From LATE to ATE

A Bayesian Approach

Isaac M. Opper

RAND Education and Labor

WR-A2499-1
November 2022

RAND working papers are intended to share researchers’ latest findings and to solicit informal peer review. They have been approved for circulation by RAND Education and Labor but have not been formally edited or peer reviewed. Unless otherwise indicated, working papers can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. RAND’s publications do not necessarily reflect the opinions of its research clients and sponsors. RAND® is a registered trademark.
From LATE to ATE:
A Bayesian Approach

Isaac M. Opper†

November 7, 2022

Abstract
We develop a Bayesian model that produces a posterior distribution of the marginal treatment effect (MTE) function. The method can be used even when the MTEs are not identified – as is the case in RCTs with imperfect compliance – thereby producing posterior distributions of unidentified estimands such as the overall average treatment effect (ATE) or the average effect on the always takers. While we focus on the case of RCTs with imperfect compliance, the model is general and allows for non-binary instruments and/or additional exogenous variables. Using the model, we show for the Oregon Health Insurance Experiment that the main source of the uncertainty in the ATE is not uncertainty due to the non-identifiability of the full MTE function and instead traditional statistical uncertainty, i.e., uncertainty in the true values of the observed moments due to the finite sample-size.

* I’ve very thankful to Aarya Suryavanshi, who took the lead in developing the R package, and to Joshua Eagan who helped QA the code. The code review is currently ongoing, but an R package to implement the method will soon be available at github.com/RANDCorporation. 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. Funding for this research was provided by gifts from RAND supporters and income from operations.
† RAND Corporation. Email: iopper@rand.org.
I Introduction

Imperfect compliance is pervasive in randomized control trials (RCTs): some individuals assigned to the treatment group will inevitably not follow through with the treatment and some individuals assigned to the control group often find a way to receive the treatment despite their initial assignment. A key question in the evaluation of such RCTs is therefore how to handle such imperfect compliance and there are, unsurprisingly, multiple approaches that applied researchers use. Perhaps the most commonly used approach is to use the treatment assignment as an instrument for treatment status and employ an instrumental variables (IV) design.

One downside of the IV design is that – even when the identification assumptions are all met – 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). Local in this context refers to the set of individuals who treatment status is impacted by instrument; in an RCT with imperfect compliance, for example, that means that the resulting estimates represent the average treatment effect on the set of individuals that will enroll in the treatment if and only if they are assigned to it.

In many cases, the LATE is precisely the main parameter of interest. In the RCT it is, for example, the average effect of the treatment assignment on the set of individuals who are affected at all. Even in these cases, however, it is rarely the only parameter of interest (Heckman and Vytlacil, 2001). It would also be valuable, for example, 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, with a binary instrument these averages are not identified without additional strong assumptions on how individuals select into treatment (e.g., Brinch et al. (2017); Kowalski (Forthcoming)), which has led to a large literature on ways to estimate bounds on the estimands of interest (e.g., Manski (1990); Balke and Pearl (1997); Bhattacharya et al. (2008); Mogstad et al. (2018)).

Rather than attempt to identify or bound the estimands, we develop a Bayesian model in this paper. Like many existing approaches, we start with a generalized Roy model in which individuals have different propensities to self-select into treatment.\footnote{Other recent work that uses a Bayesian model in the RCT context has focused on combining estimates from multiple contexts or estimating distributional effects, rather than extrapolating from the LATE to ATE, e.g., Meager (2019, 2022); Gechter and Meager (2022).}
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). We then use the Gaussian process model to place a prior distribution over these functions which captures the belief – prevalent in data science and machine learning – that smooth functions are more likely than functions that oscillate wildly.

We then show that the generalized Roy model, the Gaussian process statistical model, and the observed moments can be combined in a straightforward way to output a posterior distribution of the MTE function and therefore most estimands of interest. After specifying the general model we shift our focus to the case where there is a single binary instrument and no additional covariates, as is the case in RCTs with imperfect compliance. We illustrate how the hyperparameters of the model govern the prior distribution over functions and specify our preferred hyperprior, i.e., prior distribution over these hyperparameters.

Having specified the model, we use data from the Oregon Health Insurance Experiment (OHIE) to estimate the model. One advantage of the model is that it allows us to explicitly quantify how uncertain the estimates are for traditionally unidentified treatment effects, such as the average treatment effect (ATE), always taker average treatment effect (AT ATE), and never taker average treatment effect (NT ATE). Interestingly, we find that posterior standard deviation of the (unidentified) ATE is only moderately larger than the posterior standard deviation of the (identified) LATE, despite the uncertain extrapolation required to calculate the ATE.

Motivated by this, we define “statistical uncertainty” as the uncertainty in the estimates due solely to uncertainty in the true value of the observed moments due to the finite sample size and “extrapolation uncertainty” as the uncertainty in the estimates due solely to uncertainty in how one should extrapolate away from the observed moments to the inframarginal individuals. We show that under our preferred hyperprior the extrapolation uncertainty is consistently less than the statistical uncertainty, even in an RCT as large as the OHIE.
II Building Blocks

We start by describing the two building blocks of our approach: the generalized Roy model and Gaussian Processes. Both have been studied extensively and our goal in this section is to ensure the reader starts with the necessary background and to clarify our notation, rather than 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. In the next section, we then discuss how the two building blocks can be combined to easily compute Bayesian posteriors of the marginal treatment effect function.

II.A Generalized Roy Model

We consider the effect of a binary treatment on a single outcome, estimated via a binary instrument. 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 $\mu_i$, $\tau_i$, $\eta_i$, respectively. In other words, we use $\mu_i$ to denote individual $i$’s outcome in the absence of treatment and $\tau_i$ 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, an instrument, and (potentially) a set of additional exogenous covariates; we denote these as, $Y_i$, $T_i$, $Z_i$, and $X_i$, respectively. 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 $\eta_i$, $Z_i$, and $X_i$ to treatment status via a threshold-crossing representation, i.e.,:

$$T_i = 1(\nu(Z_i, X_i) \geq \eta_i)$$

for some (unknown) function of the instrument and covariates $\nu(Z_i, X_i)$ As is common, we will assume that $\eta_i$ is continuously distributed conditional on $X_i$, which means without loss of generality we can normalize this distribution to be uniform between zero and one. Note that individuals with higher $\eta_i$ are less likely to enroll in the treatment, i.e., have a higher implied cost of enrolling in the treatment or equivalently
have a lower propensity to enroll in the treatment. In doing so, it then follows that
\( \nu(Z_i, X_i) = \mathbb{E}[T_i | Z_i, X_i] \), which hints at how we can estimate \( \nu \). Finally, we will
assume that \((\mu_i, \tau_i, \eta_i) \perp \perp Z_i | X_i\). 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.\(^2\)

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

\[ \tau(\eta, X) = \mathbb{E}[\tau_i | \eta_i = \eta, X_i = X] \quad \text{and} \quad \mu(\eta, X) = \mathbb{E}[\mu_i | \eta_i = \eta, X_i = X] \] (2)

The first function \( \tau(\eta, X) \), 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 \( \eta \) and \( X \), we do not need
to condition on \( Z \). More specifically, the assumption is that \( \mathbb{E}[\tau_i | \eta_i = \eta, X_i = X] = \mathbb{E}[\tau_i | \eta_i = \eta, X_i = X, Z_i = Z] \) for all \( Z \), with a similar expression for \( \mu_i \).

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

\[ m_0(\eta, X) = \mathbb{E}[\mu_i | \eta_i > \eta, X_i = X] \quad \text{and} \quad m_1(\eta, X) = \mathbb{E}[\mu_i + \tau_i | \eta_i \leq \eta, X_i = X] \] (3)

Defining four sets of conditional moments is redundant, in that the functions defined
in Equation (2) would imply the value of the functions in Equation (3) and vice versa.
Most useful for our purposes, we can relate \( m_0 \) and \( m_1 \) to \( \tau \) as follows:

\[ \mathbb{E}[\tau(\eta, X) | \eta \in (\bar{\eta}, \bar{\eta} + \Delta), X_i = X] = \frac{1}{\Delta} \cdot \begin{bmatrix} \eta + \Delta \\ \bar{\eta} \\ 1 - (\bar{\eta} + \Delta) \\ -(1 - \bar{\eta}) \end{bmatrix} \cdot \begin{bmatrix} m_1(\eta + \Delta, X) \\ m_1(\bar{\eta}, X) \\ m_0(\bar{\eta} + \Delta, X) \\ m_0(\bar{\eta}, X) \end{bmatrix} \] (4)

Finally, note that these functions differ from both the functions above and the
marginal treatment responses \( m_k(u, X) \) in Mogstad et al. (2018), in that here \( m_0(\eta, X) \)
conditions on \( \eta_i \) being larger than \( \eta \) rather than equal to \( \eta \). Also of note, while
\( m_0(\eta, X) \) conditions on \( \eta_i \) being larger than \( \eta \), \( m_1(\eta, X) \) conditions on \( \eta_i \) being smaller

\(^2\)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.
than \( \eta \). The reasons for this peculiar conditioning is that it makes \( m_0 \) and \( m_1 \) more directly reflect the values we actually observe in the data, which – as we will explain below – will be quite useful.

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 \( \tau(\eta,X) \) 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 \( \mathcal{T} \). Of course, any function \( \tau \) is simply defined by its value at every point \((\eta,X)\) in its domain and so an equivalent formulation is to consider each point \( \tau(\eta,X) \) 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 \( \tau(\eta,X) \). A consequence of this is that defining how the random variables \( \tau(\eta,X) \) themselves co-vary is an alternative way of defining how likely it is to draw a particular function \( \tau \), 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 notation, let \( x = (\eta,X) \) and then denote as \( m(x) \) and \( k(x,x') \) 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 \( m(x) \) and \( k(x,x') \), 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 \( y_i = f(x_i) \). Then from the definition of a GP, it clearly follows that:

\[
f(x')|y_i = N\left(\mu, \sigma^2\right) \quad \text{with} \quad \mu = m(x') + \frac{k(x_i, x')}{k(x_i, x)}(f(x_i) - m(x_i)) \quad \text{and} \quad \sigma^2 = k(x', x') - \frac{k(x_i, x')^2}{k(x_i, x)}
\]

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 \( f \).

### III Bayesian Posterior of the MTEs

#### III.A Bayesian Hierarchical Model

We now combine the two building blocks to specify the full Bayesian hierarchical model. For simplicity, we focus here on the case without any additional \( X_i \) covariates.

We start with by defining the middle level of a Bayesian hierarchical model, which specifies that the functions \( \mu(\eta) \) and \( \tau(\eta) \) follow a Gaussian process. Specifically, we have that:

\[
\begin{align*}
\text{Gaussian process for } \mu: & \quad \mu(\eta)|\theta_\mu \sim \mathcal{GP}(0, k_\mu(\eta, \eta'|\theta_\mu)) \\
\text{Gaussian process for } \tau: & \quad \tau(\eta)|\theta_\tau \sim \mathcal{GP}(0, k_\tau(\eta, \eta'|\theta_\tau))
\end{align*}
\]

for (known) covariance functions \( k_\mu(\eta, \eta'|\theta_\mu) \) and \( k_\tau(\eta, \eta'|\theta_\tau) \) with hyperparameters \( \theta_\mu \) and \( \theta_\tau \). For the covariance functions, we will use the squared exponential covariance term, in which:

\[
k_\mu(\eta, \eta'|\theta_\mu) = \sigma_{\mu}^2 \exp\left(-\frac{(\eta - \eta')^2}{2l_{\mu}^2}\right) \quad \text{and} \quad k_\tau(\eta, \eta'|\theta_\tau) = \sigma_{\tau}^2 \exp\left(-\frac{(\eta - \eta')^2}{2l_{\tau}^2}\right)
\]

This is a common choice when modeling Gaussian processes. Each covariate function has two hyperparameters: \( \sigma^2 \) which controls the amplitude and \( l \) which is referred to as the lengthscale. We discuss these in more detail in the next subsection.

As discussed above, to generate posterior predictions we have to grapple with the fact that we do not observe \( \tau(\eta) \) and \( \mu(\eta) \) and instead observe \( m_0(\eta) \) and \( m_1(\eta) \). Luckily, if \( \tau(\eta) \) and \( \mu(\eta) \) are both GPs, then together \( m_0(\eta) \) and \( m_1(\eta) \) form one
III.A Bayesian Hierarchical Model

large GP. Specifically, define $m_1(\eta)$ and $m_0(\eta)$ as in Equation 3 and let:

$$m(t, \eta) = tm_1(\eta) + (1-t)m_0(\eta) \quad (8)$$

for $t \in \{0, 1\}$. Then $m$ is a Gaussian process with mean function of zero and a known covariance function – denoted $k_m$ – which depends on $k_\mu(\eta, \eta'|\theta_\mu)$ and $k_\tau(\eta, \eta'|\theta_\tau).$³

Specifically, as we prove in the appendix, we get that:

$$k_m((t, \eta), (t', \eta'))|\theta_\mu, \theta_\tau) = 
\begin{cases} 
\mathbb{E}[k_\mu(\bar{\eta}, \bar{\eta}'|\theta_\mu) + k_\tau(\bar{\eta}, \bar{\eta}'|\theta_\tau)|\bar{\eta} \leq \eta, \bar{\eta}' \leq \eta'] & \text{if } t = t' = 1 \\
\mathbb{E}[k_\mu(\bar{\eta}, \bar{\eta}'|\theta_\mu)|\bar{\eta} > \eta, \bar{\eta}' > \eta'] & \text{if } t = t' = 0 \\
\mathbb{E}[k_\mu(\bar{\eta}, \bar{\eta}'|\theta_\mu)|\bar{\eta} > \eta, \bar{\eta}' \leq \eta'] & \text{if } t = 0 \neq t' \\
\mathbb{E}[k_\mu(\bar{\eta}, \bar{\eta}'|\theta_\mu)|\bar{\eta} \leq \eta, \bar{\eta}' > \eta'] & \text{if } t = 1 \neq t' 
\end{cases} \quad (9)$$

We further assume that the outcomes also contain a normally distributed error term. Specifically, $Y_i = m(t, \eta) + \epsilon_i$ and letting $\epsilon$ be the vector of all individuals’ error terms we get that $\epsilon \sim N(0, \Sigma)$ for some positive semi-definite matrix $\Sigma$. It is common to assume that the error term is distributed i.i.d., in which case $\Sigma = \sigma^2 I$ where $I$ 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.

We next define $\nu$ as the function that maps the value of the instrumental variable ($Z_i$) to the cutoffs. In the case where the instrumental variable is binary, this can be expressed simply by a two-dimensional vector – i.e., $\nu = (\eta_0, \eta_1)$ – but one could imagine in more general cases $\nu$ is instead an $N$-dimensional vector or infinite-dimensional function. It then follows that: $T_i|\nu, Z_i \sim \text{Bernoulli}(\nu(Z_i))$.

Finally, to complete the Bayesian hierarchical model, we also must specify the

³On a very technical aside, we can infer $\mu(\eta)$ and $\tau(\eta)$ are almost surely measurable – and therefore that $\tilde{m}$ is well-defined – by invoking the Kolmogorov continuity theorem to show that under our covariance assumption they are sample-continuous processes. This also allows us to use the Riemann definition of the integral, which simplifies the proof.
hyperpriors, i.e., the priors over the hyperparameters. These are specified as follows:

\begin{align}
\text{Hyperprior over kernel hyperparameters:} & \quad \theta \sim p_{\theta}(\theta) \quad (10) \\
\text{Hyperprior over cutoffs:} & \quad \nu \sim p_{\eta}(\eta) \quad (11) \\
\text{Hyperprior over error variance:} & \quad \Sigma \sim p_{\Sigma}(\Sigma) \quad (12)
\end{align}

### III.B Hyperpriors and Estimation Approach with a Binary Instrument

In the above section, we defined the Bayesian hierarchical model that is flexible enough to handle various forms of the instrumental variable(s) and – with a bit more notational complexity – additional exogenous variables. We now restrict our attention to the case where there is a single binary instrument, e.g., using treatment assignment as an instrument for treatment enrollment in an RCT, and discuss estimation details and our preferred specification of the hyperpriors in this case.

Of particular note, when \( Z_i \) is a binary instrument and there are no additional \( X_i \) covariates we observe precisely four data points on \( \tilde{m} \) – two for \( m_1 \) and two for \( m_0 \) – regardless of how large the sample grows. Specifically, we observe estimates of 

\[
m_1(\eta_0) = \mathbb{E}[Y_i|T_i = 1, Z_i = 0]; \quad m_1(\eta_1) = \mathbb{E}[Y_i|T_i = 1, Z_i = 1]; \quad m_0(\eta_0) = \mathbb{E}[Y_i|T_i = 0, Z_i = 0]; \quad \text{and} \quad m_0(\eta_1) = \mathbb{E}[Y_i|T_i = 0, Z_i = 1].
\]

This has a number of important implications for the estimation approach. One advantage is that by pooling observations with the same treatment status \( (T_i) \) and instrument value \( (Z_i) \) it is extremely efficient to estimate the Gaussian process for any choice of hyperparameters.\(^4\) Furthermore, since we have repeated observations of \( m_1 \) and \( m_0 \) at two specific cutoffs, the hyperparameters \( \nu \) and \( \Sigma \) are well-identified – in the sense that the posterior distribution is highly localized even when we specify a diffuse hyperprior. We therefore estimate these parameters, rather than integrate over the hyperpriors. Specifically, letting \( \hat{\mathbb{E}} \) denote the empirical average we get that:

\[
\hat{\eta}_k = \hat{\mathbb{E}}[T_i|Z_i = k] \quad \text{for} \quad k \in \{0, 1\} \quad (13)
\]

\(^4\)This is because the computational complexity of estimating a Gaussian process is usually dominated by inverting an \( N \times N \) matrix, where \( N \) is the number of unique points where one observes the functions. Here \( N = 4 \) regardless of the sample size, so the computational complexity instead scales linearly with the sample size.
Similarly, if we assume the observations are all independent and identically distributed, we get that $\hat{\Sigma}$ is a diagonal matrix with the diagonals equal to:

$$\hat{\Sigma}_{k,k} = \frac{1}{N_k} \mathbb{E}\left[\left(Y_i - \mathbb{E}[Y_i|Z_i, T_i]\right)^2\right]$$

for $k \in \{1, 2, 3, 4\}$ (14)

where $N_k$ is the number of observations we observe with the relevant $(Z_i, T_i)$ pair.

Conversely, the fact that we only observe four points means that the data is mostly uninformative about the kernel hyperparameters. That is, loosely speaking, there will generally be a range of kernel hyperparameters that are consistent with the observed data. This means that the hyperprior $p_\theta$ can have a meaningful impact on the resulting posteriors and should therefore be chosen with care.

One 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.

This logic is illustrated in Figure 1. In it, each panel illustrates three random 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. Our preferred approach is to therefore to choose a hyperprior that suggests lengthscales in the middle row are more likely than the lengthscales in either the top or bottom row.
III.B  Hyperpriors and Estimation Approach with a Binary Instrument

Figure 1: Random Functions with Different Lengthscales

Note: This figure shows three random functions sampled from a Gaussian process with varying lengthscales and the same output variance, equal to 0.05.

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

\[
\log(l_\mu) \sim N(0.183, 0.25^2) \quad \text{and} \quad \log(\sigma_\mu) \sim N(0.5\sigma, 1.5^2) \quad (15)
\]

\[
\log(l_\tau) \sim N(0.183, 0.25^2) \quad \text{and} \quad \log(\sigma_\tau) \sim N(0.5\sigma, 1.5^2) \quad (16)
\]

where \(\sigma\) is the standard error of the residuals. The choice of 0.183 to be the mean of the hyperprior for \(l_\mu\) and \(l_\tau\) was chosen because when \(\log(l_\mu) = 0.183\) the correlation between \(\mu(\eta = 0)\) and \(\mu(\eta = 1)\) is approximately 0.5, which seems reasonable.\(^5\)

Note that we choose a more diffuse prior for the \(\sigma\) hyperparameters than the \(l\) hyperparameters since, conditional on the lengthscales, \(\sigma_\mu\) and \(\sigma_\tau\) are relatively well

\(^5\)Specifically, if the \(l = \left(\frac{1}{\log(\rho)}\right)^2\) then the correlation between \(\mu(\eta = 0)\) and \(\mu(\eta = 1)\) is \(\rho\).
identified. See Appendix B for evidence of this and Appendix C for more discussion of how the estimates depend on the hyperparameters.

IV  Empirical Example

We now explore how the method works in practice, by focusing on a specific empirical example. Our main example will be 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. See also Finkelstein et al. (2012); Taubman et al. (2014); Finkelstein et al. (2016).

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 (Forthcoming), 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 (Forthcoming), namely the likelihood that an individual will go the emergency room (ER).

IV.A  Base Results

To start, we focus on a single value for each of the the hyperparameters. The hyperparameters we choose is the maximum a posteriori (MAP), i.e., the value of $\theta$ that maximizes the $p(\theta|Y,Z,T)$.\(^6\) This approach is often referred to as empirical Bayes approach and results in the following values for the hyperparameters: $\sigma_\mu = 0.14$; $l_\mu = 1.21$; $\sigma_\tau = 0.15$, and $l_\tau = 1.21$.

\(^6\)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 we view it preferable to use an informed prior on the hyperparameters.
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.

Holding these hyperparameters fixed, we then simulate ten random functions $m_1$ and $m_0$. These simulations are shown in Figure 2; 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 2 illustrates nicely that conditioning on the four observed moments has a large impact on the plausible functions, even if it does not fully pin down the entirety of the functions.

A nice feature of Gaussian processes is that there exists a closed form solution which we can use to determine the posterior mean and variance of the process after conditioning on the four moments. Specifically, use $\overline{Y}$ to denote the four observed moments and define a $2 \times (M + 1)$ dimensional vector $\hat{m}$ as follows:

$$\overline{Y} = [\hat{m}_0(\eta_0), \hat{m}_0(\eta_1), \hat{m}_1(\eta_0), \hat{m}_1(\eta_1)]$$

$$\hat{m}' = [m_0(0), m_0(1/M), ..., m_0(1 - 1/M), m_0(1), m_1(0), m_1(1/M), ..., m_1(1 - 1/N), m_1(1)]$$
We can then define a $2(M+1) \times 2(M+1)$ matrix $K_{22}$ using the definition of $k_{\tilde{m}}$ and the vector $\tilde{m}$ such that if the $i^{th}$ row of $\tilde{m}$ is $m_t(\eta)$ and the $j^{th}$ row of $\tilde{m}'$ is $m_{t'}(\eta')$, then the $(i,j)^{th}$ element of $K_{22}$ is $k_{\tilde{m}}((t,\eta), (t',\eta'))$. We can similarly define a $4 \times 4$ matrix $K_{11}$ using the definition of $k_{\tilde{m}}$ and the vector $\bar{Y}$, as well as a $2M \times 4$ matrix $K_{21}$ using the definition of $k_{\tilde{m}}$ and the vectors $\tilde{m}'$ and $\bar{Y}$. Letting $\hat{\Sigma}$ be the covariance matrix of the error term, defined in Equation (14), we get that the posterior distribution of the $2 \times M$ dimensional vector is:

$$\hat{m}|\bar{Y} \sim N(\mu_{\hat{m}}, \Sigma_{\hat{m}})$$

with

$$\mu_{\hat{m}} = K_{21}(K_{11} + \hat{\Sigma})^{-1}\bar{Y}$$

$$\Sigma_{\hat{m}} = K_{22} - K_{21}(K_{11} + \hat{\Sigma})^{-1}K'_{21}$$

Figure 3a uses the above equations to calculate the posterior distribution of $m_1$ and $m_0$ on a 101 point grid. The red solid lines illustrate the posterior mean of $m_0$ and the 95% credible interval, while the blue dashed lines show illustrate the posterior mean of $m_1$ 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 $m$ 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. However, as would be expected, the posterior variance increases significantly away from the observed moments.

The posterior distributions of $m_1$ and $m_0$, 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 $m_1$ and $m_0$. 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 $m_1$ and $m_0$. This is illustrated in Figure 3b, 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 (Forthcoming). As illustrated in the dashed lines, however, we cannot be
Figure 3: 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 \( m_0 \) (in the red solid lines) and \( m_1 \) (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, with the solid indicating the posterior mean and the dashed lines indicating the 95% credible interval.
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 (AT ATE), never taker average treatment effect (NT ATE), and the local average treatment effect (LATE). While there is a lot of uncertainty in MTEs – as would be expected given that we only observe four moments – 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 are unidentified and therefore 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 NT ATE 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 NT ATE posterior. In addition as suggested by Figure 3b, the mean posterior of the AT ATE is greater than the LATE, which is in turn greater than the NT ATE.

An important caveat is that results in the first two columns of Table 1 take the hyperparameter value as fixed. An alternatively, and arguably preferable approach, is to incorporate uncertainty in the hyperparameters by integrating over the posterior distribution of hyperparameters. We therefore repeat analysis for a range of hyperparameters that are consistent with the data using a simple accept/reject algorithm. In doing so, we accepted approximately 18% 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, AT ATE, and NT ATE. 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.

---

7 Specifically, we first randomly draw 10,000 potential values of \( \theta \) from the hyperprior distribution defined in the section above and calculate the marginal likelihood for each of these values of \( \theta \). We then normalize these marginal likelihoods by dividing each marginal likelihood by the maximum observed value of the marginal likelihood. We next randomly draw 10,000 variables from a uniform distribution \( U(0, 1) \) and accept the value of \( \theta \) if the normalized marginal likelihood is greater than the value of the random variable.

8 Specifically, the overall mean is simply the average of the posterior means for each value of \( \theta \), while the overall variance is the average posterior variance plus the variance of the posterior means.
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 AT ATE and LATE – while the posterior standard deviations of the NT ATE increases by 40% when accounting for uncertainty in the hyperparameters.

Table 1: Posterior Means and Standard Deviations

<table>
<thead>
<tr>
<th></th>
<th>Empirical Bayes</th>
<th>Full Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std.</td>
</tr>
<tr>
<td>Avg. Treat Effect (ATE)</td>
<td>0.054</td>
<td>0.035</td>
</tr>
<tr>
<td>Always Taker Avg. Treat Effect (AT ATE)</td>
<td>0.108</td>
<td>0.027</td>
</tr>
<tr>
<td>Local Avg. Treat Effect (LATE)</td>
<td>0.083</td>
<td>0.023</td>
</tr>
<tr>
<td>Never Taker Avg. Treat Effect (NT ATE)</td>
<td>0.030</td>
<td>0.051</td>
</tr>
</tbody>
</table>

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 hyperparameter, while the last two columns integrates of the hyper-posterior distribution.

IV.B Statistical vs. Extrapolation Uncertainty

Broadly speaking, there are two sources of uncertainty in the estimates – “statistical uncertainty” in the true values of the observed moments due to the finite sample size and “extrapolation uncertainty” in how to extrapolate away from these moments to generate estimates of various unidentified average effects such as ATE, AT ATE, and NT ATE.\(^9\) We now formally define “statistical uncertainty” and “extrapolation uncertainty” to understand which (if any) source dominates the overall uncertainty measures.

Conceptually, we view extrapolation uncertainty as the variance of the posteriors in the (hypothetical) case where the four observed moments are known with certainty. In contrast, we view statistical uncertainty as the variance of the maximum a posteriori (MAP) estimates due to uncertainty in the observed moments. Using Equations

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 complete distribution of the full posterior.

\(^9\)In fact, one way to view the Bayesian approach developed here is as a way to generate a continuous measure of extrapolation uncertainty, rather than rather than having it be essentially infinite for unidentified averages such as ATE, AT ATE, and NT ATE.
(18) - (19) and the fact that $\text{Var}(\overline{Y}) = \Sigma$, we can calculate these as follows:

$$
\Sigma_{\text{extrap}} = K_{22} - K_{21}(K_{11})^{-1}K'_{21}
$$

$$
\Sigma_{\text{stat}} = K_{21}(K_{11} + \Sigma)^{-1}\Sigma(K_{11} + \Sigma)^{-1}K'_{21}
$$

While we believe that as defined $\Sigma_{\text{extrap}}$ and $\Sigma_{\text{stat}}$ provide insight into which source of uncertainty is more important, we emphasize that they do not provide a true decomposition of the overall uncertainty. Instead, the two sources interact and the overall variance is always greater than the sum of the statistical and extrapolation variances.$^{10}$

With that caveat, we present the results in Table 2. In the first three columns, we show the standard deviation of the posterior of four average effects: ATE, AT ATE, LATE, and NT ATE. Even the NT ATE, which requires the most extrapolation away from the observed moments the statistical uncertainty is 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 theoretically unidentified even with an infinite sample.$^{11}$

It’s worth emphasizing that the OHIE was much, much larger than most RCTs, with a sample size of more than 19,000 participants even 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. As an example, 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 how the uncertainty scales with the sample size, see Appendix D.

$^{10}$Formally, we prove that $\Sigma^\top_m - \Sigma_{\text{extrap}} - \Sigma_{\text{stat}}$ is a positive definite matrix if $\Sigma \neq 0$ and $\Sigma_{\text{extrap}} \neq 0$. It follows that overall uncertainty is greater than the sum of the statistical and extrapolation uncertainty for any treatment effect. We prove this in Appendix A.

$^{11}$It also means that building intuition on how the process works by assuming the moments are known for certain – which the author will admit he initially did – can be misleading.

$^{12}$The minor changes in the extrapolation uncertainty between the full sample and the sample of 1,000 in Table 2 are because the estimated cutoffs – and hence the extrapolation uncertainty – will vary slightly depending on the sample.
Table 2: Statistical vs. Extrapolation Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Full Sample</th>
<th>N = 1,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Treat Effect (ATE)</td>
<td>0.035 0.025 0.015</td>
<td>0.066 0.036 0.016</td>
</tr>
<tr>
<td>Always Taker ATE (AT ATE)</td>
<td>0.027 0.021 0.008</td>
<td>0.057 0.036 0.007</td>
</tr>
<tr>
<td>Local ATE (LATE)</td>
<td>0.023 0.021 0.000</td>
<td>0.058 0.037 0.000</td>
</tr>
<tr>
<td>Never Taker ATE (NT ATE)</td>
<td>0.051 0.033 0.025</td>
<td>0.079 0.036 0.025</td>
</tr>
</tbody>
</table>

Note: This table shows the standard deviations of the posterior distributions. The definition of statistical uncertainty, abbreviated as “Stat.” – and extrapolation uncertainty – abbreviated as “Extrap.” – are in the paper. To ease the comparison, we use the same set of hyperparameters for all six estimates, which are the hyperparameters that maximize the hyper-posterior in the full sample.

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 average treatment effect (ATE), always taker average treatment effect (AT ATE), or never taker average treatment effect (NT ATE). While we believe that the Bayesian approach developed here can be a valuable tool for researchers, we want to close by emphasizing that the model is simply one way to combine the researchers’ subjective beliefs with the observed data. Since the main application is likely to be instances where the estimands of interest are not identified – as in RCTs with imperfect compliance – the quality of the output rests on the quality of the prior.

We therefore conclude by emphasizing three decisions we made regarding the prior. First, we believe that researchers have the most intuition on what the functions $\mu(\eta)$ and $\mu(\eta)$ look like and so place the priors on these functions, rather than functions $m_1(\eta)$ and $m_0(\eta)$. Second, we believe that a Gaussian process prior with squared exponential covariance terms is a reasonable prior to place on the functions $\mu(\eta)$ and $\tau(\eta)$ and can capture a range of plausible functions. Finally, given the previous two choices, we believe that log-lengthscales outside the range of approximately $-1$ to $1$ generate random functions that either oscillate too much (if log-lengthscale is less than $-1$) or not enough (if the log-lengthscale is more than $1$) to serve as reasonable priors. While we natural believe these decisions are better than the alternatives, we
also do not believe this paper should be the final word. Instead, we hope that future work can refine the specification of the priors and view this paper as motivation for further research on the optimal specification of the priors.
References


_ and _ , “Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate


A Proofs

Theorem 1. Given the definitions in the paper, \( \tilde{m} \) is a Gaussian process with mean function of zero and a covariance function that depends on \( k_\mu(\eta, \eta'|\theta_\mu) \) and \( k_\tau(\eta, \eta'|\theta_\tau) \) as follows:

\[
k_{\tilde{m}}((t, \eta), (t', \eta')|\theta_\mu, \theta_\tau) = \begin{cases} 
\mathbb{E}[k_\mu(\tilde{\eta}, \tilde{\eta}'|\theta_\mu) + k_\tau(\tilde{\eta}, \tilde{\eta}'|\theta_\tau)|\tilde{\eta} \leq \eta, \tilde{\eta}' \leq \eta'] & \text{if } t = t' = 1 \\
\mathbb{E}[k_\mu(\tilde{\eta}, \tilde{\eta}'|\theta_\mu)|\tilde{\eta} > \eta, \tilde{\eta}' > \eta'] & \text{if } t = t' = 0 \\
\mathbb{E}[k_\mu(\tilde{\eta}, \tilde{\eta}'|\theta_\mu)|\tilde{\eta} > \eta, \tilde{\eta}' \leq \eta'] & \text{if } t = 0 \neq t' \\
\mathbb{E}[k_\mu(\tilde{\eta}, \tilde{\eta}'|\theta_\mu)|\tilde{\eta} \leq \eta, \tilde{\eta}' > \eta'] & \text{if } t = 1 \neq t'
\end{cases}
\]

Proof. Consider any finite linear combination \( \sum a_k \tilde{m}(t_k, \eta_k) \). Each \( \tilde{m}(t_k, \eta_k) \) can be arbitrarily approximated by a finite sum of jointly Gaussian variables and so \( \sum a_k \tilde{m}(t_k, \eta_k) \) can be arbitrarily approximated by a finite sum of jointly Gaussian variables, and is therefore Gaussian. Thus \( \tilde{m} \) is a Gaussian process.

We then consider two points \( \tilde{m}(1, \eta) \) and \( \tilde{m}(1, \eta') \). From the definition of \( \tilde{m} \), we get that:

\[
\text{Cov}(\tilde{m}(1, \eta), \tilde{m}(1, \eta')) = \left( \frac{1}{N} \lim_{N \to \infty} \sum_{i=0}^{N} \mu(\eta \cdot \frac{i}{N}) + \tau(\eta \cdot \frac{i}{N}) \right)
\left( \frac{1}{N} \lim_{N \to \infty} \sum_{i=0}^{N} \mu(\eta' \cdot \frac{i}{N}) + \tau(\eta' \cdot \frac{i}{N}) \right)
\]

\[
= \lim_{N \to \infty} \sum_{i=0}^{N} \left( \mu(\eta \cdot \frac{i}{N}) + \tau(\eta \cdot \frac{i}{N}) \right) \cdot \sum_{i=0}^{N} \left( \mu(\eta' \cdot \frac{i}{N}) + \tau(\eta' \cdot \frac{i}{N}) \right)
\]

\[
= \lim_{N \to \infty} \sum_{i=0}^{N} \mu(\eta \cdot \frac{i}{N}) \cdot \sum_{i=0}^{N} \mu(\eta' \cdot \frac{i}{N}) + \lim_{N \to \infty} \sum_{i=0}^{N} \tau(\eta \cdot \frac{i}{N}) \cdot \sum_{i=0}^{N} \tau(\eta' \cdot \frac{i}{N})
\]

\[
= \mathbb{E}[k_\mu(\tilde{\eta}, \tilde{\eta}'|\theta_\mu)|\tilde{\eta} \leq \eta, \tilde{\eta}' \leq \eta'] + \mathbb{E}[k_\tau(\tilde{\eta}, \tilde{\eta}'|\theta_\tau)|\tilde{\eta} \leq \eta, \tilde{\eta}' \leq \eta']
\]

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 \( t \neq 1 \) or \( t' \neq 1 \) are derived similarly.

Theorem 2. If \( \Sigma \neq 0 \) and \( \Sigma_{\text{extrap}} \neq 0 \), then \( \Sigma_{\tilde{m}} - \left( \Sigma_{\text{extrap}} + \Sigma_{\text{stat}} \right) \) is a positive definite matrix.
Proof. From the paper, we have that:

\[
\Sigma_m = K_{22} - K_{21}(K_{11} + \Sigma)^{-1}K'_{21}
\]

\[
\Sigma_{extrap} = K_{22} - K_{21}(K_{11})^{-1}K'_{21}
\]

\[
\Sigma_{stat} = K_{21}(K_{11} + \Sigma)^{-1}(K_{11} + \Sigma)^{-1}K'_{21}
\]

We therefore get that:

\[
\Sigma_m - \left(\Sigma_{extrap} + \Sigma_{stat}\right) = K_{21}(K_{11})^{-1}K'_{21} - K_{21}(K_{11} + \Sigma)^{-1}(K_{11} + \Sigma)^{-1}K'_{21} - K_{21}(K_{11} + \Sigma)^{-1}K'_{21}
\]

\[
= K_{21} \cdot \left[ (K_{11})^{-1} - (K_{11} + \Sigma)^{-1} \Sigma(K_{11} + \Sigma)^{-1} (K_{11} + \Sigma)^{-1} \right] \cdot K'_{21}
\]

\[
= K_{21} \cdot \left[ (K_{11})^{-1}(K_{11} + \Sigma) - (K_{11} + \Sigma)^{-1} \Sigma - I \right] \cdot (K_{11} + \Sigma)^{-1}K'_{21}
\]

\[
= K_{21} \cdot \left[ I + (K_{11})^{-1} \Sigma - (K_{11} + \Sigma)^{-1} \Sigma - I \right] \cdot (K_{11} + \Sigma)^{-1}K'_{21}
\]

\[
= K_{21} \cdot \left[ (K_{11})^{-1} \Sigma(K_{11} + \Sigma)^{-1} - (K_{11} + \Sigma)^{-1} \Sigma(K_{11} + \Sigma)^{-1} \right] \cdot K'_{21}
\]

We can therefore show that \(\Sigma_m - \left(\Sigma_{extrap} + \Sigma_{stat}\right)\) is positive definite by showing that the middle term, \((K_{11})^{-1} \Sigma(K_{11} + \Sigma)^{-1} - (K_{11} + \Sigma)^{-1} \Sigma(K_{11} + \Sigma)^{-1}\), is positive definite.

To do that, we use the fact that if \(A - B\) is positive definite, then \(B^{-1} - A^{-1}\) is positive definite. Thus, we will show that \((K_{11} + \Sigma) \Sigma^{-1}(K_{11} + \Sigma) - (K_{11} + \Sigma) \Sigma^{-1}(K_{11})\)
is positive definite. We do so by expanding each term to get that: these to get that:

\[
(K_{11} + \Sigma)^{-1}(K_{11} + \Sigma) - (K_{11} + \Sigma)^{-1}(K_{11}) = \\
K_{11}(K_{11} + \Sigma)^{-1}(K_{11} + \Sigma) + \Sigma(K_{11} + \Sigma)^{-1}(K_{11}) - K_{11}\Sigma^{-1}K_{11} - \Sigma\Sigma^{-1}K_{11} = \\
K_{11}(K_{11} + \Sigma)^{-1}(K_{11} + \Sigma) + (K_{11} + \Sigma) - K_{11}\Sigma^{-1}K_{11} - K_{11} = \\
K_{11}\Sigma^{-1}K_{11} + K_{11}\Sigma^{-1}\Sigma + K_{11} + \Sigma - K_{11}\Sigma^{-1}K_{11} - K_{11} = \\
K_{11}\Sigma^{-1}K_{11} + K_{11} + \Sigma - K_{11}\Sigma^{-1}K_{11} - K_{11} = \\
K_{11} + \Sigma
\]

which is clearly positive definite since both \(\Sigma\) and \(K_{11}\) are positive definite. \(\square\)

## B Understanding the Marginal Likelihood

In Section III.B we made two claims regarding the identification of the hyperparameters with a binary instrument. First, we noted that the two hyperparameters governing the lengthscales - \(l_\tau\) and \(l_\mu\) - are not well identified; in contrast, we claimed that conditional on \(l_\tau\) and \(l_\mu\) the other two hyperparameters - \(\sigma_\mu\) and \(\sigma_\tau\) - are relatively well identified. In Section III.B, as in this appendix, we consider something well-identified in the informal sense that the marginal likelihood function has a well-defined peak. Given this informal definition, we will not prove this statement formally and instead illustrate it empirically using the OHIE sample.

To better illustrate the how the marginal likelihoods depends on the hyperparameters, we will simplify by shrinking the hyperparameter space by setting \(l_\tau = l_\mu \equiv l\) and \(\sigma_\mu = \sigma_\tau \equiv \sigma\). We can then see the overall result in Figure 4, which shows how the marginal likelihood depends on the two hyperparameters: \(l\) and \(\sigma\). As can be seen, while the marginal likelihood is low when \(\sigma\) is either less than \(e^{-2.5}\) or more than \(e^{-2}\), any lengthscale has a relatively high marginal likelihood when paired with the right \(\sigma\). While this is in part due to the range of values that we show in the figure, in other exploration we find that although lengthscale less than \(e^{-0.5}\) have a low marginal likelihood, the data cannot meaningfully distinguish between lengthscales above \(e^{-0.5}\).

This can also be seen by comparing the hyperprior to the hyperposterior, i.e., comparing the prior distribution of the hyperparamters to the posterior distribution.
Figure 4: Marginal Likelihood

Note: This shows the average marginal likelihood for within each hexagonal region, with the lighter colors indicating a higher likelihood. Here, the lengthscale corresponds to both $l_\mu$ and $l_\tau$, while $\sigma$ corresponds to both $\sigma_\mu$ and $\sigma_\tau$. 
This is shown in Figure 5. From the figure it is clear that the for \( l_r \) and \( l_\mu \) the posterior distributions are quite similar to the prior distributions, reflecting the fact that the lengthscales are poorly identified. In contrast, for \( \sigma_r \) and \( \sigma_\mu \) the posterior distributions are much narrower than the prior distributions.

C Varying the Hyperparameters

As discussed in the Section III.B and Appendix B the hyperparameters – and particularly the lengthscales – are not well identified. 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 give a large number of plausible hyperparameters. 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 AT ATE, LATE, and NT ATE for each hyperparameter drawn from the hyper-posterior. The results are shown in Figure 6, which shows that the posterior mean and variance of the (unidentified) AT ATE and NT ATE depend more on the value of the hyperparameter than the (identified) LATE. How the variances depend on the hyperparameters is illustrated in Figure 7, which shows that over the range of our hyperprior the posterior variances are decreasing with the the lengthscales.

D 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 8, we show how the posterior variance for the four estimands of interest – the ATE, AT ATE, LATE, and NT ATE – 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 idiosyncracies in the random sample,
Figure 5: Prior Vs Posterior Distribution of the Hyperparameters

Note: This shows the prior distributions of each of the four hyperparameters (shown in grey) and the posterior distribution of the four hyperparameters (shown in blue). The four hyperparameters are “lengthscale mu” \((l_\mu)\); “lengthscale tau” \((l_\tau)\); “standard deviation mu” \((\sigma_\mu)\); and “standard deviation tau” \((\sigma_\tau)\).
Figure 6: Mean/Variance Estimates for Each Set of Hyperparameters 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 hyperparameters 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 7: Posterior Variances as a Function of the Lengthscales

(a) As a function of $l_\mu$

(b) As a function of $l_\tau$

Note: This figure shows the posterior variances of the four main estimates of interest as a function of the assumed lengthscales. The values on the y-axis correspond to conditional average of the estimands’ posterior variance, with the expectation taken over the posterior distribution of the hyperparameters conditional on the value of the lengthscale shown on the x-axis. 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).
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 8, 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 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 IV.B. As seen in Figure 8, 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.
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).