^{-1}\Big]^{-1} - \Big[K_{Y^{obs}}^{-1}\Sigma(K_{Y^{obs}} + \Sigma)^{-1}\Big]^{-1}$$

is positive definite. Using the rules of matrix inversion and matrix multiplication, we can then get that:

$$\begin{aligned}\Big[\big(K_{Y^{obs}} + \Sigma\big)^{-1}\Sigma(K_{Y^{obs}} + \Sigma)^{-1}\Big]^{-1} - \Big[K_{Y^{obs}}^{-1}\Sigma(K_{Y^{obs}} + \Sigma)^{-1}\Big]^{-1} &=\\ \big(K_{Y^{obs}} + \Sigma\big)\Sigma^{-1}(K_{Y^{obs}} + \Sigma) - (K_{Y^{obs}} + \Sigma)\Sigma^{-1}K_{Y^{obs}} &=\\ \big(K_{Y^{obs}}\Sigma^{-1} + I\big)\big(K_{Y^{obs}} + \Sigma\big) - \big(K_{Y^{obs}}\Sigma^{-1} + I\big)K_{Y^{obs}} &=\\ \big(K_{Y^{obs}}\Sigma^{-1} + I\big)\Sigma &=\\ K_{Y^{obs}} + \Sigma&\end{aligned}$$

which is positive definite since both $K_{Y^{obs}}$ and Σ are positive definite.

□

Remark 5. *Suppose that Z_i is a binary instrument and that ϵ_i is independent across individuals. Then:*

- *Let $k_\tau(\eta, \eta'|\theta) = k_\mu(\eta, \eta'|\theta) = 1 + \eta\eta'$. Then the resulting Bayesian marginal treatment effect curve converges to the same marginal treatment effect function as a linear extrapolation of the observed moments, e.g., Kowalski (2023a) and Brinch et al. (2017).*

- *Let $k_\tau(\eta, \eta'|\theta) = k_\mu(\eta, \eta'|\theta) = 1 + \Phi^{-1}(\eta)\Phi^{-1}(\eta')$, where Φ^{-1} is the inverse of a standard normal CDF. Then the resulting Bayesian marginal treatment effect curve converges to the same marginal treatment effect function as a traditional Heckit, e.g., Kline and Walters (2019).*

Proof. Consider first the case in which $k_\tau(\eta, \eta'|\theta) = k_\mu(\eta, \eta'|\theta) = (1 + \eta\eta')$. Since this can be re-written as $[1, \eta] \cdot [1, \eta']'$, it follows from the discussion of how the Gaussian process can be thought of as a basis expansion of a linear model – which itself is illustrated clearly in Chapter 2.1 of Rasmussen and Williams (2006) – that this is equivalent to the model where $y_k = \alpha_k + \beta_k\eta_i + \epsilon$ where the error terms are normally distributed and the priors on α_k and β_k are both $N(0, 1)$. The fact that the estimates then converge to the same marginal treatment effect function as the linear extrapolation approach thus follows from the fact that Bayesian linear regression converges to the same parameter estimates as OLS regression. The proof is identical for the case in which $k_\tau(\eta, \eta'|\theta) = k_\mu(\eta, \eta'|\theta) = 1 + \Phi^{-1}(\eta)\Phi^{-1}(\eta')$. □

B Sketch of the Algorithm

We now describe the full algorithm we use to estimate the Bayesian MTEs, which we have outlined in psuedocode as Algorithm 1.

In the first step `Collapse Data` we calculate the average Y_i for every value of the instrument Z_i and every treatment status, i.e., to collapse the data into the observed moments. While the algorithm does not exactly require a discrete instrument, it assumes that there are multiple observations for each value of the instrument and for each treatment status because in the second step `Estimate Error`, it calculates $\hat{\Sigma}$ as a diagonal matrix with the diagonals equal to:

$$\hat{\Sigma}_{k,k} = \frac{1}{N_k} \sum_{\forall i \text{ s.t. } (Z_i,T_i)=(Z,T)} \Big(Y_i - \hat{\mathbb{E}}[Y_i|Z_i = Z, T_i = T]\Big)^2 \tag{29}$$

where N_k is the number of observations we observe with the relevant (Z, T) pair and $\hat{\mathbb{E}}$ is the empirical average. Note that this converges to the true Σ matrix under the assumption that the observations are all independent and as the number of observations increases to infinity for all observed (Z, T) pairs. Similarly, in the step where we `Estimate` $\nu(Z)$, we do so by calculating:

$$\hat{\nu}(Z) = \hat{\mathbb{E}}[T_i|Z_i = Z] \tag{30}$$

which agains converges to the true value of $\nu(Z)$ as the number of observations for each value of the instrument increases. Thus, while nothing about the method described above necessitates a discrete instrument, the specific algorithm described here requires many observations for every value of the instrument.[15]

Algorithm 1: Bayesian MTEs

```
Data: $Y, T, Z$
Output: $\mu_{\tilde{Y}}, \Sigma_{\tilde{Y}}$

Pseudocode:
    Collapse Data;
    Estimate $\Sigma$;
    Estimate $\nu(Z)$;
    if Empirical Bayes then
        Estimate Hyperparameters;
        Calculate K;
        Calculate $\mu_{\tilde{Y}}$ and $\Sigma_{\tilde{Y}}$;
    end
    if Full Bayes then
        Sample Vector of Potential Hyperparameters;
        for $i \leftarrow 1$ to $N_{samples}$ do
            Accept or Reject $\theta_i$;
            if Accept $\theta_i$ then
                Calculate K;
                Calculate $\mu_{\tilde{Y}}$ and $\Sigma_{\tilde{Y}}$;
                Save $\mu_{\tilde{Y}}$ and $\Sigma_{\tilde{Y}}$;
            end
        end
        Calculate Overall $\mu_{\tilde{Y}}$ and $\Sigma_{\tilde{Y}}$;
    end
    Transform $\mu_{\tilde{Y}}$ and $\Sigma_{\tilde{Y}}$ to $\hat{\tau}(\eta)$ and $Var(\hat{\tau}(\eta))$;
```

At this point, the algorithm branches depending on whether the user specifies

[15] An alternative approach is to treat the variance of the error term and the parameters of ν as an additional hyperparameters, which would allow for more general settings.

an *Empirical Bayes* or *Full Bayes* approach. If they opt for an *Empirical Bayes* approach, the algorithm continuous to `Estimate Hyperparameters` by maximizing $p(\theta|Y, T, Z)$ as described in the section above. Given these parameters, the algorithm then can `Calculate K`. To do so, it uses Equation (18) to calculate the covariances for μ and τ on a discrete grid of η values and then uses Remark 1 to transform these into the $K_{\tilde{Y}}$, $K_{\tilde{Y},Y^{obs}}$, and $K_{Y^{obs}}$, replacing the integrals implicit in Equation (8) with discrete approximations. Given the calculations of K – or really of $K_{\tilde{Y}}$, $K_{\tilde{Y},Y^{obs}}$, and $K_{Y^{obs}}$ – and $\hat{\Sigma}$, it then uses Equations (13) and (14) to `Calculate` $\mu_{\tilde{Y}}$ `and` $\Sigma_{\tilde{Y}}$.

If the user instead chooses the *Full Bayes* approach, the algorithm then uses simple rejection sampling scheme to sample potential hyperparameters. Specifically, to `Sample Vector of Potential Hyperparameters` it randomly draws 10,000 potential values of θ from the hyperprior distribution defined in the section above. For each potential vector of hyperparameters, it then determines whether to `Accept or Reject` θ_i by calculating the marginal likelihood for this value of θ_i using Equation (19), normalizing it by dividing the marginal likelihood by the maximum value of the marginal likelihood obtained over the 10,000 samples, and then accepting the value of θ if (and only if) the normalized marginal likelihood is greater than the value of a random variable sampled $U(0, 1)$.[16] Such a sampling scheme is a relatively inefficient way to sample from the posterior, but it is sufficient for this algorithm given the small number of hyperparameters.

For each accepted vector of hyperparameters, it then proceeds as in the *Empirical Bayes* approach, e.g., the first step is to `Calculate K` and then to `Calculate` $\mu_{\tilde{Y}}$ `and` $\Sigma_{\tilde{Y}}$ as described above. Unlike the *Empirical Bayes* approach, however, it then needs to conclude by combining each of the resulting estimates to `Calculate Overall` $\mu_{\tilde{Y}}$ `and` $\Sigma_{\tilde{Y}}$. For this, we can use the fact that the resulting posterior is a mixture of normally distributed random variables, which means there is a closed form solution for the mean and variance of the posterior. Specifically, the overall mean is the average of the posterior means for each value of θ, while the overall variance is the average posterior variance plus the variance of the posterior means.[17]

[16] To see that this approach works, note that we can draw from the hyperprior distribution $p(\theta)$ and want to draw from the hyperposterior, i.e., $p(\theta|Y, ZT)$. But since the hyperposterior is proportional to $p(Y|\theta, Z, T) \cdot p(\theta)$, we get that the ratio of the hyperposterior to the hyper prior is just the marginal likelihood, i.e., $\frac{p(\theta|Y,Z,T)}{p(\theta)} = p(Y|\theta, Z, T)$.

[17] Note, however, that a mixture of Gaussians is not itself normally distributed, so we would need to further simulate the draws if we wanted to determine the distribution of the posterior rather than

Regardless of whether the user specified to use an *Empirical Bayes* or *Full Bayes* approach, the algorithm concludes with a step to `Transform` $\mu_{\tilde{Y}}$ `and` $\Sigma_{\tilde{Y}}$ `to` $\hat{\tau}(\eta)$ `and` $Var(\hat{\tau}(\eta))$. This is done using Equation (4) and is necessary to transform the posteriors of $y_0(\eta)$ and $y_1(\eta)$ to the posterior of $\tau(\eta)$, which is generally the object of interest.

Finally, we note that the publicly available code uses the algorithm outlined above, but includes as its output some additional results, such as some of the graphs produced in Section IV.B and measures of the average treatment effect, always taker average treatment effect, never taker average treatment effect, and the local average treatment effect. It also allows the user to input their own hyperprior and to specify if they care only about extrapolation, statistical, or frequentist uncertainty. See https://github.com/isaacopper/BayesianMTEs for more information about the code.

C Including Covariates

In theory, it is simple to extend the model in Section II to include other covariates, i.e., we can simply include X_i in every conditional expectation and extend the GP covariances to be functions of both η and X. This, however, raises important implementation questions about how to handle the new covariance functions. Namely, do we assume that the functions $\tau(\eta, X)$ and $\mu(\eta, X)$ are separable in η and X or allow for some interaction? If some interactions are allowed, how much interaction is allowed and how does the choice of hyperpriors reflect these decisions? These questions are not easy to answer and we feel like the question of how best to extend the model to include covariates is worth much further exploration.

Rather than provide definitive answers on how to best include covariates, we focus here on a simple case to illustrate how covariates can be used to extend and estimate the model. To do so, we will again focus on the OHIE and consider an additional characteristic of each participant: whether they have used SNAP in the previous 12 months. We will then make the further assumption that $\tau(\eta, X)$ and $\mu(\eta, X)$ are

just the mean and variance.

separable in η and X, or formally that

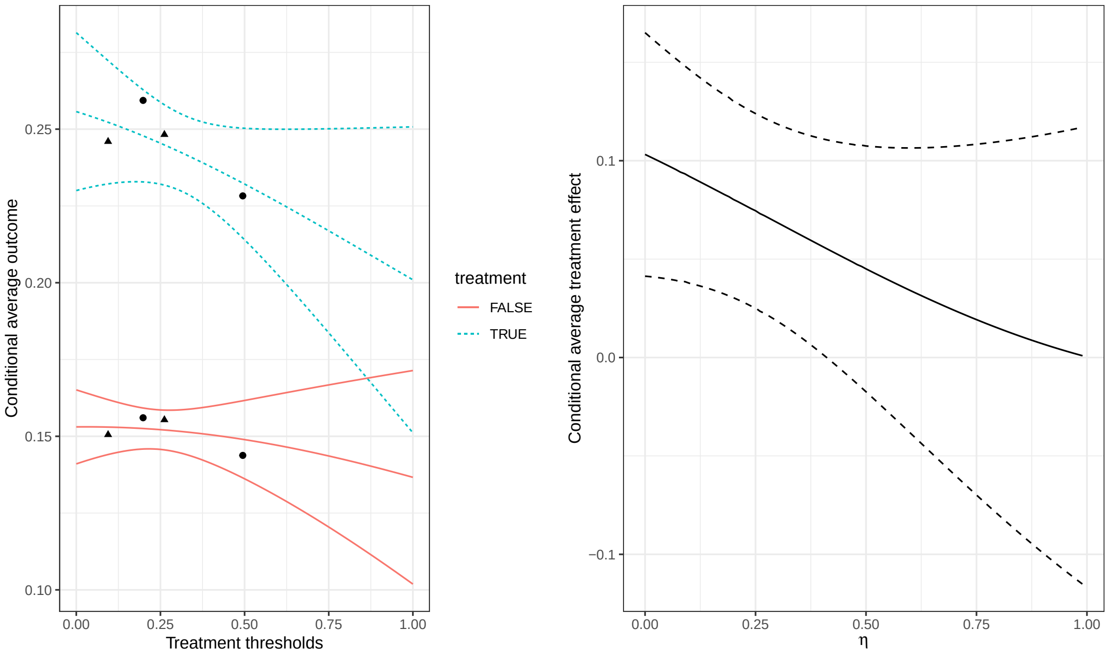
$$\text{Gaussian process for } \mu: \qquad \mu(\eta, X)|\theta_\mu \sim \mathcal{GP}\big(h_\mu(X), k_\mu(\eta, \eta'|\theta_\mu)\big) \tag{31}$$

$$\text{Gaussian process for } \tau: \qquad \tau(\eta, X)|\theta_\tau \sim \mathcal{GP}\big(h_\tau(X), k_\tau(\eta, \eta'|\theta_\tau)\big) \tag{32}$$

for known covariance functions (or kernels) $k_\mu(\eta, \eta'|\theta_\mu)$ and $k_\tau(\eta, \eta'|\theta_\tau)$ with hyperparameters θ_μ and θ_τ and mean functions of the covariates $h_\mu(X)$ and $h_\tau(X)$. We can then view the parameters of $h_\mu(X)$ and $h_\tau(X)$ as another set of hyperparameters, which we can estimate as part of the `Estimate Hyperparameters` step in the *Empirical Bayes* approach or to estimate them before the `Calculate K` step in the *Full Bayes* approach. The details of how to one can do so is laid out in Chapter 2 of Rasmussen and Williams (2006).

Beyond the fact that it is useful to understand how the treatment effects vary based on the covariates, the inclusion of covariates is also useful in that it aids in the estimation of $\tau(\eta)$ and $\mu(\eta)$. As illustrated in Figure 7, the inclusion of the covariate doubles the number of observed moments: e.g., we now observe for those assigned to the treatment and who enrolled in treatment for those who also enrolled in SNAP in the previous 12 months as well as for those who did not enroll in SNAP in the previous 12 months. This can be seen by the fact that in Figure 7a, there are four dots: two black circles representing those who had previously enrolled in SNAP and those who had not. (Note that in the figure, the observed moments are adjusted for mean differences between the triangles and dots, for both those who enroll in the treatment and those who do not.) As long as those who enrolled in SNAP in the previous 12 months had different probabilities of enrolling in the treatment than those who had not enrolled in SNAP, these additional data points can be used to determine the shape of the functions. As we discuss in Section E.A, the additional moments can also be used to help determine the value of the hyperparameters and lead to more similarity between the *Empirical Bayes* and *Full Bayes* approaches.

Figure 7: Posterior Mean and 95% CIs - With a Binary Covariate



(a) Conditional Moment Functions

(b) Marginal Treatment Effects

Note: Panel (a) shows the posterior mean and 95% credible interval of the function y_0 (in the red solid lines) and y_1 (in the blue dashed lines) as functions of η. The black dots represent the estimated moments observed in the data for those who enrolled in SNAP in the previous 12 months and the black triangles represent the estimated moments observed in the data for those who did not enroll in SNAP in the previous 12 months. Panel (b) shows the posterior distribution of the MTE function, i.e. $\tau(\eta)$, with the solid indicating the posterior mean and the dashed lines indicating the 95% credible interval. In both cases, we hold fixed the value of the covariate at the overall mean.

D Additional Results using the OHIE

D.A Results under Different Hyperparamters

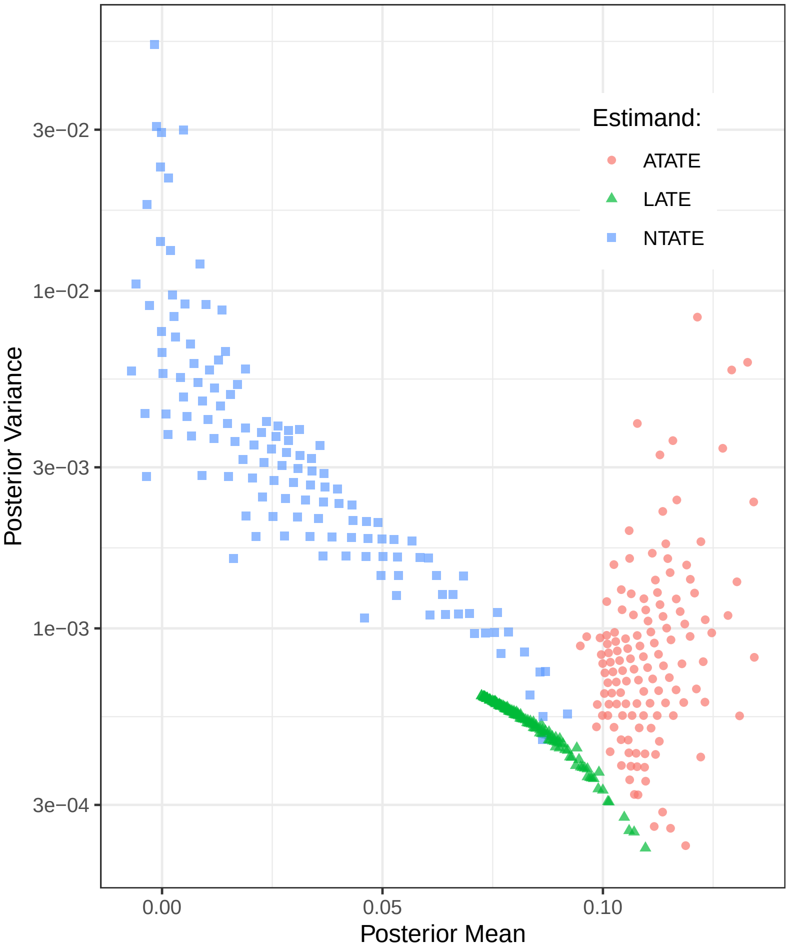
As discussed in the Section IV.B we cannot precisely estimate the hyperparameters – and particularly the lengthscales – using the observed data. Thus, our preferred approach is to integrate over the posterior distribution of the the hyperparameters. In practice, we do so via an accept/reject algorithm which gives a large number of plausible hyperparamters. For each hyperparameter, we use the method described above to estimate the mean and variance of the resulting posteriors. For our main results, we then combine these estimates by calculating the overall mean and variance of the full posterior; here, we further explore how the mean and variance of the posteriors depend on the hyperparameters.

We start by first showing the mean and variance of the Gaussian posterior of the ATATE, LATE, and NTATE for each hyperparameter drawn from the hyperposterior. The results are shown in Figure 8, which shows that the posterior mean and variance of the ATATE and NTATE depend more on the value of the hyperparameter than the LATE.

How the variances depend on the hyperparameters is illustrated in Figure 9a, which shows that over the range of our hyperprior the posterior variances are decreasing with the lengthscale.[18] The reasoning is clear when considering the random functions presented in Figure 1 – for large lengthscales, the only realistic functions are (nearly) linear ones and so the two observed points are sufficient to pin down the entire function, while for small lengthscales the functions could be highly non-linear and so there is large amounts of uncertainty about how the functions appear away from the observed points. Note that the relationship between the lengthscale and the posterior variances are not monotonic and for very short lengthscales the estimands tend to decrease, as illustrated in Figure 9b. This is because for short enough lengthscales the oscillations of MTE function are likely to average out when integrating over a range of η values, as we do for the never taker average treatment effect. It is also worth noting again that the assumed lengthscale does not have a meaningful effect on the variance of the LATE and only matters for the other estimands, all of which

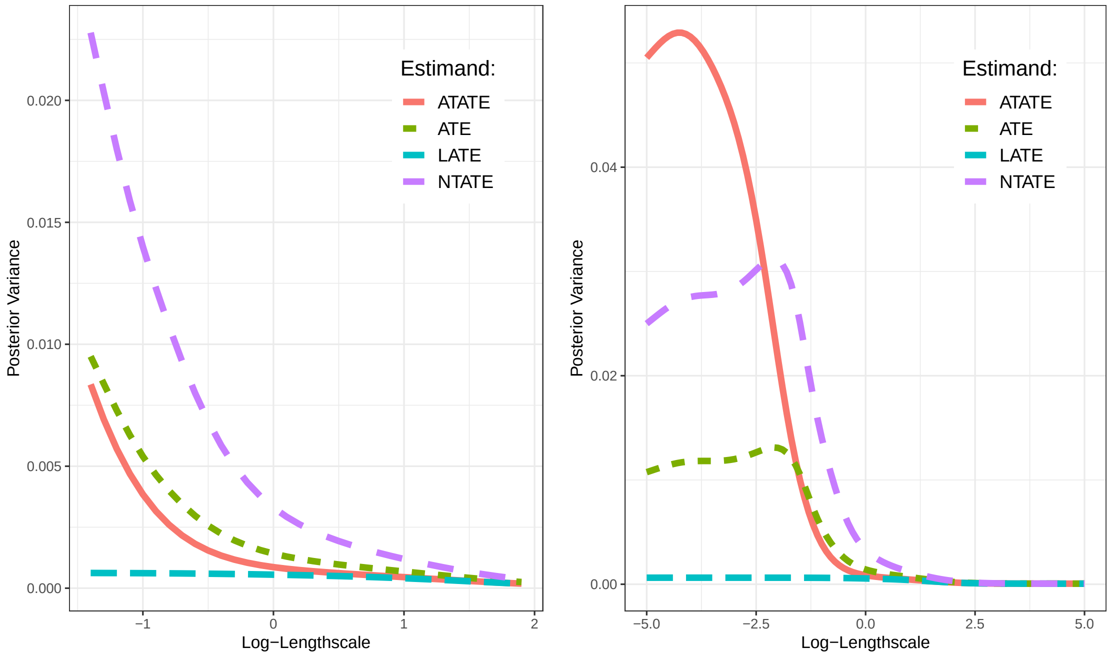
[18] As before, we show the results as a function of l by showing the posterior variance under the hyperparameters l and $\sigma^*(l)$ where $\sigma^*(l)$ is the value of σ that maximizes the hyperposterior for l, i.e., $\sigma^*(l) = \arg\max_\sigma p(l, \sigma | Y, T, Z)$.

Figure 8: Mean/Variance Estimates for Each Set of Hyperparemters Sampled



Note: Each mark shows the mean and variance of the posterior distribution for three of the estimates of interest – the always treated average treatment effect (ATATE); the complier average treatment effect (LATE); and the never treatment average treatment effect (NTATE) – with each dot representing the estimates for a particular set of hyperparmaters randomly drawn from the hyperposterior using an accept/reject algorithm. To keep the graph from being too cluttered, we do not include the overall average treatment effect (ATE); however, the ATE is simply a weighted average of the three estimands shown and so naturally the cluster of dots lies in between the three clusters shown.

Figure 9: Posterior Variances as a Function of the Lengthscales



(a) $log(l) \in (-2, 2)$

(b) $log(l) \in (-5, 5)$

Note: This figure shows the posterior variances of the four main estimates of interest as a function of the lengthscale indicated on the x-axis, with the value of σ being the one that maximizes the hyperposterior for l, i.e., $\sigma^*(l) = \arg\max_\sigma p(l, \sigma|Y, T, Z)$. The four estimands shown are: the always treated average treatment effect (ATATE), the average treatment effect (ATE); the complier average treatment effect (LATE); and the never treatment average treatment effect (NTATE). The two panels differ only in the range of l that is shown.

require some extrapolation.

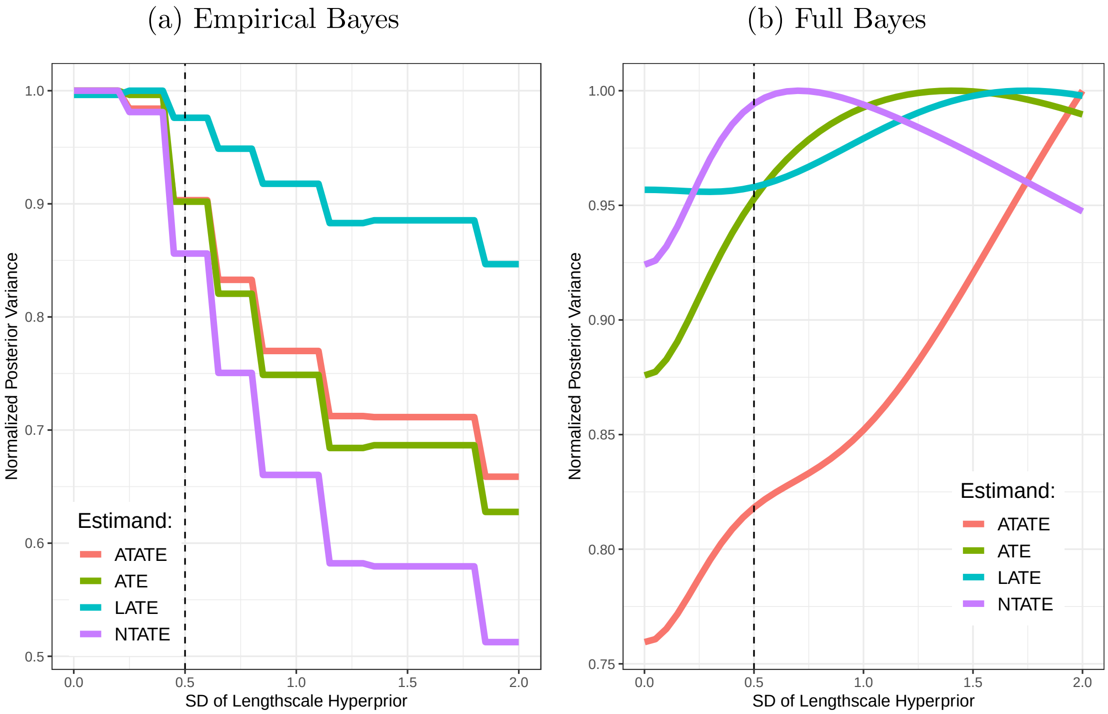
Finally, we can also look at how the assumed standard deviation of the lengthscale's hyperprior affects the posterior variance. As shown in Figure 10, the results depend in large part on whether one uses chooses to use an empirical Bayes or full Bayes approach. In Figure 10a, we see that increasing the standard deviation of the lengthscale's hypeprior tends to reduce the estimated posterior variance. In contrast, when using the full Bayes approach, the posterior variance tends to be maximed around a value of 1; however, in the full Bayes approach the prior SD of the lengthscale does not have a major impact on the estimated posterior variance. To see this, note that in both Figure 10a and Figure 10b, we normalize the posterior variance so it equals one for the largest estimated posterior variance. From the scale of the y-axis, we can see that the choice of the hyperprior's standard deviation doesn't affect the posterior variance by more than 10% for any of the estimands except for the always taker average treatment effect, which stems from the fact that in our empirical example the always takers are just a sliver of the population. In contrast, as implicit in the scale of the y-axis of Figure 10a, the assumed standard deviation of the hyperprior has a relatively large impact on the estimated posterior variance if one uses an empirical Bayes approach.

D.B Simulating the Asymptotic Uncertainty

In Section IV.B we show the mean and variance of the posterior distributions when the sample size is $1,000$ and when the sample size is $19,000$. In Figure 11, we show how the posterior variance for the four estimands of interest – the ATE, ATATE, LATE, and NTATE – adjusts as the sample size increases. To do so, we randomly sample with replacement from the original dataset, varying whether the number of observations sampled is: 100; 500; $1,000$; $5,000$; $10,000$; $100,000$; or $1,000,000$. To ensure that the results are not driven by idiosyncratcies in the random sample, we repeat the process 10 times for each sample size and show the average posterior variance over those 10 simulations.

As can be seen in Figure 11, the log posterior variance of the LATE scales linearly with the log sample size. This reflects that the LATE requires no extrapolation and so all the uncertainty stems from statistical uncertainty. Of course, from the law of large numbers the uncertainty in the moment averages is asymptotically proportional

Figure 10: Posterior Variances as a Function of the Lengthscale Hyperprior Standard Deviation

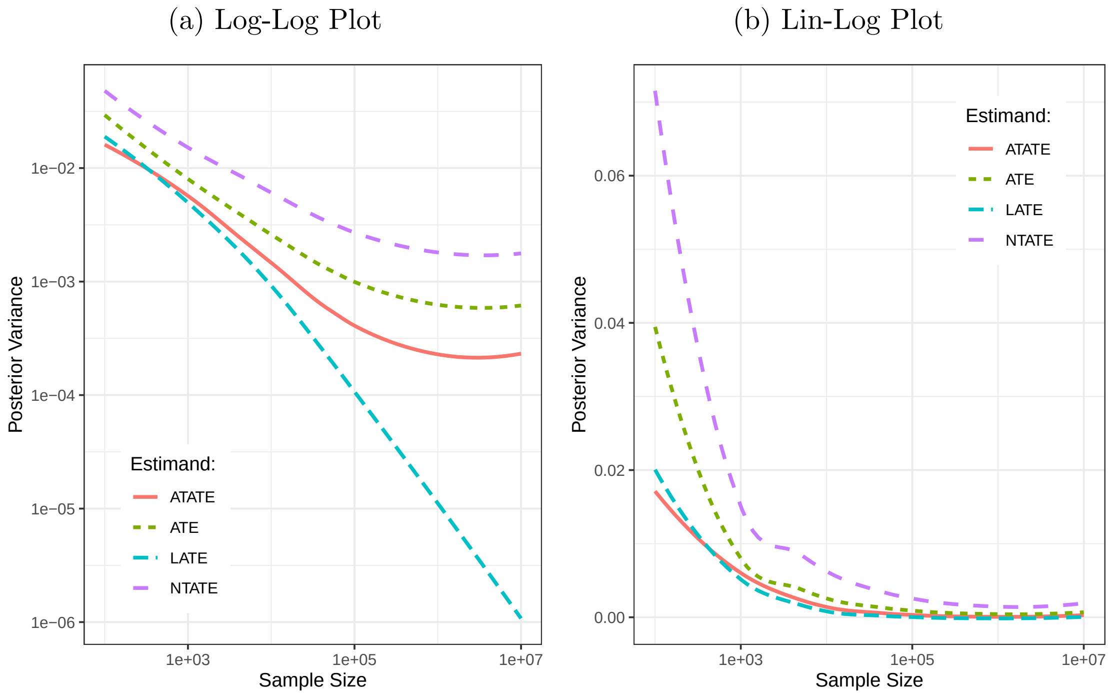


(a) Empirical Bayes

(b) Full Bayes

Note: This figure shows normalized posterior variances of the four main estimates of interest as a function of the assumed standard deviation of the hyperprior of the lengthscale; it uses the hyperprior of σ as defined in Section IV.A. We normalize the posterior variance by dividing the implied posterior variance by the maximum posterior variance for each estimand. The four estimands shown are: the always treated average treatment effect (ATATE), the average treatment effect (ATE); the complier average treatment effect (LATE); and the never treatment average treatment effect (NTATE). The vertical black dashed line corresponds to the standard deviation used for the main empirical results. The two panels differ on whether one uses an empirical Bayes approach or a full Bayes approach, as discussed in Section B.

Figure 11: Asymptotic Variances


(a) Log-Log Plot

(b) Lin-Log Plot

Note: This figure shows the variance of the posterior distribution for the four main estimates of interest and how that varies with the sample size. Both panels show the same data and simulations, with the only difference being that the y-axis on panel (a) uses a log-scale. The four estimands shown are: the always treated average treatment effect (ATATE), the average treatment effect (ATE); the complier average treatment effect (LATE); and the never treatment average treatment effect (NTATE).

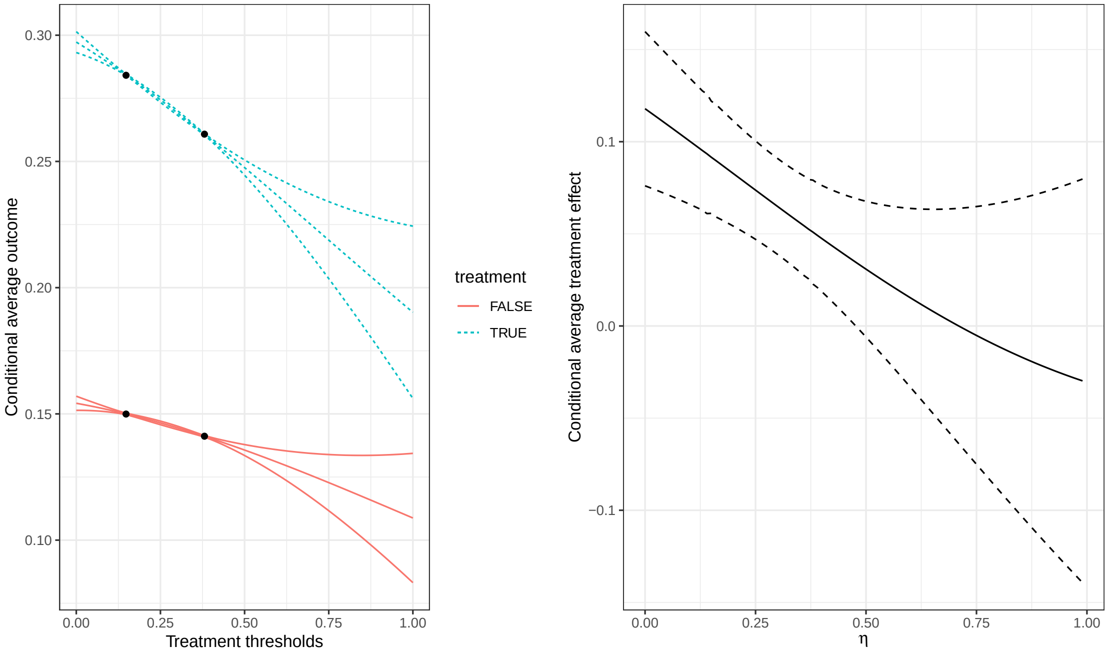
to the inverse of the sample size, hence the log posterior variance of the LATE is proportional to the log of the sample size. In contrast, the other averages do require extrapolation and so the linear relationship between the log posterior variance and the log sample size no longer holds. Instead, the posterior variances asymptotes to some value above zero; this value is the extrapolation uncertainty we define in Section III.B. As seen in Figure 11, however, for most reasonable sample sizes, uncertainty in the true values of the observed moments is large enough that increasing the sample size meaningfully reduces the posterior variance.

We can also consider the asymptotic behavior in more detail by studying the extrapolation uncertainty, as defined in Section III.B. To do so, we reproduce Figure 4 while using only the extrapolation uncertainty instead of total uncertainty, which provides a window into how the resulting estimates asymptote. (This assumes that

Figure 12: Posterior Mean and 95% CIs – Extrapolation Uncertainty Only

(a) Conditional Moment Functions

(b) Marginal Treatment Effects



Note: Here, we reproduce Figure 4, while ignoring uncertainty in the estimated moments. The Panel (a) shows the posterior mean and 95% credible interval of the function y_0 (in the red solid lines) and y_1 (in the blue dashed lines). The black dots represent the estimated moments observed in the data. Panel (b) shows the posterior distribution of the MTE function, i.e. $\tau(\eta)$, with the solid indicating the posterior mean and the dashed lines indicating the 95% credible interval.

the observed moments remain the same as more data is added.) The results are illustrated in Figure 12 below.

Finally, we can better understand how the method handles changing sample sizes by comparing the MTE estimates that result from the proposed estimator to those that result from using a linear extrapolation or a Heckit model. As seen by comparing the three panels of Figure 13, when the sample size is small the Bayesian approach results in a flatter MTE function than the other two approaches (due to what is essentially Bayesian shrinkage), which in turn results in posterior standard deviations that are smaller than the standard errors of the other approaches. When the sample size is large, in contrast, the fact that the Bayesian approach includes uncertainty in how one should extrapolate away from the observed moments means that the posterior standard deviations are larger than the standard errors of the other approaches.

E Simulations

E.A An Example of a Continuous Instrument with Limited Support

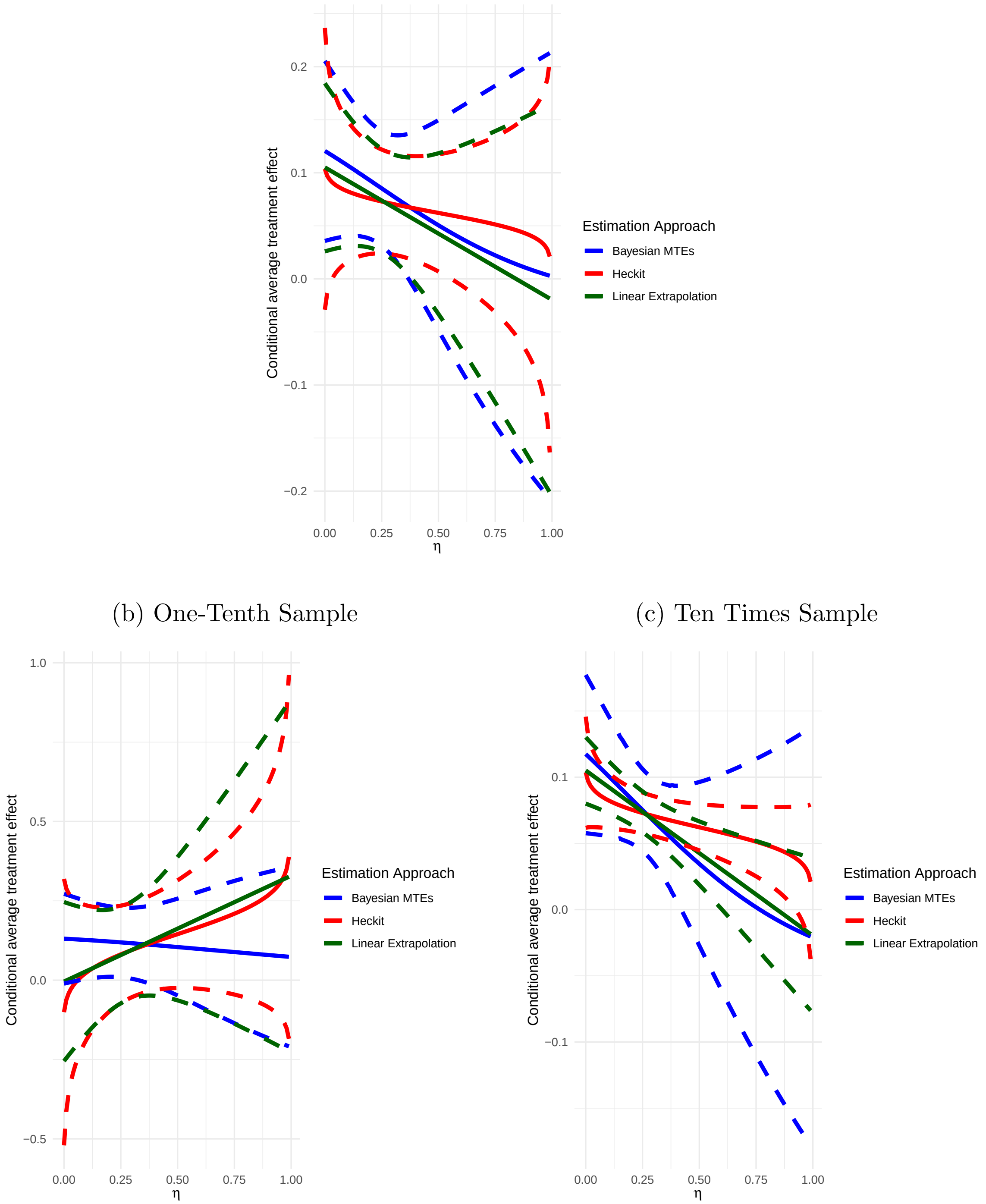
In our main example, we consider the case in which there is a single binary instrument. The method can also be used in cases with a continuous (or near continuous) instruments. For example, one can easily imagine applying it to judge/examiner IV designs. Here, we often find that there are a large number of cases in which all the judges/examiners agree, with a limited set of cases in which the decision depends on which judges is assigned the case. In our model, we can express this as the researcher observing $y_1(\eta)$ and $y_0(\eta)$ at a large number of points, all of which fall in some range $(\underline{\eta}, \overline{\eta}) \in (0, 1)$. We can therefore use the Bayesian MTE model to derive estimates of what the impact would be if, for example, we made the most lenient judge even more lenient or the most strict judge even more strict. This would be important if we were considering a major policy change, for example, that would change the behavior of all judges or examiners.

To give a sense of how the model would handle such a case, we conducted a simple simulation in which individuals are randomly assigned to one of twenty judges. The judges have different levels on leniency, but all choose to convict anywhere from 20% to 40% of cases, e.g., $\nu(Z_i) \in \{0.20, 0.21, 0.22, ..., 0.38, 0.39\}$. We then determine $\tau(\eta)$ and $\mu(\eta)$ by drawing from a mean-zero Gaussian process with a squared-exponential covariance function with $log(l) = 0$ and $log(\sigma) = -.2$. Given the two functions, each individuals' outcome is then generated according to Section II.A, with $\epsilon_i \sim N(0, 1)$.

The results are shown in Figure 14, which highlight that the method can be used in instances where the instrument is not a single binary instrument. In fact, as we discuss in the next subsection, in many ways it is easier to apply the model in this can: with a continuous instrument it is much easier to estimate the hyperparameters of the Gaussian process than when there is a single binary instrument. This also means that the the decision of whether to use an empirical Bayes or full Bayes approach is less important where there is a (nearly) continuous instrument than when there is a binary instrument.

Figure 13: Bayesian MTEs vs. Linear Extrapolation vs. Heckit

(a) Full OHIE Sample (Sample Size: 17,000)

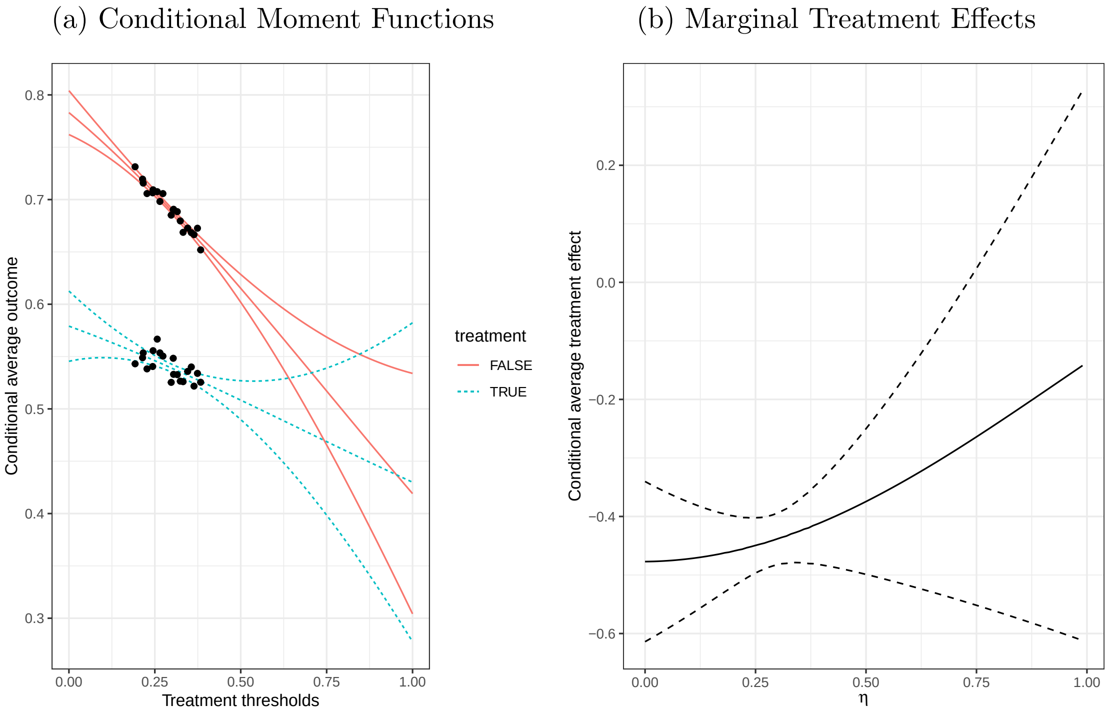


(b) One-Tenth Sample

(c) Ten Times Sample

Note: This table shows the mean and standard deviation of three different estimation approaches: the Bayesian MTE approach developed here, a linear extrapolation approach (e.g., Brinch et al. (2017) and Kowalski (2023a)), and a Heckit model (e.g., Kline and Walters (2019)). For the Bayesian MTEs, the solid line corresponds to the mean of the posterior and dashed lines correspond to a 95% credible interval; for the other two approaches, the solid line corresponds to the point estimate and the dashed line to a 95% confidence interval.

Figure 14: Posterior Mean and 95% CIs – Continuous Instrument with Limited Support



(a) Conditional Moment Functions

(b) Marginal Treatment Effects

Note: Here, we reproduce Figure 4 using a simulation in which individuals are randomly assigned to one of two instruments (e.g., judges or examiners) with $\nu(Z_i) \in \{0.20, 0.21, 0.22, ..., 0.38, 0.39\}$. Panel (a) shows the posterior mean and 95% credible interval of the function y_0 (in the red solid lines) and y_1 (in the blue dashed lines). The black dots represent the estimated moments observed in the data. Panel (b) shows the posterior distribution of the MTE function, i.e. $\tau(\eta)$, with the solid indicating the posterior mean and the dashed lines indicating the 95% credible interval.

That said, we should also note that most judge IV designs include a range of controls in their preferred specifications, especially a number of fixed effects that remove potentially endogenous sorting of cases to districts, time slots, etc. As mentioned in the conclusion, it is theoretically simple to extend the model in Section II to include other covariates, but doing so raises some additional implementation questions that we do not address in this paper. In addition, as discussed briefly in the paper the algorithm outlined in Section B assumes that the function $\nu(Z_i)$ and the matrix Σ are known with enough precision that we can treat them as known; such assumptions are often plausible in cases with a binary, but more suspect in cases where the instrument (or instruments) are not binary. As discussed briefly in the conclusion, we view these as an important avenues for future research, but will leave it for other papers to explore in depth.

E.B Estimating the Hyperparameters

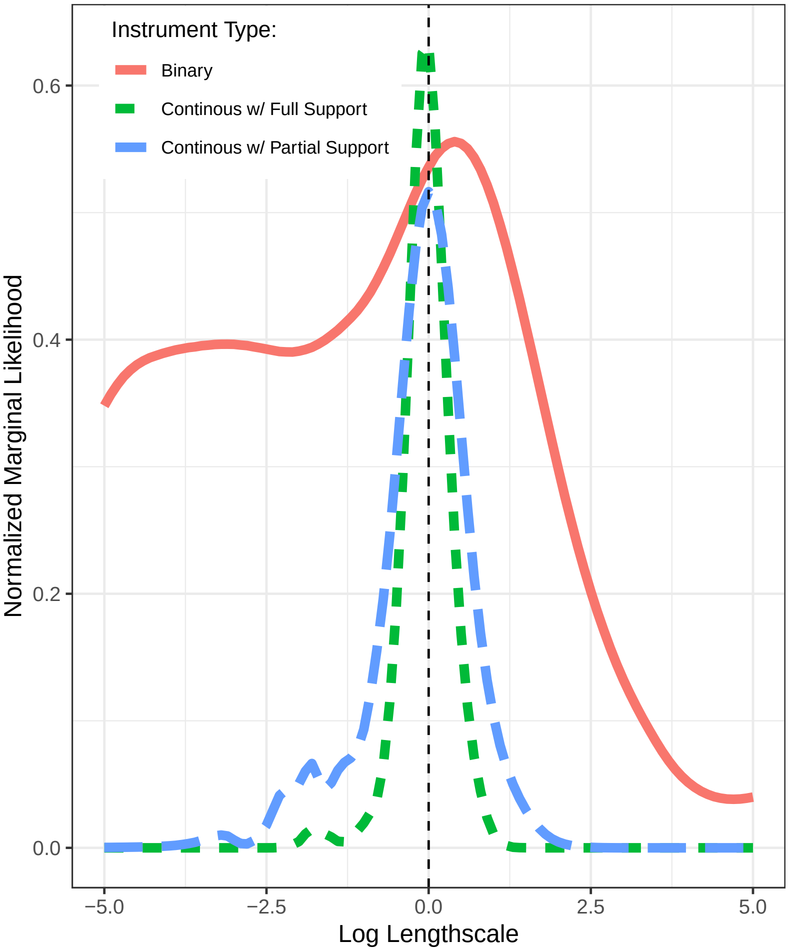
As we discuss in Section B an important decision for researchers is whether to use an empirical Bayes or full Bayes approach. The results in Figure 2, and in particular the fact that the marginal likelihood of l in the OHIE example, suggest that the lengthscale is not precisely estimated. Interpreting this result is a bit challenging, however, since we do not know the "true" lengthscale to judge the results again. Here we further explore estimation of the lengthscale via a simulation, in which we can judge the marginal likelihood against the true lengthscale we use in the simulations.

In this simulation, we assume that $\tau(\eta)$ and $\mu(\eta)$ are generated via a mean-zero Gaussian process with a squared-exponential covariance function with $log(l) = 0$ and $log(\sigma) = -.2$. For each of 100 simulations, we draw new functions $\tau(\eta)$ and $\mu(\eta)$, randomly assign individuals to the instrument, and then assign individuals to the treatment depending on their (randomly determined) value of η_i and which instrument they were assigned to. Each of the $100,000$ individuals' outcome is then generated according to Section II.A, with $\epsilon_i \sim N(0, 0.1)$, so there is little uncertainty in the true values of the observed moments.

We then consider three types of instruments: a binary instrument with $\nu(Z_i) \in \{.25, .75\}$; a "continuous instrument with partial support" in which $\nu(Z_i) \in \{0.1, 0.15, ..., 0.5, 0.55\}$; and a "continuous instrument with full support" in which $\nu(Z_i) \in \{0.025, 0.075, ..., 0.925, 0.975\}$. For each simulation, we compute the marginal likelihood for a range of values of l and

σ, plot the marginal likelihood as a function of l by choosing the value of σ for each l with the highest marginal likelihood, and then normalize the estimated marginal likelihoods so that the highest value is equal to one. We then average this marginal likelihood over all of the simulations, which we show in Figure 15. As can be seen, when there is a binary instrument the marginal likelihood is quite flat for a range of lengthscales; in contrast, when one has a nearly continuous instrument – and especially one with a large range of support – the lengthscale is more precisely estimated. This implies both that an empirical Bayes approach is likely to give similar results as the full Bayes approach and that the specification of the hyperprior will have a smaller impact on the resulting estimates when there is a continuous instrument than when there is a binary one.

Figure 15: Estimating the Lengthscale



Note: To illustrate the results on a two-dimensional graph, we show the marginal likelihood as a function of l by calculating $p(Y|l, Z, T) = \max_{\sigma} p(Y|l, \sigma, Z, T)$ for each of the simulations. We then normalize the calculations by dividing the resulting marginal likelihood by the maximum value of the marginal likelihood and the plot the average value and the plot the average value over the 100 simulations. The instruments are defined as follows: the binary instrument has $\nu(Z_i) \in \{.25, .75\}$; the continuous instrument with partial support has $\nu(Z_i) \in \{0.1, 0.15, ..., 0.5, 0.55\}$; and the continuous instrument with full support has $\nu(Z_i) \in \{0.025, 0.075, ..., 0.925, 0.975\}$.