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 Please do not cite without permission.

Structured Meta-Analysis via Hierarchical Gaussian Processes and Potential Outcomes

Wei Wang
Columbia University
ww2243@columbia.edu

Meta-Analysis, the synthesis of evidence from multiple study sources, has become increasing popular in fields such as education, psychology and public health. The major obstacle for meta-analysis is the interpretation and proper handling of study-by-study heterogeneity. Based on (Sobel, Madigan, and Wang 2015), which proposed a potential outcome framework for a formal causal approach to meta-analysis, I develop a high-dimensional Gaussian Process model that explicitly handles heterogeneities across studies, allows flexible modeling of response functions and admits fully probabilistic inference.

INTRODUCTION

Meta-Analysis, the synthesis of evidence from multiple study sources, has become increasing popular in fields such as education, psychology and public health. The major obstacle for meta-analysis is the interpretation and proper handling of study-by-study heterogeneity. Based on (Sobel, Madigan, and Wang 2015), which proposed a potential outcome framework for a formal causal approach to meta-analysis, I develop a high-dimensional Gaussian Process model that explicitly handles heterogeneities across studies, allows flexible modeling of response functions and admits fully probabilistic inference.

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 gain traction in wider academic disciplines 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 $s \in \{1, \dots, S\}$ and S is the number of studies. To handle effect size heterogeneity, a random effect model is typically used (DerSimonian and Laird 1986), in which all study effect sizes are random samples of a underlying 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-analyses based on study-level summary is still effective when the effects are homogeneous and different studies sample from similar population, they nevertheless 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-Analysis is becoming increasing common (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 [cite]. 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 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. [cite lme4, coxme and stan]

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. Is the effect estimate averaging over teachers? Schools? Or both? What can we say about the effect for a new teacher, or a new school? 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 and facilitate potential cheery-picking. Non-parametric modeling allows flexible functional form and requires little manual tuning from the researchers.

POTENTIAL OUTCOMES FOR META-ANALYSIS

(Sobel, Madigan, and Wang 2015) put meta-analysis on a concrete causal foundation by introducing a extended potential outcome framework. This section is joint work with Michael Sobel and David Madigan.

Potential Outcomes

Potential Outcomes Framework [] 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)

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

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 a meta-analysis, consisting S studies and Z treatment variants, 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, we can interpret some commonly discussed meta-analytical estimates 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 we 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 $Y_i(s_1, z) \overset{\mathrm{d}}{=} Y_i(s_2, z)$ for all pair of $s_1, s_2 \in \{1, \ldots, S\}$, where $\overset{\mathrm{d}}{=}$ 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 $Y_i(s_1, z) \overset{\mathrm{d}}{=} Y_i(s_2, z)$ holds.

Two assumptions from (Sobel, Madigan, and Wang 2015) codify these two sources of heterogeneities.

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

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

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 frequentist point of view, these two assumptions cannot be distinguished from one other. Thus (Sobel, Madigan, and Wang 2015) suggests that meta-analysts 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, we 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 (See Rubin 2005 for more details). 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, inspired by the advances in Bayesian non-parametric models (Hjort et al. 2010), (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 (C. 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

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

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

$$\begin{aligned}\boldsymbol{y}^\star \mid \boldsymbol{X}^\star, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}(&K_{\boldsymbol{X}^\star,\boldsymbol{X}}(K_{\boldsymbol{X},\boldsymbol{X}} + \sigma^2 I_N)^{-1}\boldsymbol{y},\\ &K_{\boldsymbol{X}^\star,\boldsymbol{X}^\star} - K_{\boldsymbol{X}^\star,\boldsymbol{X}}(K_{\boldsymbol{X},\boldsymbol{X}} + \sigma^2 I_N)^{-1}K^\intercal_{\boldsymbol{X}^\star,\boldsymbol{X}})\end{aligned}$$

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

$$\begin{aligned}\log p(\boldsymbol{y} \mid \boldsymbol{X}, \theta) &= \log \int p(\boldsymbol{y} \mid f, \boldsymbol{X}, \theta) p(f)\, df\\ &\propto -\big[\boldsymbol{y}^\intercal(K_{\boldsymbol{X},\boldsymbol{X}}(\theta) + \sigma^2 I_N)^{-1}\boldsymbol{y} + \log|K_{\boldsymbol{X},\boldsymbol{X}}(\theta) + \sigma^2 I_N|\big]\end{aligned}$$

and plug in the fitted value $\hat{\theta}$ into the predictive distribution of new points $\boldsymbol{X}^\star$.

Generative Models

Gaussian Processes Vector- and Matrix-Valued Functions

In the context of meta-analysis, we can naturally extend the function f to be a matrix-valued function, i.e., $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^{S\times Z}$. The finite dimensional marginal of f should follow a high-dimensional equivalent of Gaussian distribution. First consider one dimensional marginal, $\boldsymbol{f}(x_0) \in \mathbb{R}^{S\times Z}$, we can define its distribution as a matrix normal distribution

$$\boldsymbol{f}(x_0) \sim \mathcal{MN}(M_{S\times Z}, U_{S\times S}, V_{Z\times Z}) \tag{1}$$

where M is the mean matrix, U is the between-row covariance and V is the between-column covariance. Then the consistency assumption (A1) for study s_1, s_2 and treatment z indicates $\boldsymbol{f}_{s_1,z}(x_0) \overset{\text{d}}{=} \boldsymbol{f}_{s_2,z}(x_0)$.

Instead of working with random matrices, however, it is easier to work with random vectors by exploiting the connection between matrix normal distribution and multivariate (vector) normal distribution. In fact, it is well know that

$$\text{vec}\boldsymbol{f}(x_0) \sim \mathcal{N}(\text{vec}M, U \otimes V)$$

In machine learning literature, vector-valued functions are known as multi-task learning problems, where $U \otimes V$ represents between-task similarity and allows borrowing-strength among tasks (Bonilla, Chai, and Williams 2008; Yu, Tresp, and Schwaighofer 2005). (Alvarez, Rosasco, and Lawrence 2011) gives comprehensive reviews of the typical kernels used for vector-valued functions. In particular, in the context of meta-analysis involving multiple treatment arms, between-task can be separated into two dimensions, between-study, denoted by U, and between-treatment, denoted by V. Eq. 1 specifies a structure where between-study and between-treatment similarities are assumed to be separable.

A matrix-valued Gaussian process can be defined as

$$\begin{aligned} \boldsymbol{y}_i \mid \boldsymbol{f} &\sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma^2) \\ \mathrm{vec}\boldsymbol{f} &\sim GP(0, k_x \otimes k_s \otimes K_T) \end{aligned}$$

where k_x is the kernel for individual-level covariate x, k_s is for study-level covariates c, and K_T is a positive semidefinite matrix encoding between-treatment similarities. The reason that I chose a kernel to represent between-study similarity but a free-form to represent between-treatment similarity is because study-level covariates are often available, whereas treatments usually are discrete choices without descriptors associated with them. More discussion on kernel versus free-form positive semidefinite matrix modeling of between-task similarity can be found in (Bonilla, Chai, and Williams 2008).

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, we only observe a small fraction of the array of matrices $\boldsymbol{y}_{i_i} = 1^\infty$, 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 in my opinion makes the models unnecessarily complex and is detrimental to intuitive understanding. The framework outlined in this thesis, however, naturally deal with network meta-analysis, since it considers all possible treatment at the same time and thus have those natural constraints built in.

MODELS DESCRIPTIONS

INFERENCE

REAL DATA (STAR)

The demonstrative example I use is the STAR (for Student-Teacher Achievement Ratio) project, which 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 in 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 both standardized and curriculum-based tests 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. Unsurprisingly, 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 [cite], and persistent effects, e.g., test scores several years after the intervention (Krueger 1997) 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 11,598 observations on 47 variables, including information on the student, test scores and treatment assignment over the intervention years, information of the teacher, and school id et al. Due to its richness, STAR project data can be investigated in many different facets. For the sake of simplicity, we choose to look at just one outcome, standardized math test score, in one intervention year, the 1st grade. Thus we 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 our discussion.

BIBLIOGRAPHY

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

Alvarez, Mauricio A, Lorenzo Rosasco, and Neil D Lawrence. 2011. “Kernels for Vector-Valued Functions: A Review.” *arXiv preprint arXiv:1106.6251*.

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.

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.

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).

Hjort, Nils Lid, CC Holmes, Peter Müller, and Stephen G Walker. 2010. "Bayesian Nonparametrics." *AMC* 10: 12.

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).

Sobel, Michael E, David B Madigan, and Wei Wang. 2015. "Meta-Analysis: A Causal Framework, with Application to Randomized Studies of Vioxx." *Technical Report, Department of Statistics, Columbia Univeristy*.

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.

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

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.