Bayesian Analysis (2022) **17**, Number 1, pp. 127–164

Joint Bayesian Analysis of Multiple Response-Types Using the Hierarchical Generalized Transformation Model

Jonathan R. Bradley*

Abstract. Consider the situation where an analyst has a Bayesian statistical model that performs well for continuous data. However, suppose the observed dataset consists of multiple response-types (e.g., continuous, count-valued, Bernoulli trials, etc.), which are distributed from more than one class of distributions. We refer to these types of data as "multiple response-type" datasets. The goal of this article is to introduce a reasonable easy-to-implement all-purpose method that "converts" a Bayesian statistical model for continuous responses (call this the preferred model) into a Bayesian model for multiple response-type datasets. To do this, we consider a transformation of the multiple response-type data, such that the transformed data can be reasonably modeled using the preferred model. What is unique with our strategy is that we treat the transformations as unknown and use a Bayesian approach to model this uncertainty. The implementation of our Bayesian approach to unknown transformations is straightforward, and involves two steps. The first step produces posterior replicates of the transformed multiple response-type data from a latent conjugate multivariate (LCM) model. The second step involves generating values from the posterior distribution implied by the preferred model. We demonstrate the flexibility of our model through an application to Bayesian additive regression trees (BART) and a spatio-temporal mixed effects (SME) model. We provide a thorough joint multiple response-type spatio-temporal analysis of coronavirus disease 2019 (COVID-19) cases, the adjusted closing price of the Dow Jones Industrial (DJI), and Google Trends data.

Keywords: Bayesian hierarchical model, big data, multiple response-types, Markov chain Monte Carlo, non-Gaussian, nonlinear, Gibbs sampler, log-linear models.

MSC2020 subject classifications: Primary 62H11; secondary 62P12.

1 Introduction

Suppose you have a Bayesian statistical model for continuous responses that you believe works extremely well in several settings. Refer to this statistical model as the "preferred model." Also suppose you have observed a dataset consisting of multiple response-types (e.g., continuous, count-valued, Bernoulli trials etc.). These response-types may be "mismatched" with the response-types of the preferred model. For example, the dataset may consist of count-valued observations, but this preferred model may be derived

arXiv: 2002.09983

*Department of Statistics, Florida State University, 117 N. Woodward Ave., Tallahassee, FL, 32306-4330, jrbradley@fsu.edu

 https://doi.org/10.1214/20-BA1246

only for Gaussian data. The primary goal of this article is to introduce a reasonable easy-to-implement all-purpose method that "converts" a Bayesian statistical model for continuous responses into a Bayesian model appropriate for the analysis of multiple response-type datasets.

There are several methods for jointly modeling data consisting of multiple response-types, however these approaches require one to either abandon the preferred model, or it requires you modify it in a manner that creates computational difficulties. For example, Markov models Yang et al. (2014), copulas (Liu et al., 2009; Xue and Zou, 2012; Dobra and Lenkoski, 2011; Liu et al., 2012), multi-task learning models (Argyriou et al., 2007; Kim and Xing, 2009; Yang et al., 2009), regression trees, and random forests (Hastie et al., 2009; Fellinghauer et al., 2013) have been adapted to this multiple response-type setting. However, these methods do not immediately incorporate a data scientist's preferred model. An important goal of this article is to allow our model to be flexible enough that it can be adapted to other data scientist's preferred model. While our proposed model allows for this flexibility, it can be interpreted as a simple combination of two existing methods: generalized linear mixed effects models (GLMM; e.g., see McCulloch et al., 2008, for a standard reference) and LCMs (Bradley et al., 2020+).

The GLMM is a standard approach to model non-Gaussian data. For example, Poisson data is modeled hierarchically, where the log mean parameter can be analyzed using a data scientist's preferred model. GLMMs can lack conjugacy, which creates noticeable difficulty when implementing a GLMM on a modern high-dimensional dataset. There are several examples of where this approach has been used to analyze multiple response-type datasets (e.g., see Christensen and Amemiya, 2002; Schliep and Hoeting, 2013; Wu et al., 2015; Clark et al., 2017; Todd et al., 2018, among several others). A more recent alternative is the LCM. Basic theoretical results and empirical analyses in Bradley et al. (2018), Hu and Bradley (2018), Yang et al. (2019), Bradley et al. (2019b), and Bradley et al. (2020+) suggest that one can outperform a standard GLMM (specifically Latent Gaussian Process (LGP) models) in terms of prediction error. However, both the GLMM and LCM requires the preferred model to be a mixed effects model, and the LCM requires one to modify the distribution of random effects to follow the appropriate distribution based on conjugacy.

A classical approach is to *transform* the data, so that the transformed data can be reasonably modeled using the distribution assumed by the preferred model. In the non-Bayesian settings this literature is extremely well-developed and includes the Box-Cox transformations (Box and Cox, 1964), the alternating conditional expectations (ACE; Breiman and Friedman, 1985) algorithm, graphical techniques (McCulloch, 1993), and the Yeo-Johnson power transformation (Yeo and Johnson, 2000), among other techniques. More recently developments in rank based algorithms (Servin and Stephens, 2007; McCaw et al., 2019; Beasley et al., 2009) and quantile-matching (McCullagh and Tresoldi, 2020) have also been proposed in the non-Bayesian setting. It is important to note that Bayesian models for transformations have been proposed as well, but focus on the case where continuous non-normal data are observed and the preferred model assumes normality. In particular, these Bayesian models put a prior on the free parameter within the Box-Cox transformation or the Yeo-Johnson power transformation (Kim et al., 2013; Charitidou et al., 2015; Bean et al., 2016; Charitidou et al., 2018). No such

Bayesian model has been developed to analyze multiple response-type data using any preferred model for a continuous response.

There are three distributions that define our hierarchical generalized transformation (HGT) model: (a) the distribution of the data given a transformation, (b) the prior distribution of the transformation, and (c) both the distribution of the transformation given the latent process and the distribution of the process of interest (i.e., multiplicative terms in the aforementioned preferred model). Here, cross-correlations are explicitly modeled using the preferred model. In this article, we model the data given a transformation (a) using members from the exponential family. Specifically, given a transformation, continuous data follows the normal distribution, categorical data follows the binomial distribution, and count-data follow the Poisson distribution. These distributions are conjugate with the normal, the logit-beta (Gao and Bradley, 2019; Bradley et al., 2019b) and the log-gamma distributions (Bradley et al., 2018; Hu and Bradley, 2018; Bradley et al., 2020+; Yang et al., 2019), which are special cases of the Diaconis-Ylvishaker (DY) distribution (e.g., see Diaconis and Ylvisaker, 1979; Chen and Ibrahim, 2003, for key references). Consequently, the prior distribution of the transformation (b) is modeled with a DY distribution, which defines an LCM model for the transformations. To combine (a), (b), and (c) we also have to introduce a multiplicative term to the preferred model to ensure propriety and to avoid contradictions when computing the marginal distribution of the transformations. While we include this multiplicative term in the expression of the preferred model, implementation of the preferred model is un-altered when learning the values of the latent process and associated parameters. The implementation of our approach can be done using composite sampling. In particular, the first step is to sample from the posterior distribution of the transformation. Then the second step is to sample from the conditional distribution of the latent process of interest given the transformation.

The first step of the composite sampler is computationally straightforward because the DY distribution is conjugate (and easy to sample from) with the exponential family. Additionally, the first step of this algorithm is important for the purpose of analyzing multiple response-types. Specifically, at the end of the first step we obtain a replicate from the posterior distribution of the transformation (which is continuous valued). Thus, the first step of the composite sampling algorithm "transforms" the multiple response-type data into a continuous-valued quantity appropriate for the preferred model.

Implementation of the preferred model is unchanged (from the standard use of the preferred model) in the second step of our composite sampling algorithm. This is particularly noteworthy, as many of the Bayesian statistical models derived for Gaussian data are not immediately computationally efficient in the non-Gaussian data setting (e.g., see Bradley et al., 2019a; Kang and Cressie, 2011; Katzfuss and Cressie, 2012, for examples in the spatial setting). This is because GLMMs in the non-Gaussian setting have full-conditional distributions that are not Gaussian, and can not be sampled from immediately. Bayesian methods that do not have easy to sample from full-conditional distributions require difficult to tune Metropolis-Hastings algorithms (e.g., see Bradley et al., 2020+, for an example), inefficient rejection samplers (e.g., see Damien et al., 1999), or significant reparameterization to make approximate Bayesian methods (that

are only appropriate for small parameter spaces) practical (Rue et al., 2009; Neal, 2011). The second step of our composite sampling algorithm allows one to circumvent this issue entirely, and simply use the computational strategies that were developed for the preferred model.

The two steps of our composite sampler can be seen as sequential smoothing. By "smoothing" we mean a function of the data that attempts to discover important features in the data (e.g., see Simonoff, 2012, for a standard reference). Multiple layers of smoothing may lead to estimates that are "oversmooth," in the sense that many features of the data are not captured. To avoid oversmoothing we specify the model so that the posterior distribution of the transformation is "saturated." Recall a saturated model is one in which there exists at least as many parameters as there are data points, and fitting this model allows you to exactly recover the original dataset. Hence, saturated models are often an extreme example of overfitting. Thus, in the first step of our composite sampler we choose to overfit the multiple response-type data, and in the second step we smooth overfitted values (again this is done to avoid oversmoothing).

In the classical log-linear model literature, saturated models are useful for selecting more parsimonious models (e.g., see Agresti, 2007, for a standard reference). Specifically, the most parsimonious reduced model that is not significantly different (in terms of the deviance or chi-square statistic) from the saturated model is used for statistical inference. Consequently, specifying the data model to be saturated allows us to assess the goodness-of-fit of the preferred model in a fully Bayesian manner that is similar to what is done in classical residual analysis.

We use our method to analyze spatio-temporal COVID-19 incidences and social distancing related variables. This COVID-19 dataset is observed daily over large sparse regions (i.e., countries and provinces). Social distancing can be described as an effort to maintain a physical distance between individuals and has become a necessary public health measure to combat COVID-19 (CDC, 2020). Social distancing is known to weaken incidences and deaths due to COVID-19, however, there are detrimental economic and psychological effects. This motivates us to analyze incidences (and deaths) of COVID-19 along with a measure of the health of the US economy (i.e., the adjusted closing price of the Dow Jones Industrial), and a measure of the public interest in COVID-19 through Google Trends data. The HGT model is developed to be easily adapted to a data scientist's preferred method for continuous data, which aids future analyses of this important dataset. It has recently been shown that forecasts regarding COVID-19 require sophisticated models. Following the results of Donnat and Holmes (2020), we include spatio-temporal random effects through the use of basis function expansions (e.g., see Cressie and Wikle, 2011, for a standard reference). Additionally, to improve the performance of forecasting we adopt the training, validation, and testing data framework that has become standard among the machine learning literature (e.g., see Hastie et al., 2009, for a standard reference). In particular, we incorporate a Bayesian version of ResNet (He et al., 2016), where validation data is used to make linear adjustments to improve forecasts. While we are partially motivated by COVID-19 and the detrimental impacts of social distances, the HGT developed in this manuscript is of independent

interest, since this is a new way in Bayesian statistics to model non-Gaussian processes using models for continuous data. Furthermore, our methodology also allows one to analyze a single non-Gaussian response-type in a straightforward manner.

The remainder of this article is organized as follows. In Section 2, we describe how standard modeling procedures are not appropriate for this multiple response-type dataset. Then, we introduce the HGT model to analyze multiple response-type data with unknown transformations in Section 3. Additionally, we provide an example model specification. Then in Section 4, we provide details on using training, validation, and testing data for statistical inference. This allows us to perform linear adjustments to our predictions in a similar manner to He et al. (2016), which aids in forecasting. A summary of all the Bayesian models used in our analysis is also provided. In Section 5, we give simulation studies to illustrate that our approach has been developed in a manner that one can incorporate their preferred statistical model. In particular, we apply our approach to BART models and a spatio-temporal mixed effects (SME) model. Section 6 contains our joint analysis of COVID-19 mortality, incidences and recoveries, along with Google Trends data, and DJI data. Section 7 contains a discussion and derivations are provided in the Supplementary Materials (Bradley, 2020).

2 Motivation

Denote the multiple response-type data with Z_{ij}, where i indexes replicates and j indexes response-type such that $i = 1, \ldots, I_j$ and $j = 1, 2, 3$. We consider the setting where for each i, Z_{i1} is continuous-valued, Z_{i2} is integer-valued ranging from $0, \ldots, b_i$, and Z_{i3} is count valued. There are many "off-the-shelf" approaches that one might consider to analyze multiple response-type data. For example, one might define the following linear model,

$$Z_{i1} = \mathbf{x}_{i1}'\boldsymbol{\beta}_1 + \beta_{Z2}Z_{i2} + \beta_{Z3}Z_{i3} + \xi_{i1}; \;\; i = 1, \ldots, I_1,$$

where ξ_{i1} is normally distributed with mean zero and variance σ_ξ^2, $\beta_{Zk} \in \mathbb{R}$, $\boldsymbol{\beta}_1$ is an unknown p-dimensional vector, and $\mathbf{x}_{i1}$ is a p-dimensional covariate vector. However, this conditionally specified model enforces a strong assumption of linearity between the different response-types. Furthermore, the variability (and dependence) of Z_{i2} and Z_{i3} is not explicitly modeled. Additionally, it assumes that the responses are paired, which may not be the case; that is, it is assumed that we observe the triplet (Z_{i1}, Z_{i2}, Z_{i3}).

To incorporate the variability across response-types (i.e., across j) and allow for non-linear relationships, one might also consider the following hierarchical model:

$$\begin{aligned} Z_{i1} &\overset{\text{ind}}{\sim} \text{Normal}(Y_{i1}, v), \\ Z_{i2} &\overset{\text{ind}}{\sim} \text{Binomial}\left\{ b_i, \frac{\exp(Y_{i2})}{1+\exp(Y_{i2})} \right\}, \\ Z_{i3} &\overset{\text{ind}}{\sim} \text{Poisson}\left\{\exp(Y_{i3})\right\}; \;\; i = 1, \ldots, I_j, \end{aligned} \tag{2.1}$$

where we assume conditional independence of Z_{ij} given Y_{ij}, Y_{ij} is an unobserved latent process, $\text{Normal}(Y_{i1}, v)$ is a shorthand for the normal distribution with mean $Y_{ij} \in \mathbb{R}$

and variance $v > 0$, Binomial(b_i, p) is a shorthand for the binomial distribution with $b_i > 1$ number of trials and probability of success $p \in (0,1)$, Poisson(μ_{ij}) is a shorthand for the Poisson distribution with mean μ_{ij}. The covariance between observations is determined by the model for Y_{ij}:

$$\begin{aligned} cov(Z_{ij}, Z_{k\ell}) &= E\left\{cov(Z_{ij}, Z_{k\ell})|Y_{ij}, Y_{k\ell}\right\} + cov\{E(Z_{ij}|Y_{ij}), E(Z_{k\ell}|Y_{k\ell})\} \\ &= cov\left\{E(Z_{ij}|Y_{ij}), E(Z_{k\ell}|Y_{k\ell})\right\} = cov\left\{c_{ij}g_j^{-1}(Y_{ij}), c_{k\ell}g_\ell^{-1}(Y_{k\ell})\right\}, \quad (2.2) \end{aligned}$$

for $k \neq m$ and $\ell \neq j$, where the functions $g_1(x_i) = x_i$, $g_2(x_i) = log(x_i/1 - x_i)$, and $g_3(x_i) = log(x_i)$ are referred to as "link functions," and $c_{i1} = c_{i3} = 1$ and $c_{i2} = b_i$. Similarly, predicted values are determined by the model for Y_{ij}:

$$E(Z_{ij}) = E\left\{E(Z_{ij}|Y_{ij})\right\} = E\left\{c_{ij}g_j^{-1}(Y_{ij})\right\}. \quad (2.3)$$

Thus, cross-dependence and predictions are modeled through the statistical model assumed for the process Y_{ij}, and a standard choice in this context is the GLMM:

$$Y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_j + \boldsymbol{S}_{ij}'\boldsymbol{\eta} + \xi_{ij}, \quad (2.4)$$

where $\mathbf{x}_{ij}$ is a known p-dimensional vector of covariates and $\mathbf{S}_{ij}$ is a pre-specified r-dimensional vector of basis functions, $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{pj})'$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)'$, $\beta_{kj} \overset{\text{ind}}{\sim}$ Normal$(0, \sigma_\beta^2)$, $\eta_k \overset{\text{ind}}{\sim}$ Normal$(0, \sigma_\eta^2)$, $\xi_{ij} \overset{\text{ind}}{\sim}$ Normal$(0, \sigma_\xi^2)$, $\sigma_\beta^2 > 0$, $\sigma_\eta^2 > 0$, and $\sigma_\xi^2 > 0$. Then, the cross-response spatio-temporal covariance implied by this model is $cov\left(Y_{ij}, Y_{k\ell}|\sigma_\eta^2\right) = \sigma_\eta^2 \mathbf{S}_{ij}'\mathbf{S}_{k\ell}$, which propagates through and enforces dependence in the multiple response-type data through (2.2). The relationship between the different response-types can be found by estimating the unknown function $\mathbf{S}_{ij}'\boldsymbol{\eta}$ (e.g., using posterior means and credible intervals).

Computationally, the GLMM is difficult to implement in a Bayesian context. For example, a Gibbs sampler requires one to simulate from the following full-conditional distributions (Gelfand, 2000), and in this setting these distributions do not have a known form that is straightforward to simulate from. There are several approximate Bayesian computational tools available, however, for moderate sizes of p and r these approaches are not feasible. In particular, Hamiltonian Monte Carlo (HMC; Neal et al., 2011) and the integrated nested Laplace approximation (INLA; Rue et al., 2009) are only appropriate for small parameter spaces (e.g, Martino and Riebler, 2019, suggests no more than 15 parameters when implementing INLA). Additionally, INLA only allows for marginal inference (Kristensen et al., 2015). The computational issues of the generalized linear mixed model in (2.1) and (2.4) may become even more cumbersome when considering a different model for Y_{ij}. This is especially pertinent for our multiple response-type dataset in Section 6, since the US government has put out a call to action (Office of Science and Technology Policy, 2020) for data scientists to analyze COVID-19 datasets, and it would be preferable to have an approach that is flexible enough for others to specify their own model for Y_{ij} without major changes to implementation.

There are also parameters that are fixed and require specification in the GLMM. For example, the dimension of $\boldsymbol{\eta}$ (i.e., r) is chosen by the data analyst. In practice, measures

of out-of-sample variability can be used to specify such parameters. This is because estimates are often biased towards the training data, since estimators are defined to be "close" to the training data (Hastie et al., 2009). Consequently, the use of validation data (held separate from the training data) to estimate r can aid in accounting for out-of-sample error. In particular, define a validation dataset to be $\mathbf{z}_{val} = (Z_{ij} : i = I_j + 1, \ldots, I_j^{val}, j = 1, 2, 3)'$, which is observed over the indices $i \in \{I_j + 1, \ldots, I_j^{val}\}$ and $j = 1, 2, 3$, where $I_j < I_j^{val} < I$. Then a posterior predictive step (e.g., see Gelman et al., 2013) can be used to predict at $\mathbf{z}_{val}$. For example, suppose we obtain an estimate of $E(\mathbf{z}_{val}|\mathbf{z}_{trn}, r)$ using MCMC, where $n = \sum_{j=1}^{3} I_j$ and the n-dimensional data vector $\mathbf{z}_{trn} = (Z_{ij} : i = 1, \ldots, I_j, j = 1, 2, 3)'$. Then one could use $E(\mathbf{z}_{val}|\mathbf{z}_{trn}, \widehat{r})$ for inference, where for illustration

$$\widehat{r} = \arg\min \left\{ (\mathbf{z}_{val} - E(\mathbf{z}_{val}|\mathbf{z}_{trn}, \widehat{r}))'(\mathbf{z}_{val} - E(\mathbf{z}_{val}|\mathbf{z}_{trn}, \widehat{r})) \right\},$$

and the arg min is computed over r. Additionally, to assess the overall predictive performance of the selected r a testing dataset could be used. For example, let the testing data Z_{ij} be defined over the indices $i = I_j^{val} + 1, \ldots, I$ for $I_j^{val} < I$. Then the metric $\sum_{i=I_j+1}^{I} (Z_{ij} - E(Z_{ij}|\mathbf{z}_{trn}, \widehat{r}))^2$ can be used to assess the out-of-sample error of predictions based on the selected r.

Computationally, the posterior predictive steps to estimate $E(\mathbf{z}_{val}|\mathbf{z}_{trn}, \widehat{r})$ can be done efficiently once replicates of Y_{ij} from the posterior distribution have already been generated. That is, denote the b-th posterior replicate of Y_{ij} with $Y_{ij}^{[b]}$. Then a posterior replicate of Z_{ij} for $i > I_j$ can be generated by simulating from the distribution for $Z_{ij}|Y_{ij}^{[b]}$. Consequently, we will use the HGT to fit training data (developed in Section 3), but continue to use the GLMM to fit validation and testing data (see Sections 4.2–4.4).

Instead of using validation data to estimate existing parameters, in our implementation, we introduce new parameters to model the validation data to adjust for out-of-sample error. That is, we introduce an additional linear model for $\{Y_{ij} : i > I_j\}$, and then use the validation data to estimate the parameters in the linear model. This is similar to an approach in machine learning called ResNet (He et al., 2016), and will aid in adjusting for biases that occur when forecasting COVID-19 incidences and deaths. More details and justification are given in Section 4.2.

3 The Hierarchical Generalized Transformation Model

As discussed in the Introduction, there are three main components that define the HGT: (a) the distribution of the data given a transformation, (b) the prior distribution of the transformation, and (c) both the distribution of the transformation given the latent process and the distribution of the process of interest (i.e., terms in the aforementioned preferred model). Before making explicit specifications of Items (a), (b), and (c), in Section 3.1, we provide a general discussion on allowing transformations to be an unknown process to estimate. Then, in Section 3.2, we describe general Bayesian implementation using any proper generic specifications of (a), (b), and (c). Then in Section 3.3, we provide an explicit specification of Items (a) and (b) that are used in this manuscript.

Finally, in Section 3.4, we provide an example specification of densities that define the preferred model in (c). In Supplementary Appendix A, we provide a table of terminology to aid in keeping track of both terminology and notation.

3.1 Unknown Transformations of Multiple Response-Types

One classical strategy to model non-Gaussian data is to impose a transformation such that,

$$h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta} \overset{\text{ind}}{\sim} f(h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta}), \qquad i = 1, \ldots, I_j, j = 1, 2, 3, \tag{3.1}$$

where $h_j(\cdot)$ is a transformation of the datum Z_{ij}, the $h_j(Z_{ij})$'s are conditionally independent given $\{Y_{ij}\}$ and $\boldsymbol{\theta}$, f is a short-hand used for a probability density function (pdf) or a probability mass function (pmf), one should read "$h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta} \overset{\text{ind}}{\sim} f(h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta})$" as "$h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta}$ is distributed according to the pdf $f(h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta})$," $g_j\{E(Z_{ij})\} = Y_{ij} \in \mathbb{R}$, and $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. Additionally, Y_{ij} is defined for $i = 1, \ldots, I$ and $j = 1, 2, 3$, where recall $I \geq max(I_1, I_2, I_3)$. Here, $f(h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta})$ represents a multiplicative term in the aforementioned preferred model. In what remains, inference on $\{Y_{ij}\}$ and $\boldsymbol{\theta}$ is the primary goal. To aid in our exposition we drop the functional notation for $h_j(\cdot)$ and write $h_{ij} = h_j(Z_{ij})$. As an example of $f(h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta})$, suppose we assume

$$h_{ij} = Y_{ij} + \epsilon_{ij}, \tag{3.2}$$

where $\epsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$ and $\sigma_\epsilon^2 > 0$, and the mixed effects model on Y_{ij} in (2.4) is assumed.

Transformations convert a multiple response-type dataset (e.g., $\{Z_{ij}\}$) to a single response-type dataset (e.g., $\{h_{ij}\}$), since h_{ij} follows a single distribution with a continuous support. Consequently, transformations have become a standard tool in analyzing multiple response-types. Recall, transformations such as these have a long history including the box-cox transformations (Box and Cox, 1964), graphical techniques (McCulloch, 1993), the alternating conditional expectations (ACE; Breiman and Friedman, 1985) algorithm, and the Yeo-Johnson power transformation (Yeo and Johnson, 2000, among others).

In this paper, we introduce a Bayesian solution to the problem of an unknown transformation through the use of the following pdfs and pmfs:

(a) The distribution of the data given a transformation is denoted with $f(Z_{ij}|h_{ij})$. We refer to the distribution $f(Z_{ij}|h_{ij})$ as a "data model."

(b) The prior distribution of the transformation is denoted with $f(h_{ij}|\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is a real vector-valued hyperparameter. We call $f(h_{ij}|\boldsymbol{\gamma})$ a "transformation prior" and $f(\boldsymbol{\gamma})$ a "transformation hyperprior."

(c) Consider the density $f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) = \prod_i \prod_j f(h_{ij}|Y_{ij}, \boldsymbol{\theta})$ and the "process model" $f(\mathbf{y}|\boldsymbol{\theta})$, where the n-dimensional transformed data vector $\mathbf{h} = (h_{11}, \ldots, h_{I_3 3})'$, $N = 3I \geq n$, and the N-dimensional latent process $\mathbf{y} = (Y_{11}, \ldots, Y_{I1}, Y_{12},$

$\ldots, Y_{I2}, Y_{13}, \ldots, Y_{I3})'$. Notice, that $I_j \leq I$, which allows for missing values of Z_{ij}.

In Section 3.2, we describe general Bayesian implementation when Items (a), (b), and (c) have been specified.

3.2 General Bayesian Implementation

In this section, we describe Bayesian implementation of the HGT. The preferred model is represented in terms of a hierarchical model:

$$\begin{aligned} &\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) m(\mathbf{h}|\boldsymbol{\gamma}), \\ &\mathbf{y}|\boldsymbol{\theta} \sim f(\mathbf{y}|\boldsymbol{\theta}), \\ &\boldsymbol{\theta} \sim f(\boldsymbol{\theta}), \end{aligned} \tag{3.3}$$

where the product of each density in (3.3) defines the conditional density $f(\mathbf{h}, \mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\gamma})$ and $\boldsymbol{\gamma}$ is conditionally independent of $(\mathbf{y}', \boldsymbol{\theta}')'$ given $\mathbf{h}$. Following the terminology used in Cressie and Wikle (2011), we call $f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) m(\mathbf{h}|\boldsymbol{\gamma})$ the "transformed data model," $f(\mathbf{y}|\boldsymbol{\theta})$ the "process model," and $f(\boldsymbol{\theta})$ the "prior" for $\boldsymbol{\theta}$.

To guarantee that our choice of the transformation prior and transformed data model are consistent with each other we set

$$m(\mathbf{h}|\boldsymbol{\gamma}) = f(\mathbf{h}|\boldsymbol{\gamma}) / \int \int f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta}.$$

To illustrate the need for $m(\mathbf{h}|\boldsymbol{\gamma})$ consider the incorrect specification of $m(\mathbf{h}|\boldsymbol{\gamma}) = 1$. Then the preferred model in (3.3) would imply a different marginal distribution for $\mathbf{h}$ than the transformation prior $f(\mathbf{h}|\boldsymbol{\gamma})$, since

$$\int \int f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta} \neq \int f(\mathbf{h}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma}) d\boldsymbol{\gamma} = f(\mathbf{h}),$$

where recall that $f(\boldsymbol{\gamma})$ is the "transformation hyperprior." However, setting $m(\mathbf{h}|\boldsymbol{\gamma}) = f(\mathbf{h}|\boldsymbol{\gamma}) / \int \int f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta}$ guarantees that the marginal distribution of $\mathbf{h}$ stays same when computing using either the preferred model or the transformation prior. That is,

$$\begin{aligned} &\int \int \int f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) m(\mathbf{h}|\boldsymbol{\gamma}) d\boldsymbol{\gamma} d\mathbf{y} d\boldsymbol{\theta} \\ &= \int f(\mathbf{h}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma}) d\boldsymbol{\gamma} \int \int f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta} \frac{1}{\int \int f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta}} \\ &= f(\mathbf{h}). \end{aligned}$$

Notice that the presence of $m(\mathbf{h}|\boldsymbol{\gamma})$ in the preferred model does not change the preferred model proportionally as a function of $\mathbf{y}$ and $\boldsymbol{\theta}$. Consequently, the presence of $m(\mathbf{h}|\boldsymbol{\gamma})$ will not have an effect on updating $\mathbf{y}$ and $\boldsymbol{\theta}$ in an MCMC.

Algorithm 1 Implementation of the HGT Model.

1: Set $b = 1$ and initialize $\mathbf{h}$, $\boldsymbol{\gamma}$, $\mathbf{y}$, and $\boldsymbol{\theta}$ with $\mathbf{h}^{[0]}$, $\boldsymbol{\gamma}^{[0]}$, $\mathbf{y}^{[0]}$, and $\boldsymbol{\theta}^{[0]}$.
2: Sample $\mathbf{h}^{[b]}$ from $f(\mathbf{h}|\mathbf{z}_{trn}, \boldsymbol{\gamma}^{[b-1]})$.
3: Sample $\boldsymbol{\gamma}^{[b]}$ from their full-conditional distributions. We use the slice sampler (Neal et al., 2003) if the full-conditional does not have a closed form.
4: Sample $\mathbf{y}^{[b]}$ and $\boldsymbol{\theta}^{[b]}$ from $f(\mathbf{y}, \boldsymbol{\theta}|\mathbf{h}^{[b]})$, which is the posterior distribution associated with the preferred model described in (3.4).
5: Set $b = b + 1$.
6: Repeat Steps 2−5 until $b = B$ for a prespecified value of B.

Bayes rule can be used to produce the following conditional distribution (e.g., see Gelman et al., 2013, for a standard reference),

$$f(\mathbf{y}, \boldsymbol{\theta}|\mathbf{h}) = \frac{f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})\, f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{\int\int f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})\, f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta}}. \tag{3.4}$$

Similarly, one can use Bayes rule to produce the posterior distribution of the transformed data. That is,

$$f(\mathbf{h}|\mathbf{z}_{trn}) = \frac{\int f(\mathbf{z}_{trn}|\mathbf{h}) f(\mathbf{h}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma})\, d\boldsymbol{\gamma}}{\int\int f(\mathbf{z}_{trn}|\mathbf{h}) f(\mathbf{h}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma})\, d\mathbf{h}\, d\boldsymbol{\gamma}}. \tag{3.5}$$

Equations (3.4) and (3.5) can then be used to produce a posterior distribution for $\mathbf{y}$ and $\boldsymbol{\theta}$. That is, suppose $f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})$, $f(\mathbf{y}|\boldsymbol{\theta})$, $f(\boldsymbol{\theta})$, $f(\mathbf{z}_{trn}|\mathbf{h})$, $f(\mathbf{h}|\boldsymbol{\gamma})$, and $f(\boldsymbol{\gamma})$ are proper. Suppose $\mathbf{z}_{trn}$ is conditionally independent of $\boldsymbol{\gamma}$ given $\mathbf{h}$, $\boldsymbol{\gamma}$ is conditionally independent of $(\mathbf{y}', \boldsymbol{\theta}')'$ given $\mathbf{h}$, and $\mathbf{z}_{trn}$ and $(\mathbf{y}', \boldsymbol{\theta}')'$ are conditionally independent given $\mathbf{h}$. Then:

$$f(\mathbf{y}, \boldsymbol{\theta}|\mathbf{z}_{trn}) = \int f(\mathbf{y}, \boldsymbol{\theta}|\mathbf{h}) f(\mathbf{h}|\mathbf{z}_{trn}) d\mathbf{h}. \tag{3.6}$$

The derivation of the posterior distribution of the latent process and parameters stated in (3.6) can be found in Supplementary Appendix B.

The posterior distribution of the latent process and parameters can easily be simulated from using a composite sampling scheme, provided that it is easy to simulate from $f(\mathbf{y}, \boldsymbol{\theta}|\mathbf{h})$. Algorithm 1 gives the step-by-step implementation of how to simulate from the posterior distribution in (3.6). Here, we see that the implementation of the HGT model is similar to the bootstrap implementation, where we have replaced a resampling step with sampling from $f(\mathbf{h}|\mathbf{z}_{trn})$ and the full-conditional distributions associated with $\boldsymbol{\gamma}$. This similarity emphasizes the flexibility of allowing for unknown transformations in a Bayesian context, since the bootstrap algorithm is an established flexible approach in the literature (e.g., see Efron, 1992, for an early reference). Of course, the bootstrap algorithm produces replicates from a different distribution than that of Algorithm 1. Specifically, the bootstrap method results in an approximate sample from the sampling distribution of a statistic. Whereas, the composite sampling approach in Algorithm 1 can be seen as a means to sample from the posterior distribution in (3.6). This is also different from the Bayesian bootstrap (Rubin, 1981), which does not restrict the samples to be from a posterior distribution of the form in (3.6).

3.3 Data Models, Transformation Priors, and Transformation Hyperpriors

Consider the following specifications of the data models:

$$\begin{aligned} Z_{i1}|h_{i1}, v &\overset{\text{ind}}{\sim} \text{Normal}(h_{i1}, v), \\ Z_{i2}|h_{i2} &\overset{\text{ind}}{\sim} \text{Binomial}\left\{b_i, \frac{\exp(h_{i2})}{1+\exp(h_{i2})}\right\}, \\ Z_{i3}|h_{i3} &\overset{\text{ind}}{\sim} \text{Poisson}\left\{\exp(h_{i3})\right\}; \;\; i = 1, \ldots, I_j, j = 1, 2, 3, \end{aligned} \tag{3.7}$$

which is different from the GLMM in (2.1). Specifically, instead of conditioning on the latent process of interest Y_{ij}, we condition on the transformation h_{ij}. Thus, the data models in (3.7) imply that the Z_{ij}'s are conditionally independent given $\{h_{ij}\}$ and v (v will be given a prior and integrated out).

With the data model $f(\mathbf{z}_{trn}|\mathbf{h})$ defined, we are left to specify a transformation prior and transformation hyperprior. We define the transformation prior to be the conjugate distributions associated with the data models in (3.7). It follows from Diaconis and Ylvisaker (1979) that the conjugate distribution for h_{ij} is given by,

$$f_{DY}(h_{ij}|\alpha_j, \kappa_j, a, b) = K(\alpha_j, \kappa_j)\exp\left\{\alpha_j h_{ij} - \kappa_j \psi_j(h_{ij})\right\}; \; i = 1, \ldots, I_j, j = 1, \ldots, J, \tag{3.8}$$

where $K(\alpha_j, \kappa_j)$ is a normalizing constant, $h_{ij} \in \mathbb{R}$, $\alpha_1 \in \mathbb{R}$, $\kappa_2 > \alpha_2$, $\alpha_q > 0$, and $\kappa_k > 0$; for $q = 2, 3$, and $k = 1, 3$. Let $\psi_1(Z) = Z^2$, $\psi_2(Z) = \log(1 + e^Z)$, and $\psi_3(Z) = \exp(Z)$. Also, we use the shorthand $\text{DY}(\alpha_j, \kappa_j; \psi_j)$ to represent the pdf of the Diaconis and Ylvisaker (1979) prior in (3.8). Finally, let $\boldsymbol{\gamma} = (\alpha_1, \alpha_2, \alpha_3, \kappa_1, \kappa_2, \kappa_3)'$ be the transformation hyperparameter. The DY distribution is a special case of the recently introduced conjugate multivariate distribution (Bradley et al., 2020+), where the matrix-valued covariance parameter is set equal to the identity matrix.

The data models and the DY priors in (3.7) and (3.8) can be used to produce a full-conditional distribution for the elements of transformations $\mathbf{h}$:

$$\begin{aligned} h_{i1}|Z_{i1}, \boldsymbol{\gamma} &\overset{\text{ind}}{\sim} \text{Normal}\left\{\left(2\kappa_1 + \frac{1}{v}\right)^{-1}\left(\frac{Z_{i1}}{v} + \alpha_1\right), \left(2\kappa_1 + \frac{1}{v}\right)^{-1}\right\}; \;\; i = 1, \ldots, I_1, \\ h_{i2}|Z_{i2}, \boldsymbol{\gamma} &\overset{\text{ind}}{\sim} \text{DY}\left(\alpha_2 + Z_{i2}, \kappa_2 + b_i; \psi_2\right); \;\; i = 1, \ldots, I_2, \\ h_{i3}|Z_{i3}, \boldsymbol{\gamma} &\overset{\text{ind}}{\sim} \text{DY}\left(\alpha_3 + Z_{i3}, \kappa_3 + 1; \psi_3\right); \;\; i = 1, \ldots, I_3. \end{aligned} \tag{3.9}$$

The derivations of the full conditional distributions are fairly straightforward, and are given in Supplementary Appendix B. One can simulate directly from the posterior distribution of the transformations in (3.9). Posterior replicates of h_{ij} from (3.9) can be computed using the following transformation (Bradley et al., 2020+):

$$h_{i1} \overset{d}{=} \left(2\kappa_1 + \frac{1}{v}\right)^{-1}\left(\frac{Z_{i1}}{v} + \alpha_1\right) + w_1; \;\; i = 1, \ldots, I_1,$$

$$h_{i2} \stackrel{d}{=} \log\left(\frac{w_2}{1-w_2}\right); \;\; i = 1, \ldots, I_2,$$
$$h_{i3} \stackrel{d}{=} \log(w_3); \;\; i = 1, \ldots, I_3, \tag{3.10}$$

where "$\stackrel{d}{=}$" stands for equal in distribution, $w_1|Z_{i1}, \alpha_1, \kappa_1, v$ is distributed normally with mean zero and variance $\left(2\kappa_1 + \frac{1}{v}\right)^{-1}$, $w_2|Z_{i2}, \alpha_2, \kappa_2$ is distributed according to a beta distribution with shape parameters $(\alpha_2 + Z_{i2})$ and $(\kappa_2 - \alpha_2 + b_i - Z_{i2})$, and $w_3|Z_{i3}, \alpha_3, \kappa_3$ is distributed according to a gamma distribution with shape parameter $(\alpha_3 + Z_{i3})$ and rate parameter $(\kappa_3 + 1)$. Step 2 of Algorithm 1 involves simulating according to (3.10), which is straightforward.

The specification of a transformation hyperprior for $\boldsymbol{\gamma}$ is crucial to guarantee that $f(h_{ij}|Z_{ij}, \boldsymbol{\gamma})$ is proper in the event that $Z_{i3} = 0$, $Z_{i2} = 0$, or $Z_{i2} = b_i$. Thus, we assume $\alpha_1 = \kappa_1 = 0$, α_2 and α_3 are distributed according to a gamma distribution, $\kappa_2|\alpha_2$ is distributed according to a shifted (by α_2) gamma distribution, κ_3 follows a gamma distribution, and v is distributed according to an inverse gamma distribution (e.g., see Gelman, 2006, among others). These transformation hyperpriors are explicitly stated, and the full-conditional distributions for $\boldsymbol{\gamma}$ are derived in Supplementary Appendix C.1. In this context α_2 and α_3 play the role of a continuity correction for zero-valued responses. The use of continuity corrections for zero-valued responses has a long history. For example, see Cox (1970)'s use in contingency tables and Yates (1934) who suggest adding a 1/2 to zero-valued responses so that log-odds ratios are defined. The value 1/2 forces an average approximation error to zero (Cox, 1970). Different (smaller) values have been proposed, including more data driven techniques (Mosteller and Tukey, 1977; Fienberg, 1969; Sweeting et al., 2004, among others). A major difference between these continuity corrections and the proposed HGT is that, we add a random amount for the correction, rather than a pre-determined fixed amount.

The general Bayesian implementation described in Section 3.2 is flexible enough to allow for a transformation prior that implies cross-dependence among the elements of $\mathbf{h}$, but we do not consider this case in this article. The main reason for this choice is that transformations are used in place of the original multiple response-type dataset when implementing the preferred model (Step 4 of Algorithm 1). That is, the transformed values are used as a proxy for (or in place of) the multiple response-type data in the preferred model. Consequently, we would like to specify $\mathbf{h}$ to "overfit" the data so that $\mathbf{h}$ can reasonably be thought of as a proxy for the data.

Our choice of the DY prior in (3.8) leads to posterior replicates that overfit the data. In particular, it is straightforward to verify that

$$\lim_{\kappa_1 \to 0} \lim_{\alpha_1 \to 0} E\left\{h_{i1}|Z_{i1}, \boldsymbol{\gamma}\right\} = Z_{i1},$$
$$\lim_{\kappa_2 \to 0} \lim_{\alpha_2 \to 0} E\left\{b_j g_2^{-1}(h_{j2})|Z_{j2}, \boldsymbol{\gamma}\right\} = Z_{j2}, \tag{3.11}$$
$$\lim_{\kappa_3 \to 0} \lim_{\alpha_3 \to 0} E\left\{g_3^{-1}(h_{k3})|Z_{k3}, \boldsymbol{\gamma}\right\} = Z_{i3}; \;\; i = 1, \ldots, I_1, j = 1, \ldots, I_2, k = 1, \ldots, I_3.$$

See Supplementary Appendix B for the derivation of (3.11). Thus, the posterior mean of $\mathbf{h}$ (on the original scale of the multiple response-type data) is exactly the observed data

$\{Z_{ij}\}$ as the hyperparameters go to zero. This suggests that estimates from $f(\mathbf{h}|\mathbf{z}_{trn})$ overfits the data, however, it is not necessarily true that $f(\mathbf{y}, \boldsymbol{\theta}|\mathbf{z}_{trn})$ overfits the data.

3.4 Example of Bayesian Implementation

Consider the following mixed effects model for the transformed data (e.g., see Cressie and Johannesson, 2008, among others):

$$\begin{aligned}
&\text{Transformed Data Model: } \mathbf{h}|\boldsymbol{\beta}, \boldsymbol{\eta}, \xi_{ij}, \boldsymbol{\lambda} \sim m(\mathbf{h}|\boldsymbol{\lambda}) \prod_i \prod_j \phi\left(h_{ij}|\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{S}_{ij}'\boldsymbol{\eta} + \xi_{ij}, \sigma^2\right);\\
&\text{Process Model 1: } \boldsymbol{\eta}|\sigma_\eta^2 \sim \text{Normal}\left(\mathbf{0}_r, \sigma_\eta^2 \mathbf{I}_r\right);\\
&\text{Process Model 2: } \xi_{ij}|\sigma_\xi^2 \overset{\text{ind}}{\sim} \text{Normal}\left(0, \sigma_\xi^2\right);\\
&\text{Prior 1: } \sigma^2 \sim \text{IG}\left(\alpha_v, \beta_v\right);\\
&\text{Prior 2: } \boldsymbol{\beta} \sim \text{Normal}\left(\mathbf{0}_p, \sigma_\beta^2 \mathbf{I}_p\right);\\
&\text{Prior 3: } \sigma_\xi^2 \sim \text{IG}\left(\alpha_\xi, \beta_\xi\right);\\
&\text{Prior 4: } \sigma_\eta^2 \sim \text{IG}\left(\alpha_\eta, \beta_\eta\right),
\end{aligned} \tag{3.12}$$

where $\phi(\cdot|\mu, v)$ is the pdf of a normal distribution with mean μ and variance v, $\mathbf{x}_{ij}$ is a p-dimensional vector of known covariates, $\mathbf{I}_r$ is a $r \times r$ identity matrix, $\mathbf{0}_r$ is an r-dimensional vector of zeros, $\alpha_v = \alpha_\eta = \alpha_\xi = 1$, $\beta_v = \beta_\eta = \beta_\xi = 1$, $\sigma_\beta^2 = 100$, and $\boldsymbol{\xi} = (\xi_{11}, \ldots, \xi_{I_3 3})'$. The hyperparameters are chosen so that the prior is relatively "flat" and we find that our results are robust to these specifications. In Algorithm 1, we set $Y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{S}_{ij}'\boldsymbol{\eta} + \xi_{ij}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2, \sigma_\xi^2, \sigma_\eta^2)'$. The choice of basis functions and specification of r are important. In Supplementary Appendix C.2, we give these details.

The full conditional distributions for $\mathbf{y}$ and $\boldsymbol{\theta}$ are well-known (e.g.,see Cressie and Wikle, 2011, for a standard reference) and are listed in Supplementary Appendix C.2. Thus, Step 2 of Algorithm 1 involves simulating according to (3.10), which produces MCMC replicates $\{\mathbf{h}^{[b]}\}$ and $\{\boldsymbol{\gamma}^{[b]}\}$. Then Step 4 of Algorithm 1 involves the following steps:

1. Sample $\boldsymbol{\theta}^{[b]}$ from its' full conditional distribution $f(\boldsymbol{\theta}|\boldsymbol{\eta}^{[b-1]}, \{\xi_{ij}^{[b-1]}\}, \mathbf{h}^{[b]})$, and note that one can easily allow for block-wise updates. Notice that the full conditional distribution has the conditioning event $\mathbf{h}^{[b]}$, but does not condition on $\mathbf{z}_{trn}$ and $\boldsymbol{\gamma}^{[b]}$ due to the conditional independence assumptions described in Section 3.2. Let $\boldsymbol{\eta}^{[0]}$ and $\{\xi_{ij}^{[0]}\}$ be a pre-defined initialization.

2. Sample $\boldsymbol{\eta}^{[b]}$ from the full-conditional distribution $f(\boldsymbol{\eta}|\{\xi_{ij}^{[b-1]}\}, \boldsymbol{\theta}^{[b]}, \mathbf{h}^{[b]})$.

3. Sample $\{\xi_{ij}^{[b]}\}$ from the full-conditional distribution $f(\{\xi_{ij}\}|\boldsymbol{\eta}^{[b]}, \boldsymbol{\theta}^{[b]}, \mathbf{h}^{[b]})$.

4. Set $Y_{ij}^{[b]} = \mathbf{x}_{ij}'\boldsymbol{\beta}^{[b]} + \mathbf{S}_{ij}'\boldsymbol{\eta}^{[b]} + \xi^{[b]}$ and let $\mathbf{y}^{[b]} = (Y_{ij}^{[b]} : i = 1, \ldots, I_j, j = 1, 2, 3)'$.

These standard full-conditional distributions and other details are given in Supplementary Appendix C.1 and C.2.

4 Statistical Inference

Estimation and prediction over the training set can be done by computing summary statistics using the quantities generated in Step 4 of Algorithm 1. However, to forecast values (e.g., future cases or deaths due to COVID-19) we make use of validation and testing datasets.

4.1 Goodness-of-Fit Using Training Data

Assessment of the goodness of fit can be done similar to residual analyses of transformed data in traditional regression analyses. We compute the residuals $\boldsymbol{\delta} = (\delta_{ij} : i = 1, \ldots, I_j, j = 1, 2, 3)'$, $\delta_{ij} = h_{ij} - Y_{ij}$, and compute a credible region associated with $\boldsymbol{\delta}$ (e.g., see Gelman et al., 2013, for a standard reference). For example, for each i and j, find the values $q_{L,ij}$ and $q_{U,ij}$, where

$$\int_{q_{L,ij}}^{q_{U,ij}} f(\delta_{ij}|\mathbf{z}_{trn})d\delta_{ij} = 1 - \alpha; i = 1, \ldots, I_j, j = 1, \ldots, J, \tag{4.1}$$

where $1 - \alpha$ defines the "level" of the credible interval, is prespecified, and is different from the hyperparameters of the DY distribution. A default choice is $\alpha = 0.05$. In practice, it is rather straightforward to approximate $q_{L,ij}$ and $q_{U,ij}$. Let $h_{ij}^{[b]}$ and $Y_{ij}^{[b]}$ be the b-th posterior replicate of h_{ij} and Y_{ij} so that $\delta_{ij}^{[b]} = h_{ij}^{[b]} - Y_{ij}^{[b]}$ is the b-th posterior replicate of δ_{ij}. Then $q_{L,ij}$ and $q_{U,ij}$ can be approximated with the $\alpha/2$ and $1 - \alpha/2$ percentiles of the set $\{\delta_{ij}^{[b]} : b = 1, \ldots, B\}$, respectively. If the value of zero lies within this interval (e.g., $q_{L,ij} < 0 < q_{U,ij}$) for many values of i and j, then this suggests that the model for $\mathbf{y}$ provides a reasonable fit to this dataset.

Equation (3.11) shows that the posterior mean of the transformation overfits the data, which we motivated as a way to avoid oversmoothing estimates of $\mathbf{y}$ and $\boldsymbol{\theta}$ in Algorithm 1. However, the fact that the posterior distribution of the transformations overfit is also important from the point-of-view of diagnostics. In particular, in the goodness-of-fit literature, overfitted values are often used as a proxy for the data. For example, in log-linear models the most parsimonious reduced model that is not significantly different (in terms of the deviance or chi-square statistic) from the saturated model (an overfitted model) is used for statistical inference (e.g., see Agresti, 2007, for a standard reference). This is exciting because it provides a new way to conduct classical residual analysis in a Bayesian multiple response-type data context. In particular, in Section 5 we give an example of plotting the (posterior median) residuals versus a useful covariate not included in the analysis to assess whether or not it should be included in a model.

Since h_{ij} is unknown it is also of interest to determine the goodness-of-fit of our model for h_{ij}. That is, we use δ_{ij} to assess the goodness of fit of Y_{ij}, but one should consider the goodness of fit of h_{ij} as well. One approach is to "back-transform." However, since our model for the transformation h_{ij} is unknown, the back-transformation is also unknown. The Bayesian perspective is flexible enough to do inference on the back

transformation. In particular, we define the unknown back-transformation of h_{ij} to be the replicate from the data model $f(Z_{ij}|h_{ij})$, since replicates from $f(Z_{ij}|h_{ij})$ correspond to the transformation h_{ij}. Thus, posterior predictive data (denoted with Z^*_{ij}) can be used to estimate the unknown back-transformed value, and subsequently, be used to assess the performance of the transformation. In particular, if Z_{ij} is consistently (according to the posterior distribution) close to Z^*_{ij}, we obtain a transformation model that overfits, which is a goal for our model for h_{ij}. That is, the traditional posterior predictive p-value (e.g., Meng et al., 1994; Gelman et al., 1996) can be used to assess the goodness-of-fit of the back-transformation, and hence, the transformation itself. In practice, one might compute the proportion of times, over replicates of Z^*_{ij}, where

$$\sum_i \sum_j \frac{\left\{Z^*_{ij} - E(g_j^{-1}(h_{ij})|\mathbf{z}_{trn})\right\}^2}{E(g_j^{-1}(h_{ij})|\mathbf{z}_{trn})},$$

is larger than the observed chi-square statistic (or sum of squared Pearson residuals). Outside of the HGT setting, a preferable posterior predictive p-value would be 0.5 (e.g.,see Gelman, 2013, for a more complete discussion and other considerations when interpreting the posterior predictive p-value), since values close to one and zero indicate overfitting and oversmoothing, respectively. However, since we purposefully overfit the HGT when estimating $\{h_{ij}\}$, we prefer the traditional posterior predictive p-value to be "close" to one, which indicates a good fit for the model for the transformation.

4.2 Estimating Hyperparamers Using a Validation Dataset

In machine learning, one often adjusts the model for being biased towards the training data by holding aside a dataset to estimate hyperparameters. This hold-out dataset is referred to as a validation dataset (Hastie et al., 2009), where recall, the validation dataset $\mathbf{z}_{val} = (Z_{ij} : i = I_j + 1, \ldots, I_j^{val}, j = 1, 2, 3)'$ is observed over the indices $i \in \{I_j + 1, \ldots, I_j^{val}\}$ and $j = 1, 2, 3$. Additionally, let Y^*_{ij} be the posterior predictive replicate of Y_{ij}. We can not replace Y^*_{ij} with Y_{ij} in our analysis of the validation data, otherwise, the validation data would be included with the training data when updating Y_{ij}. Then, we assume

$$\begin{aligned}
Z_{i1}|Y^*_{i1} &\overset{\text{ind}}{\sim} \text{Normal}(k_1(Y^*_{i1}, \boldsymbol{\kappa}), v),\\
Z_{i2}|Y^*_{i2} &\overset{\text{ind}}{\sim} \text{Binomial}\left[b_i, g_2^{-1}\left\{k_2(Y^*_{i2}, \boldsymbol{\kappa})\right\}\right],\\
Z_{i3}|Y^*_{i3} &\overset{\text{ind}}{\sim} \text{Poisson}\left[g_3^{-1}\left\{k_3(Y^*_{i3}, \boldsymbol{\kappa})\right\}\right];\ \ i = I_j + 1, \ldots, I_j^{val},\\
\boldsymbol{\kappa} &\sim f(\boldsymbol{\kappa}),
\end{aligned} \tag{4.2}$$

where $\boldsymbol{\kappa}$ is a generic d-dimensional vector of real-valued parameters and $f(\boldsymbol{\kappa})$ is the prior distribution of this parameter. We parameterize the unknown function k_j with $\boldsymbol{\kappa}$. In this article, we define k_j to be a line so that Y^*_{ij} can be adjusted linearly,

$$k_j(Y, \boldsymbol{\kappa}) = \kappa_{j0} + \kappa_{j1}\, Y;\ \ j = 1, 2, 3,\ Y \in \mathbb{R}, \tag{4.3}$$

Algorithm 2 Steps Needed for Fitting the Validation Data.

1: Set $b = 1$ and initialize Y_{ij}^{*} and $\boldsymbol{\kappa}$ with $Y_{ij}^{*[0]}$ and $\boldsymbol{\kappa}^{[0]}$.
2: Sample $Y_{ij}^{*[b]}$ using Algorithm 1.
3: Sample $\boldsymbol{\kappa}^{[b]}$ from it's full-conditional distribution. We use the slice sampler (Neal et al., 2003) since the full-conditional distribution does not have a closed form.
4: Set $b = b + 1$.
5: Repeat Steps 2−5 until $b = B$ for a prespecified value of B.

and $\boldsymbol{\kappa} = (\kappa_{10}, \kappa_{20}, \kappa_{30}, \kappa_{11}, \kappa_{21}, \kappa_{31})'$. When we consider $\boldsymbol{\kappa}$ unknown, we choose the improper flat prior $f(\boldsymbol{\kappa}) = 1$. We also consider setting $\boldsymbol{\kappa} = (0, 0, 0, 1, 1, 1)'$ so that k_j is simply the identity function. Algorithm 2 describes implementation of our model for validation data.

There are two main reasons why we consider introducing k_j. The first, as discussed at the end of Section 2, is that often times estimates are biased towards the training data, since estimators are defined to be "close" to the training data. Consequently, the use of validation data to estimate $\boldsymbol{\kappa}$ can aid in accounting for out-of-sample error. A second reason for the use of k_j and $\boldsymbol{\kappa}$ is to model a certain feature in our COVID-19/social distancing dataset, where the validation and testing data are the second to last and last day of available data, respectively. Specifically, there is potential for dramatic day-to-day changes in our application, which may suggest that the training data is centered at a completely different value than the validation and testing. In this setting, κ_{j0} is needed to re-center the last two days. Additionally, one might redefine the link function to include k_j directly when implementing the HGT for training data; however, this specification implies that the training data will inform $\boldsymbol{\kappa}$ and as a result the out-of-sample error is not controlled. This post-processing step has been done before in the machine learning literature, where successive linear models are applied to residuals and are referred to as a "ResNet" (He et al., 2016). What differs in our implementation is that (1) we enforce only a single linear model $k_j(\cdot, \boldsymbol{\kappa})$ (i.e., we do not do multiple composites of a linear model), and (2) we are Bayesian in our implementation.

4.3 Forecasting

We produce next day forecasts for the variables in our study, which we treat as testing observations indexed over $i = I_j^{val} + 1, \ldots, I$ for $j = 1, 2, 3$. We let $\boldsymbol{\kappa}^*$ and Y_{ij}^{**} be distributed according to $f(\boldsymbol{\kappa}|\mathbf{z}_{trn}, \mathbf{z}_{val})$ and $f(Y_{ij}|\mathbf{z}_{trn})$, respectively. Again, we can not let Y_{ij}^{**} equal Y_{ij}, since otherwise, the testing data would be included when updating Y_{ij} based on the training data. Then, we assume that

$$
\begin{aligned}
&Z_{i1}|Y_{i1}^{**}, \boldsymbol{\kappa}^* \overset{\text{ind}}{\sim} \text{Normal}(k_1(Y_{i1}^{**}, \boldsymbol{\kappa}^*), v), \\
&Z_{i2}|Y_{i2}^{**}, \boldsymbol{\kappa}^* \overset{\text{ind}}{\sim} \text{Binomial}\left[b_i, g_2^{-1}\left\{k_2(Y_{i2}^{**}, \boldsymbol{\kappa}^*)\right\}\right], \\
&Z_{i3}|Y_{i3}^{**}, \boldsymbol{\kappa}^* \overset{\text{ind}}{\sim} \text{Poisson}\left[g_3^{-1}\left\{k_1(Y_{i3}^{**}, \boldsymbol{\kappa}^*)\right\}\right]; \quad i = I_j^{val}, \ldots, I. \qquad (4.4)
\end{aligned}
$$

Algorithm 3 Algorithm 3: Steps Needed for Forecasting.

1: Set $b = 1$ and initialize Y_{ij}^{**} and $\boldsymbol{\kappa}^*$ with $Y_{ij}^{**[0]}$ and $\boldsymbol{\kappa}^{*[0]}$.
2: Sample $Y_{ij}^{**[b]}$ using Algorithm 1.
3: Sample $\boldsymbol{\kappa}^{*[b]}$ using Algorithm 2.
4: Sample $Z_{ij}^{[b]}$ from (4.4).
5: Set $b = b + 1$.
6: Repeat Steps 2−5 until $b = B$ for a prespecified value of B.
7: Compute the sample mean and variance (across the index b) of $Z_{ij}^{[b]}$.

Predictions of the data process and estimation of cross-covariances can be found in a similar manner as in the GLMM example in (2.3) and (2.2). That is, the posterior mean and covariance of Z_{ij} and $Z_{k\ell}$ is, $E(Z_{ij}|\mathbf{z}_{trn})$ and $cov(Z_{ij}, Z_{k\ell}|\mathbf{z}_{trn})$, where recall, if we assume the process model in Section 3.4 $cov(Z_{ij}, Z_{k\ell}|\mathbf{z}_{trn}) = \mathbf{S}_{ij}' cov(\boldsymbol{\eta}|\mathbf{z}_{trn})\mathbf{S}_{k\ell}$, which is not necessarily zero. Implementation is summarized in Algorithm 3. When k_j is the identity function, the predictions and covariances are simply

$$\begin{aligned} E(Z_{ij}|\mathbf{z}_{trn}) &= E\big\{c_{ij}g_j^{-1}(Y_{ij})|\mathbf{z}_{trn}\big\}, \\ cov(Z_{ij}, Z_{k\ell}|\mathbf{z}_{trn}) &= cov(c_{ij}g_j^{-1}(Y_{ij}), c_{k\ell}g_\ell^{-1}(Y_{k\ell})|\mathbf{z}_{trn}). \end{aligned} \tag{4.5}$$

In the context of observing daily data, once the testing data is observed (i.e., once tomorrow's data is realized), we can assess the performance of our forecasts (e.g., through the root mean squared error, etc.).

4.4 Summaries of the Models used for Inference

We have defined models for the training data, validation data, and testing data, which we now summarize. The joint distribution of the training data, processes, and parameters is written as the product of the following conditional distributions:

$$\begin{aligned} &\text{Training Data Model 1}: Z_{i1}|h_{i1}, v \overset{\text{ind}}{\sim} \text{Normal}(h_{i1}, v); \\ &\text{Training Data Model 2}: Z_{i2}|h_{i2} \overset{\text{ind}}{\sim} \text{Binomial}\left\{b_i, \frac{\exp(h_{i2})}{1+\exp(h_{i2})}\right\}; \\ &\text{Training Data Model 3}: Z_{i3}|h_{i3} \overset{\text{ind}}{\sim} \text{Poisson}\{\exp(h_{ij})\}; \ \ i = 1, \ldots, I_j, j = 1, 2, 3; \\ &\text{Transformed Data Model}: \mathbf{h}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta})m(\mathbf{h}|\boldsymbol{\gamma}); \\ &\text{Process Model}: \mathbf{y}|\boldsymbol{\theta} \sim f(\mathbf{y}|\boldsymbol{\theta}); \\ &\text{Prior 1}: \boldsymbol{\theta} \sim f(\boldsymbol{\theta}); \\ &\text{Prior 2}: v \sim \text{IG}(1, 1); \\ &\text{Transformation Hyperprior}: \boldsymbol{\gamma} \sim f(\boldsymbol{\gamma}). \end{aligned} \tag{4.6}$$

The model in (4.6) is the aforementioned HGT model. This is a well defined proper model (see Supplementary Appendix B for these details), provided that $f(h_{ij}|\boldsymbol{\theta})$, $f(\mathbf{y}|\boldsymbol{\theta})$, and $f(\boldsymbol{\theta})$ are proper.

Recall that one motivation for the model in (4.6) is that one can incorporate their preferred model for continuous data directly into our framework, since Algorithm 1 does not require one to change the implementation of their preferred model. This flexibility arises in the data scientist's specification of $f(h_{ij}|\boldsymbol{\theta})$, $f(\mathbf{y}|\boldsymbol{\theta})$, and $f(\boldsymbol{\theta})$. In Section 3.4 we specify $f(h_{ij}|\boldsymbol{\theta})$, $f(\mathbf{y}|\boldsymbol{\theta})$, and $f(\boldsymbol{\theta})$ using a mixed effects model, and in Section 5 we also consider using BART to illustrate this flexibility. Although we only consider Bayesian specifications of the preferred model, Step 4 can easily be substituted with replicates/estimates of $\mathbf{y}$ and $\boldsymbol{\theta}$ (computed using $\mathbf{h}^{[b]}$) from empirical Bayesian models, approximate Bayesian models, or frequentist models.

The LCM is explicitly used in the HGT model in (4.6) through the term $m(\mathbf{h}|\mathbf{y})$, where recall

$$m(\mathbf{h}|\mathbf{y}) = \frac{\prod_{i,j} f_{DY}(h_{ij}|\alpha_j, \kappa_j, a, b)}{\int\int f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta}},$$

$\boldsymbol{\gamma} = (\alpha_1, \alpha_2, \alpha_3, \kappa_1, \kappa_2, \kappa_3, a, b)'$, f_{DY} denotes the DY prior, and the prior for $\boldsymbol{\gamma}$ is defined in Supplementary Appendix C.1. Recall that Algorithm 1 is a collapsed Gibbs sampler, where we update transformation $\mathbf{h}$ and $\boldsymbol{\gamma}$ using the marginal distribution of the HGT model in (4.6) that is found by integrating out the process $\mathbf{y}$ and parameters $\boldsymbol{\theta}$. Specifically, when integrating out $\mathbf{y}$ and $\boldsymbol{\theta}$ in (4.6), we obtain

$$\begin{aligned}
&\text{Training Data Model 1}: \; Z_{i1}|h_{i1} \overset{\text{ind}}{\sim} \text{Normal}(h_{i1}, v);\\
&\text{Training Data Model 2}: \; Z_{i2}|h_{i2} \overset{\text{ind}}{\sim} \text{Binomial}\left\{b_i, \frac{\exp(h_{i2})}{1+\exp(h_{i2})}\right\};\\
&\text{Training Data Model 3}: \; Z_{i3}|h_{i3} \overset{\text{ind}}{\sim} \text{Poisson}\{\exp(h_{ij})\}; \;\; i = 1, \ldots, I_j, j = 1, 2, 3;\\
&\text{Transformation Prior}: \; \mathbf{h}|\alpha_1, \alpha_2, \alpha_3 \kappa_1, \kappa_2, \kappa_3, a, b \sim \prod_{i,j} f_{DY}(h_{ij}|\alpha_j, \kappa_j, a, b);\\
&\text{Transformation Hyperprior}: \; \boldsymbol{\gamma} \sim f(\boldsymbol{\gamma}),
\end{aligned}$$

which leads to the computationally simple updates of the transformations $\mathbf{h}$ and parameters $\boldsymbol{\gamma}$ developed in Section 3.3 to be used in Step 2 of Algorithm 1.

The joint distribution of the validation data, processes, and parameters is written as the product of the following conditional distributions:

$$\begin{aligned}
&\text{Validation Data Model 1}: \; Z_{i1}|Y^*_{i1}, \boldsymbol{\kappa} \overset{\text{ind}}{\sim} \text{Normal}(k_1(Y^*_{i1}, \boldsymbol{\kappa}), v); && \text{(4.7a)}\\
&\text{Validation Data Model 2}: \; Z_{i2}|Y^*_{i2}, \boldsymbol{\kappa} \overset{\text{ind}}{\sim} \text{Binomial}\left[b_i, g_2^{-1}\{k_2(Y^*_{i2}, \boldsymbol{\kappa})\}\right]; && \text{(4.7b)}\\
&\text{Validation Data Model 3}: \; Z_{i3}|Y^*_{i3}, \boldsymbol{\kappa} \overset{\text{ind}}{\sim} \text{Poisson}\left[g_3^{-1}\{k_3(Y^*_{i3}, \boldsymbol{\kappa})\}\right]; && \text{(4.7c)}\\
&\text{Posterior Process Model}: \; Y^*_{ij}|\mathbf{z}_{trn} \sim f(Y^*_{ij}|\mathbf{z}_{trn});\\
&\text{Prior}: \; \boldsymbol{\kappa} \sim f(\boldsymbol{\kappa}); \;\; i = I_j + 1, \ldots, I^{val}_j, j = 1, 2, 3,
\end{aligned}$$

where recall that the goal of this model is to estimate $\boldsymbol{\kappa}$ from its posterior $f(\boldsymbol{\kappa}|\mathbf{z}_{val}, \mathbf{z}_{trn})$, which is a parameter that allows one to avoid overfitting the training data. The distribution $f(Y^*_{ij}|\mathbf{z}_{trn})$ is the posterior distribution implied by the HGT model in (4.6). Model

(4.7) can be implemented through Algorithm 2. When $f(\mathbf{y}|\boldsymbol{\theta})$ is specified according to a linear model (i.e., Equation (2.4)) then Equations (4.7a) through (4.7c) can be thought of as a GLMM (McCulloch et al., 2008). GLMMs also arise in our model for testing data. The joint distribution of the testing data, processes, and parameters is written as the product of the following conditional distributions:

$$\begin{aligned}
&\text{Testing Data Model 1}: Z_{i1}|Y_{i1}^{**},\boldsymbol{\kappa}^* \overset{\text{ind}}{\sim} \text{Normal}(k_1(Y_{i1}^{**},\boldsymbol{\kappa}^*),v);\\
&\text{Testing Data Model 2}: Z_{i2}|Y_{i2}^{**},\boldsymbol{\kappa}^* \overset{\text{ind}}{\sim} \text{Binomial}\left[b_i, g_2^{-1}\left\{k_2(Y_{i2}^{**},\boldsymbol{\kappa}^*)\right\}\right];\\
&\text{Testing Data Model 3}: Z_{i3}|Y_{i3}^{**},\boldsymbol{\kappa}^* \overset{\text{ind}}{\sim} \text{Poisson}\left[g_3^{-1}\left\{k_3(Y_{i3}^{**},\boldsymbol{\kappa}^*)\right\}\right];\\
&\text{Posterior Process Model}: Y_{ij}^{**}|\mathbf{z}_{trn} \sim f(Y_{ij}^{**}|\mathbf{z}_{trn});\\
&\text{Posterior Parameter Model}: \boldsymbol{\kappa}^*|\mathbf{z}_{val},\mathbf{z}_{trn} \sim f(\boldsymbol{\kappa}^*|\mathbf{z}_{val},\mathbf{z}_{trn});\\
&\qquad\qquad i = I_j^{val}+1,\ldots,I, j = 1,2,3,
\end{aligned} \tag{4.8}$$

where the goal is to predict the validation data Z_{ij} at $i = I_j^{val}+1,\ldots,I$ and $j = 1,2,3$. The distribution $f(Y_{ij}^{**}|\mathbf{z}_{trn})$ is the posterior distribution implied by the HGT model in (4.6) and $f(\boldsymbol{\kappa}^*|\mathbf{z}_{val},\mathbf{z}_{trn})$ is the posterior distribution from (4.7). The model for the testing data in (4.8) can be implemented through Algorithm 3.

5 Simulations

The goals of this simulation study are to provide a standard demonstration that the HGT model produces reasonable predictions in a computationally efficient manner. Another goal is to illustrate the flexibility of the HGT model to specify a data scientist's preferred model for continuous data. To do this we apply the HGT (4.6) to the spatio-temporal mixed effects model in Section 3.4 and BART (details in Supplementary Appendix C.3).

5.1 Simulation Setup

Friedman (1991) introduced a simulation design, which has become a useful benchmark study (e.g., see Chipman et al., 2010, among others). Let

$$\begin{aligned}
w(x_{1,ij},\ldots,x_{10,ij}) = 10\sin(\pi x_{1,ij}x_{2,ij}) + 20(x_{3,i}-0.5)^2 + 10x_{4,ij} + 5x_{5,i};\\
i = 1,\ldots,I, j = 1,2,3,
\end{aligned} \tag{5.1}$$

which includes two non-linear terms, two linear terms, and a non-linear interaction. We consider the following specifications of the distributional assumptions associated with the data:

$$\begin{aligned}
&Z_{i1} \overset{\text{ind}}{\sim} \text{Normal}(w(x_{1,i1},\ldots,x_{10,i1}),1),\\
&Z_{k2} \overset{\text{ind}}{\sim} \text{Binomial}\left\{300, \frac{\exp\left(w(x_{1,k2},\ldots,x_{10,k2})\right)}{1+\exp\left(w(x_{1,k2},\ldots,x_{10,k2})\right)}\right\},\\
&Z_{\ell 3} \overset{\text{ind}}{\sim} \text{Poisson}\left\{\exp\left(w(x_{1,\ell 3},\ldots,x_{10,\ell 3})\right)\right\},
\end{aligned} \tag{5.2}$$

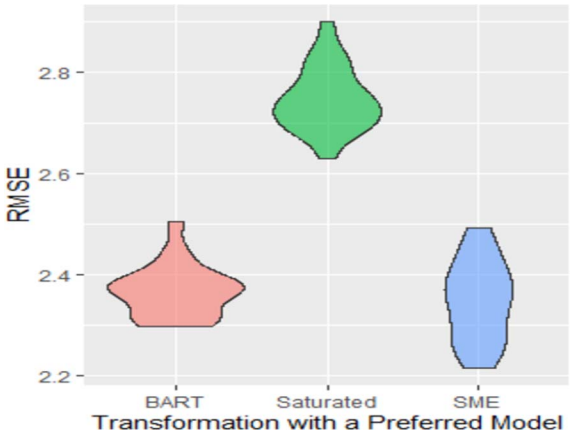

Figure 1: A violin plot of the RMSE (y-axis) by HGT method (x-axis) over 20 independent replicates of the data. The data are simulated as described in Section 5.1. Each HGT method is implemented using Algorithm 1, except the method "Saturated."

for $i = 1, \ldots, I_1$, $k = I_1 + 1, \ldots, I_1 + I_2$, and $\ell = I_1 + I_2 + 1, \ldots, I_1 + I_2 + I_3$. Methods are compared using the root mean squared error (RMSE),

$$\left(\frac{\sum_{i=1}^{I} \sum_{j=1}^{3} \left[\widehat{Y}_{ij} - w(x_{1,ij}, \ldots, x_{10,ij}) \right]^2}{3I} \right)^{1/2},$$

where $\widehat{Y}_{ij}$ is estimated using Monte-Carlo integration using 2,000 iterations with a burn-in of 1,000. For each Bayesian method, we let $\widehat{Y}_{ij}$ be the pointwise posterior mean of $g_j^{-1}(h)$. We fit the preferred model using covariates $x_{1,ij}, x_{3,ij}, x_{4,ij}, \ldots, x_{10,ij}$, and hence, we consider the case were an important covariate is not observed (i.e., $\{x_{2,ij}\}$) and several unneeded covariates are included (i.e., $\{x_{6,ij}, \ldots, x_{10,ij}\}$ are not present in (5.1)). The omissions of $\{x_{2,ij}\}$ when implementing our method is a slight departure from the original setup in Friedman (1991). However, we feel that it is more realistic to assume that not all covariates are observed in practice, and will be a helpful choice for illustration. We specify $x_{k,ij} \sim \text{Uniform}(0, 1)$, where $\text{Uniform}(0, 1)$ is a shorthand for the uniform distribution over the interval $[0, 1]$ and $k = 1, \ldots, 10$. The preferred models are spatio-temporal mixed effects and BART (and an extension), whose implementation are described in Supplementary Appendix C.2 and Supplementary Appendix C.3, respectively. Additionally, the choice of basis functions is described in Supplementary Appendix C.4. In the implementation of each preferred method, we allow each response-type to have different regression coefficients.

5.2 Simulations: Joint Analysis of Multiple Response-Types

In this section, we evaluate the predictive performance of our Bayesian model with unknown transformations in the multiple response-type data setting. In particular, we

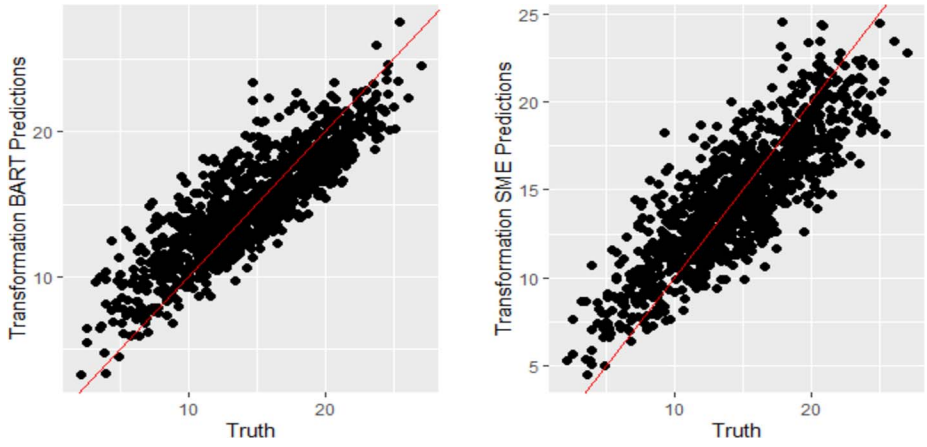

Figure 2: Estimates versus the truth for a single replicate dataset. The data are simulated as described in Section 5.1. The estimate is labeled on the y-axis. The red line indicates the line $y = x$.

set the preferred model equal to BART (Chipman et al., 2010) and a Bayesian version of the spatio-temporal mixed effects model (Cressie and Johannesson, 2008) using basis functions introduced by (Hughes and Haran, 2013). The posterior mean of h_{ij} (referred to as the saturated model) is included as a default poor estimator, since it is known to overfit the data.

The data are simulated according to (5.2), with $I = 1000$, $I_1 = 350$, $I_2 = 350$, and $I_3 = 200$. We do not include a validation dataset so that $k_j \equiv g_j$. We repeat this simulation study 20 times, and we provide violin plots of the RMSE over the 20 replicates by method in Figure 1. In Figure 2 we also plot the true function versus the estimated function for a single replicate dataset. Figures 1 and 2 suggest that the HGT-based spatio-temporal mixed effects model (and HGT-BART) performs well in terms of predictive performance. For the replicate in Figure 2 the HGT-based spatio-temporal mixed effects (and HGT-BART) model had 97% (94%) of the point-wise credible intervals of the elements of $\boldsymbol{\delta}$ containing zero. The patterns observed in Figure 1 mimic the goodness-of-fit diagnostics, which is notable because the goodness-of-fit diagnostics are data driven (and hence can be used in practice) while Figure 1 is based on the unknown truth. The posterior predictive p-value is ≈ 1, which suggests that our model for transformation overfits as desired (see Section 4.1). These results suggest that the Bayesian transformations can be used to obtain predictions in the non-Gaussian setting using two standard models, and also has a useful built-in goodness-of-fit diagnostic.

Now, suppose we have observed the values of $\{x_{2,ij}\}$, and recall these covariates are not included in the analysis. In Figure 3, we plot the posterior median of the residuals versus the covariate $\{x_{2,ij}\}$ across the indexes i and j for a single replicate of the dataset. The plot clearly indicates a sinusoidal or possibly quadratic pattern, which suggests that this behavior is not captured in our model for $\mathbf{y}$. We know this to be

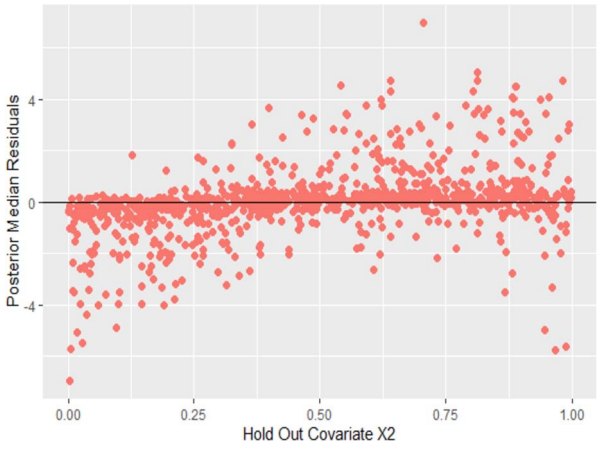

Figure 3: We simulate a single replicate of $\{Y_{ij}\}$ according to Section 5.1. Then a spatio-temporal mixed effects model is implemented using the specifications in Section 3.4. This plot displays the posterior median of $\{\delta_{ij}\}$ (see Section 4.1) versus $\mathbf{x}_{2,ij}$, which is not included in our implementation of the spatio-temporal mixed effects model. A systematic pattern in this plot suggests that including $\mathbf{x}_{2,ij}$ would improve our analysis of $\mathbf{y}$.

true because $\{x_{2,ij}\}$ is not included in our implementation, but the data was generated using $\{x_{2,ij}\}$. This is an illustration of how our approach provides a Bayesian analog to a graphical technique from classical regression analysis (i.e., systematic patterns in residuals from a multiple regression versus a covariate suggest that the covariate should be included in the analysis).

5.3 Simulations: Robustness to Departures from Model Assumptions

In this simulation study we compare the predictive performance of our HGT approach to predictions from the preferred model itself. A straightforward way to do this is to restrict ourselves to the continuous data-only setting, in which both modeling paradigms can be implemented more readily. The data are simulated according to (5.2), with $I_1 = 800$, $I = 1000$, and $I_2 = I_3 = 0$. We do not include a validation dataset.

We repeat this simulation study 20 times, and we provide violin plots of the RMSE over the 20 replicates by method in Figure 4. In this section, we include an additional predictor: soft BART (SBART; Linero and Yang, 2018, see Supplementary Appendix C.3 for more details). We again see that the HGT versions of BART and the spatio-temporal mixed effects model outperform the saturated model, with the spatio-temporal mixed effects model clearly outperforming BART. Additionally, the HGT version of BART and spatio-temporal mixed effects model perform only slightly better than or the same as their non-transformed counterparts. Here we see that SBART performs worse than the saturated model in terms of RMSE. The HGT version of SBART does

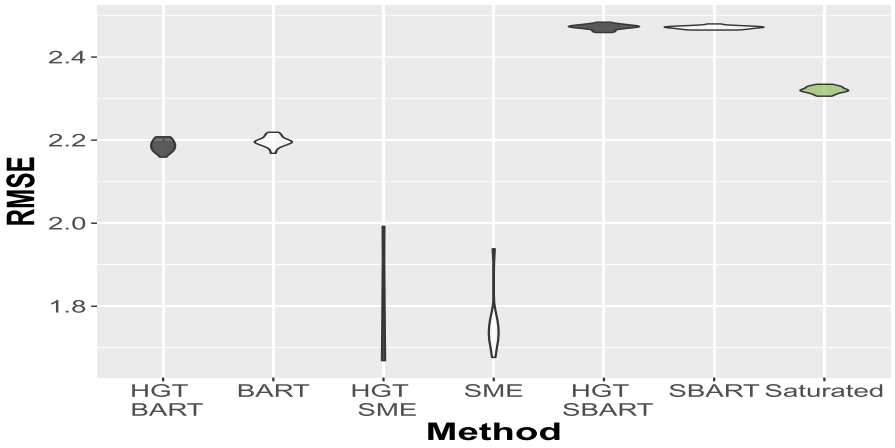

Figure 4: A violin plot of the RMSE (y-axis) by method (x-axis) over 20 independent replicates of the data. The data are simulated as described in Section 5.1. Each method is implemented using Algorithm 1, except the method "Saturated." The observed dataset are used as the predicted values for the method "Saturated." The results for the HGT method are highlighted in black, the results for the original method are highlighted white, and the saturated model is highlighted green (these models are also indicated on the x-axis).

not perform noticeably different than SBART in terms of RMSE. This suggests that in the continuous only setting, if the preferred model performs well (or poorly) one should expect the Bayesian transformation approach to perform well (or poorly). Recall that we can use the goodness-of-fit approach in Section 4.1 to assess when a method performs poorly in practice. For example, for a single replicate dataset, we found that the percent of credible intervals of the elements of $\boldsymbol{\delta}$ that contain zero (by method) are as follows: 99.8% (spatio-temporal mixed effects model), 77.4% (BART), and 58.1% (SBART). This produces the same rankings of the method in terms of RMSE. Additionally, the posterior predictive p-value is ≈ 1, which suggests that our model for transformation overfits as desired (see Section 4.1).

5.4 Simulations: Computational Considerations

In the continuous-only data setting Algorithm 1 requires more computation, and hence more computation time. This is because Step 2 in Algorithm 1 is not needed to implement the preferred model without transformation, but is required to implement the HGT. Thus, in the continuous-only data setting the main motivation for the HGT is that it allows for the diagnostics in Section 4.1, and the un-transformed preferred model does not. However, for Poisson-only, binomial-only, and multiple response-type data Algorithm 1 can lead to clear computational speed-ups compared to existing GLMMs. To demonstrate this, we compare the HGT and the GLMM implementation of the SME

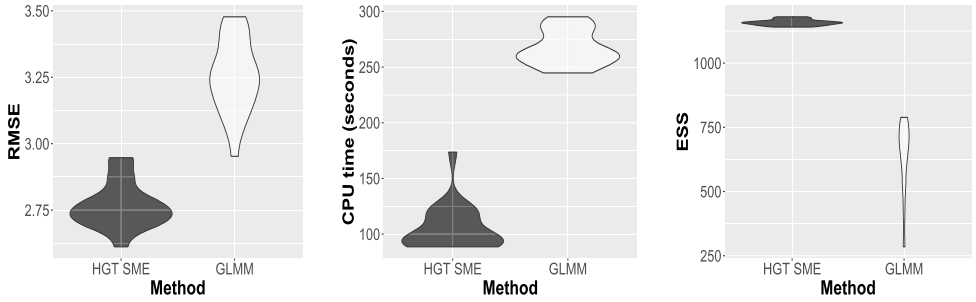

Figure 5: In the left panel we plot a violin plot of the RMSE (y-axis) by method (x-axis) over 20 independent replicates of the data, the middle plot replaces the RMSE with Central Processing Units (CPU) in seconds, and in right plot we replace the RMSE with the ESS. The data are simulated as described in Section 5.1. The HGT is implemented using Algorithm 1, and the GLMM is implemented using a Gibbs sampler with Metropolis Hastings updates. The results for the HGT method are highlighted in black, and the results for the GLMM are highlighted white (these models are also indicated on the x-axis).

model for a Poisson-only dataset (i.e., $I_1 = 0$, $I_2 = 0$, and $I_3 = 350$). The GLMM is implemented using the same covariates/basis functions as the HGT, and the R package `MCMCglmm`, based on Metropolis-Hastings updates, is used (Hadfield, 2010). Algorithm 1 is implemented using 2,000 replicates and a burn-in of 1,000. Trace plots were used to informally investigate convergence with no lack of convergence detected. The Bayesian GLMM is implemented with 13,000 replicates, a burn-in of 3,000, and thinning rate of 10. The effective sample size (ESS) of Y_{ij} is computed to better compare the computational performance of HGT and the GLMM (e.g., see Kass et al., 2016, among others). The ESS is the total MCMC replicates times a ratio of the within chain variance and between chain variance. We average component-wise ESS instead of implementing a multivariate version of ESS (Vats et al., 2019), since our goal is to simply compare the HGT and the GLMM.

In Figure 5, we see that the GLMM took more time (roughly 275 seconds for the GLMM, and roughly 100 seconds for the HGT) to produce fewer average effective samples (roughly 800 for the GLMM and roughly 1,300 for the HGT) than the HGT. Furthermore, the predictive performance of HGT is better than the GLMM, as measured by RMSE, in this Poisson-only data setting. This is somewhat different from what is seen in the continuous-only data setting in Section 5.3, where the difference in RMSE between the HGT and the preferred model was negligible.

There are of course several other methods/computational approaches besides the GLMM with Metropolis-Hastings updates that are more computationally efficient (e.g., LCM or the use of INLA, etc.). Thus, a fair conclusion would be the following: if the continuous-only preferred model (e.g., the preferred model is proportional to SME) is considerably more computationally advantageous than the non-Gaussian version of the preferred model (e.g., the GLMM), then the HGT is a more computationally practical

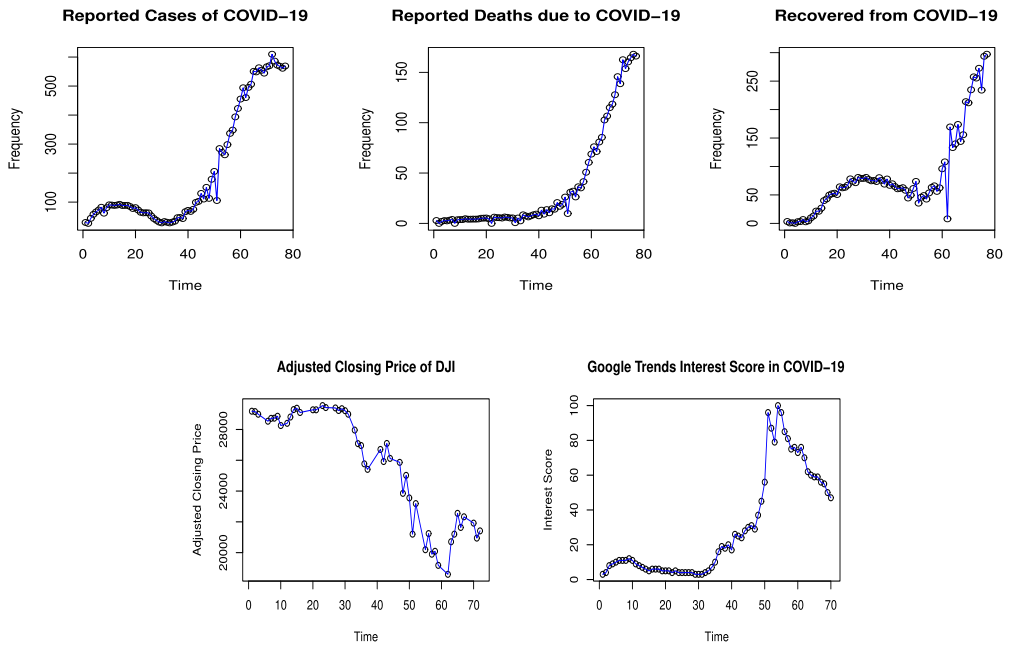

Figure 6: We plot the number of reported COVID-19 infections (top left), reported COVID-19 deaths (top middle), the reported recoveries from COVID-19 (top right), the DJI adjusted closing price (bottom left), and the Google Trends interest score for searches of "coronavirus" (bottom right). Note that the DJI price data is not available on Saturday and Sundays. The black circles are the observed data, and blue lines connecting these points are added as a reference. The top row represents only a summary of available data, since we also observe these counts over 184 countries and 82 provinces.

choice in the non-Gaussian response-type setting (as illustrated in the above example). This is because HGT can use the same implementation of the continuous-only preferred model in Step 4, and the added computations in Step 2 are "small" as compared to, say, Metropolis-Hastings updates in a GLMM.

6 Joint Analysis of Reported COVID-19 Occurrences, the Adjusted Closing Price of the Dow Jones Industrial, and Google Trends Data

We now present our joint analysis of occurrences of COVID-19, the adjusted closing price of the DJI, and the Google Trends interest score in searches of "coronavirus" (see time series displays of this data in Figure 6). The MCMC is implemented according to Algorithms 1 through 3 with 10,000 replicates and a burn-in of 1,000. Convergence was assessed visually through the use of trace plots and through Gelman-Rubin diag-

nostics (Gelman et al., 2013) with no indications of a lack of convergence. All of our analyses were implemented on Windows 10 with the following specifications: Intel(R) CORE(TM) i5-8250U CPU with 1.60Gh.

6.1 The Dataset

COVID-19 was first detected in a live animal market in Wuhan City within the Hubei Province of China (Guarner, 2020). This virus spreads easily from person to person, and there are cases of this virus where an individual is unsure of how they became infected (i.e., community transmission, Dowd et al., 2020; Guarner, 2020). As of this writing, no vaccine has been approved for broad distribution in nearly all countries (Corum et al., 2020; Wangping et al., 2020). As such, many governmental organizations, including the Centers for Disease Control and Prevention (CDC), have advised placing distance between yourself and other individuals (i.e., social distancing, CDC, 2020). Social distancing is an important public health measure that reduces close contact with people that may be infected by maintaining physical distance between all individuals (Wilder-Smith and Freedman, 2020; Zhang et al., 2020; CDC, 2020). However, social distancing comes as a cost, and can be detrimental to economies and cause psychological distress (Long, 2020; Park et al., 2020).

The data on reported deaths and cases of COVID-19 were obtained from the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) Coronavirus repository (publicly available at `https://github.com/CSSEGISandData/COVID-19`), a subset of which, is made available in the R package `coronavirus` (Krispin, 2020). Cases, recoveries and mortality counts are available over regions (i.e., country or province) and discrete time (daily). In this article, we model these counts using a Poisson distribution, and our main interest lies in estimating the mean number of reported deaths and cases of COVID-19, and estimating its dependence with interest in COVID-19 and DJI data.

The number of Google searches of "coronavirus" is indicative of the high interest on COVID-19 and can act as a loose proxy for the public interest in COVID-19. This search information is made available through Google Trends data (Google, 2020). Google Trends provide daily time series of an "interest" measure of searches on Google. This interest measure is defined on a scale from zero to one hundred with 100 indicating high interest and zero indicating low interest. In this article, we model the Google Trends interest score for the search "coronavirus" as binomial with sample size 100, since this response is a non-negative, integer-valued response that is bounded above by 100.

The DJI follows 30 publicly owned blue chip (i.e., nationally recognized and financially secure) companies that trade on the New York Stock Exchange (NYSE) and the National Association of Securities Dealers Automated Quotations (NASDAQ). It is a benchmark for blue-chip stocks and is often treated as a measure of the economic health of the US. This data was obtained through Yahoo Finance (Yahoo, 2020). We model the adjusted daily closing price with a Gaussian distribution, since it is continuous valued. Our main interest in DJI is in determining and summarizing the relationship between the adjusted closing price with both interest in COVID-19 and reported cases and deaths due to COVID-19.

Let Z_{i1} represent the negative adjusted closing price of DJI per \$10,000, Z_{i2} be the integer-valued interest score for COVID-19 searches as computed by Google Trends (with $b_i \equiv 100$), and i indexes the days ranging from January 22, 2020 to April 8, 2020. We analyze Z_{i1} on the negative scale so that we see an increasing trend over time among all three response-types. The data Z_{i3} represents the i-th replicate of the number of COVID-19 cases, where for each i there is an associated region (e.g., China) $A_i \subset\in [-180, 180] \times [-90, 90]$, day t_i (between January 22, 2020 to April 8, 2020), and an indicator d_i of whether or not the count consists of reported deaths. Let $d_i = 1$ if Z_{i3} represents reported deaths and $d_i = 0$ otherwise. Likewise, let u_i represent an indicator of whether or not the count consists of reported recoveries. Also let $t_i = 1, \ldots, T = 78$ represent each day between January 22, 2020 to April 8, 2020. In Figure 6, we plot the number of reported COVID-19 infections, reported COVID-19 deaths, the DJI adjusted closing price, and the Google Trends interest score for searches of "coronavirus." This plot is based on data reported as of this writing, and this data is continually being monitored and updated.

Our specifications of the basis functions are defined in Supplementary Appendix C.4, and covariates include an indicator for the region and response-types, d_i, and u_i. The data from January 22, 2020 to April 6, 2020 are the training data ($n = 10,600$), the data on April 7, 2020 is held-out as a validation dataset (373 observations), and the data on April 8 is held-out as a testing dataset (374 observations).

6.2 Goodness of Fit

In Figure 7 we plot the posterior mean death, confirmed cases, recovered cases, adjusted closing price, and Google Trends interest score. Here, we see that the predicted values are reasonably close to their observed values with the observed data close contained within a pointwise 95% credible interval. These results suggest that the in-sample error is small, and that the predicted values reflect the general patterns of the data. Goodness of fit can be formally investigated according to the approaches in Section 4.1. Roughly 99.4% percent of the credible intervals, as defined in (4.1), contain zero. This provides additional evidence the model provides a reasonable fit to the data. The posterior predictive p-value is ≈ 1, which suggests that our model for transformation overfits (see Section 4.1). In the bottom right panel of Figure 7 we plot the posterior median residual (i.e., $\boldsymbol{\delta}$) versus the time the observation was recorded. Here we see roughly no pattern over time, which suggests that our specification of the basis functions were reasonable.

6.3 Estimation and Prediction

We did not include the data on April 8-th, 2020, which was the most current value available at the time of the analysis. We use the HGT model to predict the number of deaths, number of confirmed recoveries, and number of confirmed cases according to Algorithm 3. In Figure 8 we provide the posterior means associated with these values versus the testing data. The use of Algorithm 2 aided in producing more accurate forecasts as the posterior means of κ_{02} and κ_{12} were roughly equal to 3.5 and 1.5, respectively. In general, the posterior means from Algorithm 3 trends the testing data,

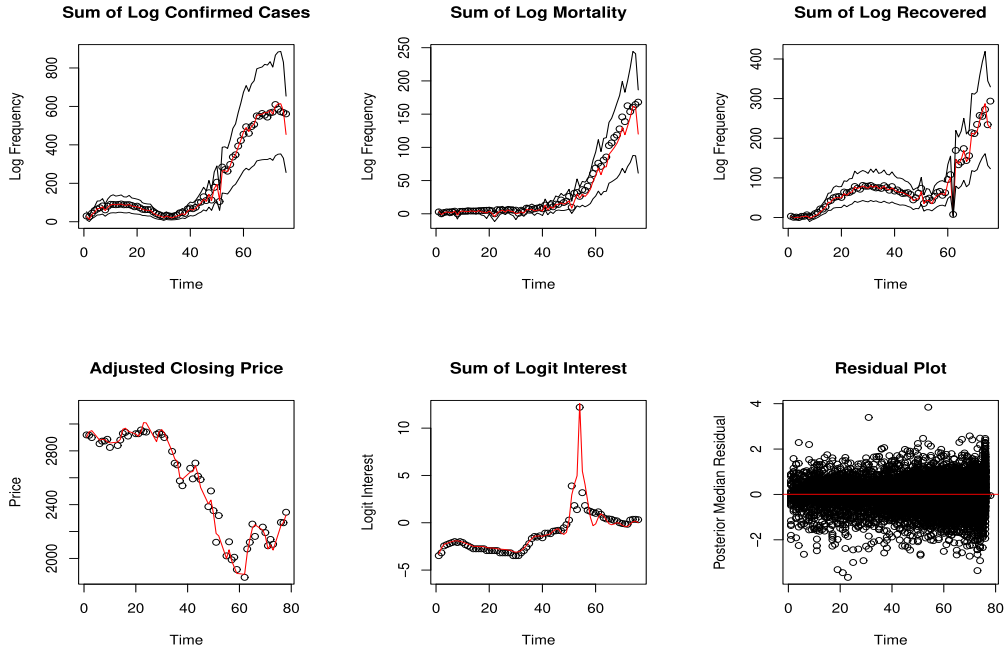

Figure 7: Goodness of Fit: We plot the sum (over regions) of log number of reported COVID-19 infections (top left), sum (over regions) log number of reported COVID-19 deaths (top middle), sum (over regions) log number of reported COVID-19 recoveries (top right), the DJI adjusted closing price (bottom left), and the logit ($log(Y_{i2}/100 - Y_{i2})$) Google Trends interest score for searches of "coronavirus" (bottom middle). Note that the DJI price data is not available on Saturday and Sundays. The red lines represent the predicted values from our model, and the black circles represent the observed values. The black lines are pointwise 95% credible intervals. The credible intervals are left out in the bottom panels for visualization purposes (credible intervals are large), and in this panel each datum falls within their respective credible interval. The posterior median residuals versus time is given in the bottom right panel.

except for smaller testing values, where there is a tendency to overestimate the log count. However, the percentage (over the testing data) of pointwise credible intervals that contain the testing data is 98.4%, which suggest that the uncertainty of these estimates are captured in the model. This property of the model is also seen in the plot of the posterior variance versus the posterior mean, also displayed in Figure 8. Here, smaller predicted values tend to be over-dispersed, and larger predicted values appear to be equi-dispersed. Thus, we appear to have accurate predictions of the areas with the largest confirmed cases, recoveries, and deaths. Being able to accurately estimate large values of (log) occurrences is particularly important. That is, if we know *where* there are large occurrences of confirmed cases, then additional testing of individuals in these regions allows one to isolate all those who test positive in this region, which ultimately reduces the spread of COVID-19 from this region to others (Ai et al., 2020).

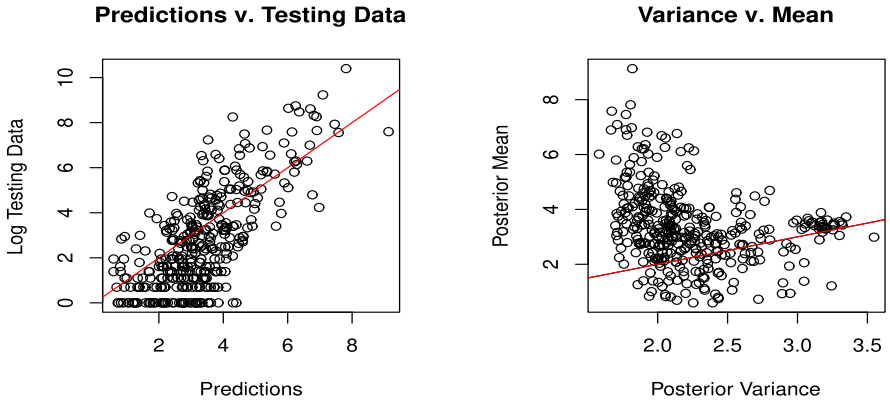

Figure 8: Forecasting: In the left panel we plot the forecasted testing data using Algorithm 3. Here the testing data represents all confirmed cases, recoveries, and deaths on April 8, 2020. The right panel plots the posterior variance of the predicted testing data versus the posterior mean.

Consequently, models such as ours can be useful at stopping the spread of COVID-19. However, finer-scale regional data would be necessary for this model to be helpful in narrowing in on potential "hot-spots" in practice.

In Figure 9, we plot the posterior mean of the random effects that is shared across response-type along with pointwise 95% credible intervals (see Section 4.1). The time period $t_i = 1$ to $t_i = 33$ (corresponding to January 22, 2020 and February 23, 2020) was particularly crucial, since this time range saw the strongest direct effects between COVID-19 cases, the negative adjusted closing price, and Google Trends interest-score in the Google search "coronavirus." Furthermore, the fact that zero does not tend to fall within the credible intervals suggests that our incorporation of dependence across response-types, spatial regions, and days was reasonable. Time point $t_i = 33$ (corresponding to February 23, 2020) marks a time in which the adjusted closing price initially started to decrease (see Figures 1 and 6), and the Google Trends interest score increases. After $t_i = 33$ the random effect appears to be negative-valued, which suggests an indirect relationship among these responses.

7 Discussion

We introduce the HGT model, which is derived from a straightforward combination of the LCM and the GLMM. This combination is motivated as a means to aid other researchers to analyze multiple response-type datasets such as the one considered in this article. In particular, our approach provides several contributions to Bayesian statistics. First, we have developed a general all-purpose Bayesian model to analyze multiple responses (e.g., continuous, Binomial counts, and Poisson counts). Our approach allows one to directly incorporate their preferred Bayesian model to analyze multiple response-

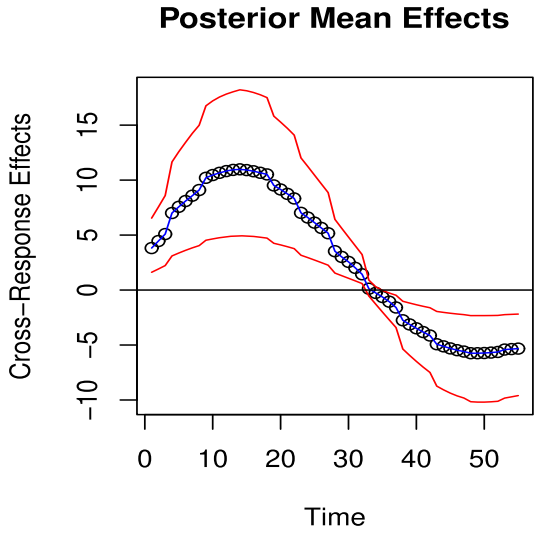

Figure 9: We plot the posterior mean of $\sum_{T_i=t} \mathbf{S}'_{ij}\boldsymbol{\eta}$. The red line indicates pointwise 95% credible intervals.

type data without completely abandoning their approach to the implementation of their preferred model. Second, we developed a general Bayesian analog to the classical comparison between a saturated model and a reduced model. This results in the use of classical residual analysis for assessing goodness-of-fit in Bayesian models for multiple response-type data. Third, we develop an approach to forecasting by introducing new parameters in a model for validation data. Code and tutorials on how to adapt the HGT to your preferred model can be found at https://github.com/JonathanBradley28/CM.

We illustrate the HGT on a dataset of COVID-19 and social distancing related variables. COVID-19 is a global epochal health disaster, and social distancing has become a necessary public health measure to protect the health of individuals. In particular, we investigate the relationship between COVID-19 cases, the US economy (specifically the negative adjusted closing price of DJI), and interest on Google (specifically Google Trends interest score for the search "coronavirus"). The data and model suggests that the relationship among these three values had the strongest positive relationship during a majority of February 2020, which suggests that this was an important time period. Additionally, there are clear cross-dependencies among response-types, regions, and days. It is important to comment that correlation does not imply causation, and to make explicit causal conclusions one needs to adopt methods among the causal inference literature (Rubin, 2005). Finally, the HGT model produces reasonable forecasts of the log frequency of cases, deaths, and recoveries from COVID-19. This suggests that with finer-scale regional data, this model could potentially be useful for targeting future hot-spots of COVID-19.

In our simulations, an illustration was given of non-linear functional analysis of multiple response-types using BART as the preferred model. Additionally, an illustration

was given of a joint spatial analysis of multiple response-types using a spatio-temporal mixed effects model as the preferred model. These results suggest that the prediction error of our approach is small (in terms of RMSE), and we can develop multiple response-type data versions of two different preferred models seamlessly. Computational benefits in the non-Gaussian setting was also illustrated. Additionally, data driven goodness-of-fit diagnostics were able to lead to the same conclusion as the RMSE criterion (based on the latent process) that is unobserved in practice.

Supplementary Material

Supplementary Materials: Joint Bayesian analysis of multiple response-types using the hierarchical generalized transformation model (DOI: 10.1214/20-BA1246SUPP; .pdf).

References

Agresti, A. (2007). *Categorical data analysis, 2nd Ed.* Springer. MR2293447. doi: https://doi.org/10.1002/0470114754. 130, 140

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., and Xia, L. (2020). "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases." *Radiology*, 200–642. 154

Argyriou, A., Evgeniou, T., and Pontil, M. (2007). "Multi-task feature learning." *Advances in neural information processing systems*, 19. 128

Bean, A., Xu, X., and MacEachern, S. (2016). "Transformations and Bayesian density estimation." *Electronic Journal of Statistics*, 10(2): 3355–3373. MR3572853. doi: https://doi.org/10.1214/16-EJS1158. 128

Beasley, T. M., Erickson, S., and Allison, D. B. (2009). "Rank-based inverse normal transformations are increasingly used, but are they merited?" *Behavior genetics*, 39(5): 580. 128

Box, G. E. P. and Cox, D. R. (1964). "An analysis of transformations." *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2): 211–243. MR0192611. 128, 134

Bradley, J. R. (2020). "Supplementary Materials: Joint Bayesian analysis of multiple response-types using the hierarchical generalized transformation model." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1246SUPP. 131

Bradley, J., Holan, S., and Wikle, C. (2018). "Computationally Efficient Distribution Theory for Bayesian Inference of High-Dimensional Dependent Count-Valued Data." *Bayesian Analysis*, 13: 253–302. MR3773410. doi: https://doi.org/10.1214/17-BA1069. 128, 129

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2020+). "Bayesian Hierarchical Models with Conjugate Full-Conditional Distributions for Dependent Data from the Natural

Exponential Family." *Journal of the American Statistical Association*. doi: https://doi.org/10.1080/01621459.2019.1677471. 128, 129, 137

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2019a). "Hierarchical Models for Spatial Data with Errors that are Correlated with the Latent Process." *Statistica Sinica*. doi: https://doi.org/10.5705/ss.202016.0230. 129

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2019b). "Spatio-temporal models for big multinomial data using the conditional multivariate logit-beta distribution." *Journal of Time Series Analysis*, 40(3): 363–382. 128, 129

Breiman, L. and Friedman, J. H. (1985). "Estimating optimal transformations for multiple regression and correlation." *Journal of the American statistical Association*, 80(391): 580–598. MR0803258. 128, 134

CDC (2020). "Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020." *Morbidity and Mortality Weekly Report*, 69(12): 343–346. 130, 152

Charitidou, E., Fouskakis, D., and I. Ntzoufras, I. (2018). "Objective Bayesian transformation and variable selection using default Bayes factors." *Statistics and Computing*, 28(3): 579–594. MR3761343. doi: https://doi.org/10.1007/s11222-017-9749-3. 128

Charitidou, E., Fouskakis, D., and Ntzoufras, I. (2015). "Bayesian transformation family selection: Moving toward a transformed Gaussian universe." *Canadian Journal of Statistics*, 43(4): 600–623. MR3433678. doi: https://doi.org/10.1002/cjs.11261. 128

Chen, M. H. and Ibrahim, J. G. (2003). "Conjugate priors for generalized linear models." *Statistica Sinica*, 13(2): 461–476. MR1977737. 129

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 4(1): 266–298. MR2758172. doi: https://doi.org/10.1214/09-AOAS285. 145, 147

Christensen, W. F. and Amemiya, Y. (2002). "Latent variable analysis of multivariate spatial data." *Journal of the American Statistical Association*, 97: 302–317. MR1947288. doi: https://doi.org/10.1198/016214502753479437. 128

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P., and Zhang, S. (2017). "Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data." *Ecological Monographs*, 87: 34–56. 128

Corum, J., Wee, S., and Zimmer, C. (2020). "Coronavirus Vaccine Tracker." *The New York Times*. https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html. (accessed October 23, 2020). 152

Cox, D. (1970). "The continuity correction." *Biometrika*, 217–219. MR0671999. doi: https://doi.org/10.2307/2335435. 138

Cressie, N. and Johannesson, G. (2008). "Fixed rank kriging for very large spatial data

sets." *Journal of the Royal Statistical Society, Series B*, 70: 209–226. MR2412639. doi: https://doi.org/10.1111/j.1467-9868.2007.00633.x. 139, 147

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley. MR2848400. 130, 135, 139

Damien, P., Wakefield, J., and Walker, S. (1999). "Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61. MR1680334. doi: https://doi.org/10.1111/1467-9868.00179. 129

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families." *The Annals of Statistics*, 17: 269–281. MR0520238. 129, 137

Dobra, A. and Lenkoski, A. (2011). "Copula Gaussian graphical models and their application to modeling functional disability data." *The Annals of Statistics*, 5: 969–993. MR2840183. doi: https://doi.org/10.1214/10-AOAS397. 128

Donnat, C. and Holmes, S. (2020). "Modeling the Heterogeneity in COVID-19's Reproductive Number and its Impact on Predictive Scenarios." *arXiv preprint* arXiv:2004.05272. 130

Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Ding, X., Liu, Y., and Mills, M. C. (2020). "Demographic science aids in understanding the spread and fatality rates of COVID-19." *Proceedings of the National Academy of Sciences of the United States of America*, 117(18): 9696–9698. 152

Efron, B. (1992). "Bootstrap methods: another look at the jackknife." In *Breakthroughs in statistics*, 569–593. Springer. MR0659849. 136

Fellinghauer, B., Buhlmann, P., Ryffel, M., Rhein, M. V., and Reinhardt, J. D. (2013). "Stable graphical model estimation with random forests for discrete, continuous, and mixed variables." *Computational Statistics amd Data Analysis*, 64: 132–152. MR3061894. doi: https://doi.org/10.1016/j.csda.2013.02.022. 128

Fienberg, S. E. (1969). "Preliminary graphical analysis and quasi-independence for two-way contingency tables." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 18(2): 153–168. MR0247728. doi: https://doi.org/10.2307/2346257. 138

Friedman, J. H. (1991). "Multivariate adaptive regression splines." *The Annals of Statistics*, 19(1): 1–67. MR1091842. doi: https://doi.org/10.1214/aos/1176347963. 145, 146

Gao, H. and Bradley, J. R. (2019). "Bayesian analysis of areal data with unknown adjacencies using the stochastic edge mixed effects model." *Spatial Statistics*. MR3955693. doi: https://doi.org/10.1016/j.spasta.2019.100357. 129

Gelfand, A. E. (2000). "Gibbs sampling." *Journal of the American statistical Association*, 95(452): 1300–1304. MR1825281. doi: https://doi.org/10.2307/2669775. 132

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1: 515–533. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 138

Gelman, A. (2013). "Two simple examples for understanding posterior p-values whose distributions are far from uniform." *Electronic Journal of Statistics*, 7: 2595–2602. MR3121624. doi: https://doi.org/10.1214/13-EJS854. 141

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, 3rd edn.* Boca Raton, FL: Chapman and Hall/CRC. MR3235677. 133, 136, 140, 152

Gelman, A., Meng, X., and Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica Sinica*, 6: 733–807. MR1422404. 141

Google (2020). "Google Trends." https://trends.google.com/trends/. 152

Guarner, J. (2020). "Three Emerging Coronaviruses in Two Decades: The Story of SARS, MERS, and Now COVID-19." *American Journal of Clinical Pathology*, 153(4): 420–421. 152

Yang, H.-C., Hu, G., and Chen, M.-H. (2019). "Bayesian Variable Selection for Pareto Regression Models with Latent Multivariate Log Gamma Process with Applications to Earthquake Magnitudes." *Geosciences*, 9(4): 169. 128, 129

Hadfield, J. D. (2010). "MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package." *Journal of statistical software*, 33(2): 1–22. 150

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer. MR2722294. doi: https://doi.org/10.1007/978-0-387-84858-7. 128, 130, 133, 141

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. 130, 131, 133, 142

Hu, G. and Bradley, J. R. (2018). "A Bayesian spatial–temporal model with latent multivariate log-gamma random effects with application to earthquake magnitudes." *Stat*, 7(1): e179. MR3796093. doi: https://doi.org/10.1002/sta4.179. 128, 129

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed model." *Journal of the Royal Statistical Society, Series B*, 75: 139–159. MR3008275. doi: https://doi.org/10.1111/j.1467-9868.2012.01041.x. 147

Kang, E. L. and Cressie, N. (2011). "Bayesian inference for the spatial random effects model." *Journal of the American Statistical Association*, 106: 972–983. MR2894757. doi: https://doi.org/10.1198/jasa.2011.tm09680. 129

Kass, R. E., Carlin, B. P., and Neal, R. M. (2016). "Markov chain Monte Carlo in practice: a roundtable discussion)." *The American Statistician*, 52: 93–100. MR1628427. doi: https://doi.org/10.2307/2685466. 150

Katzfuss, M. and Cressie, N. (2012). "Bayesian hierarchical spatio-temporal smoothing for very large datasets." *Environmetrics*, 23: 94–107. MR2873787. doi: https://doi.org/10.1002/env.1147. 129

Kim, S., Chen, M. H., Ibrahim, J. G., Shah, A. K., and Lin, J. (2013). "Bayesian inference for multivariate meta-analysis Box–Cox transformation models for individual patient data with applications to evaluation of cholesterol-lowering drugs." *Statistics in Medicine*, 32(23): 3972–3990. MR3102429. doi: https://doi.org/10.1002/sim.5814. 128

Kim, S. and Xing, E. P. (2009). "Statistical estimation of correlated genome associations to a quantitative trait network." *PLos Genetics*, 5. 128

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. (2015). "TMB: automatic differentiation and Laplace approximation." *arXiv preprint* arXiv:1509.00660. 132

Linero, A. R. and Yang, Y. (2018). "Bayesian regression tree ensembles that adapt to smoothness and sparsity." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5): 1087–1110. MR3874311. doi: https://doi.org/10.1111/rssb.12293. 148

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). "High-dimensional semiparametric Gaussian copula graphical models." *The Annals of Statistics*, 40: 2293–2326. MR3059084. doi: https://doi.org/10.1214/12-AOS1037. 128

Liu, H., Lafferty, J., and Wasserman, L. (2009). "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." *The Journal of Machine Learning Research*, 10: 2295–2328. MR2563983. 128

Long, N. J. (2020). "From social distancing to social containment: reimagining sociality for the coronavirus pandemic." *Medicine Anthropology Theory* 7(2): 247–260. 152

Martino, S. and Riebler, A. (2019). "Integrated nested Laplace approximations (inla)." *arXiv preprint* arXiv:1907.01248. 132

McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S., and Lin, X. (2019). "Omnibus Inverse Normal Transformation Based Association Test Improves Power in Genome-Wide Association Studies of Quantitative Traits." *bioRxiv*, 635706. doi: https://doi.org/10.1101/635706. 128

McCullagh, P. and Tresoldi, M. F. (2020). "A likelihood analysis of quantile-matching transformations." *arXiv preprint* arXiv:2001.03709. 128

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. NJ: Wiley. MR2431553. 128, 145

McCulloch, R. E. (1993). "Fitting regression models with unknown transformations using dynamic graphics." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 42(2): 153–160. 128, 134

Meng, X.-L. et al. (1994). "Posterior predictive p-values." *The Annals of Statistics*,

22(3): 1142–1160. MR1311969. doi: https://doi.org/10.1214/aos/1176325622. 141

Mosteller, F. and Tukey, J. W. (1977). *Data analysis and regression: a second course in statistics*. Addison-Wesley Publishing Company: Philippines. 138

Neal, R. M. (2011). "MCMC Using Hamiltonian Dynamics." In Brooks, S., Gelman, A., Jones, G. L., and Meng, X. (eds.), *Handbook of Markov Chain Monte Carlo*, 113–160. Chapman and Hall. MR2858447. 130

Neal, R. M. et al. (2003). "Slice sampling." *The annals of statistics*, 31(3): 705–767. MR1994729. doi: https://doi.org/10.1214/aos/1056562461. 136, 142

Neal, R. M. et al. (2011). "MCMC using Hamiltonian dynamics." *Handbook of Markov chain Monte Carlo*, 2(11): 2. MR3185067. 132

Office of Science and Technology Policy (2020). "Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset." https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/. 132

Park, C. L., Russell, B. S., Fendrich, M., Finkelstein-Fox, L., Hutchison, M., and Becker, J. (2020). "Americans' COVID-19 Stress, Coping, and Adherence to CDC Guidelines." *Journal of General Internal Medicine*, 1. 152

Krispin, R. (2020). "Package 'coronavirus'." Retrieved April, 2020. 152

Rubin, D. B. (1981). "The Bayesian bootstrap." *The annals of statistics*, 130–134. MR0600538. 136

Rubin, D. B. (2005). "Causal inference using potential outcomes: Design, modeling, decisions." *Journal of the American Statistical Association*, 100(469): 322–331. MR2166071. doi: https://doi.org/10.1198/016214504000001880. 156

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations." *Journal of the Royal Statistical Society, Series B*, 71: 319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 130, 132

Schliep, E. M. and Hoeting, J. A. (2013). "Multilevel latent Gaussian process model for mixed discrete and continuous multivariate response data." *Journal of agricultural biological and environmental statistics*, 18: 492–513. MR3142597. doi: https://doi.org/10.1007/s13253-013-0136-z. 128

Servin, B. and Stephens, M. (2007). "Imputation-based analysis of association studies: candidate regions and quantitative traits." *PLoS genetics*, 3(7). 128

Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media. MR1391963. doi: https://doi.org/10.1007/978-1-4612-4026-6. 130

Sweeting, M. J., Sutton, A. J., and Lambert, P. C. (2004). "What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data." *Statistics in medicine*, 23(9): 1351–1375. 138

Todd, C. M., Swallow, B., Illian, J. B., and Toms, M. (2018). "A spatiotemporal multispecies model of a semicontinuous response." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67: 705–722. MR3787973. doi: https://doi.org/10.1111/rssc.12250. 128

Vats, D., Flegal, J. M., and Jones, G. L. (2019). "Multivariate output analysis for Markov chain Monte Carlo." *Biometrika*, 106(2): 321–337. MR3949306. doi: https://doi.org/10.1093/biomet/asz002. 150

Wangping, J., Ke, H., Yang, S., Wenzhe, C., Shengshu, W., Shanshan, Y., Jianwei, W., Fuyin, K., Penggang, T., and Jing, L. (2020). "Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China." *Frontiers in medicine*, 7: 169. 152

Wilder-Smith, A. and Freedman, D. O. (2020). "Isolation, quarantine, social distancing and community containment: pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak." *Journal of travel medicine*, 27(2): taaa020. 152

Wu, G., Holan, S. H., Nilon, C. H., Wikle, C. K., et al. (2015). "Bayesian binomial mixture models for estimating abundance in ecological monitoring studies." *The Annals of Applied Statistics*, 9(1): 1–26. MR3341105. doi: https://doi.org/10.1214/14-AOAS801. 128

Xue, L. and Zou, H. (2012). "Regularized rank-based estimation of high-dimensional nonparanormal graphical models." *The Annals of Statistics*, 40: 2541–2571. MR3097612. doi: https://doi.org/10.1214/12-AOS1041. 128

Yahoo (2020). "Yahoo Finance." https://finance.yahoo.com/. 152

Yang, E., Ravikumar, P., Allen, G. I., Baker, Y., Wan, Y. W., and Liu, Z. (2014). "A general framework for mixed graphical models." *arXiv preprint* arXiv:1411.0288. 128

Yang, X., Kim, S., and Xing, E. P. (2009). "Heterogeneous multitask learning with joint sparsity constraints." *Advances in Neural Information Processing Systems*, 2151–2159. 128

Yates, F. (1934). "Contingency tables involving small numbers and the χ 2 test (Supplement)." *Journal of the Royal Statistical Society*, 1(2): 217–235. MR0769998. doi: https://doi.org/10.2307/2981577. 138

Yeo, I.-K. and Johnson, R. A. (2000). "A new family of power transformations to improve normality or symmetry." *Biometrika*, 87(4): 954–959. MR1813988. doi: https://doi.org/10.1093/biomet/87.4.954. 128, 134

Zhang, J., Litvinova, M., Liang, Y., Wang, Y., Wang, W., Zhao, S., Wu, Q., Merler, S., Viboud, C., and Vespignani, A. (2020). "Age profile of susceptibility, mixing, and social distancing shape the dynamics of the novel coronavirus disease 2019 outbreak in China." *medRxiv*. doi: https://doi.org/10.1101/2020.03.19.20039107. 152

Acknowledgments

This research was partially supported by the U.S. National Science Foundation (NSF) under NSF grant SES-1853099. I'd like to that the editor, associate editor, and reviewers for their helpful comments that improved this paper. I also would like to thank Drs. Christopher Wikle and Scott Holan at the University of Missouri on their feedback on an earlier version of this article.