[Dissertations & Theses (Open Access)](https://...)   School of Public Health

Spring 4-2019

Stochastic Search Variable Selection Applied To A Bayesian Hierarchical Generalized Linear Model For Dyads

Adriana Lopez Ordonez
UTHealth School of Public Health

Recommended Citation

STOCHASTIC SEARCH VARIABLE SELECTION APPLIED TO A BAYESIAN

HIERARCHICAL GENERALIZED LINEAR MODEL FOR DYADS

by

ADRIANA LOPEZ ORDONEZ, MS

APPROVED:

_____

MICHAEL D SWARTZ, PHD

_____

LUIS LEON-NOVELO, PHD

_____

MELISSA F PESKIN, PHD

_____

ROSS SHEGOG, PHD

_____

DEAN, THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

DEDICATION

To my dear mother Maria del Pilar and my three children David, Mark and Sara.

STOCHASTIC SEARCH VARIABLE SELECTION APPLIED TO A BAYESIAN

HIERARCHICAL GENERALIZED LINEAR MODEL FOR DYADS

by

ADRIANA LOPEZ ORDONEZ

MS, San Diego State University, 2003
MS, University of Mexico, 2000

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
April 2019

Acknowledgements

I would like to express my gratitude to my advisor Dr. Swartz and my co-supervisor Dr. Leon-Novelo for their knowledge, time and commitment on this project. I would also like to thank the members of my committee, Dr. Peskin and Dr. Shegog for their time, support and valuable comments on the manuscript of this dissertation.

My sincere thanks to Dr. Agopian that allowed me to work with him as a Graduate Research. Many thanks go to my fellow students Fadi, Suman, Jitesh, Renata and Xiao for their friendship and support.

Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my mom and my sisters Alejandra, Rocio and Lety and my brother Luis whose understanding, support and guidance are with me in whatever I pursue. Thank you to my husband, Carlos, who has supported me throughout this process. I would also thank my three wonderful children, David, Mark and Sara, who provide unending inspiration.

A special thanks to the IYG-F program for letting me work with the data set used in this dissertation.

Finally, I thank God for all I have in my life.

STOCHASTIC SEARCH VARIABLE SELECTION APPLIED TO A BAYESIAN

HIERARCHICAL GENERALIZED LINEAR MODEL FOR DYADS

Adriana L Ordonez, MS, PhD
The University of Texas
School of Public Health, 2019

Dissertation Chair: Michael D Swartz, PhD

In behavioral science research, many outcomes of interest can be influenced by interpersonal relationships (Gyarmathy & Neaigus, 2007). To assess such outcomes, data can be collected using dyads. Each dyad has two elements, an actor, who responds to a stimulus and a partner, who can potentially influence the actor (Kenny, Kashy, & Cook, 2006). One popular model for analyzing dyadic data is the Actor Partner Interdependence model (APIM). In this study, we proposed a variable selection method applied to a probit Bayesian Hierarchical Generalized Linear Model (Bayesian HGLM) to fit the APIM to dyadic data.

The proposed method used stochastic search technology to identify key predictors of the Bayesian HGLM for APIM. It included a component for selecting interactions; selecting only interactions with both main effects also included. The proposed method was evaluated in two different forms, with simulated data and with real data. When we evaluated the method using simulated data, we examined its performance on 5 different simulated scenarios with varying associated predictors and two different sample sizes: a large sample size (2000 dyads) and a small sample size. And when we evaluated the method using real

data, we used baseline data from an evaluation of the program Its Your Game-Family (IYG-F). The baseline data set had the complete information of 61 dyads.

Across the 5 scenarios, the proposed variable selection method selected the correct variables over 85% of the simulated data sets in either sample size. And using the real data, the proposed variable selection method selected one construct out of 6 to be associated with the binary outcome. Thus, using the real data, we concluded that the construct of teenage *Sex Communication Self-Efficacy Relational* explains the outcome *Sexual initiation,* and the effects are equal across dyad members (teenager-parent).

In conclusion, in this study, we implemented the first variable selection procedure specifically to analyze dyadic data, based on stochastic search technology. The selection procedure can be applied in any research study that involves dyadic data from the APIM model with a binary outcome and a set of continuous covariates.

Table of Contents

LIST OF TABLES

i

LIST OF FIGURES

LIST OF APPENDICES

Literature Review

In behavioral sciences many outcomes of interest can be influenced by interpersonal relationships, and observed behaviors are often the result of interactions with more than one person. For example, the reaction that an adolescent may experience during an early sex encounter may be influenced by parents, siblings or friends opinions. Therefore, the unit of study in behavioral sciences is often not *an* individual, but *a group* of individuals. Specifically, when only a pair of individuals is involved in an interaction, the pair is called a dyad. According to social psychologists, a dyad comprises an *actor* and a *partner*. The actor is defined as the person who rates or responds to a stimulus; and the partner is someone whose characteristics influence the actor's responses (Garcia, Kenny, & Ledermana, 2014).

Classification of Research Models for Dyads

There exist three main dyad-based models (Kenny, Kashy, & Cook, 2006), depending on the research question that has been raised. The first model, which is known as the Actor Partner Interdependency model (APIM), is used when every person in an interpersonal relationship belongs to one and only one dyad. It is often applied to research on interactions between a mother and her child, for example, when they both provide self-ratings about their communication styles and interaction quality. The APIM is also considered as the standard model for dyadic data (Kenny, Kashy, & Cook, 2006). The second model is the social relations model (SRM), or round robin design, where all members of a group of participants

1

interact with one another. In the SRM, a person (actor/partner) not only serves as a member of multiple dyads, but plays a dual role in multiple interactions (Ludtke, Kenny, & Ulrich, 2013). The SRM is often used in studies on popularity among teenagers in a given high school and given grade, where each student is paired with the rest of the students in the same grade from his/her high school. The third model involves one member being paired with multiple other members of dyads, but the rest of the members are not paired, because researchers are only interested in investigating one-way relationships. Partner level model (PLM) is an example of one-way relationships or one-sided designs (Mustanski, Starks, & Newcomb, 2015). This design is often applied when young children are rated by teachers, but young children do not rate the teachers back. In short, in each dyad case, an actor and a partner are involved. In this study, we will use data arising from an APIM; thus, we will only focus on the APIM and appropriate statistical models for this design.

Actor-Partner Interdependence Model (APIM)

The APIM allows estimation of the effects from both elements of a dyad, the actor and the partner on the outcome variable. In a statistical model, both actor and partner effects are considered independent variables, and the outcome is a function of them. Furthermore, the outcome in the APIM will be a function of two groups of independent variables: one group that comes from the actor (the same person that generates the outcome), the other group consists of the same independent variables but measured from the perspective of the partner (the person that influences the actor). For example, in the APIM, when a mother and a child are asked about their communication about sex, both of them will provide answers

2

that are the outcomes. When the mother provides the outcome, the child is present in the

analysis and considered the partner. Likewise, when the child provides the outcome, the

mother is considered the partner (Figure 1). Dyadic partners in the APIM influence each

other, and this partnership has an influence on their individual lives. What is more, if it is

possible to identify both the actor effects and the partner effects in the model, the dyad

involved is a *distinguishable* dyad; if it is not possible, it is an *indistinguishable* dyad

(Kenny, Kashy, & Cook, 2006). In other words, a dyad is distinguishable if there are one or

more characteristics to identify each member in the dyad. In the case of a parent paired with a

child, it is a distinguishable dyad because each individual can be distinguished by factors or

characteristics such as gender and age.

Figure 1: Children-mother communication problem using the APIM
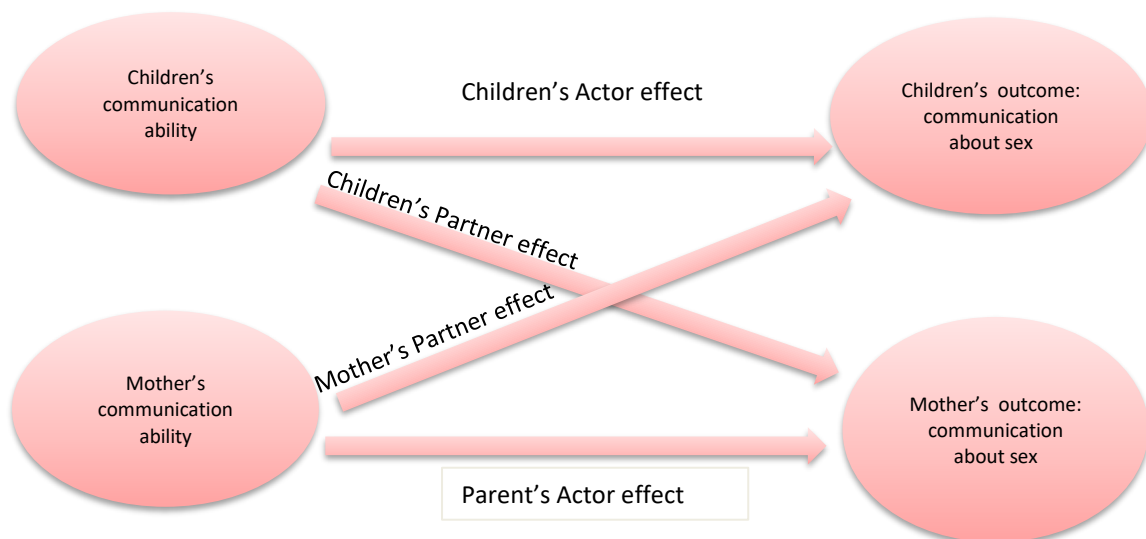
Constructs in Behavioral Science Research

In statistics, "independent variable" or "covariate" is a very general term. In most cases, it is simply a measure of a specific characteristic of the subject under study, such as sex, gender, or economic status. However, in some cases it is not so simple, and it can be a function of different measures such as body mass index (BMI). In behavioral science research, in particular, some independent variables known as "constructs" can be complex in nature, because a construct is a conceptual variable that is known to exist, but cannot be directly observed (Privitera, 2013). The construct is usually measured by a set of questions (known as items) related to an underlying individual psychological characteristic; these items are often highly interrelated. The answers to the questions are then combined to produce a single score that numerically represents the degree of presence of the construct in an individual. An APIM usually involves modeling the association between independent variables or constructs, such as gender, race and socioeconomic status with an outcome using dyadic data.

Statistical Tools for Models of Dyads

Dyadic partners affect each other, and their partnership influences their individual responses as well. From a statistical point of view, data from dyads cannot be considered a sample of independent observations. Therefore, the correlation structure cannot be ignored. For the APIM, one of the most common statistical tools to analyze dyads is the analysis of variance (ANOVA) (Gill & Swartz, 2007). One important derivation is the two-way ANOVA with random effects that decomposes the variance into actor, partner, and relationship effects. Other common methods include multilevel modeling, structural equation

modeling, and generalized estimation equations (Kenny, Kashy, & Cook, 2006). In particular, multilevel models have been extended to include Bayesian estimation to produce a hierarchical linear model (HLM) (Gill & Swartz, 2007). Under such an approach, Bayesian methods have enabled statistically reliable inferences considering variance components and correlations, even when sample sizes are small (Gill & Swartz, Statistical Analyses for Round Robin Interaction Data, 2001). Moreover, the Bayesian HLM has been extended to provide a single unified estimation method for the APIM for continuous and categorical outcomes ((Ahn, Liu, Wang, & Yuan, 2013); (Baragatti, 2011)). This integrative model is known as Bayesian Hierarchical Generalized Linear Model (Bayesian HGLM) and will be the focus of s dissertation.

Bayesian Hierarchical Generalized Linear Models (Bayesian HGLM)

The Bayesian HGLM for dyads consists of two levels: The first level models the variability *within* a dyad; and the second level models the variability *between* dyads. The first-level variables are called individual levels, and they are characteristics of the individual. An example of the individual-level covariate for a mother-and-child dyad would be educational level. Elements of the second level are called dyad-level independent variables or dyad-level constructs, and they are characteristics that equally apply to both elements of the dyad. An example of a dyad-level covariate for a mother-child dyad would be the family socioeconomic status, since in studies both the mother and child are often categorized with the same level of family socioeconomic status (Garcia, Kenny, & Ledermana, 2014).

If the research question involves finding the best set of independent variables or constructs to explain the outcome, a variable selection method should be applied to the dyadic model. The Bayesian HGLM is a type of generalized linear model, so model selection strategies can be applied to it. One such selection method that readily applies to generalized linear models under the Bayesian framework is stochastic search variable selection (SSVS) (George & McCulloch, 1993; Swartz, Mueller, & Amos, 2006; Ntzoufras, Forster, & Dellaportas, 2000). Other selection methods can be applied to Bayesian models, such as absolute shrinkage and selection operator **(**Lasso) and deviance criteria (DIC, BIC, and AIC). However, these methods have limitations when applied to an APIM for dyad data. For example, Lasso in its selection process selects only one variable out of a group of variables that are highly correlated (Zhang, et al., 2014), making this process highly restrictive. Lastly, the deviance criteria (DIC) may not be appropriate to implement as a selection method in our model due the large possible number of models to compare (2^p). Therefore, this study focuses on SSVS applied to Bayesian HGLM.

Stochastic Search Variable Selection (SSVS)

When a statistical model is built, the main objective is to capture all relevant pieces of information that explain the variability of the responses. There are different methods in Bayesian statistics to determine the subset of explanatory variables. The SSVS was originally used in linear models (George & McCulloch, 1993), but researchers have extended its use to most of the generalized linear models. For example, in conditional logistic regression models (Swartz, Mueller, & Amos, 2006), log-linear models (Ntzoufras, Forster,

& Dellaportas, 2000), in survival models (Stingo, Chen, Tadesse, & Vannucci, 2011), and in longitudinal logistic regression models (Ahn, Liu, Wang, & Yuan, 2013). For a given set of explanatory variables, the basic idea of the SSVS method is to include in a model all independent variables or constructs that balance good explanatory power with adequate estimation performance. Essentially, the SSVS framework uses a latent variable to represent the question "Does this variable belong to the model?" and describes the computational machinery to compute an answer to that question in a way that considers all possible models, or at least the most probable models.

Furthermore, Bayesian SSVS explores a set of different statistical models for a given set of covariates by limiting the posterior distribution of non-significant variables of the outcome in a small neighborhood around zero. Based on this assumption, the SSVS is easily implemented via the Gibbs sampler. Moreover, it can provide information regarding the inclusion of each variable in the final model by analyzing their corresponding posterior probability of inclusion at the end of the stochastic process (Yi, George, & Allison, Stochastic Search Variable Selection for Identifying Multiple Quantitative Trait Loci, 2003).

Researchers have been using the core idea of SSVS to solve specific problems, such as deciding when an interaction term between two or more independent variables (independent variable will be a generic name for a construct or covariate in a model) needs to be included in a statistical model. Each independent variable involved in the interaction in this case often also needs to be included in the model for interpretability. Specifically, for the Bayesian HGLM, when it is applied to the APIM, if interactions between covariates at the

7

dyad level and individual level are considered important for the research question, both the dyad-level variable and the individual-level variable should be included. Chipman, George and McCulloch (2001) showed a derivation of SSVS that takes into account such criteria for including interactions in a statistical model. Their technique consists of assigning conditional probabilities to the interaction selection indicators that are conditioned on the selection status of the main effects.

Advantages of using SSVS in Behavioral Science Research

The SSVS has been shown to be a powerful statistical method to select an accurate model in logistic regression compared with standard selection methods, regardless of how much information was specified and expressed through the priors (Swartz, Yu, & Sanjay, 2008). A Bayesian HGLM has also been applied to the APIM (Baragatti, 2011; Ahn, Liu, Wang, & Yuan, 2013). However, to date, there are limited methods that can select the interaction between constructs forcing the main effects of the interaction in the model for a small sample size. Statistical methods based on the SSVS, however, can fill this research gap and help identify constructs that are associated with an outcome/ behavior of interest and its interaction in a restricted parametric space. Therefore, in this study, we will develop a SSVS framework to select constructs as well as the interaction between constructs when analyzing dyadic data under the APIM. To test the proposed method, we also perform simulations to show that selection process is working adequately for Bayesian HGLM when applied to the APIM. We will also apply it to dyadic data from an effective health promotion program to determine its suitability for real-world data.

Public Health Significance

Many public health problems, such as teen pregnancy, are related to behaviors that are influenced by the behaviors of parents, siblings, friends, classmates, among others. The effects of parental-child relationships on health behaviors and outcomes have been widely studied, with mixed results ((Latkin & Knowlton, 2015; Looze, Constantine, Jerman, Vermeulen-Smit, & ter Bogt, 2015). These findings may help the design and delivery of health interventions since interventions may be more effective when important interactions between individuals rather than individuals themselves are identified. Dyadic analysis is emergent in the context of sexual health research because it can consider interrelated behaviors. Therefore, the proposed SSVS jointly with Bayesian HGLM is expected to provide a set of models with different combinations of independent variables (constructs) for dyadic data under the standard model (APIM). Moreover, the proposal statistical method can simplify a Bayesian HGLM by identifying the key items and interactions that are more related to the variability of the outcome of interest using a smaller sample size than that used by most statistical methods; thus, providing public health researchers a powerful tool to analyze dyadic data in the presence of small sample sizes.

Specific Aims

We will develop a stochastic search variable selection (SSVS) based method to select outcome related constructs when analyzing dyadic data from the actor-partner interdependence model (APIM). We will validate the proposed statistical method first using simulated data and second applying it to real world data. We will use data from the baseline

assessment of a randomized control trial of the It's Your Game Family (IYG-F) program. IYG-F is an internet-based intergenerational sexual health education program for adolescents and parents (Entitled the "Secret of Seven Stones") (Ceglio L. , 2015). The data set includes complete information of 61 dyads at baseline. The elements of the dyads are the parent or legal guardian, who provides most of the care, and their teen child, who is between the ages of 11-14 years old. The primary outcome of this study is parent-child communication about sex initiation. The constructs measured in the IYF-F program are intentions and beliefs about child disclosure, for example, communication about sex, quality of the communication about sex, self-efficacy for communication about sex etc. (Appendix 1 includes more details about the constructs).

Specific Aim #1: **Develop an SSVS framework to select independent variables or constructs in a Bayesian HGLM for the APIM.** We will develop and evaluate a statistical method that performs the SSVS using a probit Bayesian HGLM for the APIM. We will evaluate our method using different scenarios of simulated data with a larger sample size (greater than 1000 dyads).

Specific Aim #2: **Test and application of the SSVS framework to select independent variables or constructs in a Bayesian HGLM for the APIM using a small data set.** We will test the method from Aim #1 using simulated data for different scenarios with a small sample size (200 dyads or less), and we will apply it to the baseline data set of the IYG-F program.

CHAPTER II

Stochastic Search Variable Selection Applied to a Bayesian Hierarchical Generalized Linear Model for Dyadic Data

Journal of Statistical Computation and Simulation

ABSTRACT

In behavioral science research, many outcomes of interest can be influenced by interpersonal relationships (Gyarmathy & Neaigus, 2007). To assess such outcomes, data can be collected using dyads. Each dyad has two elements, an *actor*, who responds to a stimulus and a *partner*, who can potentially influence the actor (Kenny, Kashy, & Cook, 2006). One popular model for analyzing dyadic data is the Actor-Partner Interdependence model (APIM). In this study, we proposed a variable selection method applied to a probit Bayesian Hierarchical Generalized Linear Model (Bayesian HGLM) to fit the APIM to dyadic data. The proposed method uses stochastic search technology to identify key independent variables of the BHGLM for APIM. It includes a component for selecting interactions; selecting only interactions with both main effects also included. The proposed method was evaluated by examining its performance on 4 different simulated scenarios with varying associated independent variables. In this study, we implemented the first variable selection procedure specifically to analyze dyadic data, based on stochastic search technology. The model is able to detect associated independent variables, but requires a larger sample size to detect that the effect of a covariate is different in the partner than in the actor.

11

INTRODUCTION

In behavioral science research, many outcomes of interest (e.g., sexual behavior) can be influenced by interpersonal relationships (Gyarmathy & Neaigus, 2007). At the same time, interactions among interpersonal relationships (e.g., caregiver and care receiver) can influence the outcomes of interest. In order to truly analyze such outcomes, researchers collect information on both parties involved in the relationship, a data structure commonly referred to as dyads.   Each dyad has two elements, an *actor* and a *partner*. The actor is defined as the person who rates or responds to a stimulus, and the partner is someone whose characteristics will influence the actor's responses (Kenny, Kashy, & Cook, 2006).When it is possible to uniquely identify both the actor and partner, the dyad involved is known as a distinguishable dyad (Kenny, Kashy, & Cook, 2006). For example, the case of a primary caregiver paired with a teenager is a distinguishable dyad because each person in the dyad can be distinguished by sociodemographic factors or characteristics, such as gender and age. In the literature, there are different models for analyzing dyadic data, each of them specific to the nature of the research question of interest. One of these models is the Actor-Partner Interdependence model (APIM). The APIM allows estimation of the moderation effects from both members (partner and actor) of the dyad on the outcome variable (Maroufizadeh, Hosseini, Rahimi Foroushani, Omani-Samani, & Amini, 2018). The APIM assumes that the outcome variable is a function of two groups of independent variables: one group coming from the actor, and the other coming from the partner. When one member is assessed for the outcome, the other will be present in the analysis and considered as a partner. Finally, the

12

estimated parameters of the APIM helps to determine if the outcome is influenced by the actor only, partner only, or both in a given scenario (Kenny D. A., 1995).

In this study, we proposed a variable selection procedure applied to a probit Bayesian Hierarchical Generalized Linear Model (Bayesian HGLM) to fit the APIM to dyadic data. This method allows the inclusion of interaction effects when the main effects are included in a reduced parameter space to increase its efficiency (Leon-Novelo, Moreno, & Casella, 2012). In addition, we evaluated the performance of the proposed variable selection procedure using simulated data.

Statistical Tools for Modeling Dyads

From a statistical standpoint, dyadic data cannot be considered samples of independent observations. Members are likely to be highly correlated within dyads, so the correlation structure cannot be ignored. For studies using APIM, one of the most common statistical methods used is the analysis of variance (ANOVA) (Gill & Swartz, Bayesian Analysis of Dyadic Data, 2007), especially the two-way ANOVA with random effects that decompose the variance into actor, partner, and relationship components. Other common statistical methods include multi-level modeling, structural equation modeling, and generalized estimation equations (Kenny, Kashy, & Cook, 2006). Bayesian multi-level modeling methods could generate reliable statistical inferences with a consideration of variance components and correlations, even when sample sizes are small (Gill & Swartz, Statistical analyses for round robin interaction data, 2001). Moreover, the Bayesian HGLM

13

has been extended to provide a single unified estimation method for APIM studies for continuous and categorical outcomes (Baragatti, 2011; Ahn, Wang, & Yuan, 2013).

Bayesian Hierarchical Generalized Linear Models (Bayesian HGLM)

The Bayesian HGLM for dyadic data consists of two levels: The first level models the variability *within* a dyad; the second level models the variability *between* dyads. Elements of the first level are called dyad-level independent variables, or dyad-level constructs, and their characteristics can be equally applied to both elements of the dyad. Elements of the second level are called individual-level independent variables, and they are characteristics of each individual. When the primary research question is to identify the set of independent variables or constructs that best explains an outcome, a variable selection method should be applied to the dyadic model. The reason is that most of the time in a predictive model, a variable selection method is needed to yield the simplest model and to avoid collinearity among independent variables that may be performing the same function. Variable selection strategies can be applied to the Bayesian HGLM because it is a type of generalized linear model. In fact, one model selection strategy that can be readily applied to generalized linear models under the Bayesian framework is Stochastic Search Variable Selection (SSVS) (George & McCulloch, 1993; Swartz, Mueller, & Amos, 2006; Ntzoufras, Forster, & Dellaportas, 2000; Stingo, Chen, Tadesse, & Vannucci, 2011; Ahn, Wang, & Yuan, 2013).

Stochastic Search Variable Selection (SSVS)

SSVS was originally developed using linear models (George & McCulloch, 1993), but it has been subsequently extended to most generalized linear models, such as logistic, and

conditional logistic regression models (Swartz, Mueller, & Amos, 2006; Koslovsky, et al., 2018), log-linear models (Ntzoufras, Forster, & Dellaportas, 2000), survival models (Stingo, Chen, Tadesse, & Vannucci, 2011), and longitudinal logistic regression models (Ahn, Wang, & Yuan, 2013). The SSVS framework uses a latent indicator variable to represent the question "Does this variable belong to the model?" and it defines computational machinery to compute an answer to that question in a way that considers all possible models, or at least the most probable models.

Using a spike and slab prior on the regression coefficients, SSVS explores a set of independent variables by limiting the posterior distribution of the predictor coefficients of variables unrelated to the outcome to a small neighborhood around zero. Because of this prior specification, SSVS can be easily implemented via the Gibbs sampler. Then the variable inclusion in the final model depends on each variable's posterior inclusion probability at the end of the stochastic process (Yi, George, & Allison, 2003). However, given that there are not restrictions in the variable inclusion, the final model may include interactions without the presence of their main effects, which produces interpretability problems. To avoid the interpretability problem Chipman, George and McCulloch (2001) developed a version of SSVS which uses priors for interactions that condition on the selection status of the main effects composing that interaction. Using such priors, it is possible to assign probability zero to the models with interactions but without main effects, and ensure interpretability.

15

Although different variable selection methods have been proposed for the Bayesian HGLM (O'Hara & Sillanpaa, 2009), to our knowledge, no variable selection method for Bayesian HGLM applied to the APIM dyads has been proposed. Therefore, the present study seeks to fill this gap by offering a variable selection method for the APIM using dyads and implemented via a Bayesian HGLM framework. Our variable selection method, using the Chipman et. al(2003) version of SSVS that facilitates interpretable interactions, includes an interaction if both main effects are included in the Bayesian HGLM. Furthermore, this method will apply a restriction to the parametric space of the interaction defined by one's research question of interest. This property will help the method to converge when a large number of covariates are included in the model (Leon-Novelo, Moreno, & Casella, 2012).

METHODS

In the APIM, independent variables are correlated (X_A denotes the set of independent variables from the actor, and X_P the set of independent variables from the partner); and the error terms, ε, are allowed to be correlated to control for the sharing variance in the outcomes (e.g., due to elements of the dyad being similar on the predictor variable).

Model Specification

Probit mixed model

Let Y_{ij} denote a binary responses for the jth member ($j = 1,2$) of the ith dyad ($i = 1,2,..n$) with a probability of success equal to p_{ij} that is related to a set of independent variables or constructs through a probit model given by

$$P(Y_{ij} = 1|U,\beta) = \Phi(X_{ij}\beta + U_i), \qquad \text{Equation 1}$$

where β is a vector of coefficients for the fixed effects of dimension p; X_{ij} is a vector of independent variables (it may contain main effects and interaction effects of dimension p); U_i is an independent random effect for each dyad $i=1,...,n$; and Φ is the standard normal cumulative distribution function.

Following Albert and Chib (1993), let $L_{ij} = (L_{11},..,L_{n1},L_{12},..,L_{n2})$ denote the set of $2n$ latent variables whose distribution is given by $L_{ij} \sim N(X_{ij}\beta + U_i, 1)$, and the set is related to the original binary response variable through the relationship:

$$Y_{ij} = \begin{cases} 1 \ if \ L_{ij} > 0 \\ 0 \ if \ L_{ij} < 0 \end{cases}$$

Therefore,

$$P(Y_{ij} = 1|U_i,\beta) = P(L_{ij} > 0|U_i,\beta) = \Phi(X_{ij}\beta + U_i).$$

The latent random variable L can also be expressed as the response in a normal linear regression:

17

$$L_{ij} = X_{ij}\beta + U_i + \varepsilon_{ij},$$

where ε_{ij} are independent residuals $\forall\, i = 1,2,\dots n$ and $j = 1,2$; and U_i is defined as above.

Consider L_{ij} the response variables, and $\boldsymbol{X}_{ij} = (X_{Aij}, X_{Pij})$ a vector of independent

effect that contains information of main effects and interactions. The sub-vector X_{Aij}

corresponds to the set of independent effects recorded from the actor. Similarly, X_{Pij} is the

set of covariates recorded from the partner. The within-dyad correlation is accounted through

the dyad-specific random effect (U_i). Thus, the level 1 model of the APIM can be translated

to a Bayesian HGLM as follows:

$$L_{ij} = \beta_{0i} + X_{Aij}\beta_{Ai} + X_{Pij}\beta_{Pi} + C_{ij}\beta_{Ci} + X_{Aij}C_{ij}\beta_{ACi} + X_{Pij}C_{ij}\beta_{PCi} + \varepsilon_{ij},$$

Equation 2

Where $i = 1,2,\dots n$; $j = 1,2$; C is an indicator variable for a particular member of the dyad.

Specifically here, we use $C = 1$ when $j = 2$; and thus allow the model to incorporate a

different effect of each actor or partner covariate on the outcome for each member of the

dyad. For instance, we will be discussing child-parent dyads, and $j = 2$ denotes the

parent. $C = 1$ allows for a covariate to have a different effect on the parent's outcome,

whether that covariate is an actor covariate or partner covariate. Once we incorporate the

distinguishable indicator, we can rearrange terms in our APIM BHGLM such that the model

can be expressed as follows:

$$L_{ij} = \beta_{0i} + C_{ij}\beta_C + X_{Aij}(\beta_A + C_{ij}\beta_{AC}) + X_{Pij}(\beta_P + C_{ij}\beta_{PC}) + \varepsilon_{ij},$$

Equation 3

18

where $\varepsilon_{ij} \sim N(0,1)$.

Because a dyad can be considered as a cluster of only two elements, there is not enough information to estimate both the random slope and the random intercept. Therefore, we restricted our level 2 models to include only the random intercept. The random intercept can be expressed as a sum of an overall mean $\tilde{\eta}_{000}$ and a normal random effect u_{0i}, if the dyad-level variables and the individual-level variables are either effect-coded or grand mean-centered (Hox & Roberts, 2011). Here we have T independent variables or constructs at the dyad level. Let $\boldsymbol{D} = (D_1, .., D_T)$ be a dyad-level predictor (independent variable):

$$\beta_{0i} = \eta_{00} + D_i\eta_{0D} + u_i$$

$$\beta_{Ci} = \eta_{0C} + D_i\eta_{CD}$$

$$\beta_{Ai} = \eta_A + D_i\eta_{AD}$$

$$\beta_{Pi} = \eta_P + D_i\eta_{PD}$$

$$\beta_{ACi} = \eta_{AC} + D_i\eta_{ACD}$$

$$\beta_{PCi} = \eta_{PC} + D_i\eta_{PCD}$$

Under this new re-parametrization, Equation 3 can be re-expressed as:

$$L_{i1} = \eta_{00} + u_i + D_i\eta_{0D} + X_{Ai1}(\eta_A + D_i\eta_{AD}) + X_{Pi1}(\eta_P + D_i\eta_{PD}) + \varepsilon_{i1}$$

$$L_{i2} = \eta_{00} + \eta_{0C} + u_i + D_i(\eta_{0D} + \eta_{CD}) + X_{Ai2}((\eta_A + \eta_{AC}) + D_i(\eta_{AD} + \eta_{ACD})) + X_{Pi2}((\eta_P +$$

$$\eta_{PC}) + D_i(\eta_{PD} + \eta_{PCD})) + \varepsilon_{i2}, \qquad\qquad\qquad i = 1,2,..,n$$

Equation 4

where u_i, the level 2 random effects, also follows a normal distribution with mean zero and variance σ_u^2. Here, η_A estimates the average effect of the actor independent variables (X_A) on the response Y, and η_{AD} estimates the effect of the interaction of the dyad-level predictor and the independent actor variables DX_A on the response Y. For ease of notation, when we refer to the set of all $\eta's$ as Θ.

Let γ be a vector of indicator variables that determine the subset of η that are more important independent effects to the model, and let M_γ be the Bayesian HGLM based on the γth subset of independent variables. The overall selection was based on the posterior distribution of $P(M_\gamma|Y)$, taking into account the fact that if an interaction term is considered, each individual covariate/construct should be included in the model for interpretability. This process was done using a specific modification of SSVS (Chipman, George, & MCCulloch, 2001). In this study, we focused on performing variable selection on the independent variables or constructs and their interactions with the dyad-level covariate or construct, and the method can be easily extended to select dyad-level covariates or constructs.

Given the complexity of the APIM model parameters, for convenience, we use two vectors of coefficients to define the APIM model: One vector represents the coefficients of the main effects and interactions with the dyad-level covariate associated with the actor denoted by $\boldsymbol{\eta_A} = (\eta_A, \eta_{AC}, \eta_{AD}, \eta_{ACD})$; and the other vector represents the coefficient of the main effects and interaction with the dyad-level covariate associated with the partner denoted by $\boldsymbol{\eta_P} = (\eta_P, \eta_{PC}, \eta_{PD}, \eta_{PCD})$. Each of these vectors, $(\boldsymbol{\eta_A}, \boldsymbol{\eta_P})$, has a dimension d_1 whose

value is $d_1 = 2(Z + (T * Z))$, where $2Z$ is the number of independent variables or constructs from the partner and the actor, and T is the number of independent dyad-level variables or constructs. In general, the total number of interactions $(2 * T * Z)$ exceeds the number of interactions in which a researcher will be interested. Therefore, we use the method of controlled-dimension stochastic search proposed by Leon-Novelo et al (Leon-Novelo, Moreno, & Casella, 2012), to reduce the parameter space of the interactions to make our algorithm more efficient. After the number of interactions of interest is set to a number (q), the dimension of the latent indicator vector $\boldsymbol{\gamma}$ will be limited to $d = 4 * Z + q$.

In this study, we applied the SSVS method for the Bayesian HGLM for the APIM. This method identifies the subset of independent variables that are most important to the model invoking a Gibbs sampler. The SSVS will visit the models with highest posterior probability. In the case of binary outcomes, the posterior distribution of this HGLM is described as

$$P(M_\gamma|Y,\eta) \propto l(\eta,y)\pi(\eta|\gamma)\pi(\gamma),$$

where $l(\eta,y)$ is the likelihood function of the probit model. Specifically,

$$l(L_{ij}, \Theta, \sigma_U^2) \propto \prod_{j=1}^{2}\prod_{i=1}^{n} f(L_{ij}|\Theta,\sigma_y^2) \prod_{\{i,j:Y_{ij}=1\}} I(L_{ij} < 0) \prod_{\{i,j:Y_{ij}=0\}} I(L_{ij} > 0),$$

Equation 5

where $f(L_{ij}|\Theta, \sigma_y^2)$ is the kernel of a normal pdf with variance 1 and mean given by Equation 4, and $\pi(\eta|\gamma)$ and $\pi(\gamma)$ are the coefficients of the HGLM model and inclusion indicator priors.

Prior distributions

To complete the specification of the Bayesian HGLM, we first defined the priors. For ease of notation, we denoted all combination of indexes by ξ; for example, when $\xi = A1$, ξ is the index that corresponds to the coefficient of the first actor predictor. For the coefficients of the regression (η_ξ), we used a continuous spike-lab normal prior distribution, which models the inclusion or exclusion of covariates in the model:

$$\pi(\eta_\xi|\gamma_\xi) = (1 - \gamma_\xi)N(0, \tau_\xi^2) + \gamma_\xi N(0, c_\xi^2\tau_\xi^2). \qquad \text{Equation 6}$$

Here, γ_ξ is a Bernoulli indicator with probability π_ξ. The parameters c_ξ^2 and τ_ξ^2 control the variable selection by concentrating the prior probability on possible values of the coefficient around zero (spike) when the corresponding independent variable is not selected; and by dispersing the variance to distribute the probability over a wider range of possible values (slab) for the coefficient when it is important to the model (George & McCulloch, 1993). Determining how to choose c_ξ^2 and τ_ξ^2 is crucial in the SSVS algorithm, and effective strategies can be found in Swartz, Mueller and Amos (2006). Specifically, in our case we choose $\tau_\xi^2 = .001$ and $c_\xi^2 = 36/0.001$ to force the range of the η_ξ to be between -6 and 6 with a probability of 99% when the covariate is associated and shrunk to zero those coefficients of covariates non-associated.

The set of latent variables can then be divided into two groups: one for independent variables and one for interactions. For interpretability, we identify variable indicators separately for main effects and interactions. Our indicators for main effects are denoted as γ_{wr} and defined as

$$\gamma_{wr} = \begin{cases} 1 \ if \ \eta_{wr} \neq 0 & i.\,e.\,, \text{independent variable is selected} \\ 0 \ if \ \eta_{wr} = 0 & i.\,e.\,, \text{the independent variable is not selected,} \end{cases}$$

where $r = 1,..,Z$ with $w \in \{A, AC, P, PC\}$, and each γ_{wr} was independent and followed a Bernoulli distribution with probability of success $0 < p_{wr} \leq 1$. We assumed p_{wr} was equal across all wr ($p_{wr} = \pi$). The other group of indicators for interaction effects are denoted as $\gamma_{wr'}$, where $w \in \{AD, ACD, PD, PCD\}$ $and \ r' = 1..Z * T$ and they are defined as

$$\gamma_{wr'=}$$

$$\begin{cases} Bernoulli\ (\pi), & if \ \gamma_{w_{\text{element1}}r'} = \gamma_{w_{\text{element2}}r'} = 1 & i.\,e, \text{both main effects in the interaction are selected} \\ 0, & if \ \gamma_{w_{\text{element1}}r'} = 0 \ \ or \ \ \gamma_{w_{\text{element2}}r'} = 0 & i.\,e, \text{at least one main effect in the interaction is not selected} \end{cases}$$

Furthermore, the random intercepts $U = (u_{01},..,u_{0n})$ are independent variables with a normal prior distribution with mean zero and variance σ_U^2

$$u_{0i} \ \overset{iid}{\sim} \ N(0, \sigma_U^2) \quad \text{with } i = 1,..,n \text{ and } \sigma_U^2 \sim IG(a,b).$$

As mentioned previously, we restricted our space of possible models to include interactions when their main effects are present, yet they do not have more than 3 interaction terms.

23

Gibbs Sampler implementing SSVS for APIM

It is known that the SSVS is a Gibbs sampler whose stationary distribution is the posterior distribution of the models; and therefore the SSVS tends to visit the models with the highest posterior probability. To implement a SSVS, we simulate from the following list of full conditionals:

Given initial values for $\gamma^{(0)}, \beta^{(0)}, L_{i1}^{(0)}, L_{i2}^{(0)}, U^{(0)}, \sigma_u^{2(0)}$,

1- Generate $\gamma^{(t+1)}$ from the full conditional

$$P(\gamma|L,U) \propto \left(\frac{1}{|\Gamma||X^t X + \Gamma^{-1}|}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}[(L-U)^t(I - XV^t X^t)(L-U)]\right\} \prod_{j=1}^{P} \pi_j^{\gamma_j}(\pi_j)^{1-\gamma_j}$$

2- Simulate $\eta_\gamma^{(t+1)}$ from the full conditional $f(\eta|L_{i1}^{(t)}, L_{i2}^{(t)}, U^{(t)}, \gamma^{(t+1)})$

3- Simulate $\sigma_u^{2(t+1)}$ from the full conditional

$$f(\sigma_u^2|U) = Inv\ Gamma(a + \frac{n}{2}, b + \frac{\sum_{i=1}^{n} U_i}{2})$$

4- Simulate $L_{ij}^{(t+1)}$ from the full conditional

$L_{ij}|\beta, U, Y_{ij} = 1 \sim N(X_{ij}\eta + U_i, 1)$ left truncated at zero

$L_{ij}|\beta, U, Y_{ij} = 0 \sim N(X_{ij}\eta + U_i, 1)$ right truncated at zero

For $j = 1$, the mean is given by

$$\eta_{00} + u_{ij} + D_i\eta_{0D} + X_{Aij}(\eta_A + D_i\eta_{AD}) + X_{Pij}(\eta_P + D_i\eta_{PD})$$

<div align="right">Equation 7</div>

and for $j = 2$, the mean is given by

$$\eta_{00} + \eta_{0C} + u_{ij} + D_i(\eta_{0D} + \eta_{CD}) + X_{Aij}((\eta_A + \eta_{AC}) + D_i(\eta_{AD} + \eta_{ACD})) +$$

$$X_{Pi}j((\eta_P + \eta_{PC}) + D_i(\eta_{PD} + \eta_{PCD}))$$

<div align="right">Equation 8</div>

5- Simulate $U^{(t+1)}$ from the full conditional $f(U^{(t)}|L_{ij}^{(t+1)}, \eta_\gamma^{(t+1)}, \sigma_u^{2(t+1)})$ for $j = 1,2$

$$f(U|L_{ij}, \eta_\gamma, \sigma_u^2) = N\left(\left(2n + \frac{1}{\sigma_u^2}\right)^{-1} * \left(\frac{w_{i1} + w_{i2}}{\sigma_\varepsilon^2}\right), \left(2n + \frac{1}{\sigma_u^2}\right)^{-1}\right)$$

where

$$w_{i1} = L_{i1} - (\eta_{00} + D_i\eta_{0D} + X_{Ai1}(\eta_A + D_i\eta_{AD}) + X_{Pi1}(\eta_P + D_i\eta_{PD})$$

and

$$w_{i2} = L_{i2} - (\eta_{00} + \eta_{0C} + u_{ij} + D_i(\eta_{0D} + \eta_{CD}) + X_{Ai2}((\eta_A + \eta_{AC}) + D_i(\eta_{AD} + \eta_{ACD})) +$$

$$X_{Pi2}((\eta_P + \eta_{PC}) + D_i(\eta_{PD} + \eta_{PCD}))).$$

We repeated the algorithm until it converged and we burn-in the first 500 iterations to facilitate convergence. Convergence was checked using trace plots, coefficient sample histograms, and different starting values for the covariate coefficients. These values were: using the *lmer* function implemented in *R* (R Core Team, 2017), all zeros and using the same values used to generate the data set. The starting point for the gamma were always one for all the covariate coefficients considered in the model. Lastly, we selected all covariates where $P(\gamma_\xi = 1|data) \geq 50\%$; this method is known as the median model decision rule, which was first described by Barbieri et al. (2002).

Simulations

 To test the performance of the algorithm developed above, we conducted a simulation study. We used R to implement the proposed algorithm and to simulate the data. Our program that was built to test our proposed model used 4 main covariates and 2 dyad-level covariates. These covariates produce a model with 52 terms. The terms were: 4 main effects from the partner and 4 main effects from the actor in L_1 and L_2. In addition to 16 terms that correspond to the interactions between these 16 main effects and the 2 dyad-level covariates. Therefore, the program was tested using simulated data that included 50 terms and 2 intercepts. The simulated data was divided in five scenarios, each of these scenarios was chosen to represent real data. The data were simulated as follows: There are 4 covariates from the actor and 4 covariates from the partner; each of them was simulated from a univariate standard normal distribution. Two dyad-level covariates were simulated from a binomial distribution, with success rate of 0.54 and 0.5, respectively. These values were chosen from commonly observed percentages of dyad concordance for certain characteristics, such as sex. We set the coefficients $\eta \in [0,2]$ and generated a random error from a normal distribution centered at 0 and variance 0.2. We explored five simulation scenarios where the data-generating model (1) was under a null model; (2) included only one main effect, two intercepts and one dyad-level covariate coefficient different than zero; (3) included 2 main effects 1 of the actor and 1 of the partner, two intercepts and two dyad-level effect and one interaction with the dyad-level covariate; (4) included 2 main effects ; and (5) included 2 interactions with no main effects, two intercepts and one dyad-level covariate were included.

26

For each scenario, we simulated 200 simulated datasets (simulation replicates), where each dataset consisted of $n = 1,000$ dyads. To compute the posterior distribution of our model space, we ran the Markov Chain Monte Carlo simulation for 2,000 iterations, with a burn-in of 500 iterations. For each simulated replicate, we calculate the median model given by: the number of times a variable was selected in any model divided by total number of models visited. Furthermore, at the end of the 200 simulation replicates we calculate the percent correct. The percent correct for a given coefficient was defined as the number of correctly selected or correctly not selected divided by the number of models visited.

Discussion

Table 2.1 shows the results for each of the simulated scenarios. The table depicts 10 coefficients that were main effects, 2 intercepts, and 4 interactions that were non-zero in the simulations. These coefficients were randomly selected from all non-zero coefficients from the simulation, and the results were similar when different coefficients were selected (data not shown).

Results for Scenario 1 (null model) showed that the median model was always the null model, since the probability of being correctly selected in the model was 100%. For Scenario 2, the most frequently selected model (across the 200 simulation replicates) was very close to the simulated models, except our method seemed to have trouble identifying one of the interactions $(X_{a12}D_1)$ present in the simulation models. Intercepts and covariates with a non-zero coefficient $(B_0, B_1, D_1, X_{a11}D_1)$ and those with zero coefficients had a 100% probability to being correctly selected equal, except for the interaction $(X_{a12}D_1)$ which was

simulated with a non-zero coefficient, but it only had a 0.19 probability to be correctly

selected. This interaction is part of L_2 (*Equation 4*) and represents the relationship between

an actor covariate and a dyad-level covariate. For Scenario 3 we observe that the most

included terms in the model were very close to the simulated models. The simulated model

included only one main effect in L_1 and L_2 in addition to the dyad level variables and the

intercepts. Results show that the probability to being correctly selected for the covariates

with a non-zero coefficient was 1. However, for those covariates with a zero coefficient, the

probability to be correctly not included was 0.89 in some cases and 1 in others. Similar

results were observed for Scenario 4. Scenario 4 results show that the probability to be

correctly selected or not selected for each term in the model was at least 0.89. In Scenario 5,

the data set included the intercepts and two interactions $(X_{a11}D_1, X_{a12}D_1)$ without main

effects. The results show the probability of correctly not being included of zero and 0.32 for

the main effects related to the interactions (X_{a11}, X_{a12}). Overall, these results indicate that our

variable selection procedure performs well.

Conclusions:

In this study, we implement the first variable selection procedure specifically to

analyze dyadic data based on SSVS technology. This is a model selection procedure in the

space of a Bayesian HGLM for dyadic data from an APIM model. It allows the inclusion of

interaction effects only when the main effects are included in the model. The proposed

variable selection procedure uses the spike and slab prior that allows easier computation than

for example using the g-prior that requires the inversion of $(X'X)$. The implemented program

28

used to apply our variable selection procedure was restricted to include a maximum of 4 interactions ($q = 4$) to ensure the identifiability of the model with n=1,000 dyads but more interactions can be considered with larger sample sizes. When applying our variable selection procedure, the simulation results show that the median model included and excluded covariates with coefficients non zero and zero, respectively, of the generating model, with high probability. The simulation results also show that the median model will include an interaction effect with low probability if the generated data only included one main effect of the two involved in the interaction effect. These results were expected since our model selection algorithm is able to include an interaction only when the two main effects are included.

The proposed variable selection procedure is able to detect if a covariate has an impact on the response but requires more information (i.e., larger sample size) to detect that the effect of this covariate is different in partner and actor. Furthermore, our variable selection procedure may have a problem finding the best model if a set of highly correlated covariates are present in the model since it will violate an assumption of the median model. Finally, our variable selection procedure was tested only with continuous covariates in the model; however, there is no restriction against the inclusion of discrete covariates in the model where our variable selection will be applied. Nevertheless, future work is needed in this direction.

Table 2.1. Average percentage of term inclusion in each simulated scenario.

L			Term		Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
					Coefficient	%correct	Coefficient	%correct	Coefficient	%correct	Coefficient	%correct	Coefficient	%correct
			B_0		0	100	-2	100	-2	100	-2	100	-2	100
			B_1		0	100	2	100	2	100	2	100	6	100
		Dyad	D_1		0	100	4	100	2	100	2	100	0	0
			D_2		0	100	0	100	2	100	2	100	0	100
L1	Main effects	Actor	X_{a11}		0	100	0	100	1.2	100	0	89	0	0
			X_{a21}		0	100	0	100	0	100	1.2	100	0	99.5
		Partner	X_{p11}		0	100	0	100	0	89	0	89.5	0	99.5
			X_{p21}		0	100	0	100	0	100	0	89.5	0	99.5
	Interactions	Dyad* Actor	$X_{a11}D_1$		0	100	1.7	100	0	100	0	95	1.7	82.5
		Dyad * partner	$X_{p21}D_1$		0	100	0	100	0	89	0	94	0	100
L2	Main effects	Actor	X_{a12}		0	100	0	100	0.8	100	0	89	0	31.5
			X_{a22}		0	100	0	100	0	100	0.8	99	0	100
		Partner	X_{p12}		0	100	0	100	0	89	0	89.5	0	99
			X_{p22}		0	100	0	100	0	89	0	89.5	0	99.5
	Interactions	Dyad * Actor	$X_{a12}D_1$		0	100	0.9	19	1.7	100	0	93.5	0.9	2
		Dyad * Partner	$X_{p221}D_1$		0	100	0	0	0	89	0	94	0	100

REFERENCES:

Ahn, L. J., Wang, W., & Yuan, Y. (2013, December). Bayesian Latent-Class Mixed-Effect Hybrid Models for Dyadic Longitudinal Data with Non-Ignorable Dropouts. *Biometrics, 69*(4), 914-24.

Albert, J. H., & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association, 88*(422), 669-679.

Baragatti, M. (2011). Bayesian Variable Selection for Probit Mixed Models Applied to Gene Selection. *Bayesian Analysis, 6*(2), 209-230.

Carlin, B. P., & Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B (Methodological), 57*(3), 473-484.

Ceglio, L., Shegog , R., Markham , C., Dube, S., Song, H., Chaudhary, P., . . . McLaughlin, J. (n.d.). The Secret of Seven Stones: Development of intergenerational online game for middle-school youth to prevent HIV-STI and pregnancy using an Intervention Mapping approach. *In review*.

Chipman, H. (1996, March). Bayesian Variable Selection with Related Predictors. *Canadiand Journal of Statistics, 24*(1), 17-36.

Chipman, H., George, E., & MCCulloch, R. (2001). The practical implementation of Bayesian model selection. (P. LAHIRI, Ed.) *Lecture Notes-Monograph Series, 38*.

D'Cruz, J., Santa Maria, D., Dube, S., Markham, C., McLaughlin, J., Wilkerson, J. M., . . . Shegog, R. (2015). Promoting Parental-Child Sexual Health Dialogue with an Intergenerational Game: Parent and Youth Perspectives. *Games for Health Journal, 4*(2), 113-122.

Gelman, A. (2006). Prior distributions for variance paramenter in hierarchical . *Bayesian Analysis*, 515-533.

George, E. I., & McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *American Statistical Association, 88*(423), 881-889.

Gill, P. S., & Swartz, T. B. (2007). Bayesian Analysis of Dyadic Data. (I. American Sciences Press, Ed.) *American Journal of Mathematical and Management Sciences, 27*, 73-92.

Gill, P. S., & Swartz, T. B. (2009). Statistical analyses for round robin interaction data. (Wiley-Blackwell, Ed.) *Canadian Journal of Statistics, 29*(2), -.

Guangyu , Z., & Ying, Y. (2012, June). Bayesian Modeling Longtudinal Dyadic Data with Nonignorable Dropout, with Application to a Breast Cancer Study. *The Annals of Applied Statistics, 6*(2), 753-771.

Hox, J. J., & Roberts, K. J. (2011). *Handbook of Advanced Multilevel Analysis.* New Yor, NY: Routledge; Taylor & Francis Group.

James, A., & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association, 88*(422), 669-679.

Kenny, D. A. (1995). The effect of nonindependence on significance testing in dyadic research. *Personal Relationships, 2*(1), 65-75.

Kenny, D. L., Kashy, D. A., & Cook, W. L. (2006). *Dyadic Data Analysis.* The Guilford Press.

Koslovsky, M. D., Swartz, M. D., Wenyan, C., Leon-Novelo, L., Wilkinson, A. V., Darla, K. E., & Businelle, M. S. (2018). Using the EM algorithm for Bayesian variable selection in

logistic regression models with related covariates. *Statistical Computing Simulation, 88*(3), 575-596.

Kwon, D., Tadesse, M. G., Sha, N., Pfeiffer, R. M., & Vanucci, M. (2007). Identifying Biomarkers from Mass Spectrometry Data with Ordinal Outcome. *Cancer Informatics, 3*, 19-28.

Latkin, C. A., & Knowlton, A. R. (2015). Social Network Assessments and Interventions for Health Behavior Change: A Critical Review. Behavioral Medicine. *Behavioral Medicine, 41*(3), 90-97.

Leon-Novelo, L., Moreno, E., & Casella, G. (2012). Objective Bayes model selection in probit models. *Statistics in Medicine, 31*(4), 353-65.

Little, Y. Y., & Roderick, J. A. (2009, June). Mixed-Effect Hybrid Models for Longitudinal Data with Nonignorable Dropout,. *Biometrics*(65), 478-486.

Ludtke, O., Kenny, D., & Ulrich, T. (2013). A General and Flexible Approach to Estimate the Social Relations Model Using Bayesian Methods. *American Psychological Association, 18*(1), 101-119.

Newton, M. A. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 155-176.

Ntzoufras, I., Forster, J., & Dellaportas, P. (2000). Stochastic search variable selection for log-linear models. *Statistical Computation and Simulation, 68*, 23-37.

O'Hara, R. B., & Sillanpaa, M. J. (2009). A review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis, 4*(1), 85-118.

Privitera, G. J. (2013). *Research Methods for the Behavioral Sciences* (Second ed.). SAGE.

Schönbrodt, F. D. (2012). TripleR: An R package for social relations analyses based on round-robin designs. *Behavior Research Methods, 44*, 455–470.

Schonbrodt, F. D., Back, M. D., & Schumkle, S. C. (2015). TripleR: Social Relation Model (SRM) analyses for single or multiple groups (R package version 1.4.1).

Sedory, A. (2016). *Let's Talk About Sex: A Dyadic Analysis Using Baseline Data From "The Secret of Seven Stones" Program on Communication Between Parent and Adolescent Youth Aabout Initiation of Sex.* Biostatistics. Houston: UTSPH.

Stakartar, K. K., & Dubson, D. B. (2007, September). Fixed and Random Effects Selection in Linear and Logistic Models. *Biometrics, 63*, 690-698.

Stingo, F. C. (n.d.). *Thesis: Bayesian Methods for Data Integration with Variable Selection: New Challenges in the Analysis of Genomic Data.*

Stingo, F. C., Chen, Y. A., Tadesse, M. G., & Vannucci, M. (2011, September 1). Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. *Ann Appl Stat, 5*(3), 1978-2002.

Swartz, M. D. (2004). Stochastic Search Gene Suggestion: Hierarchical Bayesian Model Selection Meets Gene Mapping. Houston, TX: Rice Univerity.

Swartz, M. D., Cai, Y., Chan, W., Symanski, E., Mitchell, L. E., Danysh, H. E., . . . Lupo, P. J. (2015). Air toxics and birth defects: a Bayesian hierarchical approach to evaluate multiple pollutants and spina bifida. *Environmental Health*, 14:16.

Swartz, M. D., Yu, R. K., & Sanjay, S. (2008). "Finding Factors Influencing Risk: Comparing Variable Selection Methods Applied to Logistic Regression Models of Cases and Controls. *Statistics in medicine*, 6158–6174.

Swartz, M., Mueller, P., & Amos, C. (2006, June). Stochastic Search Gene Suggestion: a Bayesian Hierarchical Model for Gene Mapping. *Biometrics, 62*(2), 495-503.

Tanner, M. A., & Wong , W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association, 82*(398), 528-540.

Won, D., Tadesse, M. G., Sha, N., Pfeiffer, R. M., & Vanucci, M. (2007). Identifying biomarkers from mass spectrometry data with ordinal outcome. *Cancer informatics*.

Yi, N., George, V., & Allison, D. B. (2003, July 1). Stochastic Search Variable Selection for Identifying Multiple Quantitative Trait Loci. *Genetics , 164*(3), 1129-1138.

Yuan, M., Joseph, V. R., & Hui, Z. (2009). Structured Variable Selection and Estimation. *The Annals of Applied Statistics, 3*(4), 1738-1757.

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., & Do, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society Series C Applied Statistics, 63*(4), 595–620.

CHAPTER III

Bayesian Variable Selection for Dyadic Data: An Application to a Parent-Teenager Computer-based Sexual Health Education Program for Middle School Youth

Journal of Statistical Computation and Simulation

ABSTRACT

The analysis of dyadic data with a small data set can be challenging due to the large number of terms in the model. Based on a Bayesian hierarchical generalized linear model we propose a Bayesian variable selection method to analyze dyadic data when there is a small sample. The proposed method takes into account the interdependence between the actor and partner in a restricted parametric space. The proposed statistical technique was evaluated using 4 different scenarios and applied to a real data set.

INTRODUCTION

In behavioral science, many outcomes of interest can be influenced by interpersonal relationships, and observed behaviors are often the result of interactions with more than one person. For example, an adolescent's decision to engage in early sexual initiation can be influenced by interactions with of his/her friends, classmates, siblings, or parents, among others. Therefore, the unit of analysis in behavioral science research is often not an individual, but a group of individuals. Specifically, when only a pair of individuals is involved in an interaction, the pair is called a dyad, which comprises an *actor* and a *partner* (Garcia, Kenny, & Ledermana, 2014). The actor is the person who rates or responds to a stimulus; the partner is a person whose characteristics influence the actor's responses.

There are different dyad-based models (Kenny, Kashy, & Cook, 2006), depending on the research question. One of these models is the actor partner interdependence model (APIM), which is considered the standard model, and it is used when every person in an interpersonal relationship belongs to one and only one dyad. The APIM allows estimation of the effects from both elements of the dyad, the actor and the partner effect, on the outcome variable. In a statistical model, both actor and partner effects are considered independent variables and the outcome is a function of both. Furthermore, the outcome of the APIM can be considered a function of two groups of independent variables: one group that comes from the actor (the same person who generates the outcome), the other group consists of the same independent variables but measured from the partner (the person who influences the actor).

There are different statistical methods available to analyze data from an APIM, i.e., bivariate logistic regression (Busse, Fishbein, Bleakley, & Hennessy, 2010), percent change (Aronowitz, Ogunlade, Nwoso, & Gona, 2015), hierarchical regression model based on possible cluster effects at the classroom level (Looze, Constantine, Jerman, Vermeulen-Smit, & ter Bogt, 2015) and Bayesian hierarchical model (Ahn, Wang, & Yuan, 2013). In most of the cases, these statistical methods require a sample size of more than 100 dyads (Tambling, Johnson, & Johnson, 2011). Consequently, when the sample size is less than 100 dyads, it may be difficult to use any of the traditional methods.

Within the field of behavioral science, dyadic data are generated, for example, when a study includes parents and their children (i.e., parent-child dyad). This population is a focus of public health research because young people (13 to 24 years of age) face health problems due to their risky sexual behaviors. In 2016, young people accounted for 21% of the new HIV cases (Prevention, Morbidity and Mortality Weekly report (MMWR), 2018). Specifically, 488,700

cases of chlamydia, gonorrhea, and syphilis (Kann, et al., 2018) were reported among those aged 13 and 24 years; and 1,688 cases of HIV were reported between among those 13 and 19 (Prevention, HIV surveillance: Adolescents and young adults. Atlanta, GA: National Center for HIV/AIDS, Viral Hepatitis, STD & TB Prevention, Centers for Disease Control and Prevention, 2016). Furthermore, teenagers between 15 and 19 years old gave birth to 209, 890 babies.

The IYG-F program has as a main objective to prevent teen pregnancy and STI/HIV infections by delaying sexual activity in middle school students through the use of video game components as learning tools. Video game components have already been proven to be effective in delaying sexual activity (Shegog, et al., 2014). However, in other studies, it has been reported that communication between friends (Busse, Fishbein, Bleakley, & Hennessy, 2010) or parents and teenagers may trigger the sexual activity (Looze, Constantine, Jerman, Vermeulen-Smit, & ter Bogt, 2015). Other studies have reported that communication between mothers and daughters may play an important role in HIV-prevention behaviors (Aronowitz, Ogunlade, Nwoso, & Gona, 2015).

In this article we evaluated the small sample performance of a novel variable selection procedure using simulated data, and we apply the method to IYG-F baseline data.

METHODS

In the APIM, the set of independent variables from the actor (X_A) and the set of independent variables from the partner (X_P) are correlated. Furthermore, the error terms (ε) are allowed to be correlated to control for the sharing variance in the outcomes. Therefore, to analyze dyads under the APIM framework with a Hierarchical Bayesian Linear model (HGLM), we use a probit mixed model expressed as:

$$L_{i1} = \eta_{00} + u_i + D_i \eta_{0D} + X_{Ai1}(\eta_A + D_i \eta_{AD}) + X_{Pi1}(\eta_P + D_i \eta_{PD}) + \varepsilon_{i1}$$

$$L_{i2} = \eta_{00} + \eta_{0C} + u_i + D_i(\eta_{0D} + \eta_{CD}) + X_{Ai2}((\eta_A + \eta_{AC}) + D_i(\eta_{AD} + \eta_{ACD})) + X_{Pi2}((\eta_P +$$

$$\eta_{PC}) + D_i(\eta_{PD} + \eta_{PCD})) + \varepsilon_{i2}, \qquad\qquad\qquad i = 1,2,..,n$$

Equation 1

where u_i, are the level 2 random effects following a normal distribution with mean zero and

variance σ_u^2; n is the number of dyads, η is a vector of coefficients of dimension p; X_{zij} is a

vector of independent variables (it may contain main effects and interaction effects of dimension

p and $j = 1,2$); Here, η_A estimates the average effect of the actor independent variables (X_A) on

the response Y, and η_{AD} estimates the effect of the interaction of the dyad-level predictor and the

independent actor variables $X_A D$ on the response Y.

Our selection method consist in fitting the HGLM using a Markov Chain Monte Carlo

(MCMC) algorithm (Described in Chapter 2). We iterated the algorithm until it converged and

we burn-in the first k iterations to facilitate convergence. Convergence was checked using trace

plots, coefficient sample histograms, and different starting values for the covariate coefficients.

Initial values were taken from the *lmer* function implemented in *R*, or setting all the parameters

to zero when *lmer* could not be used due a small sample size. The implemented program used to

apply our variable selection procedure was restricted to include a maximum of 2 interactions

($q = 2$) to ensure the identifiability of the model.

Simulations

To test the performance of the algorithm in a small data set, we conducted a simulation

study. We used R (R Core Team, 2017) to implement the proposed algorithm and to simulate the

data. Our program, that was built to test our proposed model, used 4 main covariates and 2 dyad-

level covariates. These covariates produce a model with 52 terms. The terms were: 4 main

effects from the partner and 4 main effects from the actor in L_1 and L_2. In addition to 16 terms

that correspond to the interactions between these 16 main effects and the 2 dyad-level covariates.

Therefore, the program was tested using simulated data that included 50 terms and 2 intercepts.

We generated outcomes under five scenarios, each of which was chosen to represent real

data. The data were simulated as follows: There are 4 covariates from the actor and 4 covariates

from the partner; each of them was simulated from a univariate standard normal distribution.

Two dyad-level covariates were simulated from a binomial distribution, with success rate of 0.54

and 0.5, respectively. These values were chosen from commonly observed percentages of dyad

concordance for certain characteristics, such as sex. We set the coefficients $\eta \in [0,2]$ and

generated a random error from a normal distribution centered at 0 and variance 0.2. We explored

five simulation scenarios where the data-generating model (1) was under a null model; (2)

included only one main effect, two intercepts and one dyad-level covariate coefficient different

than zero; (3) included 2 main effects 1 of the actor and 1 of the partner, two intercepts and two

dyad-level effect and one interaction with the dyad-level covariate; (4) included 2 main effects ;

and (5) included 2 interactions with no main effects, two intercepts and one dyad-level covariate

were included.

For each scenario, we simulated 200 simulated datasets (simulation replicates), where

each dataset consisted of $n = 200$ dyads. To compute the posterior distribution of our model

space, we ran the MCMC simulation for 2,000 iterations, with a burn-in of 500 iterations. For

each simulated replicate, we calculate the marginal probability of inclusion given by: the number

of times a variable was selected in any model divided by total number of models visited. For

selection, we used the median model decision rule: select the model consisting of all variables

with marginal probability of inclusion greater than 50%. Furthermore, at the end of the 200

simulation replicates we calculate the percent correct. That is given by (the number of correctly selected or correctly not selected) divided by the number of data sets. In addition, when applying our variable selection procedure, we included an adjustment that restricted models to include a maximum of 2 interactions ($q = 2$) to ensure the identifiability of the model with $n = 200$ dyads. Furthermore, for only one scenario, we simulate 200 simulation replicates with 61 dyads in each data set. For this scenario specifically, we use the value of the coefficients that were observed from fitting our algorithm to the real data. Therefore, the model for this scenario included 7 main effects and only one dyad-level covariate (59 terms in total). We generated 2,500 and the first 500 iterations were burned-in, the media model decision rule was used and the percent correct was calculated.

Application

We applied the selection method to the baseline data set of the IYG-F program. The main objective was to find the best subset of covariates that best explained the relationship between early sexual initiation and communication in the parent-teenager dyad.

The proposed algorithm was implemented using 6 constructs pre-determined to be individually statistically associated (Sedory, 2016) as predictors with the binary response ("*Have you and your caregiver ever talked about when to start having sex?"*). The 6 constructs were part of the initial set of 7 constructs: *1-Quality of the Communication About Sex, 2-Sex Communication Self-Efficacy Basic, 3-Sex Communication Self-Efficacy Relational, 4-Sex Communication Outcome Expectancy Cognitive, 5-Sex Communication Outcome Expectancy Emotional, 6-Sex Communication Outcome Expectancy Social and 7-Communication Ability*. Each of these constructs was an aggregate score from different questions that are explained in more detail in Appendix 1. The final 6 constructs (2-7) were selected after identifying a high correlation

between construct 1 and construct 3 and consulting an expert regarding which is more of interest. When applying the algorithm, we use 100,000 posterior samples for inference after burning-in 5,000. The initial values were all set to zero since *lmer* could not fit the full model due to our small sample. Our analysis investigated 6 main covariates and 1 dyad-level covariate. These covariates produced a model with 51 terms. The terms were: 6 main effects from the partner and 6 main effects from the actor (L_1 and L_2) in addition to 24 terms that correspond to the interactions between these 24 main effects and 1 dyad-level covariate. Therefore, the algorithm used 49 terms and 2 intercepts. As in the simulation study, we selected the median model as our final model.

Results

Simulations

Table 3.1 displays the results for each of 5 the simulation scenarios. The results shown in Table 3.1 belong to the coefficients of 10 main effects, 2 intercepts, and 4 interactions that were non-zero in at least one of the simulations. These coefficients were randomly selected to be set different from zero. The results were similar when different coefficients were selected (not shown). Results for Scenario 1 (null model) showed that the probability of selecting the correct model was 100%, based on the median model. For Scenario 2, the most frequently selected model, across the 200 simulation replicates, was very close to the simulated models. Results showed that the probability to be correctly selected for the term X_{a22} was lower (78.9%) compared to the probability to be correctly selected for the main effect X_{a21}(100%). For Scenario 3 results showed a low probability for the term X_{a12} to be correctly selected (49.75%) compared with the interaction term $X_{a12}D_1$ (65.83%) and the main effect X_{a11} (100%). When generating the data for Scenario 4, the main effect X_{a11} had a coefficient equal to 0, but it only

had a 2.56% probability to be correctly non-selected. Similar results were observed for the

interaction term $X_{a12}D_1$, that was given a non-zero coefficient to generate the data, but it only

had a 6.41% probability to be correctly selected. This interaction is part of L_2 (*Equation 1*) and

represents the relationship between an actor covariate and a dyad-level covariate. However, the

interaction term $X_{a11}D_1$ has 88.46% probability to be correctly selected. Overall, these results

indicate that our variable selection procedure performs well in a small data set.

Table 3.2 displays the results of 1 simulation scenario. The results shown in Table 3.2

includes the results of 5 of the coefficients of the main effects and the coefficient of one

interaction that were simulated with value different than zero from a total of 59. Results showed

that based on the median model only 2 main effects were included in the model (X_{a11}, X_{a31}).

Real data

Table 3.3 and Table 3.4 display the results from analyzing the baseline data from the

IYG-F study. Using the median model criterion, results show only the main effect of the teenager

construct *Sex Communication Self-Efficacy Relational -teenager* (X_{a11}) was included in the

model (inclusion probability 89.71%). None of the interactions with the dyad-level covariate

were included in the model because their inclusion probability was less than 50%. Table 3.3

displays the posterior mean conditional on inclusion (γ=1) and the central 95% credible interval

for the construct *Sex Communication Self-Efficacy Relational-teenager* (conditional posterior

mean: 0.31; central 95% credible interval (CI): (0.01, 0.52). Furthermore, Table 3.3 also shows

the conditional posterior mean and the central 95%CI for 2 interaction terms with high posterior

inclusion probability (greater than 25%) that did not meet the 50% (interaction: *Gender & Sex*

Communication Self-Efficacy Relational -teenager and the interaction: *Gender & Sex*

Communication Self-Efficacy Relational –parent).

The results in Table 3.3 can be interpreted as: the construct of the teenager *Sex*

Communication Self-Efficacy Relational explains the outcome *Sexual initiation,* and the effects

are equal across dyad members (teenager-parent). Furthermore, the positive sign of the

coefficient of *Sex Communication Self-Efficacy Relational-teenager (*conditional posterior mean:

0.31; Central 95% CI: (0.01, 0.52)) means that a unit increase in its score will increase the

probability to have a positive response to *Have you and your caregiver ever talked about when to*

start having sex?.

Discussion

Scenarios 3 and 4 (Table 3.1) showed that the median model will include an interaction

effect with low probability if the generated data only included one main effect of the two

involved in the interaction effect. And that probability will be even lower if the term is the one

that represents the difference between the teenager acting as an actor and the parent acting as an

actor (L_2 $term$). Similar results in Table 3.2 were observed for the L_2 $term$s with coefficient

different than zero. These results were expected since our sample size is small. In addition, we

implemented a simulation study generating responses from models using the coefficient values

estimated from the IYG-F data, simulating $n = 61$ dyads and 200 replicates. Our findings were

similar as the results presented from the simulation study with 200 dyads: the terms included

were those in the response generating model with a non-zero coefficient (details not shown).

The application results showed that only one main effect was included in the model.

However, two other terms had a posterior probability substantially higher than all other excluded

terms, yet did not meet the median model inclusion criterion of 50%

$(X_{a12}D_1: 29.88\%$ and $X_{p12}D_1: 33.21\%)$. These interaction terms are formed by the dyad-level covariate Gender concordance and the construct *Sex Communication Self-Efficacy Relational* measure in the teenager and the parent $(X_{a12}D_1: teenager, \ X_{p12}D_1: parent)$, and could be worth exploring in future research using other data sets.

Collectively, our findings suggest that the proposed algorithm can adequately handle small data sets of dyadic data and that sexual communication, especially *self-efficacy relational*, is important for delaying sexual initiation.

Table 3.1. Simulation results: Average percentage of term inclusion in each simulated scenario, using 200 dyads.

			Term	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
				Coefficient	%correct	Coefficient	%correct	Coefficient	%correct	Coefficient	%correct	Coefficient	%correct
			η_{00}	0	100	-2	100	-2	100	-2	100	-2	100
			η_{0C}	0	100	2	100	2	100	2	100	6	100
		Dyad	D_1	0	100	2	100	4	100	4	100	0	100
			D_2	0	100	2	100	2	100	0	100	0	100
L1	Main effects	Actor	X_{a11}	0	100	0	100	1.2	100	2.56	100	1.2	100
			X_{a21}	0	100	1.2	100	0	100	0	100	0	0.83
		Partner	X_{p11}	0	100	0	100	0	100	0	100	0	0.82
			X_{p21}	0	100	0	100	0	100	0	100	0	0.84
	Interactions	Dyad* Actor	$X_{a11}D_1$	0	100	0	98.99	0	100	1.7	88.46	1.7	0.95
		Dyad * partner	$X_{p21}D_1$	0	100	0	100	0	100	0	100	0	0.94
L2	Main effects	Actor	X_{a12}	0	100	0	98.99	0.8	49.75	0	93.58	0.8	0.64
			X_{a22}	0	100	0.8	74.87	0	100	0	100	0	0.82
		Partner	X_{p12}	0	100	0	99.49	0	99.50	0	100	0	0.82
			X_{p22}	0	100	0	100	0	100	0	100	0	0.81
	Interactions	Dyad * Actor	$X_{a12}D_1$	0	100	0	98.99	1.7	65.83	0.9	6.41	0.9	0.31
		Dyad * Partner	$X_{p21}D_1$	0	100	0	100	0	100	0	100	0	0.93

Table 3.2. Simulation results: Average percentage of term inclusion for 61 dyads

				Scenario 1	
			Term	Coefficient	%correct
		Intercept	η_{00}	-2	100
			η_{0C}	2	100
		Dyad	D_1	1	100
L1	Main effects	Actor	X_{a11}	1.5	100
			X_{a31}	0.8	100
		Partner	X_{p11}	0	100
			X_{p31}	0.4	0.03
	Interactions	Dyad* Actor	$X_{a31}D_1$	2	0.32
		Dyad * partner	$X_{p21}D_1$	0	100
L2	Main effects	Actor	X_{a22}	1.8	0.27
			X_{a32}	1.2	0
		Partner	X_{p12}	0	100
			X_{p22}	0	100
	Interactions	Dyad * Actor	$X_{a12}D_1$	0	100
		Dyad * Partner	$X_{p21}D_1$	0	100

Table 3.3. Application to IYG-F baseline data: Probability of term inclusion for each construct included in the model.

			Term	Inclusion probability	Conditional posterior mean (central 95% credible interval)				Term	Inclusion probability	Conditional posterior mean (central 95% credible interval)
		Intercept	η_{00}	100.00%							
		Dyad level	D_1: Gender Concordance	100.00%							
L1	Main effects	Teenager	X_{a11}: sex communication self-efficacy basic	0.89%		L1	Interactions	Teenager	X_{a12}	0.94%	
			X_{a12}: sex communication self-efficacy relational	**89.71%**	**0.31 (0.01, 0.52)**				$X_{a12}D_1$	29.88%	0.02 (-0.01, 0.71)
			X_{a13}: sex communication outcome expectancy cognitive	6.02%					$X_{a13}D_1$	9.26%	
			X_{a14}: sex communication outcome expectancy emotional	8.15%					$X_{a14}D_1$	3.70%	
			X_{a15}: sex communication outcome expectancy social	0.63%					$X_{a15}D_1$	2.35%	
			X_{a16}: communication ability	0.14%					$X_{a16}D_1$	1.51%	
		Parent	X_{p11}: sex communication self-efficacy basic	0.82%				Parent	$X_{p11}D_1$	1.98%	
			X_{p12}: sex communication self-efficacy relational	8.73%					$X_{p12}D_1$	33.21%	0.002 (0.01, 0.52)
			X_{p13}: sex communication outcome expectancy cognitive	0.96%					$X_{p13}D_1$	1.22%	
			X_{p14}: sex communication outcome expectancy emotional	1.92%					$X_{p14}D_1$	1.07%	
			X_{p15}: sex communication outcome expectancy social	2.09%					$X_{p15}D_1$	2.93%	
			X_{p16}: sex communication self-efficacy basic	0.41%					$X_{p16}D_1$	0.31%	

Table 3.4. Application to IYG-F baseline data: Probability of term inclusion for each construct included in the model.

		Term	Inclusion probability				Term	Inclusion probability
	Intercept	η_{0C}	100.00%					
L2	Main effects	Parent	X_{a21}:sex communication self-efficacy basic	0.01%	L2	Interactions	Parent	$X_{a22}D_1$	0.00%

		Term	Inclusion probability				Term	Inclusion probability
	Intercept	η_{0C}	100.00%					
		X_{a21}:sex communication self-efficacy basic	0.01%				$X_{a22}D_1$	0.00%
		X_{a22}:sex communication self-efficacy relational	7.75%				$X_{a22}D_1$	0.18%
	Parent	X_{a23}:sex communication outcome expectancy cognitive	0.11%			Parent	$X_{a23}D_1$	0.10%
		X_{a24}:sex communication outcome expectancy emotional	0.11%				$X_{a24}D_1$	0.01%
		X_{a25}:sex communication outcome expectancy social	0.02%				$X_{a25}D_1$	0.00%
L2	Main effects	X_{a26}:communication ability	0.00%	L2	Interactions		$X_{a26}D_1$	0.00%
		X_{p21}:sex communication self-efficacy basic	0.00%				$X_{p21}D_1$	0.00%
		X_{p22}:sex communication self-efficacy relational	0.77%				$X_{p22}D_1$	0.20%
	Teenager	X_{p23}:sex communication outcome expectancy cognitive	0.00%			Teenager	$X_{p23}D_1$	0.00%
		X_{p24}:sex communication outcome expectancy emotional	0.01%				$X_{p24}D_1$	0.02%
		X_{p25}:sex communication outcome expectancy social	0.06%				$X_{p25}D_1$	0.02%
		X_{p26}:sex communication self-efficacy basic	0.00%				$X_{p26}D_1$	0.00%

47

REFERENCES

Ahn, J., Liu, S., Wang, W., & Yuan, Y. (2013, December). Bayesian Latent-Class Mixed-Effect Hybrid Models for Dyadic Longitudinal Data with Non-ignorable Dropouts. *Biometrics, 69*, 914-924.

Aronowitz, T. P., Ogunlade, I. J.-B.-B., Nwoso, C. M.-B., & Gona, P. N. (2015). Sexual communication intervention for African American mothers & daughters. *Applied Nursing Research, 28*, 229-234.

Busse, P., Fishbein, M., Bleakley, A., & Hennessy, M. (2010). The Role of Communication with Friends in Sexual Initiation. *Communication Research, 37*(2), 239-255.

Garcia, R. L., Kenny, D. A., & Ledermana, T. (2014). Moderation in the actor-partner inderdependency model. *Personal relationships, 22*(1), 8-29.

Kann, L., MacManus, T., Harris, W., Shanklin, S., Flint, K. H., Queen, B., . . . Ethier, K. A. (2018). Youth Risk Behavior Surveillance United States 2017. *Morbidity and Mortality Weekly Report Surveillance Summaries 2018, 67*(SS-8), 1-114.

Kenny, D. L., Kashy, D. A., & Cook, W. L. (2006). *Dyadic Data Analysis.* The Guilford Press.

Looze, M., Constantine, N. A., Jerman, P., Vermeulen-Smit, E., & ter Bogt, T. (2015). Parent-Adolescent Sexual Communicaton and Its association With Adolescent Sexual Behaviors: A Nationally Representative Analysis in the Netherlands. *Sex Research, 52*(3), 257-268.

Prevention, C. f. (2016). *HIV surveillance: Adolescents and young adults. Atlanta, GA: National Center for HIV/AIDS, Viral Hepatitis, STD & TB Prevention, Centers for Disease Control and Prevention*. Retrieved March 26, 2019, from https://www.cdc.gov/hiv/pdf/library/slidesets/cdc-hiv-surveillance-adolescents-young-adults-2016.pdf

Prevention, C. f. (2018, June 15). *Morbidity and Mortality Weekly report (MMWR)*. Retrieved 26 2019, 3, from https://www.cdc.gov/mmwr/volumes/67/ss/ss6708a1.htm#suggestedcitation

Sedory, A. C. (2016). LET'S TALK ABOUT SEX: A DYADIC ANALYSIS USING BASELINE DATA FROM "THE SECRET OF SEVEN STONES" PROGRAM ON COMMUNICATION BETWEEN PARENT AND ADOLESCENT YOUTH ABOUT INITIATION OF SEX. *Dissertation*.

Shegog, R., Markham, C. M., Peskin, M. F., Johnson, K., Cuccaro, P., & Tortolero, S. R. (2014). It's Your Game…Keep It Real: Can innovative public health prevention research thrive within a comparative effectiveness research framework? *Primary Prevention, 34*(0), 89-108.

Tambling, R. B., Johnson, S. K., & Johnson, L. N. (2011). Analyzing Dyadic Data From Small Samples: A Pooled Regression Actor-Partner Interdependence model Approach. 101-114.

CHAPTER IV

Conclusion

In this work we present a stochastic search variable selection framework applied to a probit model with random effects as a Bayesian approach for dyadic data under an APIM model. We introduced a two dimensional selection indicator to facilitate modeling both actor and partner effects. In order to make the model more identifiable in small samples, we incorporated three key concepts. First, an interaction will be selected if both main effects were selected. Second, the parametric space for the coefficients of the interactions is restricted to a number less than the total number possible of interaction in the model. Third, the use of a Spike and Slab prior for the selection variable allows an easier computation than other priors. These three properties provide an alternative selection method for dyadic data with a binary outcome, especially when classical methods have difficulty making inference due to a small sample size. To accomplish these 3 properties, the method brings together statistical theory already tested and published.

The method stochastic search variable selection applied to a Bayesian hierarchical generalized linear model for dyadic data was assessed in different scenarios. These scenarios used simulated data and real data with different sample sizes. The scenarios included different values assigned to the coefficients of the terms in the model to generate the response variable. We observe that if the term in the model was generated with a small coefficient (<0.2) and it was a differential factor (differentiating a dyad member as a partner relative to as the actor), the method will have difficulty to include it in the model when the

sample was small. However, for bigger samples this inclusion problem was not observed. Furthermore, each of the covariates included in the model was generated using standard normal distribution. Even though there are no theoretical limitations for using categorical predictors, we did not directly simulate categorical predictors.

The covariates included to generate the data to test the model were independent, and the proposed selection method worked without a problem. When we applied our method to the real data, however, we observed that our proposed selection method included a different set of covariates conditional on the number of iterations used. This was due to having highly correlated covariates in the set of possible variables, a known challenge for SSVS based methods (George & McCulloch, 1993). A practical approach to this situation is to consult content experts to identify the most important factors among those correlated, to reduce the correlated variables in the search. This will help to have a more consistent decisions regarding the final model which are independent of the number of iterations used. To our knowledge there is no selection method applicable to the APIM model. Therefore we were not able to compare our selection method with other except with the simulations performed.

This work can be extended in different ways. One immediate extension would be to restrict the parametric space of main effects only to help the use of the method in small samples. A second possibility would be to determine the minimum sample size required for covariate inclusion, especially the differentiable coefficients, in the model. A third possibility will be to use the expected maximization variable selection method instead of SSVS. Finally, this method can be extended to select elements within in constructs, when constructs consists of multiple elements per score.

50

APPENDICES

Appendix A: Constructions in the IYG-F used in this study.

Domain	Construct	Actor (child) Items	Parent (or main caregiver) item(s)	Scale response
outcomes of study	Communication about sex	*Have you and your caregiver ever talked about when to start having sex?*	*Have you and your child ever talked about when to start having sex?*	Yes; No; Refuse to Answer (0,1)
Constructs of interest	Quality of parental communication about sex	*I don't know enough about sexual topics like these to talk to my child.*	*My caregiver doesn't know enough about sexual topics like this to talk to me*	N=5, 5-point (Strongly disagree to strongly disagree)
		I want to know my child's questions about these sexual topics.	*My caregiver wants to know my questions about sexual topics like this*	
		I try to understand how my child feels about sexual topics like these.	*My caregiver tries to understand how I feel about sexual topics like this*	
		When I talk to my child about these sexual topics, I warn or threaten them about the consequences.	*When my caregiver talks to me about sexual topics, they warn or threaten me about the consequences*	
		I know how to talk to my child about sexual topics like these.	*My caregiver knows how to talk to me about sexual topics like this*	
		My child can ask me the questions they really want to know about sexual topics like these.	*I can ask my caregiver the questions I really want to know about sexual topics like this*	
		My child and I talk openly and freely about these sexual topics.	*My caregiver and I talk openly and freely about these sexual topics*	
		I tell my child things about these sexual topics that they already know.	*My caregiver tells me things about these sexual topics that I already know*	
		If my child talked to me about these sexual topics, I would think they are doing these things.	*If I talked to my caregiver about these sexual topics, they would think I'm doing these things*	

Domain	Construct	Actor (child) Items	Parent (or main caregiver) item(s)	Scale response
		I don't talk to my child about these sexual topics, I lecture my child.	*My caregiver doesn't talk to me about these sexual topics, they lecture me*	
	Sex communication, self-efficacy, basic	*How to tell if a boy or girl really loves you*	*You can explain to your child how to tell if a boy or girl really loves them.*	
		Why you need to wait until you're older before you have sex (e.g. vaginal or oral sex)	*You can explain to your child why they need to wait until they are older before they have sex.*	
		How to make a boy or girl wait until you are ready to have sex	*You can explain to your child how to make a boy or girl wait until they are ready to have sex.*	
		How to tell a boy or girl "no" if you do not want to have sex	*You can explain to your child how to tell a boy or girl if they do not want to have sex.*	
		Ways to have fun with a boy or girl without having sex (e.g. vaginal or oral sex)	*You can explain to your child ways to have fun with a boy or girl without having sex.*	
	Sex communication, self-efficacy, relational	*How sure are you that you can talk to your caregiver about:* *How to use birth control*	*You can explain to your child how to use birth control.*	N=16; 7 point (Not sure at all to completely sure)
		Where to buy or get birth control	*You can explain to your child where to buy or get birth control.*	
		How birth control keeps girls from getting pregnant	*You can explain to your child how birth control keeps girls from getting pregnant.*	
		Where to buy or get condoms	*You can explain to your child where to buy or get condoms.*	
		How to put on a condom	*You can explain to your child how to put on a condom.*	
		Why an unmarried person should use a condom when they have sex	*You can explain to your child why an unmarried person should use a condom when they have sex.*	
		Using a condom if you decide to have sex	*You can explain to your child that they should*	

Domain	Construct	Actor (child) Items	Parent (or main caregiver) item(s)	Scale response
			use a condom if they decide to have sex.	
		What is happening when a girl has her period	*You can explain to your child what is happening when a girl has her period.*	
		Why wet dreams occur	*You can explain to your child why wet dreams occur.*	
		How someone can get HIV/AIDS if they don't use a condom	*You can explain to your child how someone can get HIV/AIDS if they don't use a condom.*	
		What you think about a teen your age having sex	*You can explain to your child what you think about adolescents their age having sex.*	
	Sex communication outcome expectancy-emotional	*You will feel you did the right thing*	*If you talk with your child about sexual topics, you will feel that you did the right thing.*	
		You will be proud	*If you talk with your child about sexual topics, you will be proud.*	
		You will be embarrassed	*If you talk with your child about sexual topics, you will be embarrassed.*	
		You will feel comfortable	*If you talk with your child about sexual topics, you will feel comfortable.*	
		You will find some things difficult to talk about	*If you talk with your child about sexual topics, you will find some things difficult to talk about.*	
		It will be unpleasant	*If you talk with your child about sexual topics, it will be unpleasant.*	
		You will feel ashamed	*If you talk with your child about sexual topics, you will feel ashamed.*	

Domain	Construct	Actor (child) Items	Parent (or main caregiver) item(s)	Scale response
	Sex communication outcome expectancy-social	*You will be less likely to get pregnant or get a girl pregnant*	*If you talk with your child about sexual topics, your child will be less likely to get pregnant or get a girl pregnant.*	
		You will be less likely to have sex (e.g. vaginal or oral sex) as a young teen	*If you talk with your child about sexual topics, your child will be less likely to have sex as a young teen.*	
		You think it will do some good	*If you talk with your child about sexual topics, you think it will do some good.*	
		You will feel relieved	*If you talk with your child about sexual topics, you will feel relieved.*	
		You will do what you want no matter what they say	*If you talk with your child about sexual topics, your child will do what they want no matter what you say.*	
		You will be less likely to get pregnant or get a girl pregnant	*If you talk with your child about sexual topics, your child will be less likely to get pregnant or get a girl pregnant.*	
	Communication ability	*How would you rate your ability to communicate with your caregiver about sexual topics? Remember, sexual topics refer to issues related to when to start having sex, birth control, condoms, AIDS/HIV, pregnancy, physical/sexual development, sexually transmitted diseases (STDs), and peer pressure about sex*	*How would you rate your ability to communicate with your child about sexual topics?*	Terrible; Very Poor; Poor; Fair; Good; Very Good; Excellent; Refuse to Answer

REFERENCES

Ahn, J., Liu, S., Wang, W., & Yuan, Y. (2013, December). Bayesian Latent-Class Mixed-Effect Hybrid Models for Dyadic Longitudinal Data with Non-ignorable Dropouts. *Biometrics, 69*, 914-924.

Albert, J. H., & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association, 88*(422), 669-679.

Aronowitz, T. P., Ogunlade, I. J.-B.-B., Nwoso, C. M.-B., & Gona, P. N. (2015). Sexual communication intervention for African American mothers & daughters. *Applied Nursing Research, 28*, 229-234.

Baragatti, M. (2011). Bayesian Variable Selection for Probit Mixed Models Applied to Gene Selection. *Bayesian Analysis, 6*(2), 209-230.

Busse, P., Fishbein, M., Bleakley, A., & Hennessy, M. (2010). The Role of Communication with Friends in Sexual Initiation. *Communication Research, 37*(2), 239-255.

Carlin, B. P., & Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B (Methodological), 57*(3), 473-484.

Ceglio, L. (2015). *The Secret of Seven Stones: Development of intergenerational online game for middle-school youth to prevent HIV-STI and pregnancy using an Intervention Mapping approach.* Thesis for master of public health. Houston: University of Texas School of Public Health.

Ceglio, L., Shegog , R., Markham , C., Dube, S., Song, H., Chaudhary, P., . . . McLaughlin, J. (n.d.). The Secret of Seven Stones: Development of intergenerational online game for middle-school youth to prevent HIV-STI and pregnancy using an Intervention Mapping approach. *In review*.

Chipman, H. (1996, March). Bayesian Variable Selection with Related Predictors. *Canadiand Journal of Statistics, 24*(1), 17-36.

Chipman, H., George, E., & MCCulloch, R. (2001). The practical implementation of Bayesian model selection. (P. LAHIRI, Ed.) *Lecture Notes-Monograph Series, 38*.

D'Cruz, J., Santa Maria, D., Dube, S., Markham, C., McLaughlin, J., Wilkerson, J. M., . . . Shegog, R. (2015). Promoting Parental-Child Sexual Health Dialogue with an Intergenerational Game: Parent and Youth Perspectives. *Games for Health Journal, 4*(2), 113-122.

Garcia, R. L., Kenny, D. A., & Ledermana, T. (2014). Moderation in the actor-partner inderdependency model. *Personal relationships, 22*(1), 8-29.

Gelman, A. (2006). Prior distributions for variance paramenter in hierarchical . *Bayesian Analysis*, 515-533.

George, E. I., & McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *American Statistical Association, 88*(423), 881-889.

Gill , P., & Swartz, T. (2007). Bayesian Analysis of Dyadic Data. *American Journal of Mathematical and Management Sciences, 27*, 73-92.

Gill, P. S., & Swartz, T. B. (2001, June). Statistical Analyses for Round Robin Interaction Data. (Wiley-Blackwell, Ed.) *The Canadian Journal of Statistics, 29*(2), 321-331.

Guangyu , Z., & Ying, Y. (2012, June). Bayesian Modeling Longtudinal Dyadic Data with Nonignorable Dropout, with Application to a Breast Cancer Study. *The Annals of Applied Statistics, 6*(2), 753-771.

Gyarmathy, V., & Neaigus, A. (2007). The relationship of sexual dyad and personal network characteristics and individual attributes to unprotected sex among young injecting drug users. *AIDS and behavior, 13*(2), 196-206.

Hox, J. J., & Roberts, K. J. (2011). *Handbook of Advanced Multilevel Analysis.* New Yor, NY: Routledge; Taylor & Francis Group.

James, A., & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association, 88*(422), 669-679.

Kann, L., MacManus, T., Harris, W., Shanklin, S., Flint, K. H., Queen, B., . . . Ethier, K. A. (2018). Youth Risk Behavior Surveillance United States 2017. *Morbidity and Mortality Weekly Report Surveillance Summaries 2018, 67*(SS-8), 1-114.

Kenny, D. A. (1995). The effect of nonindependence on significance testing in dyadic research. *Personal Relationships, 2*(1), 65-75.

Kenny, D. L., Kashy, D. A., & Cook, W. L. (2006). *Dyadic Data Analysis.* The Guilford Press.

Koslovsky, M. D., Swartz, M. D., Wenyan, C., Leon-Novelo, L., Wilkinson, A. V., Darla, K. E., & Businelle, M. S. (2018). Using the EM algorithm for Bayesian variable selection in logistic regression models with related covariates. *Statistical Computing Simulation, 88*(3), 575-596.

Kwon, D., Tadesse, M. G., Sha, N., Pfeiffer, R. M., & Vanucci, M. (2007). Identifying Biomarkers from Mass Spectrometry Data with Ordinal Outcome. *Cancer Informatics, 3*, 19-28.

Latkin, C. A., & Knowlton, A. R. (2015). Social Network Assessments and Interventions for Health Behavior Change: A Critical Review. Behavioral Medicine. *Behavioral Medicine, 41*(3), 90-97.

Leon-Novelo, L., Moreno, E., & Casella, G. (2012). Objective Bayes model selection in probit models. *Statistics in Medicine, 31*(4), 353-65.

Little, Y. Y., & Roderick, J. A. (2009, June). Mixed-Effect Hybrid Models for Longitudinal Data with Nonignorable Dropout,. *Biometrics*(65), 478-486.

Looze, M., Constantine, N. A., Jerman, P., Vermeulen-Smit, E., & ter Bogt, T. (2015). Parent-Adolescent Sexual Communicaton and Its association With Adolescent Sexual Behaviors: A Nationally Representative Analysis in the Netherlands. *Sex Research, 52*(3), 257-268.

Ludtke, O., Kenny, D., & Ulrich, T. (2013). A General and Flexible Approach to Estimate the Social Relations Model Using Bayesian Methods. *American Psychological Association, 18*(1), 101-119.

Maroufizadeh, S., Hosseini, M., Rahimi Foroushani, A., Omani-Samani, R., & Amini, P. (2018). The relationship between marital satisfaction and depression in infertile

couples: an actor–partner interdependence model approach. *BMC Psychiatry, 18*(1), 310.

Mustanski, B., Starks, T., & Newcomb, M. E. (2015). Methods for the Design and Analysis of Relationship and Partner Effects on Sexual Health. *Mustanski, Brian et al. "Methods for the design and analysis of relationship and partner effects on sexual health." Archives of sexual behavior vol. 43,1 (2014): 21-33. doi:10.1007/s10508-013-0215-9, 43*(1), 21-23.

Newton, M. A. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 155-176.

Ntzoufras, I., Forster, J., & Dellaportas, P. (2000). Stochastic search variable selection for log-linear models. *Statistical Computation and Simulation, 68*, 23-37.

O'Hara, R. B., & Sillanpaa, M. J. (2009). A review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis, 4*(1), 85-118.

Prevention, C. f. (2016). *HIV surveillance: Adolescents and young adults. Atlanta, GA: National Center for HIV/AIDS, Viral Hepatitis, STD & TB Prevention, Centers for Disease Control and Prevention*. Retrieved March 26, 2019, from https://www.cdc.gov/hiv/pdf/library/slidesets/cdc-hiv-surveillance-adolescents-young-adults-2016.pdf

Prevention, C. f. (2018, June 15). *Morbidity and Mortality Weekly report (MMWR)*. Retrieved 26 2019, 3, from https://www.cdc.gov/mmwr/volumes/67/ss/ss6708a1.htm#suggestedcitation

Privitera, G. J. (2013). *Research Methods for the Behavioral Sciences* (Second ed.). SAGE.

R Core Team. (2017). R: A Language and environmental for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. Retrieved from https://www.R-project.org/: https://www.R-project.org

Schönbrodt, F. D. (2012). TripleR: An R package for social relations analyses based on round-robin designs. *Behavior Research Methods, 44*, 455–470.

Schonbrodt, F. D., Back, M. D., & Schumkle, S. C. (2015). TripleR: Social Relation Model (SRM) analyses for single or multiple groups (R package version 1.4.1).

Sedory, A. C. (2016). LET'S TALK ABOUT SEX: A DYADIC ANALYSIS USING BASELINE DATA FROM "THE SECRET OF SEVEN STONES" PROGRAM ON COMMUNICATION BETWEEN PARENT AND ADOLESCENT YOUTH ABOUT INITIATION OF SEX. *Dissertation*.

Shegog, R., Markham, C. M., Peskin, M. F., Johnson, K., Cuccaro, P., & Tortolero, S. R. (2014). It's Your Game…Keep It Real: Can innovative public health prevention research thrive within a comparative effectiveness research framework? *Primary Prevention, 34*(0), 89-108.

Stakartar, K. K., & Dubson, D. B. (2007, September). Fixed and Random Effects Selection in Linear and Logistic Models. *Biometrics, 63*, 690-698.

Stingo, F. C. (n.d.). *Thesis: Bayesian Methods for Data Integration with Variable Selection: New Challenges in the Analysis of Genomic Data*.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., & Vannucci, M. (2011, September 1). Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. *Ann Appl Stat, 5*(3), 1978-2002.

Swartz, M. D. (2004). Stochastic Search Gene Suggestion: Hierarchical Bayesian Model Selection Meets Gene Mapping. Houston, TX: Rice Univerity.

Swartz, M. D., Cai, Y., Chan, W., Symanski, E., Mitchell, L. E., Danysh, H. E., . . . Lupo, P. J. (2015). Air toxics and birth defects: a Bayesian hierarchical approach to evaluate multiple pollutants and spina bifida. *Environmental Health*, 14:16.

Swartz, M. D., Yu, R. K., & Sanjay, S. (2008). "Finding Factors Influencing Risk: Comparing Variable Selection Methods Applied to Logistic Regression Models of Cases and Controls. *Statistics in medicine*, 6158–6174.

Swartz, M., Mueller, P., & Amos, C. (2006, June). Stochastic Search Gene Suggestion: a Bayesian Hierarchical Model for Gene Mapping. *Biometrics, 62*(2), 495-503.

Tambling, R. B., Johnson, S. K., & Johnson, L. N. (2011). Analyzing Dyadic Data From Small Samples: A Pooled Regression Actor-Partner Interdependence model Approach. 101-114.

Tanner, M. A., & Wong , W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association, 82*(398), 528-540.

Won, D., Tadesse, M. G., Sha, N., Pfeiffer, R. M., & Vanucci, M. (2007). Identifying biomarkers from mass spectrometry data with ordinal outcome. *Cancer informatics*.

Yi, N., George, V., & Allison, D. B. (2003, July 1). Stochastic Search Variable Selection for Identifying Multiple Quantitative Trait Loci. *Genetics, 164*(3), 1129-1138.

Yuan, M., Joseph, V. R., & Hui, Z. (2009). Structured Variable Selection and Estimation. *The Annals of Applied Statistics, 3*(4), 1738-1757.

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., & Do, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society Series C Applied Statistics, 63*(4), 595–620.