November, 2015. Incomplete Draft. Please do not cite without permission.

Causal Inference of Meta-analysis via Gaussian Processes

Wei Wang
Columbia University
ww2243@columbia.edu

INTRODUCTION

Meta-Analysis, the synthesis of evidence from multiple study sources, has become increasing popular in fields such as education, psychology and public health (Cooper, Hedges, and Valentine 2009). The major obstacle for meta-analysis is the interpretation and proper handling of study-by-study heterogeneity, e.g., the estimated treatment effect in study 1 is different from in study 2. Traditional approaches tend to focus on developing novel ways of weighting different estimates based on certain measure of study-level uncertainty and/or quality. (Sobel, Madigan, and Wang 2016), however, approaches this problem from a formal causal inference perspective, proposing an extended potential outcome framework for meta-analysis. Although it is not a panacea for explaining and accounting for heterogeneities, this approach indeed clarifies the originations of heterogeneities and help researchers to think more clearly on underlying assumptions that are often overlooked. In this chapter, I review briefly the extended potential outcome framework discussed in (Sobel, Madigan, and Wang 2016). However, (Sobel,

Madigan, and Wang 2016) use simple linear models for analysis. In remaining part of this thesis, I develop a non-parametric model that explicitly handles heterogeneities across studies based on Gaussian Processes (GP). As it is well known (CKI Williams and Rasmussen 2006), GP allows for flexible modeling of response functions and admits fully probabilistic inference. Finally, a real educational intervention data set is analyzed with this model to illustrate. This chapter is joint work with Michael Sobel and David Madigan, and part of it is published in (Sobel, Madigan, and Wang 2016).

META-ANALYSIS

Meta-analyses combine data from distinct but related studies for higher resolution inference and more nuanced understanding of the effect under investigation. Originally hailed in medical and education research, meta-analyses are gaining traction as the awareness for open data is increasing across all scientific communities.

However, traditional meta-analyses are mostly conducted based on extracting and combining study-level effect summaries, since access to individual-participant level data tend to be inherently difficult to obtain. In this framework, researchers extract effect size estimates y_s and standard errors σ_s^2, where study index $s \in \{1, \ldots, S\}$ and S is the total number of studies. To handle effect size heterogeneity, typically a random effect model is used (DerSimonian and Laird 1986), in which all study effect sizes are assumed to be a random sample of a underlying hyper-population of effect sizes, i.e.

$$
\begin{aligned}
y_s &= \mu_s + \sigma_s^2 \\
\mu_s &\sim \mathcal{N}(\mu_0, \tau^2)
\end{aligned}
$$

Admittedly, meta-analysis based on study-level summary is still effective when the effects are homogeneous and different studies sample from similar populations; nevertheless, they are prone to well-known statistical fallacies, such as ecological bias, when the underlying populations and effects are heterogeneous, as it is often the case in real data.

Individual-Participant Meta-Analyses

Individual-Participant Data (IPD) Meta-Analyses are becoming more and more common, thanks to the increasing availability of original data (Julian Higgins

et al. 2001). It has been argued that IPD data increases the power of analysis (Cooper and Patall 2009) and more robust to heterogeneous effects sizes and populations. To account for between study heterogeneity in treatment effects, the use of covariates and/or random effects models is often recommended (Aitkin 1999; Tudur Smith, Williamson, and Marson 2005). The random effects models can be seen as Bayesian hierarchical models (Gelman and Hill 2006), based on the justification that conditioned on an appropriate set of covariates, both individual-level and study-level, the residual heterogeneities are exchangeable. There are mature softwares for fitting various types of Bayesian hierarchical models, including Generalized Linear Models and Proportional Hazard Models (Bates et al. 2015; "Stan: A C++ Library for Probability and Sampling, Version 2.8.0" 2015; Therneau 2012).

Despite its convenient form and ease of inference, traditional IPD meta-analysis based on parametric hierarchical models suffer from two problems. The first is the lack of formal causal framework. It is difficult to pinpoint the causal interpretation of the effect estimates from a traditional IPD hierarchical model. Consider the following example, education researchers try to determine the effect of a new intervention program, applied to different classroom and administered by different teachers. In this case, the heterogeneity might come from either the different teachers or the different populations of schools, or both. It is often unclear whether the effect estimate based on traditional methods are averages over the teachers, or over the schools, or both. The second problem is the inflexible form of the parametric model. Traditional parametric model requires explicit modeling assumptions from the researchers, which makes the model sensitive to model specifications and facilitate potential cheery-picking. Non-parametric modeling allows flexible functional form and requires little manual tuning from the researchers.

A POTENTIAL OUTCOME FRAMEWORK FOR META-ANALYSIS

(Sobel, Madigan, and Wang 2016) put meta-analysis on a concrete causal foundation by introducing an extended potential outcome framework. I will discuss the key ideas in this framework.

Potential Outcomes

Potential Outcomes Framework (Rubin 2011) defines causal effects as comparisons of outcomes under hypothetical counter-factual treatment assignment. For example, with binary treatment $Z \in \{0, 1\}$, the causal effect of treatment Z on individual i can be defined as $y_i(1) - y_i(0)$. Typically, researchers are interested in estimating quantities such as the population average treatment effect (PATE)

$$E(Y(1) - Y(0)),$$

and the population average treatment effect on the treated (PATT)

$$E(Y(1) - Y(0) \mid Z = 1)$$

The key assumption in causal inference is the ignorability assumption (or unconfoundedness assumption) (Rosenbaum and Rubin 1983), which states that given a set of observed covariates, the treatment assignment Z is independent of the potential outcomes $(Y(0), Y(1))$

$$Y(0), Y(1) \perp Z \mid X$$

In the case of randomized experiment, this assumption is trivially met without any covariates X. Under ignorability assumption,

$$\begin{aligned} & E(Y|X, z = 1) - E(Y|X, z = 0) \\ =& E(Y(1)|X, z = 1) - E(Y(0)|X, z = 0) \\ =& E(Y(1)|X) - E(Y(0)|X) \\ =& E(Y(1) - Y(0)|X) \end{aligned}$$

and thus causal effect can be identified from observations.

Extended Potential Outcomes

In the case of meta-analysis, consisting S studies and Z treatment arms, the potential outcomes $\boldsymbol{Y}$ for individual i can be defined as a matrix

$$\boldsymbol{Y}_i = \begin{pmatrix} y_i(1,1) & y_i(1,2) & \cdots & y_i(1,Z) \\ y_i(2,1) & y_i(2,2) & \cdots & y_i(2,Z) \\ \vdots & \vdots & \ddots & \vdots \\ y_i(S,1) & y_i(S,2) & \cdots & y_i(S,Z) \end{pmatrix}$$

With this notation, some commonly discussed meta-analytical estimates can be interpreted in a causal way. For example, assuming there are only two level of treatment (0 and 1) and the causal comparison is the difference, *study-specific* treatment effect for study s is $E(y(z,s)-y(z',s))$. Note that this is different from *study-level* treatment effect θ_s in random effects models, which is $E(y(z,s) - y(z',s) \mid S = s)$. Below I will discuss conditions that will connect these two quantities.

In the context of meta-analyses, unconfoundedness can be recast as unconfoundedness within each study, i.e.,

$$Y(0,s), Y(1,s) \perp Z \mid X, S = s$$

However, this assumption is not sufficient for identifying causal effects in meta-analysis. One added layer for complexity of meta-analysis is the confounding of study selection. Consider an example of clinical trials. If some studies sample from mostly young patients while some other studies sample from mostly elderly patients, and the treatment is more effective on younger patients, then heterogeneities in treatment effects across studies would arise. Hierarchal models without adequately addressing this selection problems would result in misleading results.

However, study selection is not the only factor contributing to heterogeneities in treatment effects across studies. One lingering question is whether the same treatment z is implemented identically in all studies, or in another word, whether $Y_i(s_1,z) \overset{\text{d}}{=} Y_i(s_2,z)$ for all pair of $s_1, s_2 \in \{1,\ldots,S\}$, where $\overset{\text{d}}{=}$ stands for equal in distribution. Consider an example of education intervention, in which interventions are carried out by teachers with various experience levels, then it is reasonable to question whether $Y_i(s_1,z) \overset{\text{d}}{=} Y_i(s_2,z)$ holds.

Two assumptions from (Sobel, Madigan, and Wang 2016) codify these two sources of heterogeneities.

A1. *Weak response consistency assumption for treatment* z: For any $z \in \{1,\ldots,Z\}$ and any pair $s_1, s_2 \in \{1,\ldots,S\}$,

$$Y_i(s_1,z) \overset{\text{d}}{=} Y_i(s_2,z)$$

A2. *Unconfounded study selection*:

$$\boldsymbol{Y} = \begin{pmatrix} y(1,1) & y(1,2) & \cdots & y(1,Z) \\ y(2,1) & y(2,2) & \cdots & y(2,Z) \\ \vdots & \vdots & \ddots & \vdots \\ y(S,1) & y(S,2) & \cdots & y(S,Z) \end{pmatrix} \perp S \mid X$$

.

From a modeling perspective, these two assumptions cannot be distinguished from one other. Thus (Sobel, Madigan, and Wang 2016) suggests that researchers first assess the plausibility of the two assumptions based on the characteristics of the studies, and typically assume one of these two to hold and then build models to see whether the heterogeneity could be accounted for by the other assumption. From a Bayesian point of view, I can use a very general model, and encode regularization through appropriate prior distributions to allow for reasonable separation of these two sources of heterogeneities. This will be the topic of the following sections.

META-ANALYSIS USING BAYESIAN NON-PARAMETRICS

Traditionally, causal inference using potential outcomes focuses on two questions. Modeling of the treatment assignment process $p(z \mid x)$, also known as the propensity score, and modeling of the scientific process of how responses relate to treatment and covariates $p(y \mid z, x)$, also known as the response surface (Rubin 2005). A myriad of methods based on the either treatment assignment mechanism (e.g., propensity score matching), or response surface modeling (e.g., regression), or combination of these two (e.g., the doubly-robust method), has been proposed for causal inference of observational data.

Recently, following the advances in Bayesian non-parametric models , (J. L. Hill 2011) proposed a model that focuses on accurately estimating the response surface using flexible Bayesian Additive Regression Trees, or BART (Chipman, George, and McCulloch 2010). Besides the well-known benefits of being robust to model misspecifications and being able to capture highly non-linear and interaction patterns, Bayesian non-parametric models provide natural and coherent posterior intervals to convey inferential uncertainty.

Gaussian Processes

Gaussian Processes (GP) have become a popular tool for nonparametric regression. A random function $f : \mathcal{X} \to \mathbb{R}$ is said to follow a GP process with kernel k if any

finite-dimensional marginal of it is Gaussianly distributed, i.e.

$$f(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}}), \forall\, \boldsymbol{x} \in \mathbb{R}^d \text{ and } d$$

where $\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}}$ is the Gram matrix of kernel k. The key component in a GP model is the kernel k, a semi-definite function defined on $\mathcal{X} \times \mathcal{X}$ that encodes the structure. Judiciously choosing K is the most important part of fitting a GP model.

A large part of its popularity is probably due to the fact it can be interpreted as a generalization of linear regression with Gaussian errors, the predominant model for parametric regression. In fact, according to Mercer's Theorem (CKI Williams and Rasmussen 2006), kernel k can be decomposed into

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i^{\intercal}(x')$$

where λ_i and ϕ_i are respective eigenvalues and eigenfunctions of kernel k with respect to a measure μ, i.e.,

$$\int k(x, x') \phi_i(x)\, d\mu(x) = \lambda \phi_i(x'),$$

Then GP can be considered as a basis expansion method that maps input x to an infinite dimensional space via the infinite series of functions $\{\phi_i(x)\}_{i=1}^{\infty}$.

Inference for Standard GP

Standard GP model for N observation pairs $(y_i, \boldsymbol{x}_i)_{i=1}^{N}$ is

$$\begin{aligned} y_i \mid f &\sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2) \\ f &\sim GP(0, k) \end{aligned}$$

For a given kernel k, the marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}(0, K_{\boldsymbol{x},\boldsymbol{x}} + \sigma^2 I_N)$$

where $K_{\boldsymbol{x},\boldsymbol{x}}$ is the Gram matrix of kernel k whose entries are $k(x_i, x_j)$. The predictive distribution at test points $\boldsymbol{X}^{\star}$ is

$$\boldsymbol{y}^\star \mid \boldsymbol{X}^\star, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}(K_{\boldsymbol{X}^\star,\boldsymbol{X}}(K_{\boldsymbol{X},\boldsymbol{X}} + \sigma^2 I_N)^{-1}\boldsymbol{y}, \\ K_{\boldsymbol{X}^\star,\boldsymbol{X}^\star} - K_{\boldsymbol{X}^\star,\boldsymbol{X}}(K_{\boldsymbol{X},\boldsymbol{X}} + \sigma^2 I_N)^{-1}K_{\boldsymbol{X}^\star,\boldsymbol{X}}^\intercal)$$

For inference on hyperparameters, e.g., parameters governing the kernels, a standard practice is to maximize log marginal likelihood

$$\log p(\boldsymbol{y} \mid \boldsymbol{X}, \theta) = \log \int p(\boldsymbol{y} \mid f, \boldsymbol{X}, \theta) p(f)\, df \\ \propto -\left[\boldsymbol{y}^\intercal (K_{\boldsymbol{X},\boldsymbol{X}}(\theta) + \sigma^2 I_N)^{-1}\boldsymbol{y} + \log \det(K_{\boldsymbol{X},\boldsymbol{X}}(\theta) + \sigma^2 I_N)\right]$$

and plug in the MAP (maximum a posteriori) $\hat{\theta}$ into the predictive distribution of new points $\boldsymbol{X}^\star$.

As it is well known, despite the simplicity of the procedure for GP inference, the main difficulty lies in the matrix inversions required for both estimating hyperparameters and predicting at new points, which involves $\mathcal{O}(N^3)$ time complexity with N being the number of observations. Data sets that have more than several thousands observations are already prohibitively expensive for computation. In those cases, a number of approximation methods such as low-rank approximations of the Gram kernel matrix, known as Nystr{"o}m method (Christopher Williams and Seeger 2001), and judicious selections of subset of observations (Banerjee et al. 2008) are often recommended.

GP with Hierarchical Structure

GP can be extended to handle group structure. A common approach from machine learning perspective is to pose this question as a multi-task learning problem (Bonilla, Chai, and Williams 2008; Yu, Tresp, and Schwaighofer 2005), in which the objective function's values are vectors, or even matrices. In fact, in the case of the potential outcome framework of meta-analysis outlined above, the outcomes are study-by-treatment matrices. In this setting, the curse of dimensionalities become even more acute, as the sample size effectively multiples by the number of outcomes under investigation. One remedy is to place restriction on the structure of kernels. For example, assuming the kernel is separable, the finite dimensional marginals of the vector-valued random function $\boldsymbol{f}$ is a matrix normal distributed

$$\mathrm{vec}\boldsymbol{f}(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}}, \otimes \boldsymbol{U}_{\text{task}})$$

Assuming separability can significantly reduce the dimensionality of the problem, and properties of Kronecker product can be used to make the inference efficient (Gilboa, Saatçi, and Cunningham 2015). However, this approach works best for the case of "complete design", meaning every combination of predictors values have been observed. This assumption is reasonable for areas such as computer experiments and robotics, where experiments can be artificially planned at pre-specified values of X's; but it is virtually impossible in fields such as education and public health, where researchers' control over data collection is very limited.

Incorporating Group Structure into Kernels However, it is relatively straightforward to directly code the group structure into kernels. Take the most popular kernel, square exponential, for example, discrete group indicator terms can be added by using delta metrics, i.e., 0 if two observations are from the same group, and 1 otherwise. Further, it is often important to add group-level predictors in hierarchical models (Gelman and Hill 2006), as it can account for the variations that cannot be explained by categorical group membership.

$$\kappa(x_i, x_j) = \sigma^2 \exp -\frac{1}{2} \sum_{k=1}^{d} \frac{(x_{ik} - x_{jk})^2}{l_k^2}$$

In square exponential kernels, the length scale l_k governs the correlation scale in input dimension k and the magnitude σ^2 controls the overall variability of the process. Thus the magnitudes of the length scale l_k can be used for feature selection; the larger the length scale, the more important the corresponding feature.

Network Meta-Analysis

The compact form of Kronecker product assumes a block-design structure, i.e., the same set of x's are observed for every combination of study and treatment. In reality, of course, this is not the case; causal inference, in particular, is about filling those holes, i.e., potential outcomes, in hypothetical combination of study and treatment. In fact, researchers only observe a small fraction of the array of matrices $\boldsymbol{y}_{ii} = 1^{\infty}$, namely $\frac{1}{ST}$ of all potential outcomes. However, this framework is general enough to handle a lot of particular problems.

Network Meta-analysis (Lumley 2002) deal with treatment pair comparisons that depend on indirect evidence. For example, if treatment A and treatment B are not assigned in any of the studies at the same time, and thus researchers have

to resort to indirect comparisons, e.g., treatment C co-occurs with treatment A and treatment B in some of the studies. Since traditional analysis tend to handles one comparison at a time, violations of natural constraints are frequent, e.g., AB+BC≠BC. More sophisticated models are proposed to handle this so-called "inconsistency" (JPT Higgins et al. 2012), which makes the models unnecessarily complex and is detrimental to intuitive understanding. The framework outlined in this chapter, however, naturally deal with network meta-analysis, since it considers all possible treatment at the same time and thus have those constraints built in organically.

REAL DATA EXAMPLE

The demonstrative example I use is the STAR (Student-Teacher Achievement Ratio) project. It was approved by Tennessee state legislature and began in 1985 to study the effect of early grades class size on student achievement in Tennessee. The study is a state-wide randomized experiment applied to over 7,000 pupils from 79 schools and last for 4 years. Each student was randomly assigned to one of three class types, class of 13 to 15 students, class of 22 to 25 students, and class of 22 to 25 students with a paid teaching aid. Outcomes of end-of-year test scores were used to assess the performance of those students in areas of math, reading and study skills. Classroom teachers were also randomly assigned to the classes they would teach. The interventions were initiated as the students entered school in kindergarten and continued through third grade, based on the common belief that early intervention has persistent effects well into later lives of the students. Due to its size and ambition, STAR is perhaps the most important education study in history. Numerous studies have been devoted to analyzing the STAR data, on both immediate effects, e.g., test scores at the end of the year of intervention (krueger1997experimental; Hanushek, Mayer, and Peterson 1999; Word and others 1990), and persistent effects, e.g., test scores several years after the intervention or even earning as an adult (Chetty et al. 2010).

The public access data set is collected from Project STAR Web site at `http://www.heros-inc.org/star.htm`, with information on the student demographics, test scores, treatment assignments over the intervention years, information of the teachers, and school situations et al. Due to its richness, STAR project data can be investigated in many different facets. For the sake of simplicity, I look at just one outcome, scaled math test score, in one intervention year, the 1st grade. Thus I can focus on the meta-analytic part of the data, without being distracted

by the longitudinal aspect of the data, which is a nuisance for the discussion. To be precise, I only study the students who participated STAR project in their first grades and use their scaled math score at the end of the first year as outcomes. Further, as mentioned previously, data sets with thousands of observations pose prohibitive computational burden on GP. So I select a subset of the data set, including only 8 biggest schools for the first grade. The sample size of this restricted data set is about 1,000.

Characteristics such as student gender, ethnicity, receiving free lunch or not are included in the data set; furthermore, I can determine the general neighborhood economic situations by calculating the proportion of students receiving free lunch, a school-level predictor. As for types of treatment, as I mentioned, there are three types of treatment, small-size classroom, regular-size classroom and regular with a paid teaching aid. Instead of combining regular and regular with an aid, an approach adopted by most of previous literature, I treat them as separate interventions. All of the above predictors are fed into a square exponential kernel, with a delta metric for discrete predictors including type of treatment and school ID. Computations are conducted via an MATLAB toolbox `GPStuff` (Vanhatalo et al. 2013).

While traditional parametric inference focuses on interpretation of some model coefficient estimates, whose validity greatly hinges on the validity of the model specifications, non-parametric inference attempts to construct the response surfaces and yields much more faithful uncertainty estimates when extrapolating. The results of inference are presented as a series of figures below. In each of the figures, predictive estimates for a student in a certain demographic subgroup, e.g., a minority girl receiving free lunch, were she in each of the 8 schools, are presented side-by-side under three different treatments. The schools are arranged in order of proportions of students receiving free lunch, a proxy of neighborhood economic situation, from most affluent to the left to the most deprived to the right. Figure 1 denotes a minority pupil with free lunch, figure 2 a minority pupil with paid lunch, figure 3 a white pupil with free lunch and figure 4 a white pupil with paid lunch.

There are several interesting points worth noting from the figures. First, in general, attending smaller class translates to a modest increaes in scaled math test score for the first year pupils, and in particular, the effects are more pronouced in schools with higher level of poverty, even after adjusting for proportions of students receiving free lunch. The heterogeneity of the small class sizes on educational outcomes have been noted with traditional parametric analysis (Krueger 1997), but most often by adding parametric interaction terms that

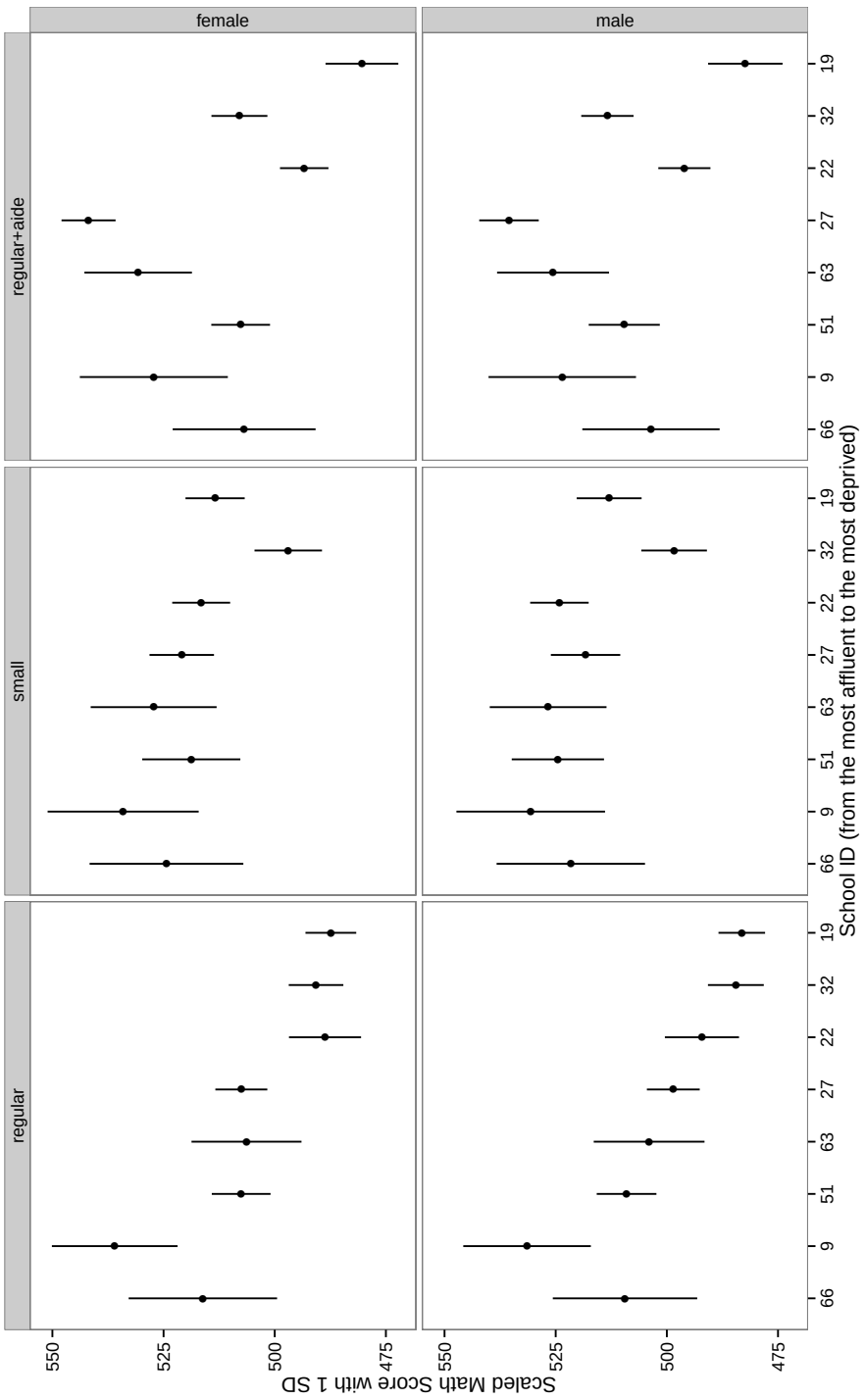

Figure 1: Above: Counterfactual scaled math scores with one standard deviation if a minority girl receiving free lunch were assigned to 8 different schools and 3 different treatments. Below: Counterfactual scaled math scores with one standard deviation if a minority boy receiving free lunch were assigned to 8 different schools and 3 different treatments. Schools are ordered from left to right by the proportions of student receiving free lunch.

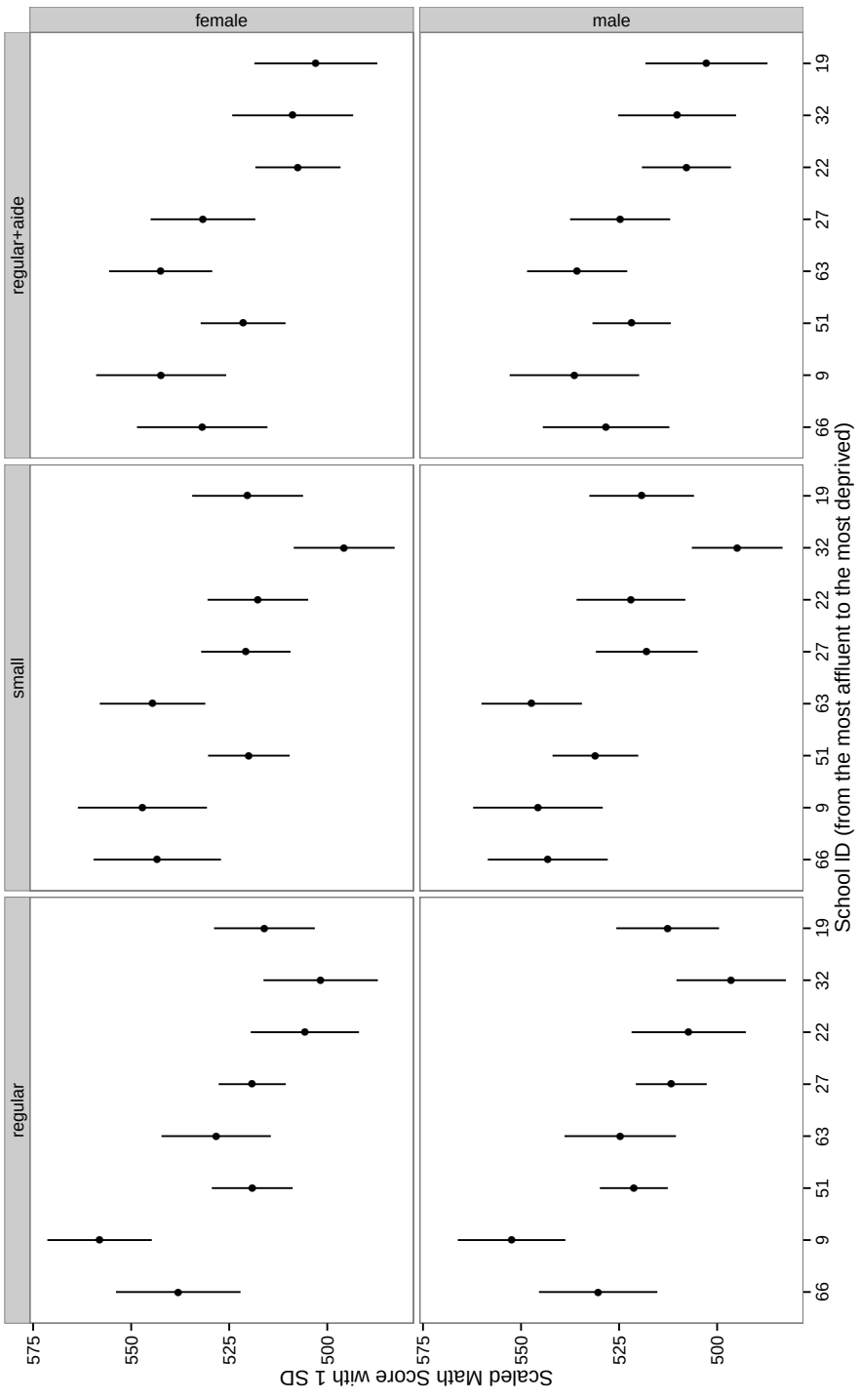

Figure 2: Above: Counterfactual scaled math scores with one standard deviation if a minority girl not receiving free lunch were assigned to 8 different schools and 3 different treatments. Below: Counterfactual scaled math scores with one standard deviation if a minority boy not receiving free lunch were assigned to 8 different schools and 3 different treatments. Schools are ordered from left to right by the proportions of student receiving free lunch.

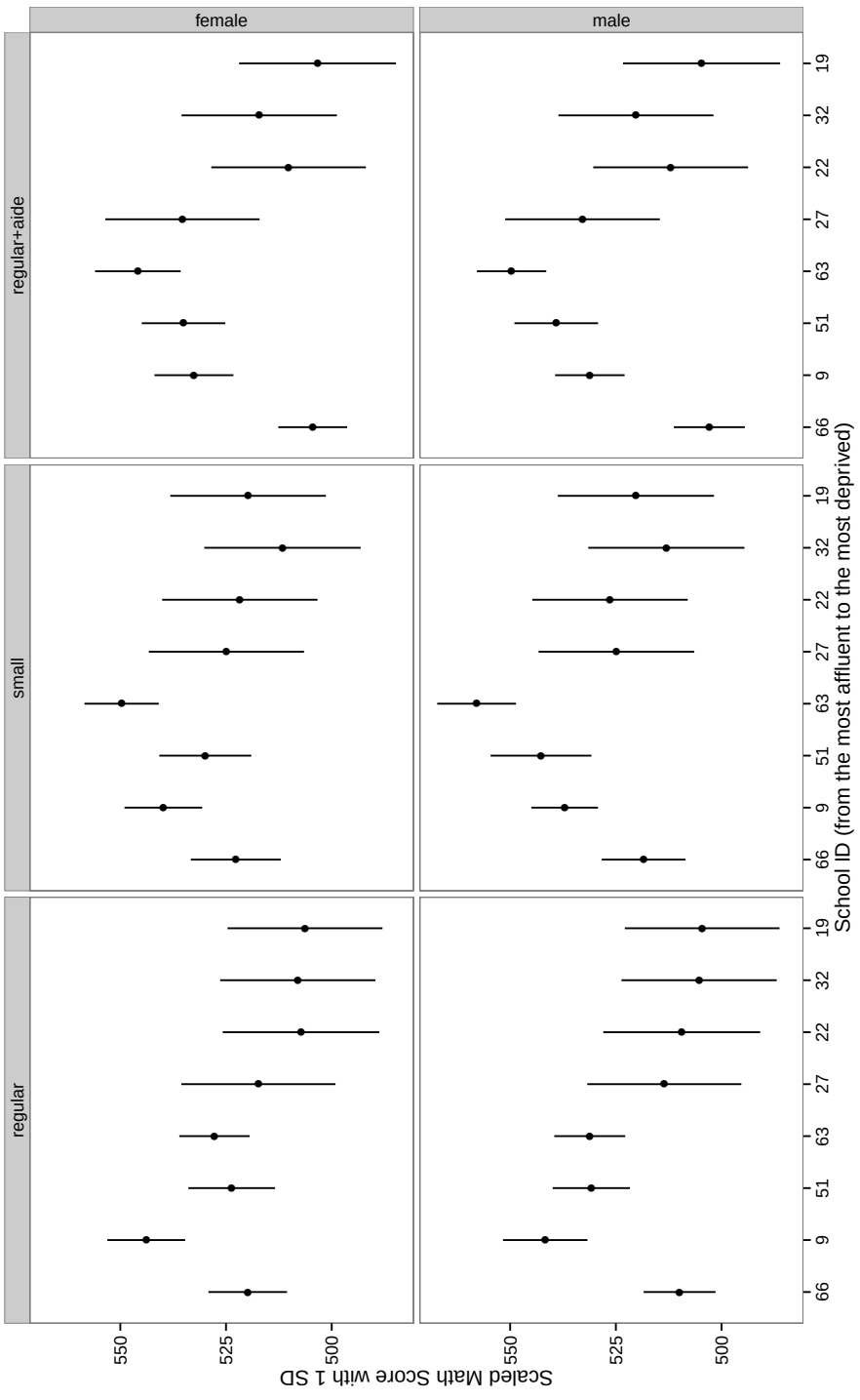

Figure 3: Above: Counterfactual scaled math scores with one standard deviation if a white girl receiving free lunch were assigned to 8 different schools and 3 different treatments. Below: Counterfactual scaled math scores with one standard deviation if a white boy receiving free lunch were assigned to 8 different schools and 3 different treatments. Schools are ordered from left to right by the proportions of student receiving free lunch.

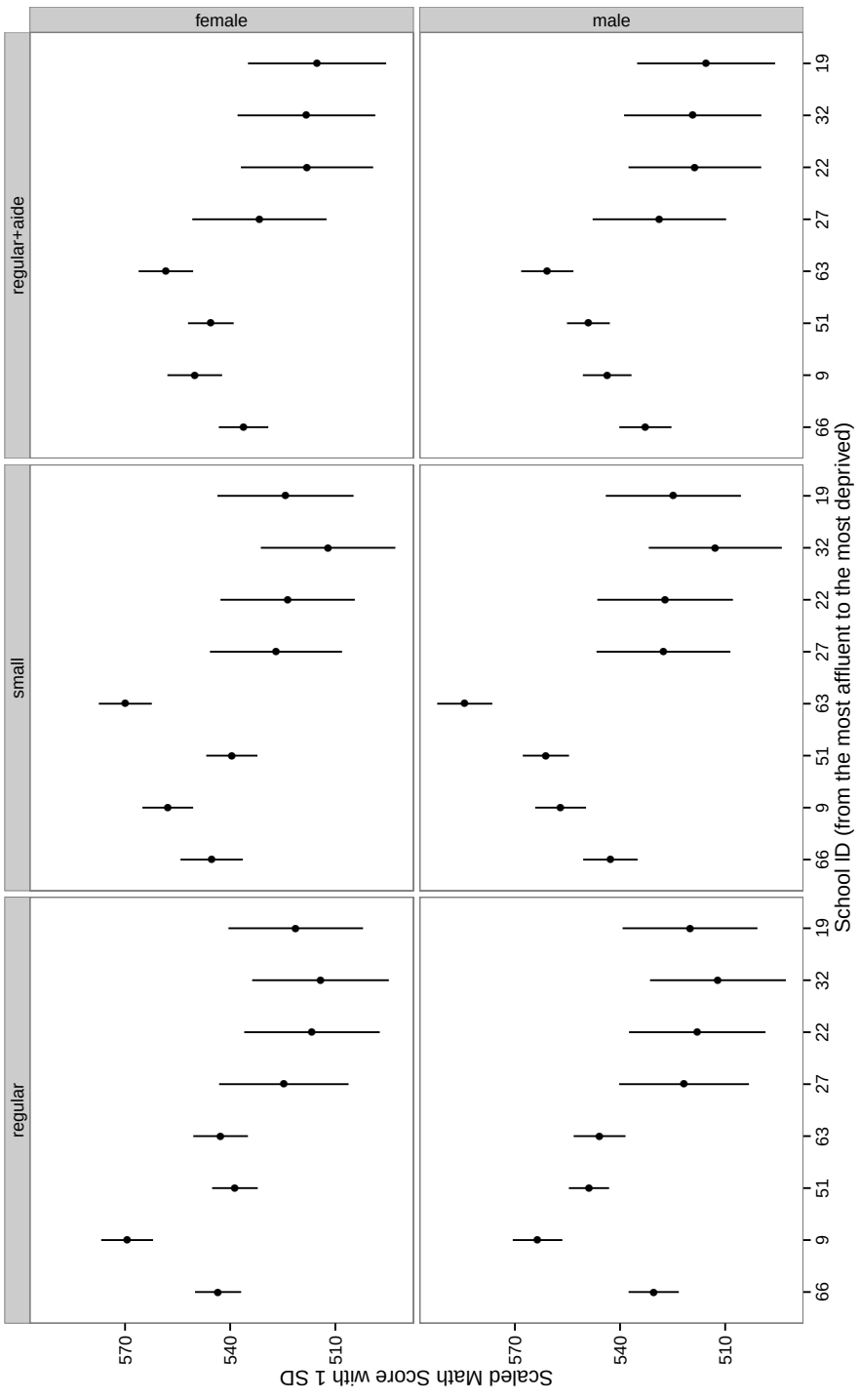

Figure 4: Above: Counterfactual scaled math scores with one standard deviation if a white girl not receiving free lunch were assigned to 8 different schools and 3 different treatments. Below: Counterfactual scaled math scores with one standard deviation if a white boy not receiving free lunch were assigned to 8 different schools and 3 different treatments. Schools are ordered from left to right by the proportions of student receiving free lunch.

requires very specific assumptions; without specifying parametric form, GP could uncover patterns of heterogeneities. Second, note that uncertainty is small whenever more observations are available. For example, in figure 1, which conrresponds to poor minority pupils that are present in schools to the right side of the figures, those schools to the right indeed carry much shorter error bar as compared to more affluent schools to the left. Meanwhile, estimates in figure 4 go in the opposite direction: schools to the left have much shorter error bars, since white, non-poor pupils are more present in those schools. Although GP are just doing what they are supposed to do, it is reassuring to know that fidelity is being preserved for uncertainty estimates. Thrid, a sizable amount of variation can be observed for a pupil receving the same treatment under differnet schools. This suggests a departure from Weak Response Consistency Assumption (A1). This is not surprising, because otehr than the same class size, different schools definitely offer education at different qualities. Take Figure 1 for example, which focuses on a minority and economically disadvantaged pupil. School 9 tends to do consistently very well across all three treatments, but school 19 is the school that would benefit the most by adopting a small class size.

DISCUSSION

In this chapter, I discuss an extended potential outcome framework built for meta-analysis. Comparing with the classic potential outcome framework, a plethora of counterfactuals with certain structures need to be created to handle meta-analysis. I then introduce a GP based approach to tackle this problem. The advantages of GP, and in general any non-parametric methods, is well-known. In particular, the fidelity of inferential uncertainty is a very desirable property. The central question to the extended potential outcome framework is how to incorporate the group structure. I discuss different ways of building group structure into Gaussian Processes. The most intuitive and computationally straight-forward approach is used to re-analyze an influential educational intervention program, the STAR project on class size and test scores. Lucid and straight-forward visualization can be used to display the inferential results, and reveal patterns that are otherwise hidden in tables of coefficients estimates often seen in traditional parametric analysis.

GP are a much sought-after field of research recently and reasonably so. Applying it to casual inference, and in particularly causal inference with group structure, can surely yield ellucidating insights. The scope of this chapter is quite

limited, and I hope it demonstrates the potential of non-parametric methods in causal inference.

BIBLIOGRAPHY

Aitkin, Murray. 1999. "Meta-Analysis by Random Effect Modelling in Generalized Linear Models." *Statistics in Medicine* 18(17-18): 2343–51.

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.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67(1): 1–48.

Bonilla, Edwin, Kian Ming Chai, and Christopher Williams. 2008. "Multi-Task Gaussian Process Prediction."

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2010. *How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR*. National Bureau of Economic Research.

Chipman, Hugh A, Edward I George, and Robert E McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *The Annals of Applied Statistics*: 266–98.

Cooper, Harris, and Erika A Patall. 2009. "The Relative Benefits of Meta-Analysis Conducted with Individual Participant Data Versus Aggregated Data." *Psychological methods* 14(2): 165.

Cooper, Harris, Larry V Hedges, and Jeffrey C Valentine. 2009. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation.

DerSimonian, Rebecca, and Nan Laird. 1986. "Meta-Analysis in Clinical Trials." *Controlled Clinical Trials* 7(3): 177–88.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge University Press.

Gilboa, Elad, Yunus Saatçi, and John P Cunningham. 2015. "Scaling Multidimensional Inference for Structured Gaussian Processes." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 37(2): 424–36.

Hanushek, Eric A, Susan E Mayer, and Paul Peterson. 1999. "The Evidence on Class Size." *Earning and learning: How schools matter*: 131–68.

Higgins, JPT, D Jackson, JK Barrett, G Lu, AE Ades, and IR White. 2012. "Consistency and Inconsistency in Network Meta-Analysis: Concepts and Models for Multi-Arm Studies." *Research Synthesis Methods* 3(2): 98–110.

Higgins, Julian, Anne Whitehead, Rebecca M Turner, Rumana Z Omar, and Simon G Thompson. 2001. "Meta-Analysis of Continuous Outcome Data from Individual Patients." *Statistics in Medicine* 20(15): 2219–41.

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

Krueger, Alan B. 1997. *Experimental Estimates of Education Production Functions*. National Bureau of Economic Research.

Lumley, Thomas. 2002. "Network Meta-Analysis for Indirect Treatment Comparisons." *Statistics in Medicine* 21(16): 2313–24.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes." *Journal of the American Statistical Association* 100(469).

———. 2011. "Causal Inference Using Potential Outcomes." *Journal of the American Statistical Association*.

Sobel, Michael E, David B Madigan, and Wei Wang. 2016. "Meta-Analysis: A Causal Framework, with Application to Randomized Studies of Vioxx." *Psychometrika*.

"Stan: A C++ Library for Probability and Sampling, Version 2.8.0." 2015. http://mc-stan.org/.

Therneau, Terry. 2012. "Coxme: Mixed Effects Cox Models." *R package version* 2(3).

Tudur Smith, Catrin, Paula R Williamson, and Anthony G Marson. 2005. "Investigating Heterogeneity in an Individual Patient Data Meta-Analysis of Time to Event Outcomes." *Statistics in medicine* 24(9): 1307–19.

Vanhatalo, Jarno, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. 2013. "GPstuff: Bayesian Modeling with Gaussian Processes." *The Journal of Machine Learning Research* 14(1): 1175–79.

Williams, Christopher, and Matthias Seeger. 2001. "Using the Nyström Method to Speed up Kernel Machines." In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, 682–88.

Williams, CKI, and CE Rasmussen. 2006. *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.

Word, Elizabeth R, and others. 1990. "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report (1985-1990)."

Yu, Kai, Volker Tresp, and Anton Schwaighofer. 2005. "Learning Gaussian Processes from Multiple Tasks." In *Proceedings of the 22nd International Conference on Machine Learning*, ACM, 1012–19.