From LATE to Great: Efficient Estimation of Heterogeneous Treatment Effects with Gaussian Markov Random Fields

Brice Green

11/1/2021

Abstract

Accurately estimating conditional average treatment effects is extremely important for extrapolating from local causal estimates to new populations. Bayesian hierarchical models with Gaussian Markov Random Field priors provide a way to do this when when treatment effects vary along a continuous variable that is both intuitive and accurate.

Introduction

Our goal in estimating causal effects is not only to understand what *has* happened, but also to divine what *might happen* should we take some action. However, at least in the social sciences, the measured effects are typically contextual. The effect of a given policy will vary with the timing, recipient, geography, and cultural context that surrounds it. In cases where a treatment effect is constant, a randomized experiment may grant us insight that persists the world over, but if the effect of treatment is highly contextual the best we can do is estimate the average effect for people in the sample who are induced to take up a treatment by the decision (Imbens and Angrist 1994); this is typically referred to as a Local Average Treatment Effect (LATE).

But what do we *do* with LATEs? It is common to see policy-makers and journalists discuss papers as "finding a causal relationship," but in a context where the magnitude of causal relationship is highly variable, it is far from a solid assumption that the measured effect will persist in a new population. A natural option is to directly model varying local average treatment effects in sub-populations we believe might have varying responses. Given these estimates, all we would need to know is the weights for each sub-group in the new population in order to forecast the treatment effects for a new population. This approach is common in the survey literature, which often deals with non-representative sample populations (e.g. Gelman and Little (1997).

Estimating a series of LATEs by sub-group gives us

$$\begin{aligned} \tau_1 &\sim N(\mu_1, \sigma_1) \\ &\vdots \\ \tau_k &\sim N(\mu_k, \sigma_k) \end{aligned}$$

where k is the number of subgroups of interest, μ_i is the estimated treatment effect for group i, and σ_i is the standard deviation of the sampling distribution for group i.

Quickly, though, this strategy runs into a mathematical reality: for each dimension we incorporate into our analysis the number of groups of interest grows geometrically, and we quickly run out of observations. Suppose we were looking at variation in treatment effects over age and income; if we have 10 categories of each, this immediately becomes 100 separate subgroups!

Often, this naive separation of groups into subsamples discards structural information we have about the problem. For example, if we were willing to estimate a local average treatment effect across the whole population, it is likely true we believe that treatment effect to have a finite expectation and variance over the target groups. When we move to estimate subgroups, the most conservative choice we can make when modeling these treatment effects which conditions on that information is to assume that the subgroup treatment effects are drawn from a normal distribution (Jaynes 2003). This gives us structure to the problem, allowing the different sub-group estimates to inform each other.

This assumption gives us the model

$$\begin{aligned} \tau_i &\sim N(\mu_i, \sigma_i) \\ \mu_i &\sim N(\mu, \gamma) \end{aligned}$$

This model is used extensively in education research (Rubin 1981), meta-analysis (Meager 2019), and the political science and survey literature that uses post-stratification (Ghitza and Gelman 2013). It is a natural way to avoid the problem of geometric growth in sub-groups–because the different estimates inform each other, the effective sample size of each estimate is higher than in the case where they are estimated independently.

This type of prior is not designed to perfectly represent the current state of available knowledge, but instead to regularize our estimates using information relevant to the structure of the problem. This type of structural regularization is prevalent in many "penalized complexity" priors, which serve to pull estimates towards their simplest nested model configuration, in this case, a model with constant treatment effects across groups (Simpson et al. 2017).

Combining these notions, we have a way to avoid the curse of dimensionality when estimating treatment effects across subgroups, and a regularization device derived from structural model assumptions that dynamically pools estimates across groups to the extent warranted, all in one.

In this paper I consider a model for treatment effects which are a function of continuous variables with a different penalized-complexity prior. The approach is analogous to the divide-and-conquer strategy displayed above: first I divide continuous variables into a lattice, and I estimate a model which penalizes the first differences between the estimates of adjacent treatment effects. This model has a nice interpretation as a piece-wise linear basis function representation of the treatment effect function, and deep connections to both the smoothing literature and stochastic partial differential equations. With simulations I show that they are far more effective than independent subsample analysis at estimating interactions, both showing less bias and variance as the number of groups gets large.

There is a large body of work dedicated to estimating heterogeneous treatment effects. In recent work, Wager and Athey (2018), use a random forest model to directly estimate heterogeneous treatment effects. Hill (2011) uses Bayesian Additive Regression Trees for this purpose. Chernozhukov et al. (2018) uses machine learning estimates of treatment effects as proxies for inference functions, giving analysts powerful post-processing tools to analyze outcomes which otherwise may not be consistently identified. Nie and Wager (2021) show that a two-pass method allows for generic inference by many algorithms, first identifying the causal signal in the data and subsequently constructing a data-specific loss function. Dorie et al. (2019) report the results of a causal inference competition, finding that BART and calCause (a combination of random forest and Gaussian Process regression) most precisely identified varying treatment effects. Abrevaya, Hsu, and Lieli (2015) shows that the conditional average treatment effect function can be estimated continuously with kernel smoothing and inverse propensity score weighting.

This paper is also related to the literature dedicated to extrapolating to representative populations from non-representative surveys. Gao et al. (2021) shows the efficacy of the Gaussian Markov Random Field prior in dealing with ordered survey covariates, and Gao, Kennedy, and Simpson (2021) continues this work, applying hierarchical priors to a randomized educational growth mindset experiment and analyzing the portability of those estimated conditional average treatment effects. Earlier work in political science has used priors to identify interactions for small cells which would otherwise be difficult to identify (Ghitza and Gelman 2013).

Priors for Varying Treatment Effects

Setting up the model

A multilevel model, in the most general sense, is a means of relating coefficients across estimates. To give a concrete example, imagine we have a regression model which associates someone's log income, Y with a set of covariates X. We observe N of people over time, which means we could estimate the model

$$\begin{aligned} Y_1 &= \alpha_1 + X_1\beta_1 + \epsilon_1 \\ &\vdots \\ Y_N &= \alpha_N + X_N\beta_N + \epsilon_N \end{aligned}$$

In this case, we might think of this model as being a series of independent regressions, or as estimating a series of individual-level interactions. A multilevel model combines this observational model with a model for the parameters. For example, we might believe that the various estimates, $\beta_i \in \{i, ..., N\}$ have a finite mean and variance. Then we could combine them (either ex-post or jointly) according to a model where $beta_i \sim Normal(\beta, \gamma)$. Equivalently, we could break each regression into a model where

$$\begin{aligned} Y_i &\sim N(\alpha_1 + X\beta + X\beta_i, \sigma_i^2) \\ \beta_i &\sim N(0, \gamma) \end{aligned}$$

This helps regularize the variation around the average cross-sectional estimate, β, since the estimates for the other samples help inform what a "reasonable" β_i is. Multilevel models are extremely useful for estimating interactions because they perform a kind of dynamic regularization. Once we allow estimates to inform each other across groups, they pull noisy or highly uncertain groups towards the gravity of a central mean.

While this is a very popular form of multilevel model, it is just one possible form. For example, imagine that we had information about the physical distance between two locations being measured, and we expected groups that were closer together to be more similar, then we could encode the spatial correlation over covariates in the hierarchical likelihood (Gao et al. 2020).

This characterization is useful in the case of heterogenous treatment effects because it is quite common to have continuous covariates over which we expect the treatment effects to vary. In this situation it may be to large an assumption to posit a global functional form, but there still may be information about the local behavior of a function. To give a concrete example, suppose an analyst wanted to understand how a causal effect varies with different levels of income. In many contexts it is plausible that the causal effect on for a person who making \$30,000 a year be more informative about the causal effect on someone making \$40,000 a year than it would for someone making \$3,000,000 a year.

One intuitive approach is to break the continuous variable into subgroups, and perform analysis within those subgroups. This general tactic is common (outside of a causal framework) in fields like finance, where stocks are sorted into characteristics and placed into portfolios, with each examined as a separate outcome variable in a time-series model. Suppose we have a continuous variable, like $log(income)$, and break it into equal subgroups, estimating a series of individual treatment effects for each group. This gives us a one-dimensional lattice of random variables (our estimated treatment effects).

We can represent a lattice as a series of neighbor relationships. Let each bin's local average treatment effect a node on a graph or lattice. The clique is the set of adjacent points on the graph along each dimension. In other words, for a treatment effect τ in cell i with neighboring cells j,

$$p(\tau_i|\tau_{j\neq i}, \tau_{-j}) = p(\tau_i|\tau_{j\neq i})$$

This models the process over which parameter values evolve along the covariate lattice as a Markov Random

Field, meaning that it only depends on the states immediately proceeding this.[1] These models have a substantial literature, dating back to Besag (1974), who motivates this approach via the Hammersley-Clifford theorem, demonstrating that even though the model is local it can recover the global structure for the variation over a lattice. This model is a workhorse approach for characterizing spatial data with areal representations and in image processing, where pixels have a similar lattice structure.

Owing to the central limit theorem, the conditional distribution is modeled as a Gaussian density, with

$$p(\phi_i|\phi_{j\neq i}, \phi_{-j}) = N(\alpha\Sigma_{i\sim j} w_{i,j}\phi_j, q_i^{-1})$$

where q^{-1} is the precision of the variance of cell i, $w_{i,j}$ represents the distance between j and i, $i \sim j$ refers to the members of the clique of i's neighbors, and α is the degree of spatial dependence. Because α is an estimated coefficient, these are often called "exact" CAR models.

A common variant, which allows for non-stationary variation around a central value, is called the Intrinsic Conditional Auto-Regressive Model (ICAR), where α is set to 1. Distance is arbitrary when working with scaled covariates on an evenly spaced lattice, and unobserved for ordinal covariates, so I make the simplifying assumption that the weights matrix is a set of binary indicators representing neighbor relationships. As a consequence of both of these assumptions, the expression becomes

$$p(\tau_i|\tau_{j\neq i}, \tau_{-j}, q_i^{-1}) \sim N\left(\frac{\Sigma_{i\sim j}\tau_i}{d_{i,i}}, \frac{1}{d_{i,i}q_i}\right),$$

where $i \sim j$ implies that i and j are neighbors.[2] We can equivalently think of this model as being a state-space model where the variation across cells is a latent random walk over neighboring cells.[3]

The first-order ICAR model (or Intrinsic Gaussian Markov Random Field model) can also be re-written in terms of first differences, where

$$\tau_i - \tau_j \sim N\left(0, \frac{1}{q}\right).$$

In this form, it becomes clear that the intrinsic GMRF model is a probabilistic penalty of the first-difference of the treatment effects between neighboring cells, in turn penalizing the first derivative of the treatment surface with respect to the dimension along which τ_i neighbors τ_j.

This prior, by itself, is not proper, as it is invariant to any added constant. However, you can use it as a prior on the estimated treatment effects by assuming that the random variation around the average treatment effect across the population sums to 0. This simplex constraint identifies the model, and while if it were directly being used as a prior on observational data it might be problematic, in the case of a varying-coefficients model that construction is quite useful.

While in this paper I work with evenly distributed lattices, extending the model to irregular grids and triangulations simply involves re-weighting the various neighbors based on the size of the cells. In other applications of this type of model, it is common to break the domain into a set of discrete objects with Delaunay triangulations rather than even lattices, especially in cases where observations are not uniform over the problem domain (Blangiardo et al. 2013).

[1] The effectiveness of this assumption is likely application specific. For incorporating non-local information, it might be good to include more distant relationships through a Gaussian process prior, (see, for example Banerjee et al. (2008)) or through a higher-order autoregressive process.

[2] This model is a slightly simplification of the prior in the specific case where the prior on the random variation in treatment effects (away from the mean). However, since we can just re-scale treatment effects.

[3] For a deeper discussion of these models, see Rue and Held (2005), which discusses the intrinsic model on a lattice in section 3.3.2.

Connections to Smoothing

Both models have deep connections to the smoothing literature. Both the normal/normal hierarchical model presented in the introduction and the intrinsic Gaussian Markov random field model admit a penalized spline representation (Wood 2017). To see this, consider the problem where we are estimating a univariate outcome

$$y_i = f(x_i) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

Generally, we estimate this by choosing a set of basis functions (say, local cubic polynomials), such that

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j.$$

In order to prevent functions from getting too "wiggly," these cubic splines are often penalized with a matrix that penalizes the second derivative of $f(x)$,

$$\int_{x_1}^{x_k} f''(x)^2 dx = \beta^T S \beta$$

and then estimating the model to minimize

$$||y - f(x)||^2 + \lambda \beta^T S \beta$$

where λ is the penalty on the smoothing parameter. Given the matrix representing the knots, S, we can re-write this smoothing penalty as a prior distribution, $f(\beta) \propto \exp(-\lambda \beta^T S \beta / 2)$, which implies

$$\beta \sim N\left(0, \frac{S^-}{\lambda}\right)$$

on the smoothing terms of the spline model (after subtracting the mean). This recovers the normal/normal multilevel model presented at the beginning of the section in the case of a linear function where S is an identity matrix.

Gaussian Markov Random Fields can also be interpreted in the context of smoothing splines, with a penalty matrix that penalizes the squared differences between neighboring coefficients. As shown in the prior section, the intrinsic GMRF prior can be re-written in terms of pairwise differences. We can then think of a penalty matrix which sums the squared differences between neighboring coefficients, which gives us a penalty matrix

$$\beta^T S \beta = \sum_{j=1}^{m} \sum_{i \sim j} (\beta_j - \beta_i)^2,$$

where S is equal to -1 if i and j are neighbors and the diagonal elements of S are simply the number of neighbors that the associated district has. In this case, we also recover $\beta \sim N\left(0, \frac{S^-}{\lambda}\right)$.

Finally, we can represent GMRF priors as a set of penalized piecewise linear basis functions. This directly relates GMRF models to stochastic partial differential equations, with these models recovering stationary solutions to the SPDE derived from a Matern covariance matrix (Lindgren, Rue, and Lindström 2011).

The connections between spline models and Bayesian multilevel models are helpful because they allow us to build intuitions for *why* these models might be good at approximating the space of treatment effects. If we represent the models as penalties, in the form

$$y = ||y - X\beta||^2 + \beta^T S\beta + \epsilon$$

we can think of them as regularizers which dynamically down-weight treatment effect functions which are a priori unlikely, but which have support across the full space of possible functions. While the normal/normal model penalizes the linear function solely in terms of completely independent draws (since S is the identity matrix), the GMRF model penalizes the differences between neighboring coefficients, emphasizing a kind of smoothness by probabilistically weighting potential models based on the average first difference from their neighbors. Perhaps even more appealing is the way that they penalize these aspects by weighting them over a probability distribution, and decrease in their informativeness in proportion to the available data.

Considerations for priors in a causal setting

Priors are often a sticking point for people who are unfamiliar with Bayesian modeling. Often the fear is that these types of models require so many assumptions. In choosing a prior for how treatment effects vary, there are several considerations. First, we want the model to be able to recover the relevant range of possible models. Second, to the extend that a model is regularized by a prior distribution, we should understand the "base" model that the model pulls towards (Simpson et al. 2017). Finally, we must check our model both by simulating before model fitting and by checking the fitted model against observational data to see if it is mis-specified (Gelman et al. 2020).

Let us first consider the normal multilevel model, where

$$\begin{aligned} y &= X\beta + \epsilon \\ \beta &\sim N(\mu, \sigma). \end{aligned}$$

The clear *baseline model* or the average model we recover, is one with no varying estimates of β. In the case of treatment effects, this is the local average treatment effect we currently estimate over the pooled dataset. However, while this model prefers this baseline before seeing any data, as $\sigma \to \infty$ we recover complete subgroup independence (this also happens as the number of observations $n_i \to \infty$ for each group i).

The GMRF prior also has the same behavior. After differencing out the average effect, we have a set of random walks with mean 0 over the effects of the data. While the model prefers a first difference between neighboring subgroups to be equal to 0, e.g.

$$\beta_i - \beta_j \sim N\left(0, \frac{1}{q}\right),$$

where q is the precision, this prior has has support over the whole real line, and as $n \to infty$ and $q \to 0$ it recovers completely independent subgroup analyses. In other words, both models have an asymptotic regime which converges to the same group-wise mean as two-stage least squares. The GMRF model also converges to a continuously indexed Gaussian field as the number of cells in the lattice approaches infinity (Rue and Held 2005).

It is also important to consider what we consider a "conservative" model. Often the fear which drives completely independent analyses of subgroups is that we don't want to posit a model for how treatment effects vary. But discarding information associated with our parameters can inject unecessary noise into our estimate. If we consider estimating a completely independent set of coefficients, we can structure it as a multilevel model with an improper uniform prior over coefficients, e.g.

$$y = \beta_i X_i + \epsilon, \beta_i \sim Unif(-\infty, \infty),$$

which has a completely unbounded mean and variance. This model is *less* conservative than the others, in the sense that it equally weights all possible functional forms for how β_i varies.

Simulation Results

Here I present a series of simulation results for estimated treatment effects, comparing the bias, mean squared error, and R^2 of models fit with the intrinsic GMRF prior relative to independent subgroup analysis. Both models are estimated via two stage least squares in a context with a strong binary instrument and binary treatment, with effects that vary across subgroups according to a smooth function.

Causal Setting

Causal identification requires assumptions, and estimating heterogeneous treatment effects require require even more. Because of the popularity of the instrumental variables estimator, I demonstrate the efficacy of this estimation technique in that setting. Consider the situation with a binary instrument, Z, associated with a binary treatment status, D, and potential outcomes $Y_i(0)$ and $Y_i(1)$ which, for individual i are a function of treated status D. Going forward, I presume the standard IV conditions for identification of local average treatment effects are satisfied,

1. SUTVA: Given two vectors of treatments and instruments, D and D' and Z and Z', $D_i(Z) = D_i(Z')$ if $Z_i = Z_i'$ and $Y_i(D) = Y_i(D')$ if $D_i = D_i'$.
2. Independence: Y_i and D_i are jointly independent of Z_i.
3. Exclusion: $Y_i(D_i, Z = 0) = Y_i(D_i, Z = 1)$, or in other words, Z only affects Y through D
4. First-Stage: Z must be correlated with D.
5. Monotonicity: Either $E(D_i(Z = 1)) > E(D_i(Z = 0))\ \forall i$ or $E(D_i(Z = 1)) < E(D_i(Z = 0))\ \forall i$

While I primarily work with instrumental variables designs, there is nothing specific about the models discussed in this paper which requires this approach. The same principles for estimating varying treatment effects could apply to random experiments, difference in differences, regression discontinuity, or other designs of interest where an analyst is interested in understanding high-dimensional interactions.

Simulation Setup

I first test the Gaussian Markov Random Field prior, taking a K-dimensional space of continuous random variables drawn from a normal distribution, and constructing true treatment effects as a function of those variables. When estimating both subgroups and the GMRF model, I divide the continuous covariate space into an equally spaced lattice.

I consider three functions for the treatment effects, all of which are smooth. For simplicity, I first consider a linear model with weights. Then I consider a squared function, to see whether the increased curvature of the function causes the performance of the GMRF prior to deteriorate. Finally, to capture potentially cyclical treatment effects, I use a sin function. In each case

$$\tau_i = \beta_1 f(X_{1,i}) + \beta_2 f(X_{2,i}) + ... + \beta_k f(X_{k,i})$$

and recover the treatment effect estimate. The Bayesian model is estimated via Markov Chain Monte Carlo, using the Stan modeling language and the Stan implementation of Hamiltonian Monte Carlo with No-U-Turn Sampling (Carpenter et al. 2017).[4]

The local average treatment estimation with instrumental variables only recovers treatment effects for the complier population. In order to differentiate between the general causal effect and the complier effect, I assume that people in the top 10% of all treated individuals are "always-takers" and are not in the complier population.

I estimate each model several times on lattices with different numbers of splits. There is a natural tension here, since functions with greater curvature will benefit from having more subgroups, but the number of cells

[4]My implementation of the GMRF (ICAR) model drew extensively from the Stan case-study on implementing intrinsic conditional autoregressive models. For the associated paper, see Morris et al. (2019), and see here for the case study in question: https://mc-stan.org/users/documentation/case-studies/icar_stan.html.

grows geometrically with K, the dimension of the regressors being used. As a consequence, estimating the treatment effects for the subgroups has nearly infinite variance as the number of subgroups per dimension and the number of dimensions grows, while the model with the GMRF prior becomes more accurate because the piece-wise linear interpolant is more accurate.

I first consider a medium-scale problem with 1,000 observations. I take posterior expectations as the point prediction for the treatment effect. When examining the GMRF prior, I run several simulations for each data generating process, and share measures of the fitted distribution for the model relative to the independent 2SLS model. When examining the normal multilevel model I do not vary the dimension in the same way as the GMRF model, simply because way that two stage least squares for independent subgroups degenerates is fairly obvious from the GMRF simulations.

Results: In Sample

I first look at the GMRF model, examining the absolute value of the bias, mean squared error, and R^2 of the predicted estimates relative to the treatment effects, comparing to a comparable estimate fit with two-stage-least-squares by subgroup. Since results are simulated, estimates are measured with the true underlying treatment effect. The y-axis in each graph is plotted on a log scale due to the very large variance of 2SLS with few observations.

The Gaussian Markov Random Field prior strictly dominates independent 2SLS in the settings considered, even at moderate dimension with a small number of subgroups. Even in moderate dimension (say, 4), with a moderate number of subgroups (say, 4) independent 2SLS runs out of observations, and its variance explodes. The geometric explosion of the number of possible subgroups makes it very difficult to get accurate estimates. However, even at sample sizes where independent 2SLS cannot be estimated, the GMRF prior has extremely low mean squared error. Because of the model's ability to flexibly model the functions, adding additional groups actually decreases the mean squared error. We can even see that the model consistently estimates treatment effects when then number of subgroups exceeds the number of observations, with the groups without samples being informed by their neighbors.

In-sample R^2 has a similar story, improving when subgroups are added because this allows the function to be more flexible, while not degenerating as the number of observations per group approaches 0. We might expect the discrepancies between the 2SLS and GMRF models to grow in an out of sample setting, since some groups that have very few observations in the starting sample may have large weights in the subsequent domain. When interactions are important, getting a better approximation of the treatment effect function improves portability. Especially for non-linear functions, using a finer mesh substantially improves R^2 when using the GMRF prior model, while independent two-stage-least-squares quickly becomes less effective due to the small sample sizes of each individual cell.

A typical fear of using priors is that they will introduce asymptotic bias into the model. However, in this case we see that average bias is comparable to the independent 2SLS estimates. Perhaps with more simulations we might see more consistent differences in the average bias between the two models. However, when we think interactions are important and we want to project our estimates onto new populations, we risk omitting important variables.

To see a specific example, consider a case where we have 1,000 observations, each dimension is broken into 5 subgroups, and we have 3 covariates of interest. This figure plots posterior predictive intervals in red and the group average effect in blue, with the shade of blue representing the number of observations in each group. For ease of comparison, I've sorted the plot by the true conditional average treatment effects, shown in black. Even in the most extreme parts of the distribution, and in cases where we observe almost no data directly for the given group, we get good coverage of the true treatment effects

Projecting Treatment Effects onto New Populations

Intuitively, the fidelity of the projection onto a new population depends on the accuracy of both our estimates and our approximation of the treatment effect function. For example, if we were to approximate the treatment

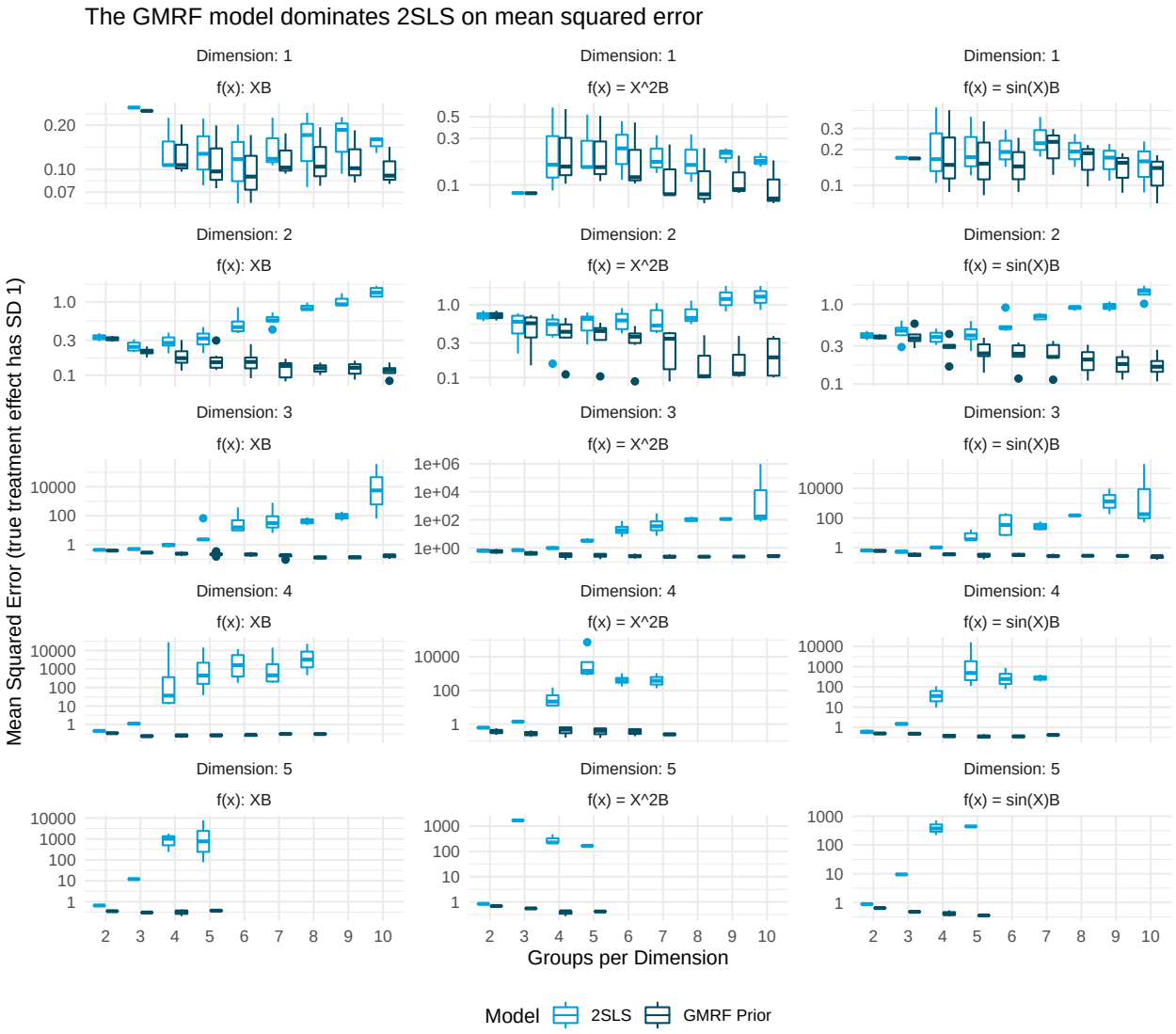

Figure 1: This figure shows the mean squared error of the treatment effects estimated via independent two stage least squares by group and two stage least squares with a Gaussian Markov random field prior. Mean squared error is measured via $(E(T|M) - T)^2$, where M is either the independent 2SLS model or the model estimated with the GMRF prior and T is the true treatment effect.

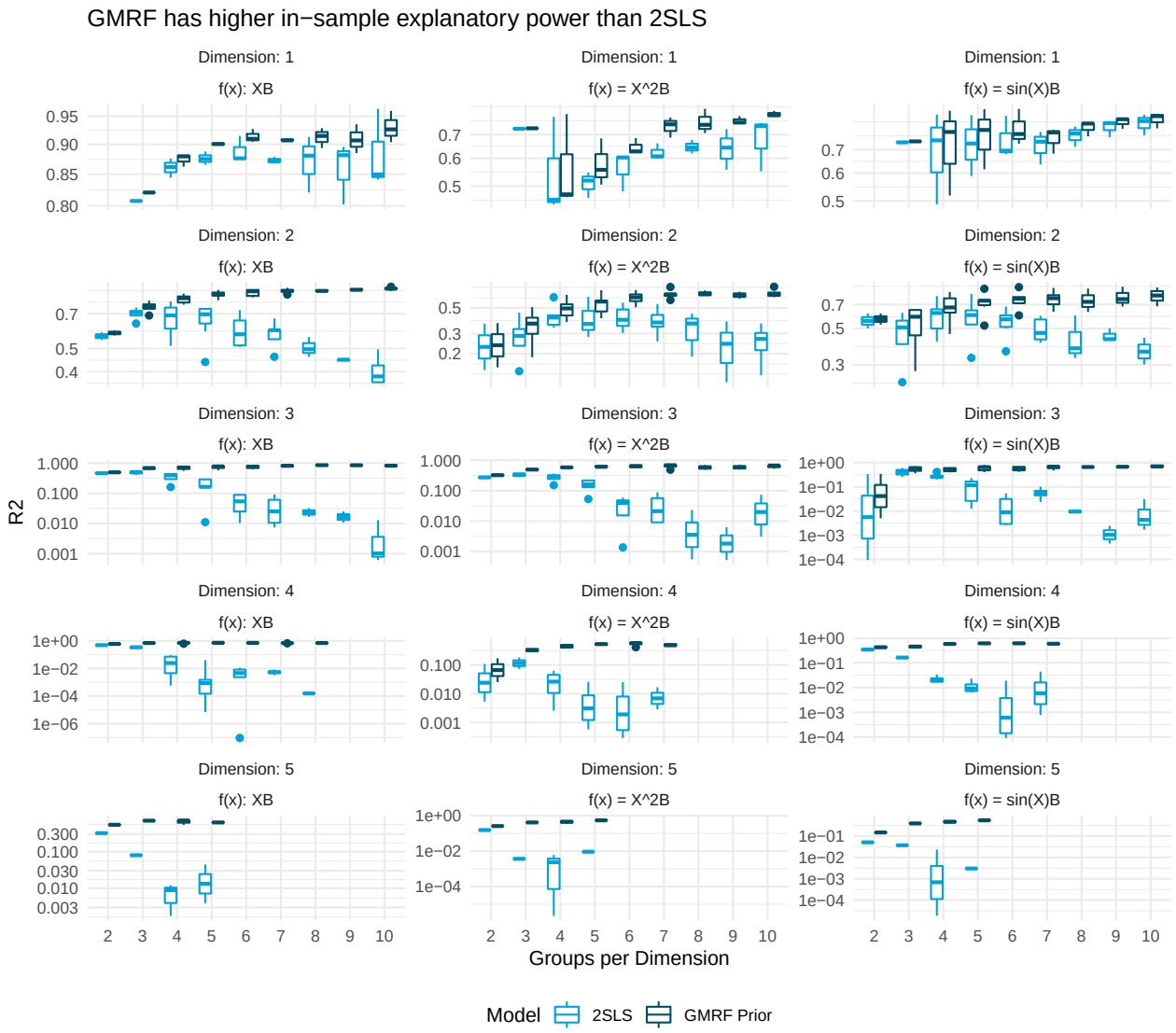

Figure 2: This figure shows the R^2 of the treatment effects estimated via independent two stage least squares by group and two stage least squares with a Gaussian Markov random field prior. The differences are enormous, with the GMRF model having comparable or better explanatory power across the board. R^2 is measured via $cor(E(T|M), T)^2$, where M is either the independent 2SLS model or the model estimated with the GMRF prior and T is the true treatment effect.

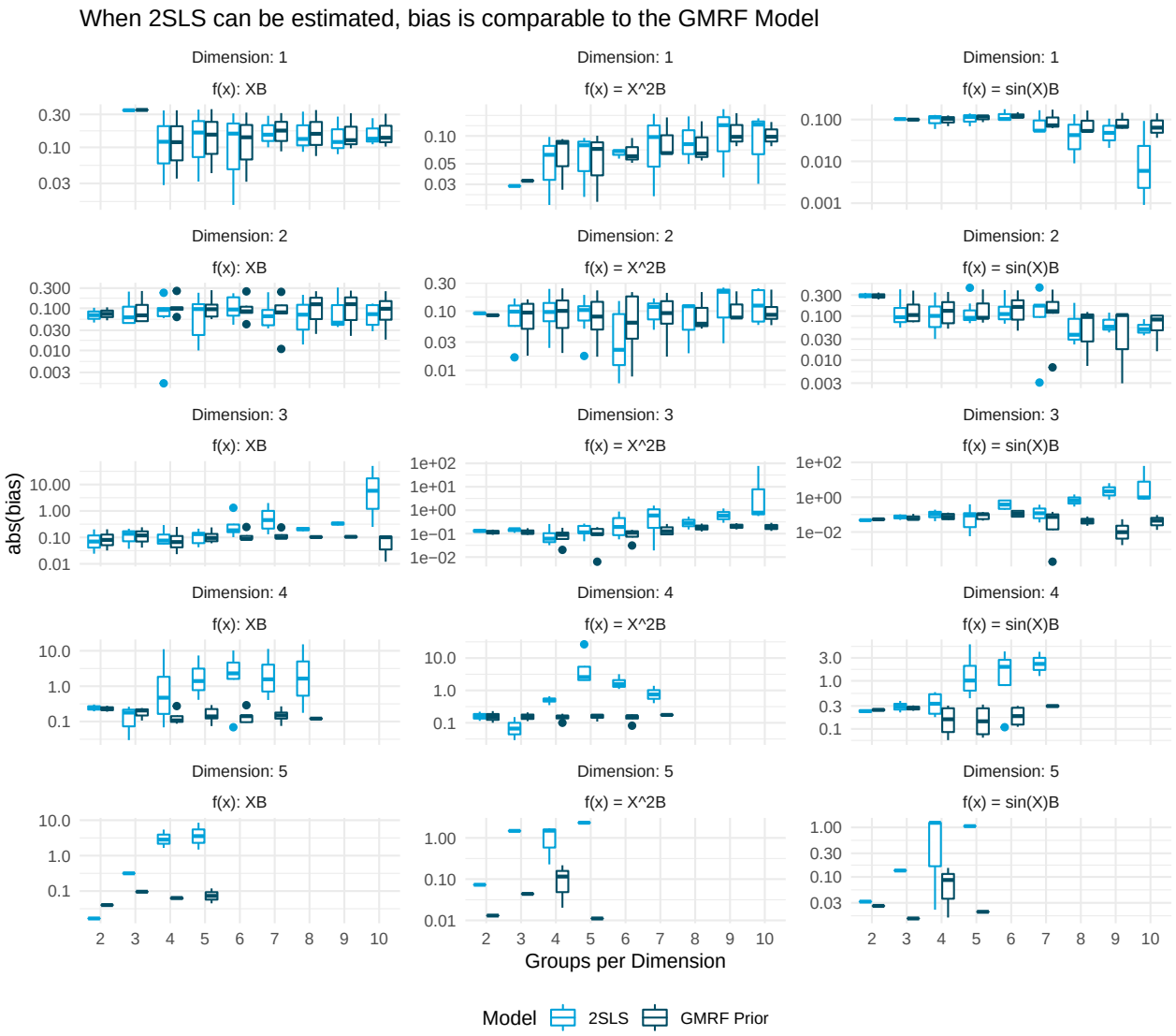

Figure 3: This figure shows the bias of two stage least squares by group and group-wise two stage least squares with a Gaussian Markov random field prior. While two stage least squares has no asymptotic bias, the difference in simulated finite sample bias between the models in small dimensions is not large. Bias is measured as $E(T|M) - T$, where M is either the independent 2SLS model or the model estimated with the GMRF prior and T is the true treatment effect.

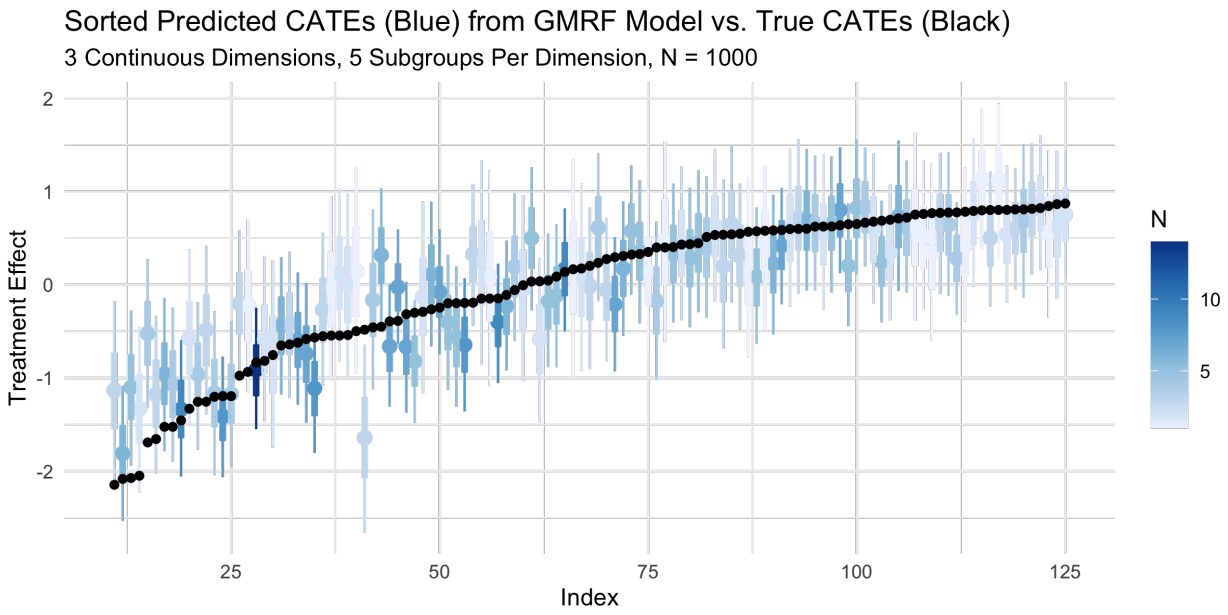

Figure 4: This figure shows the coverage of true treatment effects in high dimensional case using GMRF model. The model is estimated with three dimensions and 5 subgroups per dimension. Thick bands correspond to a 50% posterior predictive interval, and thin bands correspond to a 95% posterior predictive interval. The color of the bars correspond to the number of observations in the associated cell of the lattice.

effect function with a linear model, this might work well for a relatively local shift, but it will get farther and farther away from an accurate estimate except in the case where the true function is linear.

Since the GMRF model can be written as a series of piece-wise linear basis functions, we can consider the treatment effect for a new draw from a continuous covariate by linearly interpolating between neighboring values for β. In other words, say we observe an individual between two neighboring cells. For a simplicity, let us consider a single continuous variable, X, and a single observation x_i. Say we have two cells which are closest to X, x_j and x_q. Then we have estimates of conditional average treatment effects τ_j and τ_q. The prediction for τ_i is simply what would be the prior for that cell, e.g.

$$\tau_i \sim N(\frac{\tau_j \delta_j + \tau_q \delta_q}{\delta_j + \delta_q}, \frac{\sigma_j^2 \delta_j + \sigma_q^2 \delta_q}{\delta_j + \delta_q}),$$

where $\delta_z = |x_i - x_z|$, which is simply a weighted average of neighboring cells. The locally linear functional form allows for a continuous set of predictions for the continuous variable, even though we have only estimated a series of average effects.

Targeting Policy Functions

Optimizing policy functions is very straightforward with the model, as a fully Bayesian model is generative, in that it can directly simulate expected values for any quantity implied by the model. The cost/benefit analysis can be modeled formally be specifying a utility or loss function which represents preferences over outcomes implied by the model. If an agent is planning on taking some action, say providing a treatment to an individual, we can directly simulate the consequences of that action in terms of its cost and model-implied treatment effect. Because of the setting, I'll assume that potential benefits depend on the treatment effect associated with the action. Then, preferences over some act a implied by a model M are given by

$$U(a|M) = E(u(a)|M) = \int u(a|\tau) P_M(\tau) d\tau.$$

which is a standard von Neumann-Morgenstern (vNM) utility function (see, e.g. Savage (1972)).

To give a practical example, an agent may be risk average and have a per-unit cost, C, associated with a rollout. In this case, one potential functional form for the inner utility function would be

$$u(a) = log(\tau) - C.$$

Given the ability simulate from the model, we can get an arbitrarily good approximation of the integral by generating a number of Monte Carlo samples for τ given the information available. In other words, we have a conditional predictive density for τ, then we just approximate the integral with

$$U(a|M) = E(log(\tau) - C|M) \approx \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} \left(\log(\tau_i^{sim}) - C\right)$$

The posterior predictive distribution density will depend on the trained model, but in general (given an arbitrary model M with parameters θ, outcome y, and information set I), it is given by

$$P(y^{new}|M) = \int p(y^{new}|\theta, I) p(\theta|I) d\theta.$$

Suppose there are n possible people to give the treatment to. First, generate m samples, where m is chosen to lower monte carlo error below some tolerance. Then τ^{sim} is an $n \times m$ matrix, with columns corresponding to draws from the posterior predictive distribution for the individual in row n. U^{sim} is a matrix where $U_{i,j}^{sim} = \log(\tau_{i,j}^{sim}) - C$. Let w be a vector filled with m values of $\frac{1}{m}$. Then the utility maximization problem becomes

$$\max_a \quad a U^{sim} w,$$

which in this example is solved by setting $a_i = 1$ if $U^{sim} w > 0$, and 0 otherwise. If there is a cap on the budget to be consumed, then this simply involves rank-ordering expected utility and cutting off when the budget is exhausted. In a case where a is continuous, for example representing a percentage allocation to various subgroups, this becomes a standard convex optimization problem.

Portability: Simulation Results

In order to demonstrate how much explanatory power we gain by using this model, I benchmark it against a situation where we know the true LATE for an initial population, keep the same functional form for the treatment effects, and draw a new population. This benchmark is useful for two reasons: first, at the sample sizes available for subgroups in high dimensions, it is clear that independent estimates have extremely high variances, and second it helps identify the explanatory power we gain from the more flexible functional form. The "constant treatment effect" model is never argued for explicitly in an academic context, but it is implicit in a number of policy settings where politicians or bureaucrats use an aggregate causal estimate as evidence for whether to move forward with a specific program.

To test the portability of the results, I simulate a model where treatment effects are a non-linear function of a set of normal random variables, drawn from a random covariance matrix. Then, for each estimate of the model, I simulate 10 new populations, and predict the treatment effects using the linear basis representation in the prior section. I fit the model with a varying number of subgroups in the lattice in order to examine the potential benefits of using a finer mesh.

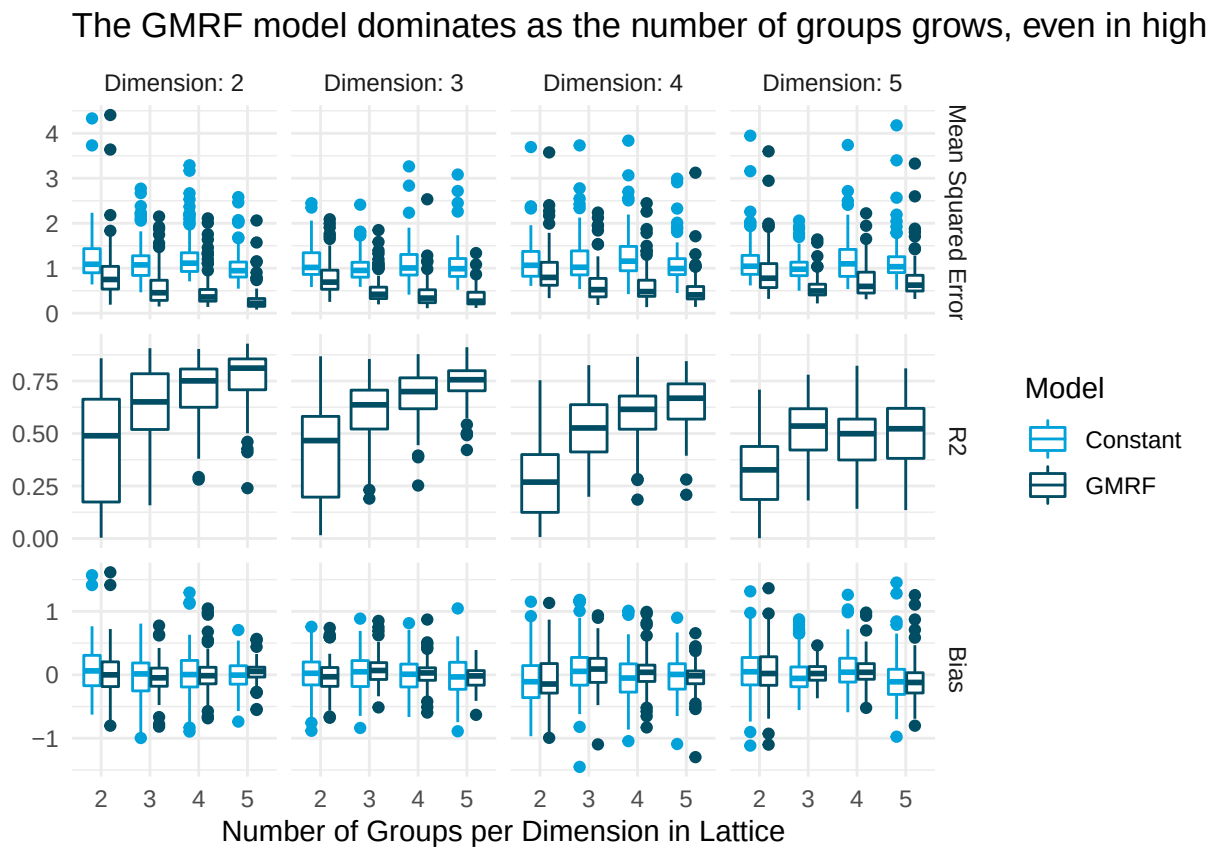

Figure 5: This figure shows the performance of a constant-treament-effect model and the GMRF prior model. The GMRF model has lower bias and lower mean squared error, as well as very high R^2, with magnitude of the improvement growing with the granularity of the lattice, even in high dimensions. The model is trained on a sample of 1,000 individuals, and extrapolates to a new population drawn from a random covariance matrix of the same dimension with 100 individuals.

The GMRF model has extremely high out of sample R^2 as well as mean squared error and lower bias than the constant treatment effects model, consistently improving as the number of groups gets large. Note that treatment effects in the initial population are scaled to have a standard deviation of 1, and new populations are drawn from a distribution with comparable variance. This means that we should expect the constant treatment effects model to have average mean squared error of 1, and helps put into context the magnitude of the improvements of the model. The average MSE across new populations is often more than halved by the GMRF prior, even through the model does not have perfect information about the treatment effects in the original population!

There are cases where the GMRF model does exceptionally well, and others where it still performs but not quite as well. Noticeably, the explanatory power of the same lattice shrinks as the number of dimensions grows. This is because in higher dimensions, with the same number of subgroups, proportionally more cells sit at the boundary, and are not informed as much by neighboring estimates. The boundary conditions of the model reduce to a constant treatment effects assumption (with growing variance) and so the predictions are not as accurate for new populations that are likely to have many observations outside of the original lattice. This problem may be mitigated through a more efficient construction of a lattice, either through a recursive tree-based method or through optimal tessellation of the covariate space (Lindgren, Rue, and Lindström 2011), but will be present in some form for any model where we have to extrapolate well beyond the bounds of what has previously been observed.

Conclusion

Gaussian Markov random fields are a flexible class of models which can effectively model conditional average treatment effects, approximating treatment effect functions of a continuous variable through a clever linear basis representation. It maintains the intuition of a divide-and-conquer strategy, while not running into the curse of dimensionality like independent group-wise estimation because of the partial pooling inherent in the model. The model is a principled prior, which penalizes overly complex functional forms in cases where a researcher expects adjacent treatment effect estimates to be informative for each other, while still asymptotically recovering arbitrary functional forms.

GMRF priors in multilevel models have lower mean squared error, higher R^2, and comparable bias in-sample compared to independent group-wise estimation, and have lower mean-squared error, lower bias, and high out of sample R^2 when projecting those effects to new populations. The model is not a black-box method, and is easily interpreted by analysts as a means to penalize candidate treatment effect functions when estimating LATEs by subgroup. The model assumptions can also be checked through prior and posterior predictive simulations (Gelman et al. 2020).

In future work, it would be useful to compare the performance of the model to other standard approaches, like Bayesian additive regression trees or causal forests. One interesting question is whether embedding this type of prior in a tree-like model could make such methods more efficient in cases with continuous variables, perhaps combining a series of weak learners with Bayesian model stacking (Yao et al. 2018).

References

Abrevaya, Jason, Yu-Chin Hsu, and Robert P Lieli. 2015. "Estimating Conditional Average Treatment Effects." *Journal of Business & Economic Statistics* 33 (4): 485–505.

Banerjee, Sudipto, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. 2008. "Gaussian Predictive Process Models for Large Spatial Data Sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (4): 825–48.

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society: Series B (Methodological)* 36 (2): 192–225.

Blangiardo, Marta, Michela Cameletti, Gianluca Baio, and Håvard Rue. 2013. "Spatial and Spatio-Temporal Models with r-INLA." *Spatial and Spatio-Temporal Epidemiology* 4: 33–49.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018. "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India." National Bureau of Economic Research.

Dorie, Vincent, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. 2019. "Automated Versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition." *Statistical Science* 34 (1): 43–68.

Gao, Yuxiang, Lauren Kennedy, and Daniel Simpson. 2021. "Treatment Effect Estimation with Multilevel Regression and Poststratification." *arXiv Preprint arXiv:2102.10003*.

Gao, Yuxiang, Lauren Kennedy, Daniel Simpson, Andrew Gelman, and others. 2020. "Improving Multilevel Regression and Poststratification with Structured Priors." *Bayesian Analysis*.

———. 2021. "Improving Multilevel Regression and Poststratification with Structured Priors." *Bayesian Analysis*.

Gelman, Andrew, and Thomas C Little. 1997. "Poststratification into Many Categories Using Hierarchical Logistic Regression."

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. "Bayesian Workflow." *arXiv Preprint arXiv:2011.01808*.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57 (3): 762–76.

Hill, Jennifer L. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20 (1): 217–40.

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Jaynes, Edwin T. 2003. *Probability Theory: The Logic of Science*. Cambridge university press.

Lindgren, Finn, Håvard Rue, and Johan Lindström. 2011. "An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (4): 423–98.

Little, Roderick JA. 1993. "Post-Stratification: A Modeler's Perspective." *Journal of the American Statistical Association* 88 (423): 1001–12.

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

Morris, Mitzi, Katherine Wheeler-Martin, Dan Simpson, Stephen J Mooney, Andrew Gelman, and Charles DiMaggio. 2019. "Bayesian Hierarchical Spatial Models: Implementing the Besag York Mollié Model in Stan." *Spatial and Spatio-Temporal Epidemiology* 31: 100301.

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-Oracle Estimation of Heterogeneous Treatment Effects." *Biometrika* 108 (2): 299–319.

Rubin, Donald B. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics* 6 (4): 377–401.

Rue, Havard, and Leonhard Held. 2005. *Gaussian Markov Random Fields: Theory and Applications*. CRC press.

Savage, Leonard J. 1972. *The Foundations of Statistics*. Courier Corporation.

Simpson, Daniel, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn H Sørbye. 2017. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science* 32 (1): 1–28.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with r*. CRC press.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2018. "Using Stacking to Average Bayesian Predictive Distributions (with Discussion)." *Bayesian Analysis* 13 (3): 917–1007.