ORIGINAL RESEARCH ARTICLE

**Modeling the effects of multiple exposures with unknown group memberships: a Bayesian latent variable approach**

Alexis Zaveza, Emeir M. McSorleyb, Alison J. Yeatesb, and Sally W. Thurstona,c

aDepartment of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA; bNutrition Innovation Centre for Food and Health (NICHE), Ulster University, Coleraine, Northern Ireland; cDepartment of Environmental Medicine, University of Rochester Medical Center, Rochester, NY, USA

**ARTICLE HISTORY**

Compiled September 28, 2020

**ABSTRACT**

We propose a Bayesian latent variable model to allow estimation of the covariate- adjusted relationships between an outcome and a small number of latent exposure variables, using data from multiple observed exposures. Each latent variable is as- sumed to be represented by multiple exposures, where membership of the observed exposures to latent groups is unknown. Our model assumes that one measured ex- posure variable can be considered as a sentinel marker for each latent variable, while membership of the other measured exposures is estimated using MCMC sampling based on a classical measurement error model framework. We illustrate our model using data on multiple cytokines and birth weight from the Seychelles Child Devel- opment Study, and evaluate the performance of our model in a simulation study. Classification of cytokines into Th1 and Th2 cytokine classes in the Seychelles study revealed some differences from standard Th1/Th2 classifications. In simulations, our model correctly classified measured exposures into latent groups, and estimated model parameters with little bias and with coverage that was similar to the oracle model.

**KEYWORDS**

Immune response; Inflammation; Latent variables; Markov chain Monte Carlo; Multiple exposures; Seychelles Child Development Study

# Introduction

Applications that examine associations between multiple correlated observed expo- sures and an outcome arise in many areas of scientific research [11, 15, 21, 32, 33]. Often it is assumed that the observed exposures can be classified into one or more meaningful groups. One approach, weighted quantile sum (WQS) regression, is appro- priate when all observed exposures are classified into one group [3]. WQS develops a weighted and additive sum of exposure quantiles that can be used as a single variable in a regression model. For applications involving one or more exposure groups, the sum (or mean) of multiple exposures in a group can be modeled as the single relevant expo- sure metric. Another approach applies principal components analysis (PCA) to each group of observed exposures, and models the resulting principal components as pre-

CONTACT Alexis Zavez. Email: alexis [zavez@urmc.rochester.edu](mailto:zavez@urmc.rochester.edu)

dictor variables. However, these two approaches require exposure group memberships to be known, and neither approach provides inference at the level of the individual measured exposures.

An alternative approach that allows for exposure-level inference uses structural equation models (SEMs), which permit flexible modeling of multiple, potentially cor- related predictors [23]. SEMs can include univariate or multivariate outcomes and are often utilized for applications involving latent variables, although SEMs may also in- clude only observed variables. SEM mixture models [13, 19] have also been proposed to allow SEM parameter estimates to differ across population subgroups. However, both traditional SEMs and SEM mixture models assume a fixed SEM model structure, and assign exposures to latent groups *a priori*. One approach that does not require expo- sure group memberships to be known is exploratory factor analysis (EFA). EFA has similarities to PCA but assumes that the observed exposures are manifestations of unmeasured factors [14]. EFA identifies latent factors among the measured exposures, but does not consider the relationship between the measured exposures and a speci- fied outcome. Additionally, results from EFA are not always reproducible, as different methods of rotation may give different results [6].

A further approach to modeling the effect of a complex exposure mixture on an outcome is Bayesian kernel machine regression (BKMR) [2]. This very flexible method allows for a high dimensional exposure-response surface in which components of the mixture may have non-linear and interactive effects on the outcome. BKMR may also be useful for variable selection when exposures are highly correlated within groups. In this case, one exposure (at most) is selected from the group, making for a more parsimonious model.

While WQS and BKMR focus on modeling the effect of a set (or subset) of specific exposures, some applications warrant an expanded SEM-like approach that estimates membership of exposures to different latent groups (i.e. exposure membership). **This is** relevant not only for applications involving multiple measures of toxicant exposures, but also for applications in which interest focuses on understanding the relationship between multiple correlated variables and an outcome. For example, there is a large body of literature surrounding T-helper type 1 (Th1) and T-helper type 2 (Th2) cytokines as measures of the immune response, and estimating their association with a downstream outcome, such as birth weight. Th1 cytokines are produced by Th1 cells and are associated with predominantly pro-inflammatory responses, while Th2 cytokines are associated with predominantly anti-inflammatory responses [1]. However, the classification of individual cytokines to Th1 or Th2 is not always well established. **In practice, classification may be more straightforward for some cytokines**

# compared to others. For example, the Th1 pathway is heavily reliant on IFN-*γ* [17] while Th2 cells originate under the influence of interleukin-4 (IL-4) [29], making it straightforward to classify IFN-*γ* as a Th1 cytokine and IL-4 as a Th2 cytokine. In contrast, IL-10 is often classified as a Th2 cytokine [9, 29], but is also produced by Th1 cells [22]. As a result, it may be harder to determine if IL-10 should be classified with IFN-*γ* or with IL-4.

**Three approaches discussed above (modeling the sum/mean as the single relevant exposure metric, WQS regression, and PCA regression) require that each cytokine be classified as either Th1 or Th2 *a priori*. Given the uncertainty in classification for some cytokines (e.g., IL-10), these methods may be too restrictive. In addition,** it may be overly simplistic to assume that a given cytokine is always pro- (or anti-) inflammatory, as cytokine properties are

influenced by numerous biological and experimental factors [5].

**For these reasons,** we propose a Bayesian latent variable model that does not require the relationships between all observed exposures and the latent variables to be known. Instead, our model estimates the latent variable membership for most of the observed exposures (e.g. cytokines, in our example) in examining the covariate- adjusted association between the latent variables and an outcome. We apply our model to data collected from the Seychelles Child Development Study (SCDS) Nutrition Cohort 2 (*n* 1500), where maternal cytokines were measured at 28 weeks gestation. We limit our focus to seven Th1/Th2 cytokines previously analyzed by [21] and [33] in this cohort: interferon (IFN)-*γ*, interleukin (IL)-2, IL-1*β*, tumor necrosis factor (TNF)- *α*, IL-4, IL-5, and IL-10, and two latent variables: “Th1” and “Th2.” We employ the Th1/Th2 literature to assign one observed exposure to each latent variable. This ensures model identifiability and prevents label switching [26]. MCMC sampling in a Bayesian framework is used to determine the membership (either Th1 or Th2) for the five remaining cytokines and to estimate all model parameters.

*≈*

**Our model is intended for applications (such as the Th1/Th2 cytokines) in** which the primary focus is to better understand the relationship between an out- come and multiple exposures, where the exposures are believed to separate into more than one group and the group memberships are not fully known. We aim to estimate the membership of exposures to each group, and to estimate the association between each group and the outcome, such that, like WQS and SEMs, each exposure within a group contributes to this effect. Our approach therefore differs from BKMR in that we are interested in understanding the grouping of exposures and the effects of the latent exposure groups, rather than modeling the effect of a subset of specific exposures.

In Section 2, we introduce our model and sampling procedure, and in Section 3 we evaluate the performance of our model in a simulation study. In Section 4, we present results from fitting our proposed model to data from the Seychelles Child Development Study. We close with a discussion in Section 5.

# Methodology

We treat the measured cytokines as the observed exposures, and assume that each observed exposure belongs to one latent variable. While we can extend the model to situations with more than two latent variables (see Appendix A), we illustrate the model with only two latent variables for clarity and to match the motivating exam- ple. Since measurements may be on different scales, we center and scale all observed exposures, covariates, and the outcome variable. These steps eliminate the need to estimate intercept parameters in the models presented below.

## Notation and Models

For subject *i*, *i* = 1*, ..., n*, let *yi* denote the observed centered and scaled outcome, *x*1*,i* and *x*2*,i* denote the two latent exposure variables, and *wi,j* denote the *j*th centered and scaled measured exposure variable, *j* = 1*, ..., J*. We assume the availability of covariates ***Zy,i*** = (*Zy*1*,i, Zy*2*,i, ..., Zyp,i*)*,* a *p*-dimensional vector of centered and scaled covariates associated with the outcome, and ***Zx,i*** = (*Zx*1*,i, Zx*2*,i, ..., Zxq,i*), a *q*-dimensional vector of centered and scaled covariates associated with the latent exposures. We use ***y*** to denote the *n*-length column vector of outcomes, ***wj*** to denote the *n*-length column vector of the *j*th exposure, and ***x*1** and ***x*2** to denote the *n*-length column vectors of

the first and second latent variables, respectively. We let ***X*** be the *n ×* (2 + *p*) matrix with columns corresponding to ***x*1**, ***x*2**, ***Zy*1** *, ...,* ***Zyp*** and let ***Zx*** be the *n × q* matrix created by stacking ***Zx,i***, *i* = 1*, ..., n*. For *j* = 1*, ..., J*, our models are

*yi|x*1*,i, x*2*,i,* ***Zy,i****, βx , βx ,* ***βy****, σ*2 *∼ N* (*βx x*1*,i* + *βx x*2*,i* + ***Zy,iβy****, σ*2) *x*1*,i|****Zx,i****,* ***γ*1***, η*2 *∼ N* (***Zx,iγ*1***, η*2)

1 2 1 2

1

1

*x*2*,i|****Zx,i****,* ***γ*2***, η*2 *∼ N* (***Zx,iγ*2***, η*2)

2

2

*wi,j|x*1*,i, x*2*,i, α*1*,j, α*2*,j, τ* 2

1*,j*

2

2*,j*

*, τ*

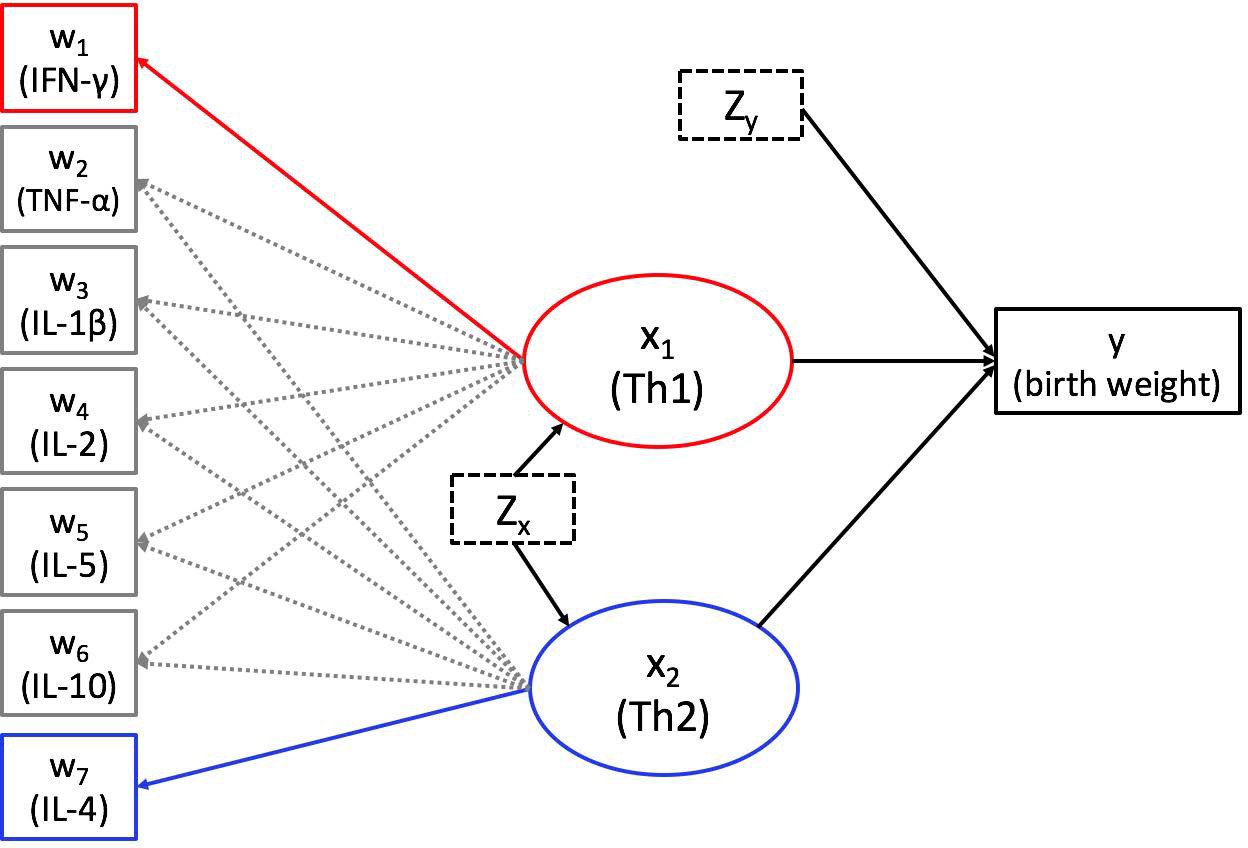
*∼ N* (*zjα*1*,jx*1*,i* + (1 *− zj* )*α*2*,jx*2*,i, zjτ*1*,j*

+ (1 *− zj*)*τ* 2 )

where *zj ∼* Bernoulli(*pj*). If *zj* = 1, the *j*th measured exposure is classified in the first latent group. If *zj* = 0, the *j*th measured exposure is classified in the second latent group. An illustration of our model is given in Figure 1.

2

2*,j*



**Figure 1.** Illustration of our proposed model with the structural equation model notation used by [23]. Solid arrows indicate fixed associations; dashed arrows indicate unknown group memberships.

2*,j*

Our model allows *τ* 2

1*,j*

and *τ* 2

, the conditional variances of the *j*th measured ex-

posure, *wi,j*, around *x*1*,i* or *x*2*,i* to differ for each observed measured exposure. When group membership of all measured exposure variables is known, this set of models is equivalent to the classical measurement error paradigm [4], where the latent variables in our model are analogous to true exposure in the measurement error case. What is different about our model is that we do not assume group memberships are known.

## Identifiability, Priors, and Posteriors

One way to ensure model identifiability is to satisfy two general conditions [23]. First, one observed exposure must be assigned to each latent variable. Without loss of gener- alizability, we fix the first exposure (***w*1**) in the first latent group and the last exposure

(***wJ*** ) in the second latent group. Thus, *z*1 is fixed at 1 and *zJ* is fixed at 0. For iden- tifiability, we also need the slope parameter for one exposure in each latent group to be fixed. Since all observed exposures and covariates are centered and scaled, we fix the slope parameter for each of the two assigned exposures at one (i.e *α*1*,*1 = 1 and *α*2*,J* = 1).

We assume that all slope parameters are independent from each other and are normally distributed. Specifically, we assume ***β*** = (*βx*1 *, βx*2 *,* ***βy***) *∼ N* (***β*0***,* **Σ*β***)*,* ***γ*1** *∼ N* (***γ*0***,* **Σ*γ*** )*,* ***γ*2** *∼ N* (***γ*0***,* **Σ*γ*** )*, α*1*,j ∼ N* (*α*0*, σ*2 )*,* and *α*2*,j ∼ N* (*α*0*, σ*2 )*, j* = 1*, ..., J*. We

*α*

*α*

assume that all variance parameters are independent, independent from all slope pa-

rameters, and follow Inverse-Gamma (IG) distributions, thus, *σ*2

2

*∼ IG*(*aσ*2 *, bσ*2 )*, τ ,j ∼*

*IG*(*aτ*2 *, bτ*2 )*,* and *η ∼ IG*(*aη*2 *, bη*2 ), */!* = 1*,* 2; *j* = 1*, ..., J*. We also assume that *pj ∼* Beta(*c, d*) for *j* = 2*, ..., J −* 1. Model results can be sensitive to hyperparameter values, so we fit our model under two priors. For both priors, we set ***β*0** = ***γ*0** = **0***, α*0 = 0,

2

**Σ*β*** = **Σ*γ*** = diag(100)*, σ*2

*α*

= 100, and *c* = *d* = 1. **A mean-zero Normal prior**

# with a diagonal covariance matrix containing large variances is a typical

**choice [10] in the absence of prior knowledge, where the specific hyperpa- rameter values have almost no influence on the posterior mean or variance. The Beta(1,1) prior is equivalent to a uniform prior, and has been recom- mended by others [20]. Under the uniform prior all** *pj* **values are equally likely, which is appropriate given that no additional information regarding the likelihood of** *pj* **values is assumed.**

For prior A, we set the IG shape hyperparameters *aσ*2 = *aτ*2 = *aη*2 = 0*.*05 and the IG scale hyperparameters *bσ*2 = *bτ*2 = *bη*2 = 0*.*01. For prior B, we set *aσ*2 = *aτ*2 = *aη*2 = 0*.*5 and *bσ*2 = *bτ*2 = *bη*2 = 0*.*1. Each IG shape parameter is proportional to the prior sample size and each scale parameter is proportional to the prior sum of squares. The selected shape values of 0.05 and 0.5 are small relative to both the Seychelles (*n* 1500) and simulated (*nsim* = 500) sample sizes. Preliminary analyses of the Seychelles and simulated data indicated that the selected scale values of 0.01 and 0.1 were small relative to observed sum of squares.

*≈*

Let ***θ*** denote the parameter vector (*x*1*,i, ..., x*1*,n, x*2*,i, ..., x*2*,n,* ***β****, σ*2*, α*1*,*1*, ..., α*1*,J , α*2*,*1*,*

*..., α*2*,J , τ* 2

*, ..., τ* 2

*, τ* 2

*, ..., τ* 2 *,* ***γ*1***,* ***γ*2***, η*2*, η*2) and let *K* denote the total number of

1*,*1

1*,J*

2*,*1

2*,J* 1 2

observed exposures assigned to the first latent variable in the current iteration of

the MCMC sampler. We briefly use *w*˜1*,i, ..., w*˜*K,i* to denote the *K* observed exposures assigned to the first latent variable, *α*˜1*,*1*, ..., α*˜1*,K* to denote the corresponding exposure

1*,K*

model slopes, and *τ*˜2

1*,*1

*, ..., τ*˜2

to denote the corresponding exposure model variances.

We use *w*˜*K*+1*,i, ..., w*˜*J,i* to denote the *J − K* observed exposures assigned to the second

latent variable, with slopes *α*˜2*,K*+1*, ..., α*˜2*,J* and variances *τ*˜2 *, ..., τ*˜2 . The latent

2*,K*+1

2*,J*

variable posterior distributions are

*x*1*,i|****yi****,* ***Zy,i****,* ***Zx,i****,* ***θ****, w*˜1*,i, ..., w*˜*K,i ∼ N* (*µx*

*x*1

1*,i*

*, σ*2 )

*x*2*,i|****yi****,* ***Zy,i****,* ***Zx,i****,* ***θ****, w*˜*K*+1*,i, ..., w*˜*J,i ∼ N* (*µx*

*x*2

2*,i*

*, σ*2 )*,* where

*β*2

*σ*2

=

*x*1 + 1*,*1 + *...* + 1*,K* +

*σ*2

*x*1

*α*˜2

2

1*,*1

*τ*˜

*α*˜2

2

1*,K*

*τ*˜

1 )*−*1

*η*

2

1

*µ* = *σ*2

*×*  *βx*1 (*yi − x*2*,iβx*2 *−* ***Zy,iβy*** ) + *w*˜1*,iα*˜1*,*1 + *...* + *w*˜*K,iα*˜1*,K* + ***Zx,iγ*1** )

*τ*˜

*τ*˜

*η*

*x*1*,i x*1

*σ*2

*β*2

*σ*2

=

*x*2 + 2*,K*+1 + *...* + 2*,J* +

*σ*2

*x*2

*α*˜2

2

2*,K*+1

*τ*˜

2

1*,*1

*α*˜2

2

2*,J*

*τ*˜

1 )*−*1

*η*

2

2

2 2

1*,K* 1

*µ* = *σ*2 *×* *βx*2 (*yi − x*1*,iβx*1 *−* ***Zy,iβy*** ) + *w*˜*K*+1*,iα*˜2*,K*+1 + *...* + *w*˜*J,iα*˜2*,J* + ***Zx,iγ*2** )

*τ*˜

*τ*˜

*η*

*x*2*,i*

*x*2

*σ*2

2

2*,K*+1

2

2*,J*

2

2

Of note, the posterior for *x*1*,i* depends on observed exposures assigned to the first latent variable, but not on exposures assigned to the second latent variable, and vice

versa for *x*2*,i*. Posteriors for ***β***, *σ*2, ***γ*1**, ***γ*2**, *η*2, *η*2, *α*1*,j*, *α*2*,j*, *τ* 2 , and *τ* 2

*, j* = 1*, ..., J*,

1 2 1*,j* 2*,j*

*−*

as well as *pj* and *zj*, *j* = 2*, ..., J* 1 are straightforward, and presented in Appendix

B.

## Sampling Algorithm

In our sampling algorithm, superscript *t* indicates the draw of the parameter at MCMC iteration *t* and *S* is a fixed positive integer value greater than 1.

**Step 0:** Initialize all parameters. We randomly sample one permutation of the *zj*’s to initialize *zj*, *j* = 2*, ..., J* 1 (see Section 2.4 for details).

*−*

**For** *t* **in** 2 : *T* **iterations:**

**Step 1:** Draw each parameter in ***θ***(*t*) from its posterior, conditional on *z*(*t−*1),

*j*

*j* = 1*, ..., J*.

**Step 2a:** Draw each *p*(*t*) and *z*(*t*), *j* = 2*, ..., J −* 1, from its posterior,

*j j*

conditional on ***θ***(*t*) .

**if** (*t* mod *S*) */*= 0 *{*

Proceed directly to **Step 3**.

# else

*} {*

**Step 2b:** Compute the posterior density under the current permutation (i.e.

*Pcurrent*) using ***θ***(*t*).

**Step 2c:** Randomly propose a new permutation of the *zj* variables. We store

the proposed permutation as ***zj***(*tt*), where ***zj***(*tt*) = (*z*(*tt*)*, z*(*tt*)*, ..., z*(*tt*) *, z*(*tt*)) (see

Section 2.4 for details).

1 2 *J−*1 *J*

**Step 2d:** Re-draw each parameter in ***θ*** from its posterior, conditional on ***zj***(*tt*), to obtain ***θ***(*tt*).

**Step 2e:** Compute the posterior density under the newly proposed permutation (i.e. *proposed*) using ***θ***(*tt*).

*P*

**Step 2f:** Accept the new permutation with probability equal to

*Pproposed/Pcurrent*.

# if the proposed permutation is accepted *{*

Update ***θ***(*t*) to be equal to ***θ***(*tt*), and update each *z*(*t*), *j* = 2*, ..., J −* 1 to be

*j*

equal to *z*(*tt*), *j* = 2*, ..., J −* 1, respectively.

*j*

*}*

*}*

**Step 3:** Reallocate each observed exposure, ***wj****, j* = 2*, ..., J −* 1 to the latent group

based on *z*(*t*), *j* = 2*, ..., J −* 1.

*j*

*}*

Steps 1-3 are repeated until the model has converged. To evaluate whether the model has reached convergence, we examine traceplots of all parameters and compute the effective sample size (*neff* ) for each parameter. The effective sample size approximates the effective number of independent MCMC draws [10].

## Details of MCMC Procedure

Our initial implementation of the sampling algorithm presented in Section 2.3 did not include Steps 2b - 2f, and convergence diagnostics indicated extremely poor mixing between *zj* = 0 and *zj* = 1 for each *zj* estimated by the model. We then implemented Steps 2b - 2f such that, every *S* iterations, a new permutation of the *zj* variables is proposed. The model estimates (*J* 2) *zj* parameters (*zj, j* = 2*, ..., J* 1), so there are 2*J−*2 = *P* possible permutations. For *J* = 7, as considered in the application and

*− −*

simulation, *P* = 32. For the application and simulation, we set *S* = 5 so that, every 5

iterations, a *zj* permutation is randomly sampled from the 32 possible permutations (with replacement). We run the MCMC with *T* iterations, where *T* is sufficiently large enough to ensure that all *P* permutations are proposed (see Appendix C). We also record *neff* for all parameters. We discard the first 1,000 iterations as “burn-in” iterations, and retain the following (*T* 1*,* 000) iterations. Only retained draws are utilized for inference.

*−*

# Simulations

## Simulation Study Design

We designed our simulation study to be representative of the Th1/Th2 inflammatory marker data collected from the Seychelles NC2 cohort, with a smaller number of co- variates for simplicity and a smaller sample size of *nsim* = 500. For each simulated individual, we simulated one outcome, three *Zy* covariates, two *Zx* covariates, two latent exposures, and seven observed exposures. Each simulated exposure is assumed to belong to one of two latent groups. The first three simulated exposures (***w*1***,* ***w*2** and ***w*3**) are assumed to belong to the first latent group, while the last four simulated exposures (***w*4***,* ***w*5***,* ***w*6** and ***w*7**) are assumed to belong to the second latent variable group.

*Simulated Exposure Models:* For the *i*th subject, i = 1,..., 500, we simulated

*x*1*,i, x*2*,i, Zx*1*,i* and *Zx*2*,i* from a multivariate normal distribution:

 *x*1*,i* 

*x*2*,i* 0

0

0*.*500 0*.*250 0*.*071 0*.*212

  *∼ N*  *,*

 

*Zx*1*,i Zx*2*,i*

0

0

0*.*250 0*.*500 0*.*283 0*.*106

0*.*071 0*.*283 1*.*000 0*.*000

0*.*212 0*.*106 0*.*000 1*.*000

We then used the two simulated latent variables (*x*1*,i* and *x*2*,i*) to simulate seven

 

 

observed exposures for each subject:

*w*1*,i ∼ N* (*x*1*,i, τ* 2 )

1*,*1

*w*2*,i ∼ N* (*α*1*,*2*x*1*,i, τ* 2 *w*3*,i ∼ N* (*α*1*,*3*x*1*,i, τ* 2 *w*4*,i ∼ N* (*α*2*,*4*x*2*,i, τ* 2 *w*5*,i ∼ N* (*α*2*,*5*x*2*,i, τ* 2 *w*6*,i ∼ N* (*α*2*,*6*x*2*,i, τ* 2

1*,*2

1*,*3

2*,*4

2*,*5

2*,*6

) where *α*1*,*2 = 0*.*6

) where *α*1*,*3 = 0*.*8

) where *α*2*,*4 = 0*.*5

) where *α*2*,*5 = 0*.*7

) where *α*2*,*6 = 0*.*85

*w*7*,i ∼ N* (*x*2*,i, τ* 2 )

2*,*7

We assumed that *V ar*(***x*1**) = *V ar*(***x*2**) = 0*.*5 and *V ar*(***wj***) = 1*, j* = 1*, ..., J*. Thus,

*,j*

2

*τ*

*,j*

= 1*−*0*.*5*α*2

, */!* = 1 or 2 (see Appendix D). Jointly drawing ***x*1***,* ***x*2***,* and ***Zx*** allowed

us to specify the correlation between latent ***x*1** and ***x*2**. We set *Corr*(***x*1***,* ***x*2**) = 0*.*5, so

that the correlations among the simulated ***wj***’s mirrored the correlations among the observed cytokines in the NC2 cohort.

*Simulated Outcome Model:* For the *i*th subject, we simulated one binary covariate, *Zy*1*,i*, with mean 0.55, and two Gaussian covariates (*Zy*2*,i*, *Zy*3*,i*). Each ***Zy*** was centered and scaled prior to drawing *yi*, where *yi* = *β*1*x*1*,i*+*β*2*x*2*,i*+*β*3*Zy*1*,i* +*β*4*Zy*2*,i* +*β*5*Zy*3*,i* +*Ei*,  ***β*** = (0*.*05*,* 0*.*03*,* 0*.*15*,* 0*.*1*,* 0*.*1), and *Ei N* (0*,* 1). The values of ***β*** were chosen to be similar to the Seychelles data, where associations between observed cytokines and the birth weight outcome were weak.

*− − ∼*

## Simulation Results

For each of 50 simulated datasets, we compared the performance of our model to two competitor models using three metrics (bias, interval length, and coverage), which we calculated for all estimable parameters. The first competitor model assumes that the true classification for each of the observed ***wj***’s is known, where ***w*1***,* ***w*2** and ***w*3** are functions of ***x*1**, while ***w*4***,* ***w*5***,* ***w*6** and ***w*7** are functions of ***x*2**. Under this competitor model, *zj* , *j* = 1*, ...,* 7 are fixed and all parameters in ***θ*** are estimated using Gibbs sampling. We call this competitor the “fixed group” model.

The second competitor model also assumes that the true classification for each of the observed ***wj***’s is known, but additionally treats the simulated ***x*1** and ***x*2** variables as if they were observed. This model estimates the slope and variance parameters in ***θ*** using the simulated ***x*1** and ***x*2**. Since it is not possible to observe ***x*1** and ***x*2** in real applications, we call this competitor the “oracle” model. We compared the performance of our model to these two competitor models under priors A and B, setting *T* = 16,000 and *S* = 5.

Let *θ*ˆ denote an estimate (either the MLE for the oracle model or the posterior

mean for our model and the fixed group model) of true parameter *θ* from which the data are generated. For bias, we report the mean of (*θ*ˆ *θ*). For interval length, we report either the mean 95% **equal-tailed** posterior interval length (our model and fixed group model) or the mean 95% confidence interval length (oracle model). For coverage, we report the proportion of intervals that cover the true parameter value.

*−*

For each prior, we present simulation results in two ways. First, we report results using only the retained draws (i.e. draws post-burn-in) that correspond to the top permutation. For comparison purposes, we also report results using all retained draws

(i.e. all draws post-burn-in) under both priors in the Appendix. The results from the top permutation are most relevant because the interpretation of each model parameter may depend on the current permutation. For example, *βx*1 estimates the effect of the first latent variable, ***x*1**, on the outcome, conditional on the second latent variable, the *Zy* covariates, and *σ*2. However, the interpretation of ***x*1** when ***w*1***,* ***w*2***,* and ***w*3** are assigned to ***x*1** is different from the interpretation of ***x*1** when, for example, ***w*1**, ***w*2** and ***w*6** are assigned to ***x*1**.

**Table 1.** Simulation results under prior A (IG(0.05, 0.01)) for our model (“Proposed”), the fixed group model (“Fixed”), and the oracle model (“Oracle”). **Mean estimates for our model are also provided (“Estimate”)**. Reported values for our model are averages over 50 simulated datasets using only retained draws in the top permutation. For identifiability, *α*1*,*1 and *α*2*,*7 are fixed at 1 in our model and in the fixed group model.

Bias Interval Length Coverage

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Truth | **Estimate** | Proposed | Fixed | Oracle | Proposed | Fixed | Oracle | Proposed | Fixed | Oracle |
| *βx*1 0.050 | 0.045 | -0.005 | -0.005 | 0.013 | 0.343 | 0.344 | 0.282 | 0.960 | 0.960 | 0.940 |
| *βx*2 -0.030 | -0.030 | -0.000 | -0.000 | -0.013 | 0.330 | 0.332 | 0.282 | 0.940 | 0.940 | 0.980 |
| *βZy*1 0.150 | 0.152 | 0.002 | 0.002 | 0.003 | 0.172 | 0.172 | 0.172 | 1.000 | 1.000 | 1.000 |
| *βZy*2 -0.100 | -0.098 | 0.002 | 0.002 | 0.003 | 0.173 | 0.172 | 0.172 | 0.940 | 0.960 | 0.940 |
| *βZy*3 0.100 | 0.096 | -0.004 | -0.004 | -0.005 | 0.173 | 0.173 | 0.172 | 0.920 | 0.920 | 0.920 |
| *σ*2 1.000 | 0.954 | -0.046 | -0.046 | -0.044 | 0.240 | 0.240 | 0.240 | 1.000 | 0.980 | 1.000 |
| *α*1*,*1 1.000 | - | - | - | 0.005 | - | - | 0.177 | - | - | 1.000 |
| *α*1*,*2 0.600 | 0.603 | 0.003 | 0.006 | -0.004 | 0.433 | 0.428 | 0.227 | 0.920 | 0.920 | 0.920 |
| *α*1*,*3 0.800 | 0.823 | 0.023 | 0.025 | -0.004 | 0.563 | 0.551 | 0.207 | 0.940 | 0.940 | 1.000 |
| *α*2*,*4 0.500 | 0.516 | 0.016 | 0.017 | -0.002 | 0.345 | 0.345 | 0.234 | 1.000 | 1.000 | 1.000 |
| *α*2*,*5 0.700 | 0.737 | 0.037 | 0.038 | 0.001 | 0.381 | 0.382 | 0.217 | 0.980 | 0.980 | 0.980 |
| *α*2*,*6 0.850 | 0.874 | 0.024 | 0.025 | -0.006 | 0.419 | 0.419 | 0.201 | 0.980 | 0.980 | 0.960 |
| *α*2*,*7 1.000 | - | - | - | 0.003 | - | - | 0.176 | - | - | 1.000 |
| *τ* 2 0.500 | 0.491 | -0.009 | -0.005 | 0.001 | 0.373 | 0.358 | 0.125 | 0.880 | 0.860 | 0.860 |
| *τ* 2 0.820 | 0.829 | 0.009 | 0.008 | 0.005 | 0.241 | 0.241 | 0.206 | 1.000 | 1.000 | 1.000 |
| *τ* 2 0.680 | 0.681 | 0.001 | -0.000 | 0.007 | 0.272 | 0.267 | 0.172 | 0.960 | 0.960 | 1.000 |
| *τ* 2 0.875 | 0.877 | 0.002 | 0.002 | 0.003 | 0.235 | 0.235 | 0.220 | 1.000 | 1.000 | 1.000 |
| *τ* 2 0.755 | 0.749 | -0.006 | -0.006 | 0.001 | 0.224 | 0.224 | 0.189 | 0.980 | 0.980 | 1.000 |
| *τ* 2 0.639 | 0.648 | 0.009 | 0.009 | 0.009 | 0.225 | 0.224 | 0.162 | 0.980 | 1.000 | 0.960 |
| *τ* 2 0.500 | 0.522 | 0.022 | 0.023 | 0.001 | 0.241 | 0.241 | 0.125 | 0.960 | 0.960 | 0.940 |
| *γ*1*,*1 0.071 | 0.065 | -0.006 | -0.006 | -0.003 | 0.152 | 0.152 | 0.117 | 0.940 | 0.940 | 0.960 |
| *γ*1*,*2 0.212 | 0.207 | -0.005 | -0.005 | -0.001 | 0.159 | 0.160 | 0.117 | 0.940 | 0.920 | 0.960 |
| *γ*2*,*1 0.283 | 0.272 | -0.011 | -0.011 | -0.001 | 0.151 | 0.151 | 0.112 | 0.940 | 0.940 | 0.920 |
| *γ*2*,*2 0.106 | 0.104 | -0.002 | -0.002 | 0.000 | 0.141 | 0.141 | 0.112 | 0.940 | 0.940 | 0.960 |
| *η*2 0.450 | 0.459 | 0.009 | 0.005 | -0.004 | 0.387 | 0.370 | 0.112 | 0.900 | 0.920 | 0.900 |
| *η*2 0.409 | 2 0.389 | -0.020 | -0.021 | -0.003 | 0.252 | 0.252 | 0.102 | 0.980 | 0.980 | 0.940 |

1*,*1

1*,*2

1*,*3

2*,*4

2*,*5

2*,*6

2*,*7

1

Simulation results were fairly consistent across priors A (Table 1) and B (Table 2). For all 50 simulated datasets, the top permutation was the true permutation. On average, 98.8% (range over datasets: 79.0% - 99.8%) of retained draws were in the top permutation under prior A, and 99.1% (89.9% - 99.8%) were in the top permutation under prior B. For each prior, the estimates produced using all retained draws were similar to estimates produced using only retained draws in the top permutation (see Table E1 and Table E2 in the Appendix for details). Although this might not always be the case, it occurred here because, for each simulated dataset, a large percentage of the retained draws were in the top permutation. For each simulated dataset, all 32 *zj* permutations were proposed. On average, 10 of the 32 permutations were accepted (prior A range: 5.0 - 18.0, prior B range: 5.0 - 17.0).

Coverage probabilities under our model were similar to both competitor models and were generally close to the 95% level. Biases and interval lengths for *σ*2 and ***β*** were similar under all three models, although posterior intervals under both Bayesian models (i.e., our model and the fixed group model) were slightly larger for *βx*1 and *βx*2 compared to the oracle model. Posterior intervals under both Bayesian models were moderately larger for the *α*, *τ* 2, and *η* parameters (and slightly larger for the *γ*

**Table 2.** Simulation results under prior B (IG(0.5, 0.1)) for our model (“Proposed”), the fixed group model (“Fixed”), and the oracle model (“Oracle”). **Mean estimates for our model are also provided (“Estimate”)**. Reported values for our model are averages over 50 simulated datasets using only retained draws in the top permutation. For identifiability, *α*1*,*1 and

*α*2*,*7 are fixed at 1 in our model and in the fixed group model.

Bias Interval Length Coverage

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Truth | **Estimate** | Proposed | Fixed | Oracle | Proposed | Fixed | Oracle | Proposed | Fixed | Oracle |
| *βx*1 0.050 | 0.045 | -0.005 | -0.005 | 0.013 | 0.344 | 0.344 | 0.282 | 0.960 | 0.960 | 0.940 |
| *βx*2 -0.030 | -0.030 | -0.000 | -0.000 | -0.013 | 0.331 | 0.332 | 0.282 | 0.940 | 0.940 | 0.980 |
| *βZy*1 0.150 | 0.152 | 0.002 | 0.002 | 0.003 | 0.172 | 0.172 | 0.172 | 1.000 | 1.000 | 1.000 |
| *βZy*2 -0.100 | -0.098 | 0.002 | 0.002 | 0.003 | 0.172 | 0.172 | 0.172 | 0.960 | 0.960 | 0.940 |
| *βZy*3 0.100 | 0.096 | -0.004 | -0.004 | -0.005 | 0.173 | 0.173 | 0.172 | 0.920 | 0.920 | 0.920 |
| *σ*2 1.000 | 0.952 | -0.048 | -0.046 | -0.044 | 0.239 | 0.240 | 0.240 | 0.980 | 0.980 | 1.000 |
| *α*1*,*1 1.000 | - | - | - | 0.005 | - | - | 0.177 | - | - | 1.000 |
| *α*1*,*2 0.600 | 0.608 | 0.008 | 0.006 | -0.004 | 0.431 | 0.428 | 0.227 | 0.940 | 0.920 | 0.920 |
| *α*1*,*3 0.800 | 0.830 | 0.030 | 0.025 | -0.004 | 0.559 | 0.551 | 0.207 | 0.940 | 0.940 | 1.000 |
| *α*2*,*4 0.500 | 0.517 | 0.017 | 0.017 | -0.002 | 0.346 | 0.345 | 0.234 | 1.000 | 1.000 | 1.000 |
| *α*2*,*5 0.700 | 0.738 | 0.038 | 0.038 | 0.001 | 0.383 | 0.382 | 0.217 | 0.980 | 0.980 | 0.980 |
| *α*2*,*6 0.850 | 0.875 | 0.025 | 0.025 | -0.006 | 0.421 | 0.419 | 0.201 | 0.980 | 0.980 | 0.960 |
| *α*2*,*7 1.000 | - | - | - | 0.003 | - | - | 0.176 | - | - | 1.000 |
| *τ* 2 0.500 | 0.498 | -0.002 | -0.005 | 0.001 | 0.357 | 0.358 | 0.125 | 0.900 | 0.860 | 0.860 |
| *τ* 2 0.820 | 0.827 | 0.007 | 0.008 | 0.005 | 0.240 | 0.241 | 0.206 | 1.000 | 1.000 | 1.000 |
| *τ* 2 0.680 | 0.677 | -0.003 | -0.000 | 0.007 | 0.270 | 0.267 | 0.172 | 0.960 | 0.960 | 1.000 |
| *τ* 2 0.875 | 0.876 | 0.001 | 0.002 | 0.003 | 0.234 | 0.235 | 0.220 | 1.000 | 1.000 | 1.000 |
| *τ* 2 0.755 | 0.747 | -0.008 | -0.006 | 0.001 | 0.224 | 0.224 | 0.189 | 0.980 | 0.980 | 1.000 |
| *τ* 2 0.639 | 0.647 | 0.008 | 0.009 | 0.009 | 0.225 | 0.224 | 0.162 | 1.000 | 1.000 | 0.960 |
| *τ* 2 0.500 | 0.522 | 0.022 | 0.023 | 0.001 | 0.241 | 0.241 | 0.125 | 0.960 | 0.960 | 0.940 |
| *γ*1*,*1 0.071 | 0.064 | -0.006 | -0.006 | -0.003 | 0.151 | 0.152 | 0.117 | 0.940 | 0.940 | 0.960 |
| *γ*1*,*2 0.212 | 0.207 | -0.005 | -0.005 | -0.001 | 0.159 | 0.160 | 0.117 | 0.900 | 0.920 | 0.960 |
| *γ*2*,*1 0.283 | 0.272 | -0.011 | -0.011 | -0.001 | 0.151 | 0.151 | 0.112 | 0.940 | 0.940 | 0.920 |
| *γ*2*,*2 0.106 | 0.104 | -0.002 | -0.002 | 0.000 | 0.141 | 0.141 | 0.112 | 0.940 | 0.940 | 0.960 |
| *η*2 0.450 | 0.451 | 0.001 | 0.005 | -0.004 | 0.369 | 0.370 | 0.112 | 0.920 | 0.920 | 0.900 |
| *η*2 0.409 | 2 0.387 | -0.022 | -0.021 | -0.003 | 0.250 | 0.252 | 0.102 | 0.980 | 0.980 | 0.940 |

1*,*1

1*,*2

1*,*3

2*,*4

2*,*5

2*,*6

2*,*7

1

parameters) compared to the oracle model. Biases for the *α* parameters were slightly larger under the Bayesian models, while biases for many *τ* 2 parameters were similar to the oracle model. Under both Bayesian models, the biases for *βx*1 and *βx*2 were smaller than under the oracle model, which may be the result of estimating each of ***x*1** and ***x*2** with multiple observed exposures. Wider confidence intervals for the exposure slopes are expected under a measurement error paradigm, and are a consequence of ***x*1** and ***x*2** not being fixed quantities. The larger biases in the *α* parameters may reflect a deficiency in both our model and the fixed group model, or they may simply be due to random variability from the 50 datasets. Simulation performance under our model was similar to performance under the fixed group model, which suggests that there is little to be gained from classifying the observed exposures *a priori*.

# Application: Seychelles Child Development Study

## Study Background

The Seychelles Child Development Study (SCDS) is an on-going research project in- vestigating the relationship between prenatal mercury exposure from fish consumption and children’s neurodevelopment [27]. The SCDS includes several large cohorts of chil- dren whose mothers consumed fish frequently during their pregnancy. Here, we restrict our attention to data collected from the Nutrition Cohort 2 (NC2), the largest and most recent cohort of the SCDS. Briefly, from 2008 - 2011, mother-child pairs (*n* = 1518 eligible mothers) were enrolled at eight health centers across the main Seychelles island of Mah´e. To meet inclusion criteria, eligible mothers needed to be native Sey-

chellois, at least 16 years of age, have a singleton pregnancy and exhibit no obvious health concerns. A further description of the NC2 cohort is given elsewhere [27].

## Model Details

For the SCDS NC2 application, our model has seven observed exposures (i.e. cytokines) and two latent variables: “Th1” and “Th2.” Since the balance of Th1 and Th2 cy- tokines may be important for a healthy pregnancy [29], we select a birth outcome (specifically birth weight measured in kilograms) as the outcome for this work.

*Cytokine selection:* In this cohort, thirteen maternal inflammatory markers, includ- ing seven Th1/Th2 cytokines, were measured. Previously, [21] modeled the linear re- lationship between each inflammatory marker and maternal methylmercury (MeHg), total n3 polyunsaturated fatty acids (PUFA), and total n6 PUFA. [33] modeled the linear relationship between each of three birth outcomes (birth weight, birth length, and birth head circumference) and each inflammatory marker. To better satisfy re- gression assumptions and account for zero-valued inflammatory marker data, [21] and

[33] transformed each inflammatory marker variable by first adding a constant of 1 to each reported value and then applying the natural logarithmic transformation. For our model, we restrict our attention to the seven cytokines studied by [21] and [33] that were classified as either Th1 (IL-1*β*, IL-2, IFN-*γ*, and TNF-*α*) or Th2 (IL-4, IL-5, and IL-10) and apply the same transformation procedure.

*Covariate selection:* As discussed in Section 2, our model permits two separate sets of covariates. First, ***Zy,i*** = (*Zy*1*,i, Zy*2*,i, ..., Zyp,i*) is a *p*-dimensional vector of covariates associated with the outcome (i.e. birth weight) for the *i*th subject. Our selection of covariates was based on [33], who analyzed birth weight in the NC2, and [30], who analyzed birth weight in the Nutrition Cohort 1 (NC1), an earlier SCDS cohort. Both

[33] and [30] adjusted for child sex, gestational age at birth, maternal age, maternal body mass index (BMI), and Hollingshead socioeconomic status [12] modified for use in Seychelles [8, 18]. [30] adjusted for pre-pregnancy BMI, which was not available for the NC2 cohort. Like [33], we adjusted for postnatal BMI as a surrogate for pre-pregnancy BMI, as NC1 data suggests a high correlation (*ρ* = 0*.*93) between pre-pregnancy and postnatal BMI [21]. Thus, we selected child sex (equal to 1 if male; 0 if female), gestational age at birth, maternal BMI, maternal age, and socioeconomic status as ***Zy*1** *,* ***Zy*2** *,* ***Zy*3** *,* ***Zy*4** and ***Zy*5** , respectively.

Second, ***Zx,i*** = (*Zx*1*,i, Zx*2*,i, ..., Zxq,i*) is a *q*-dimensional vector of covariates asso- ciated with the latent exposures for the *i*th subject. For the SCDS NC2 application, ***Zx*** should include covariates that are associated with inflammation and the immune response. In models for cytokine outcomes in the NC2 cohort, [21] adjusted for child sex, maternal age, maternal smoking status, maternal BMI, gestational age at time of blood draw, maternal MeHg, total n6 PUFA, and total n3 PUFA (or the n6:n3 PUFA ratio). In the SCDS NC2 study, total n6 PUFA is the sum of two omega-6 PUFA: linoleic acid (LA) and arachidonic acid (AA), while total n3 PUFA is the sum of three omega-3 PUFA: *α*-linolenic acid (ALA), eicosapentaenoic acid (EPA), and do- cosahexaenoic acid (DHA). All SCDS models that adjust for n3 and n6 do not adjust for n6:n3 ratio and vice versa. We adjusted for a subset of covariates considered by [21], excluding maternal age and child sex because each was weakly correlated with inflammatory markers (*ρ*’s *<* 0*.*04). We also excluded smoking status, which had very low prevalence and high missingness.

Although not included by [21], socioeconomic status may be associated with inflam-

mation [28] and is moderately correlated (*ρ* = *−*0*.*14) with IL-5 in the NC2 cohort, so we adjusted for socioeconomic status in our models. The final set of ***Zx*** covariates included maternal BMI, gestational age at time of blood draw, total n6 PUFA, total

n3 PUFA, maternal MeHg, and socioeconomic status as ***Zx*1** *,* ***Zx*2** *,* ***Zx*3** *,* ***Zx*4** *,* ***Zx,*5** and

***Zx*6** , respectively.

*Assumptions and Sampling Details:* For identifiability, we fixed IFN-*γ* (***w*1**) in the Th1 group with a slope of 1 (i.e. *α*1*,*1 = 1) and IL-4 (***w*7**) in the Th2 group with a slope of 1 (i.e. *α*2*,*7 = 1). IFN-*γ* is routinely classified as a Th1 cytokine [1, 21, 25, 29, 33] and noted to have pro-inflammatory properties [34], while IL-4 is classified as a Th2 cytokine [21, 25, 29, 33] with anti-inflammatory properties [34]. We initialized all slope and variance parameters at 1, initialized *pj, j* = 2*, ..., J* 1 at 0.5, set *T* = 80*,* 000 and *S* = 5.

*−*

## Seychelles Model Results

In Table 3, we summarize the *zj* permutation frequencies across all retained MCMC iterations for both priors. Permutations that never occurred in retained draws for either prior are not shown. For each prior, the *zj* permutation that occurred most often in the retained draws is considered the “top” permutation. The top permutation was consistent across priors A and B. TNF-*α* (***w*2**), IL-1*β* (***w*3**) and IL-10 (***w*6**) were classified in the Th1 group with IFN-*γ* (***w*1**), while IL-2 (***w*4**) and IL-5 (***w*5**) were classified in the Th2 group with IL-4 (***w*7**).

**Table 3.** Summary of Seychelles cytokine permutations for Prior A (IG(0.05, 0.01)) and Prior B (IG(0.5, 0.1)). IFN-*γ* (*z*1) and IL-4 (*z*7) were

fixed at 1 and 0 respectively, while classification of TNF-*α* (*z*2), IL-1*β* (*z*3),

IL-2 (*z*4), IL-5 (*z*5) and IL-10 (*z*6) as a Th1 cytokine (*zj* = 1) or a Th2

cytokine (*zj* = 0) was determined from MCMC sampling. The numbers in the “Prior A” and “Prior B” columns are the number of samples (out of 79,000) from the corresponding *zj* permutation.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Perm. Rank | *z*1 | *z*2 | *z*3 | *z*4 | *z*5 | *z*6 | *z*7 | Prior A | Prior B |
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 52305 | 44748 |
| 2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 20768 | 29276 |
| 3 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 3031 | 2595 |
| 4 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1048 | 680 |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1040 | 605 |
| 6 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 347 | 259 |
| 7 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 80 | 61 |
| 8 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 58 | 44 |
| 9 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 57 | 35 |
| 10 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 42 | 32 |
| 11 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 30 | 347 |
| 12 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 30 | 0 |
| 13 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 28 | 21 |
| 14 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 24 | 24 |
| 15 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 23 | 16 |
| 16 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 21 | 18 |
| 17 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 17 | 21 |
| 18 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 15 | 13 |
| 19 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 15 | 9 |
| 20 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 8 | 7 |
| 21 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 5 | 0 |
| 22 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 5 |
| 23 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 3 | 2 |
| 24 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 4 |
| 25 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 26 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 135 |
| 27 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 40 |
| 28 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |

We compute parameter estimates and **equal-tailed** posterior 95% intervals over all retained draws and subsetting to only draws in the top permutation. As mentioned in Section 3.2, the latter set of estimates is more relevant because the interpretation of each model parameter may depend on the current permutation. Top permutation parameter estimates and posterior 95% intervals under both prior A and prior B are given in Table 4. Under the top permutation, effective sample sizes were greater than 400 for all parameters, and greater than 4,000 for all ***β***, ***γ*1***,* and ***γ*2** parameters (Appendix Table E3). For both priors, neither latent Th1 nor latent Th2 was signif- icantly associated with child birth weight in the NC2 cohort. Child sex, gestational age at birth, maternal BMI, and maternal age were all positively associated with birth weight. Total n6 PUFA was positively associated with both latent Th1 and latent Th2, while Maternal MeHg was negatively associated with both latent Th1 and latent Th2. Socioeconomic status was negatively associated with latent Th2. Parameter estimates and posterior 95% intervals over all retained draws were similar (Appendix Table E4), however, socioeconomic status was not significantly associated with latent Th2.

**Table 4.** Seychelles parameter estimates and 95% posterior intervals under prior A (IG(0.05, 0.01)) and prior B (IG(0.5, 0.1)). Reported values are calculated using only retained draws that correspond to the top permutation. For identifiability, *α*1*,*1 and *α*2*,*7 are fixed at 1.

Prior A Prior B

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Posterior Mean | 95% Interval | Posterior Mean | 95% Interval |
| *βx*1 | -0.001 | (-0.131, 0.130) | -0.001 | (-0.131, 0.128) |
| *βx*2 | -0.025 | (-0.132, 0.082) | -0.026 | (-0.132, 0.082) |
| *βZy*1 | 0.142 | (0.093, 0.190) | 0.142 | (0.094, 0.190) |
| *βZy*2 | 0.487 | (0.438, 0.535) | 0.487 | (0.438, 0.535) |
| *βZy*3 | 0.095 | (0.045, 0.144) | 0.095 | (0.045, 0.145) |
| *βZy*4 | 0.071 | (0.021, 0.121) | 0.071 | (0.021, 0.121) |
| *βZy*5 | 0.040 | (-0.009, 0.089) | 0.040 | (-0.008, 0.090) |

*σ*2 0.724 (0.669, 0.785) 0.724 (0.668, 0.784)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *α*1*,*1 | 1.000 | - | 1.000 | - |
| *α*1*,*2 | 1.655 | (1.105, 1.996) | 1.655 | (1.106, 1.990) |
| *α*1*,*3 | 1.293 | (1.096, 1.518) | 1.293 | (1.097, 1.516) |
| *α*1*,*6 | 1.396 | (1.199, 1.624) | 1.396 | (1.200, 1.621) |
| *α*2*,*4 | 1.278 | (0.969, 1.688) | 1.271 | (0.953, 1.646) |
| *α*2*,*5 | 0.518 | (0.389, 0.651) | 0.517 | (0.389, 0.648) |
| *α*2*,*7 | 1.000 | - | 1.000 | - |

2

*τ*

|  |  |  |  |
| --- | --- | --- | --- |
| 0.824 | (0.752, 0.899) | 0.824 | (0.753, 0.899) |
| 0.396 | (0.291, 0.746) | 0.396 | (0.291, 0.745) |
| 0.635 | (0.572, 0.705) | 0.635 | (0.572, 0.703) |
| 0.575 | (0.505, 0.646) | 0.574 | (0.505, 0.643) |
| 0.407 | (0.228, 0.557) | 0.410 | (0.240, 0.560) |
| 0.903 | (0.828, 0.982) | 0.902 | (0.827, 0.982) |

1*,*1

2

*τ*

1*,*2

2

*τ*

1*,*3

2

*τ*

1*,*6

2

*τ*

2*,*4

2

*τ*

2*,*5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *τ* 2 | 0.635 | (0.521, 0.740) | 0.632 | (0.514, 0.736) |
| *γ*1*,*1 | 0.027 | (-0.005, 0.059) | 0.027 | (-0.006, 0.059) |
| *γ*1*,*2 | -0.015 | (-0.047, 0.016) | -0.015 | (-0.046, 0.016) |
| *γ*1*,*3 | 0.082 | (0.048, 0.119) | 0.082 | (0.048, 0.118) |
| *γ*1*,*4 | -0.006 | (-0.04, 0.028) | -0.007 | (-0.041, 0.028) |
| *γ*1*,*5 | -0.071 | (-0.104, -0.039) | -0.071 | (-0.104, -0.039) |
| *γ*1*,*6 | -0.023 | (-0.054, 0.008) | -0.023 | (-0.054, 0.008) |
| *γ*2*,*1 | 0.016 | (-0.025, 0.059) | 0.016 | (-0.026, 0.058) |
| *γ*2*,*2 | -0.028 | (-0.070, 0.013) | -0.028 | (-0.070, 0.014) |
| *γ*2*,*3 | 0.089 | (0.043, 0.140) | 0.089 | (0.043, 0.139) |
| *γ*2*,*4 | -0.040 | (-0.090, 0.007) | -0.040 | (-0.090, 0.007) |
| *γ*2*,*5 | -0.057 | (-0.101, -0.015) | -0.057 | (-0.102, -0.015) |
| *γ*2*,*6 | -0.057 | (-0.102, -0.014) | -0.057 | (-0.102, -0.014) |

2*,*7

2

*η*

|  |  |  |  |
| --- | --- | --- | --- |
| 0.210 | (0.157, 0.270) | 0.210 | (0.157, 0.270) |
| 2 0.359 | (0.255, 0.482) | 0.361 | (0.261, 0.486) |

1

2

*η*

## Seychelles Sensitivity Analyses

We performed a secondary analysis where we assigned IL-1*β* as the sentinel Th1 cy- tokine (instead of IFN-*γ*). IL-1*β* was strongly correlated with TNF-*α* (*ρ* = 0*.*51), and both IL-1*β* and TNF-*α* are often considered to be Th1 [21, 33]. This secondary analysis produced the same top cytokine grouping (see Appendix Table E5). In this secondary analysis, neither latent Th1 nor Th2 was significantly associated with birth weight. Covariate findings were similar to those discussed above.

# We performed a tertiary analysis to investigate the sensitivity of cytokine

**class membership to the selection of *Zx* covariates. We assigned IFN-***γ* **as the sentinel Th1 cytokine and included only *Zx*1 (maternal BMI) and *Zx*2 (gestational age at time of blood draw) in the model for *Zx*. This tertiary analysis also produced the same top cytokine grouping (see Appendix Table E6).**

1. **Discussion**

We introduced a Bayesian latent variable model that does not require the classifications of all observed exposures to be established *a priori*. Instead, our model estimates the latent variable membership for all non-sentinel exposures. We applied our model to data from the SCDS NC2 cohort, and examined the covariate-adjusted associations between seven observed Th1/Th2 cytokine exposures, latent Th1, latent Th2, and birth weight. Similar to [21] and [33], who defined total Th1 as the sum of IFN-*γ*, IL-1*β*, TNF-*α*, and IL-2, and total Th2 as the sum of IL-10, IL-5, and IL-4, we found no association between birth weight and either latent Th1 or Th2 in our models. Associations between birth weight and covariates were similar to those found by [33]. Covariate associations with latent Th1 and Th2 were similar to those reported by [21]. Our model generally classified TNF-*α* (***w*2**), IL-1*β* (***w*3**) and IL-10 (***w*6**) in the Th1 group with IFN-*γ* (***w*1**), and classified IL-2 (***w*4**) and IL-5 (***w*5**) in the Th2 group with IL-4 (***w*7**). These classifications are similar to those used by [21] and [33], but there are two differences. First, unlike [21] and [33], our model classified IL-10 (***w*6**) in the Th1 group. While IL-10 is often considered to be a Th2 cytokine [25], in the NC2 cohort IL-10 was highly correlated with TNF-*α* (*ρ* = 0*.*53) and IL-1*β* (*ρ* = 0*.*37), two cytokines often considered to be Th1 [21, 33]. These high correlations could be partially explained by the fact that IL-10 is produced by Th1 cells [29], or the fact that IL-10 regulates TNF-*α* [24] and IL-1*β* [16]. The second difference from [21] and

1. is that our model classified IL-2 (***w*4**) in the Th2 group. This classification may be partially explained by the fact that IL-2 had the highest correlation with IL-4 (*ρ* = 0*.*46), the sentinel Th2 cytokine, or the fact that both IL-2 and IL-4 are involved in Th2 differentiation [7]. The cytokine permutation assumed by [21] and [33] was only sampled 24 times (0*.*03%) under each prior (see row 14 of Table 3).

We have assumed that each exposure (i.e. cytokine) belongs to a single group, that the effects of Th1 and Th2 are additive and linear, and that associations between the

observed cytokines and latent Th1/Th2 are also linear. In the SCDS application, a*√*p-

proximately 64% of IL-2 and IL-4 values were undetectable, and a value of LLOD*/* 2

was imputed for these samples as previously described [21, 33]. While the assumption

of linearity does not e*√*ntirely hold for these observations, linearity is reasonable when

values at the LLOD*/* 2 are removed. For all other cytokines, the assumption of lin-

earity appears to be reasonable, and facilitates comparisons between our findings and

earlier Seychelles work.

Our model requires one cytokine to be fixed as the sentinel Th1 cytokine and one cytokine to be fixed as the sentinel Th2 cytokine. Our decision to assign IFN-*γ* and IL- 4 to Th1 and Th2 respectively was motivated by the large body of scientific literature surrounding Th1/Th2 expression. However, model results may be sensitive to choice of sentinel cytokine. Our secondary analysis with IL-1*β* as the sentinel Th1 cytokine (instead of IFN-*γ*) produced the same top permutation, although a second permutation of the Th1/Th2 group indicators did occur frequently in the retained draws. In this secondary analysis, the [33] cytokine permutation was only sampled 34 (0*.*04%) times under Prior A and 33 (0*.*04%) times under Prior B (see row 15 of Appendix Table E5). The assumption that each cytokine can be classified into one of just two latent groups may be an oversimplification for the Seychelles application. Our decision to model two latent groups was motivated by the Th1:Th2 hypothesis, which originated in 1993 [31] and postulates that successful pregnancy may align with dominance of the Th2 response and downregulation of the Th1 response [29]. However, complete membership of each cytokine into one of two latent groups may not adequately explain the relationships between cytokines and birth weight in the Seychelles data. Although more than one permutation of the cytokines was plausible in our application, subsets of the cytokines did routinely group together. IL-1*β* (***w*3**) and IL-10 (***w*6**) were classified in the same group for over 96% of retained draws, and IL-1*β* (***w*3**), IL-10 (***w*6**), and TNF-*α* (***w*2**) were classified in the same group for over 91% of retained draws. IL-2

(***w*4**) was classified with IL-5 (***w*5**) in over 96% of retained draws.

In future work, we intend to expand the proposed model to allow for more com- plicated relationships among the measured exposures. Instead of assuming that each exposure belongs to one of a fixed number of groups, we plan to allow for partial group membership, and model additional latent groups. These extensions will allow more cytokines of interest to be modeled, such as those not routinely considered Th1 or Th2. Our extended model will allow for greater understanding of cytokines in re- lation to birth outcomes in the Seychelles NC2 cohort. More broadly, results from our model may drive hypothesis generation and future scientific discovery by enabling researchers to better understand relationships between an outcome and broad classes of exposures, as well as the membership of exposure variables in each class.

# Funding

This work was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Grant T32ES007271, Grant P30ES001247, and Grant R03ES027514. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

# References

* 1. A. Berger, *Science commentary: Th1 and Th2 responses: what are they?*, Allergy 55 (2000), pp. 2–10.
  2. J.F. Bobb, L. Valeri, B. Claus Henn, D.C. Christiani, R.O. Wright, M. Mazumdar, J.J. Godleski, and B.A. Coull, *Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures*, Biostatistics 16 (2015), pp. 493–508.
  3. C. Carrico, C. Gennings, D.C. Wheeler, and P. Factor-Litvak, *Characterization of weighted*

*quantile sum regression for highly correlated data in a risk analysis setting*, Journal of Agricultural, Biological, and Environmental Statistics 20 (2015), pp. 100–120.

* 1. R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu, *Measurement Error in Nonlinear Models: a Modern Perspective*, Chapman and Hall/CRC, 2006.
  2. J.M. Cavaillon, *Pro-versus anti-inflammatory cytokines: myth or reality.*, Cellular and Molecular Biology (Noisy-le-grand, France) 47 (2001), p. 695.
  3. C. Chatfield and A.J. Collins, *Factor analysis*, in *Introduction to Multivariate Analysis*, Springer, 1980, pp. 82–87.
  4. J. Cote-Sierra, G. Foucras, L. Guo, L. Chiodetti, H.A. Young, J. Hu-Li, J. Zhu, and W.E. Paul, *Interleukin 2 plays a central role in Th2 differentiation*, Proceedings of the National Academy of Sciences 101 (2004), pp. 3880–3885.
  5. P.W. Davidson, G.J. Myers, C. Cox, C. Axtell, C.F. Shamlaye, J. Sloane-Reeves, E. Cernichiari, L. Needham, A. Choi, Y. Wang, M. Berlin, and T.W. Clarkson, *Effects of prenatal and postnatal methylmercury exposure from fish consumption at 66 months of age: the seychelles child development study*, JAMA 280 (1998), pp. 701–707.
  6. L.A. Dent, *For better or worse: common determinants influencing health and disease in parasitic infections, asthma and reproductive biology*, Journal of reproductive immunology 57 (2002), pp. 255–272.
  7. A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, *Bayesian data analysis*, CRC press, 2013.
  8. E.A. Gibson, Y. Nunez, A. Abuawad, A.R. Zota, S. Renzetti, K.L. Devick, C. Gennings,

J. Goldsmith, B.A. Coull, and M.A. Kioumourtzoglou, *An overview of methods to address distinct research questions on environmental mixtures: an application to persistent organic pollutants and leukocyte telomere length*, Environmental Health 18 (2019), p. 76.

* 1. A.B. Hollingshead, *Four factor index of social status* (1975).
  2. K. Jedidi, H.S. Jagpal, and W.S. DeSarbo, *Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity*, Marketing Science 16 (1997), pp. 39–59.
  3. I.T. Joliffe and B.J.T. Morgan, *Principal component analysis and exploratory factor anal- ysis*, Statistical Methods in Medical Research 1 (1992), pp. 69–95.
  4. H.H. Keller, *The SCREEN I (Seniors in the Community: Risk Evaluation for Eating and Nutrition) index adequately represents nutritional risk*, Journal of Clinical Epidemiology 59 (2006), pp. 836–841.
  5. A. Kelly, A. Lynch, E. Vereker, Y. Nolan, P. Queenan, E. Whittaker, L.A.J. O’Neill, and

M.A. Lynch, *The anti-inflammatory cytokine, interleukin (Il)-10, blocks the inhibitory effect of Il-1β on long term potentiation: A role for JNK*, Journal of Biological Chemistry 276 (2001), pp. 45564–45572.

* 1. P. Kidd, *Th1/Th2 balance: the hypothesis, its limitations, and implications for health and disease*, Alternative Medicine Review 8 (2003), pp. 223–46. Available at h[ttps://www.ncbi.nlm.nih.gov/pubmed/12946237.](http://www.ncbi.nlm.nih.gov/pubmed/12946237)
  2. R.W. Kobrosly, E. van Wijngaarden, S. Galea, D.A. Cory-Slechta, T. Love, C. Hong,

C.F. Shamlaye, and P.W. Davidson, *Socioeconomic position and cognitive function in the seychelles: a life course analysis*, Neuroepidemiology 36 (2011), pp. 162–168.

* 1. S.Y. Lee and X.Y. Song, *Bayesian model selection for mixtures of structural equation models with an unknown number of components*, British Journal of Mathematical and Statistical Psychology 56 (2003), pp. 145–165.
  2. D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter, *The BUGS book: A practical introduction to Bayesian analysis*, CRC press, 2012.
  3. E.M. McSorley, A.J. Yeates, M.S. Mulhern, E. van Wijngaarden, K. Grzesik, S.W. Thurston, T. Spence, W. Crowe, P.W. Davidson, G. Zareba, G.J. Myers, G.E. Watson,

C.F. Shamlaye, and J.J. Strain, *Associations of maternal immune response with mehg ex- posure at 28 weeks gestation in the seychelles child development study*, American Journal of Reproductive Immunology 80 (2018), p. e13046.

* 1. R. Sabat, G. Gru¨tz, K. Warszawska, S. Kirsch, E. Witte, K. Wolk, and J. Geginat, *Biology*

*of interleukin-10*, Cytokine & growth factor reviews 21 (2010), pp. 331–344.

* 1. B.N. S´anchez, E. Budtz-Jørgensen, L.M. Ryan, and H. Hu, *Structural equation models: a review with applications to environmental epidemiology*, Journal of the American Statis- tical Association 100 (2005), pp. 1443–1455.
  2. G.V. Shmarina, A.L. Pukhalsky, S.N. Kokarovtseva, D.A. Pukhalskaya, L.A. Shabalova,

N.I. Kapranov, and N.J. Kashirskaja, *Tumor necrosis factor-α/interleukin-10 balance in normal and cystic fibrosis children*, Mediators of Inflammation 10 (2001), pp. 191–197.

* 1. G.R. Souto, C.M. Queiroz-Junior, M.H.N.G. de Abreu, F.O. Costa, and R.A. Mesquita,

*Pro-inflammatory, Th1, Th2, Th17 cytokines and dendritic cells: a cross-sectional study in chronic periodontitis*, PloS One 9 (2014), p. e91636.

* 1. M. Stephens, *Dealing with label switching in mixture models*, Journal of the Royal Statis- tical Society: Series B (Statistical Methodology) 62 (2000), pp. 795–809.
  2. J.J. Strain, A.J. Yeates, E. van Wijngaarden, S.W. Thurston, M.S. Mulhern, E.M. McSor- ley, G. Watson, T.M. Love, T.H. Smith, K. Kelley Yost, D. Harrington, C.F. Shamlaye,

J. Henderson, G.J. Myers, and P.W. Davidson, *Prenatal exposure to methyl mercury from*

*fish consumption and polyunsaturated fatty acids: associations with child development at 20 mo of age in an observational study in the Republic of Seychelles*, The American Journal of Clinical Nutrition 101 (2015), pp. 530–537.

* 1. S. Stringhini, S. Polidoro, C. Sacerdote, R.S. Kelly, K. van Veldhoven, C. Agnoli, S. Grioni,

R. Tumino, M.C. Giurdanella, S. Panico, A. Mattiello, D. Palli, G. Masala, V. Gallo, R. Castagn, F. Paccaud, G. Campanella, M. Chadeau-Hyam, and P. Vineis, *Life-course so- cioeconomic status and DNA methylation of genes regulating inflammation*, International Journal of Epidemiology 44 (2015), pp. 1320–1330.

* 1. L. Sykes, D.A. MacIntyre, X.J. Yap, T.G. Teoh, and P.R. Bennett, *The Th1:Th2 dichotomy of pregnancy and preterm labour*, Mediators of Inflammation 2012 (2012), p. 967629. Available at h[ttps://www.ncbi.nlm.nih.gov/pubmed/22719180](http://www.ncbi.nlm.nih.gov/pubmed/22719180) [http://spiral.imperial.ac.uk/bitstream/10044/1/12876/.](http://spiral.imperial.ac.uk/bitstream/10044/1/12876/)
  2. E. van Wijngaarden, D. Harrington, R. Kobrosly, S.W. Thurston, T. O’Hara, E.M. Mc- Sorley, G.J. Myers, G.E. Watson, C.F. Shamlaye, J. Strain, and P.W. Davidson, *Prenatal exposure to methylmercury and LCPUFA in relation to birth weight*, Annals of Epidemi- ology 24 (2014), pp. 273–278.
  3. T.G. Wegmann, H. Lin, L. Guilbert, and T.R. Mosmann, *Bidirectional cy- tokine interactions in the maternal-fetal relationship: is successful pregnancy a*

*Th2 phenomenon?*, Immunology Today 14 (1993), pp. 353–6. Available at h[ttps://www.ncbi.nlm.nih.gov/pubmed/8363725.](http://www.ncbi.nlm.nih.gov/pubmed/8363725)

* 1. J. Weuve, B.N. S´anchez, A.M. Calafat, T. Schettler, R.A. Green, H. Hu, and R. Hauser,

*Exposure to phthalates in neonatal intensive care unit infants: urinary concentrations of monoesters and oxidative metabolites*, Environmental Health Perspectives 114 (2006), pp. 1424–1431.

* 1. A.J. Yeates, E.M. McSorley, M.S. Mulhern, T. Spence, W. Crowe, K. Grzesik, S.W. Thurston, G.E. Watson, G.J. Myers, P.W. Davidson, C.F. Shamlaye, E. van Wijngaar- den, and J.J. Strain, *Associations between maternal inflammation during pregnancy and infant birth outcomes in the Seychelles Child Development Study*, Journal of reproductive immunology 137 (2020), p. 102623.
  2. X. Zhou, M.S. Fragala, J.E. McElhaney, and G.A. Kuchel, *Conceptual and methodological issues relevant to cytokine and inflammatory marker measurements in clinical research*, Current Opinion in Clinical Nutrition and Metabolic Care 13 (2010), p. 541.

# Appendix A. Extension to *K >* 2 latent variables

We extend the notation given in Section 2.1 to *K >* 2 latent exposure variables, which we denote as *x*1*,i*, *x*2*,i, ..., xK,i*. As before, *yi* is the outcome, *wi,j* is the *j*th measured exposure variable, *j* = 1*, ..., J* where *J ≥ K*, ***Zy,i*** = (*Zy*1*,i, Zy*2*,i, ..., Zyp,i*) are covariates associated with the outcome, and ***Zx,i*** = (*Zx*1*,i, Zx*2*,i, ..., Zxq,i*) are covariates

associated with the latent exposures. All variables are centered and scaled. Also, ***αj*** =

= (*τ*

(*α*1*,j, ..., αK,j*), ***τ* 2**

***j***

*K,j*

2

1*,j*

*, ..., τ* 2

) and ***xi*** = (*x*1*,i, ..., xK,i*) are all *K*-length column

vectors. Our models for *yi*, *xk,i*, and *wi,j*, *k* = 1*, ..., K*; *j* = 1*, ..., J* are

*yi|x*1*,i, x*2*,i, ..., xK,i,* ***Zy,i****,* ***β****, σ*2 *∼ N* (*x*1*,iβ*1*,x* + *x*2*,iβ*2*,x* + *...* + *xK,iβK,x* + ***Zy,iβy****, σ*2) *xk,i|****Zx,i****,* ***γk****, η*2 *∼ N* (***Zx,iγk****, η*2)

*k*

*k*

*wi,j|x*1*,i, ..., xK,i,* ***αj****,* ***τ* 2** *∼ N* (***zj***(***αj*** *0* ***xi***)*,* ***zjτ* 2**)

***j***

***j***

where *0* denotes the Hadamard (element-wise) product and ***zj***, the indicator of expo- sure membership for the *j*th exposure, is the *K*-length row vector (*zj,*1*, zj,*2*, ..., zj,K*) where the *k*th element of ***zj*** is equal to 1 if the *j*th measured exposure is classified

in the *k*th latent group. All other elements of ***zj*** are equal to zero. We model

***zj*** *∼* Multinomial(1*,* ***pj*** = (*pj,*1*, ..., pj,K*))*, j* = 1*, ..., J*.

Posteriors for most model parameters follow the same form as shown in Section 2.2 and Appendix B, except with *K >* 2, the prior on each non-sentinel ***pj*** is now Dirichlet (i.e. ***pj*** *∼* Dirichlet(*a*1*, ..., aK*)). Consequently, *p*(***pj****|****zj***) *∼* Dirichlet(*a*1 + *zj,*1*, ..., aK* + *zj,K*) and each corresponding *p*(***zj*** *θ,* ***pj***) Multinomial where the denominator for

*| ∼*

the posterior as shown in Section 2.2 is a sum of *K* terms, for *K >* 2.

# Appendix B. Posteriors and select derivations for proposed model with 2 latent variables

**Posterior for the first latent variable (***x*1*,i***)**

*p*(*x*1*,i|****yi****,* ***Zy,i****,* ***Zx,i****, θ*) *∝ p*(*yi|****β****, σ*2*, x*1*,i, x*2*,i,* ***Zy,i***)*p*(*w* 1*,i|x*1*,i, α* 1*,*1*, τ*˜2

1*,*1

(*∝ −*

) *· · · p*(*w* *K,i|x*1*,i, α* 1*,K, τ*˜2

)*p*(*x*1*,i|****Zx,i****,* ***γ*1***, η*2)

1

exp 1 (*y*

2*σ*2 *i*

* *x*1*,i*

*βx*1

1*,K*

* *x*2*,i*

*βx*2

* ***Zy,iβy***

)2

1 (*w* 1*,i − x*1*,iα* 1*,*1)2

*τ*˜

(*× −*

*×* exp

*−* 2

2

1*,*1

+ *...* +

(*w* *K,i − x*1*,iα* 1*,K* )2 )

exp 1 (*x* 2*η*2

1

1*,i*

* ***Zx,iγ*1**

*τ*˜

2

1*,K*

)2

= exp ( *−* 1 (*y*2 *−* 2*y x*

2*σ*2

*i*

*i*

1*,i*

*β −* 2*y x*

*β −* 2*y* ***Z***

***β*** + *x*2

*β*2+

2*x*1*,iβx x*2*,iβx*

1

*i*

2*,i*

2

*i*

***y,i***

***z***

1*,i*

1

+ 2*x*1*,iβx* ***Zy,iβy*** + *x*2 *β*2

+ 2*x*2*,iβx* ***Zy,iβy*** + ***Zy,iβyβT Zy,i***)

1 2 1

2*,i x*2 2 ***y***

1 (*w* 1*,i − x*1*,iα* 1*,*1)2

*τ*˜

*×* exp

*−* 2

2

1*,*1

+ *...* +

(*w* *K,i − x*1*,iα* 1*,K* )2 )

exp 1 (*x*

(*× −*

2*η*2

1

1*,i*

* ***Zx,iγ*1**

*τ*˜

2

1*,K*

)2

*∝* exp (*−*  1 (*x*2 *β*2 *−* 2*x*

2*σ*2

1*,i*

*x*1

1*,i*

1

*i*

2*,i*

*x*2

***y,i***

***y***

*β* (*y − x β −* ***Z β*** )

1 *α* 2

*x*

*×* exp

*−* 2

2

1*,i*

1*,*1

*τ*˜2

*−* 2*x*1*,i*

1*,i*

*τ*˜2

+ *...* +

*K,i*

*τ*˜2

1*,K*

*α*2 ) *w α*

*w α* ) )

*×* exp (*−*  1 (*x*2

1

2*η*2

1*,*1

+ *...* + 1*,K*

*τ*˜2

1*,*1

*−* 2*x* (***Z***

1*,i*

1*,i*

1*,K*

***γ*** ))

*α*2

***x,i***

**1**

1*,*1

1*,K*

1 *β*2

*x*1

1*,*1

1*,K*

*τ*˜

1 ) 2 )

*η*

*∝* exp *−* 2

*σ τ*˜

2 + 2

1*,*1

*α*2

*τ*˜

*τ*˜

*η*

1*,i*

*σ*2

2

1*,*1

2

1*,K*

2

1

+ *...* + 2 + 2

1*,K* 1

*x*1*,i*

*×* exp 1 2*x*

*−* 2

*−*

*βx*1 (*yi − x*2*,iβx*2 *−* ***Zy,iβy*** ) + *w* 1*,iα* 1*,*1 + *...* + *w* *K,iα* 1*,K* + ***Zx,iγ*1** )

The posterior for *x*1*,i* is *N* (*µx*

1*,i*

*, σ*2 ) where

2

*x*1

*β*

2 *x*1 +

*σ*

=

*x*1 *σ*2

*α*2

+ *...* +

1*,*1

*τ*˜

2

1*,*1

*α*2

2

1*,K*

*τ*˜

1*,K*

1 *−*1

2

+

)

*η*

1

*µ* = *σ*2 *×*  *βx*1 (*yi − x*2*,iβx*2 *−* ***Zy,iβy*** ) + *w* 1*,iα* 1*,*1 + *...* + *w* *K,iα* 1*,K* + ***Zx,iγ*1** )

*τ*˜

*τ*˜

*η*

*x*1*,i*

*x*

*σ*2

2

1*,*1

2

1*,K*

2

1

# Posterior for the second latent variable (*x*2*,i*)

The posterior for *x*2*,i* is *N* (*µx*

2*,i*

*, σ*2 ) where

2

*x*2

*β*

2 *x*2 +

*σ*

=

*x*2 *σ*2

*α*2

+ *...* +

2*,K*+1

*τ*˜

2

2*,K*+1

*τ*˜

*τ*˜

*η*

*x*2

*σ*2

2

2*,K*+1

2

2*,J*

2

2

*α*2

2

2*,J*

*τ*˜

2*,J*

1 *−*1

2

+

)

*η*

2

*µ* = *σ*2

*x*2*,i*

*×*  *βx*2 (*yi − x*1*,iβx*1 *−* ***Zy,iβy*** ) + *w* *K*+1*,iα* 2*,K*+1 + *...* + *w* *J,iα* 2*,J* + ***Zx,iγ*2** )

# Posterior for *β*

*p*(***β****|****y****,* ***X****, θ*) *∼ N* (***M*** *,* ***V*** ) where

***V*** = *σ*2(***XX*** + *σ*2**Σ*β****−*1)*−*1

***M*** = ***V*** (*σ−*2***Xy*** + **Σ*β****−*1***β*0**)

# Posterior for *σ*2

2

*p*(*σ*

*|****y****,* ***X****, θ*) *∼ IG*

2

( *n* (***y*** *−* ***Xβ***)(***y*** *−* ***Xβ***) + 2*bσ*2

**Posterior for** *α*1*,j*

2 + *aσ*2 *,*

*p*(*α*1*,j|****wj****, θ*) *∼ N* (*Mα*1 *, Vα*1 ) where

*Vα* = *τ* 2

(***xi x*1** + *τ* 2

(*σ*2 )*−*1)*−*1

1 1*,j* **1** 1*,j α*

*Mα* = *Vα* ((*τ* 2 )*−*1***xi wj*** + (*σ*2 )*−*1*α*0)

1 1 1*,j* **1** *α*

**Posterior for** *α*2*,j*

*p*(*α*2*,j|****wj****, θ*) *∼ N* (*Mα*2 *, Vα*2 ) where

*Vα* = *τ* 2

(***xi x*2** + *τ* 2

(*σ*2 )*−*1)*−*1

2 2*,j* **2** 2*,j α*

*Mα* = *Vα* ((*τ* 2 )*−*1***xi wj*** + (*σ*2 )*−*1*α*0)

# Posterior for *τ* 2

1*,j*

2 2 2*,j* **2** *α*

*p*(*τ* 2 *|****wj****, θ*) *∼ IG*

1*,j*

*n*

2 + *aτ*2 *,*

(

(***wj*** *−* ***x*1***α*1*,j*)(***wj*** *−* ***x*1***α*1*,j*) + 2*bτ*2 2

# Posterior for *τ* 2

2*,j*

*p*(*τ* 2 *|****wj****, θ*) *∼ IG*

2*,j*

*n*

2 + *aτ*2 *,*

(

(***wj*** *−* ***x*2***α*2*,j*)(***wj*** *−* ***x*2***α*2*,j*) + 2*bτ*2 2

# Posterior for *γ*1

*p*(***γ*1***|****Zx****, θ*) *∼ N* (***Mγ*1** *,* ***Vγ*1** ) where

***Vγ*** = *η*2(***Zi Zx*** + *η*2**Σ*−*1)*−*1**

**1** 1 ***x*** 1 ***γ***

***Mγ*** = ***Vγ*** (*η−*2***Zi x*1 + Σ*−*1*γ*0)**

**1 1** 1 ***x γ***

# Posterior for *γ*2

*p*(***γ*2***|****Zx****, θ*) *∼ N* (***Mγ*2** *,* ***Vγ*2** ) where

## Vγ

= *η*2(***Zi Zx*** + *η*2**Σ*−*1)*−*1**

**2** 2 ***x*** 2 ***γ***

***Mγ*** = ***Vγ*** (*η−*2***Zi x*2 + Σ*−*1*γ*0)**

**2 2** 2 ***x γ***

# Posterior for *η*2

1

*p*(*η*2*|****Zx****, θ*) *∼ IG*

1

*n*

2 + *aη*2 *,*

(

(***x*1** *−* ***Zxγ*1**)(***x*1** *−* ***Zxγ*1**) + 2*bη*2 2

# Posterior for *η*2

2

*p*(*η*2*|****Zx****, θ*) *∼ IG*

2

*n*

2 + *aη*2 *,*

(

(***x*2** *−* ***Zxγ*2**)(***x*2** *−* ***Zxγ*2**) + 2*bη*2 2

# Posterior for *pj*

*p*(*pj|zj*) *∝ p*(*zj|pj*)*p*(*pj*)

*zj* 1*−zj c−*1 *d−*1

*∝ p* (1 *− pj*) *p* (1 *− pj*)

*j*

*j*

*∝ pzj* +*c−*1(1 *− pj* )1*−zj* +*d−*1

*j*

*∝ p*(*c*+*zj* )*−*1(1 *− pj* )*d*+(1*−zj* )*−*1

*j*

Since *p*(*pj|zj*) is proportional to a Beta kernel, *p*(*pj|zj*) *∼* Beta(*c* + *zj, d* + (1 *− zj*)).

# Posterior for *zj*

*p*(*z*

= 1*|****w*** *, θ, p* ) = *p*(*zj* = 1*|θ, p, j*)

*j* ***j*** *j*

*p*(*zj* = 1*|θ, p, j*) + *p*(*zj* = 0*|θ, p, j*)

(*τ* 2

*n*

)*− n* exp (*−* (*wi,j−α*1*,jx*1*,i*)2 *p*

2*τ*

*n*

=

(*τ* 2

1*,j*

)*−* 2 exp

(*− i*(*w*

*i,j*

1*,j*

*−α*1*,j*

2*τ*

2

1*,j*

2

*x*1*,i*)2 *p*

*i*

+ (*τ* 2

2*τ*

2*,j*

2

2*,j*

2

1*,j*

*j*

*j*

)*−* 2 exp

*j*

(*− i*(*w*

*i,j*

*−α*2*,j*

*x*2*,i*

)2

(1 *− p* )

# Appendix C. Proposed permutation probabilities

We would like to determine the probability that, after *T* iterations, all *P* permutations have been sampled with replacement. This is equivalent to the coupon collectors prob- lem, which is described in detail elsewhere (see *Randomized algorithms*, Cambridge University Press, 1995). Briefly, let random variable *X* represent the number of trials required to collect at least one of each type of *n* unique coupons. It has been shown elsewhere (pg. 60 of *Randomized algorithms*, Cambridge University Press, 1995) that, for any constant *c ∈* ***R*** and *m* = *n* ln *n* + *nc*,

*Pr*[*X > m*] *≈* 1 *− e−e−c*

For our problem, we let *X∗* denote the total number of proposals required to ensure that each of the 32 unique *zj* permutations has been proposed at least once. For our application, there are 32 unique *zj* permutations so *n* = 32. By running our chain for

*T* = 16*,* 000 iterations and proposing a new permutation every 5 iterations, we are proposing 16*,* 000*/*5 = 3*,* 200 permutations. Thus, we should set *m* = 3*,* 200.

*≈*

Given that *m* = 3*,* 200, we can solve for *c*:

3*,* 200 = *n* ln *n* + *nc*

= 32 ln 32 + 32*c c* = (3*,* 200 *−* 32 ln 32) 96*.*53

*→ ≈*

32

Thus, *Pr*[*X > m*] *≈* 1 *− e−e−c* = 1 *− e−e−*96*.*53 = 1*.*20 *∗* 10*−*42 *≈* 0.

Thus, the probability that we need more than 3,200 proposals to propose each of 32 unique permutations at least one time is (essentially) zero.

# Reference:

Motwani, Rajeev, and Prabhakar Raghavan. *Randomized algorithms*. Cambridge Uni- versity Press, 1995.

# Appendix D. Simulation study variance derivation

We derive *τ* 2 as an example. Recall that *V ar*(*x*1) = 0*.*5, *V ar*(*w*2) = 1, and *w*2*,i* =

1*,*2

*α*1*,*2*x*1*,i* + *E*1*,*2 where *E*1*,*2 *∼ N* (0*, τ* 2

1*,*2

). Thus,

*V ar*(*w*2*,i|α*1*,*2) = *V ar*(*α*1*,*2*x*1*,i* + *E*1*,*2)

= *V ar*(*α*1*,*2*x*1*,i*) + *V ar*(*E*1*,*2) by independence

2

= *α*

1*,*2

2

= *α*

1*,*2

1*,*2

1*,*2

*V ar*(*x*1*,i*) + *V ar*(*E*1*,*2) *V ar*(*x*1*,i*) + *τ* 2

2

1*,*2

so *τ*

1*,*2

= *V ar*(*w*2*,i|α*1*,*2) *− α*2

*V ar*(*x*1*,i*) = 1 *−* 0*.*5*α*2 .

# Appendix E. Additional Tables

**Table E1.** Simulation results under prior A (IG(0.05, 0.01)) for our model (“Proposed”), the fixed group model (“Fixed”), and the oracle model (“Oracle”). **Mean estimates for our model are also provided (“Estimate”)**. Reported values for our model are averages over 50 simulated datasets using all retained draws. For identifiability, *α*1*,*1 and *α*2*,*7 are fixed at 1 in our model and in the fixed group model.

Bias Interval Length Coverage

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Truth | **Estimate** | Proposed | Fixed | Oracle | Proposed | Fixed | Oracle | Proposed | Fixed | Oracle |
| *βx*1 0.050 | 0.045 | -0.005 | -0.005 | 0.013 | 0.343 | 0.344 | 0.282 | 0.960 | 0.960 | 0.940 |
| *βx*2 -0.030 | -0.030 | 0.000 | -0.000 | -0.013 | 0.331 | 0.332 | 0.282 | 0.940 | 0.940 | 0.980 |
| *βZy*1 0.150 | 0.152 | 0.002 | 0.002 | 0.003 | 0.172 | 0.172 | 0.172 | 1.000 | 1.000 | 1.000 |
| *βZy*2 -0.100 | -0.098 | 0.002 | 0.002 | 0.003 | 0.173 | 0.172 | 0.172 | 0.940 | 0.960 | 0.940 |
| *βZy*3 0.100 | 0.096 | -0.004 | -0.004 | -0.005 | 0.173 | 0.173 | 0.172 | 0.920 | 0.920 | 0.920 |
| *σ*2 1.000 | 0.954 | -0.046 | -0.046 | -0.044 | 0.240 | 0.240 | 0.240 | 1.000 | 0.980 | 1.000 |
| *α*1*,*1 1.000 | - | - | - | 0.005 | - | - | 0.177 | - | - | 1.000 |
| *α*1*,*2 0.600 | 0.602 | 0.002 | 0.006 | -0.004 | 0.436 | 0.428 | 0.227 | 0.920 | 0.920 | 0.920 |
| *α*1*,*3 0.800 | 0.822 | 0.022 | 0.025 | -0.004 | 0.566 | 0.551 | 0.207 | 0.940 | 0.940 | 1.000 |
| *α*2*,*4 0.500 | 0.515 | 0.015 | 0.017 | -0.002 | 0.347 | 0.345 | 0.234 | 1.000 | 1.000 | 1.000 |
| *α*2*,*5 0.700 | 0.737 | 0.037 | 0.038 | 0.001 | 0.381 | 0.382 | 0.217 | 0.980 | 0.980 | 0.980 |
| *α*2*,*6 0.850 | 0.874 | 0.024 | 0.025 | -0.006 | 0.420 | 0.419 | 0.201 | 0.980 | 0.980 | 0.960 |
| *α*2*,*7 1.000 | - | - | - | 0.003 | - | - | 0.176 | - | - | 1.000 |
| *τ* 2 0.500 | 0.490 | -0.010 | -0.005 | 0.001 | 0.374 | 0.358 | 0.125 | 0.900 | 0.860 | 0.860 |
| *τ* 2 0.820 | 0.830 | 0.010 | 0.008 | 0.005 | 0.242 | 0.241 | 0.206 | 1.000 | 1.000 | 1.000 |
| *τ* 2 0.680 | 0.681 | 0.001 | -0.000 | 0.007 | 0.274 | 0.267 | 0.172 | 0.960 | 0.960 | 1.000 |
| *τ* 2 0.875 | 0.877 | 0.002 | 0.002 | 0.003 | 0.235 | 0.235 | 0.220 | 1.000 | 1.000 | 1.000 |
| *τ* 2 0.755 | 0.749 | -0.006 | -0.006 | 0.001 | 0.225 | 0.224 | 0.189 | 0.980 | 0.980 | 1.000 |
| *τ* 2 0.639 | 0.648 | 0.009 | 0.009 | 0.009 | 0.226 | 0.224 | 0.162 | 0.980 | 1.000 | 0.960 |
| *τ* 2 0.500 | 0.522 | 0.022 | 0.023 | 0.001 | 0.241 | 0.241 | 0.125 | 0.960 | 0.960 | 0.940 |
| *γ*1*,*1 0.071 | 0.065 | -0.006 | -0.006 | -0.003 | 0.152 | 0.152 | 0.117 | 0.940 | 0.940 | 0.960 |
| *γ*1*,*2 0.212 | 0.207 | -0.005 | -0.005 | -0.001 | 0.159 | 0.160 | 0.117 | 0.940 | 0.920 | 0.960 |
| *γ*2*,*1 0.283 | 0.272 | -0.011 | -0.011 | -0.001 | 0.151 | 0.151 | 0.112 | 0.940 | 0.940 | 0.920 |
| *γ*2*,*2 0.106 | 0.104 | -0.002 | -0.002 | 0.000 | 0.141 | 0.141 | 0.112 | 0.940 | 0.940 | 0.960 |
| *η*2 0.450 | 0.460 | 0.010 | 0.005 | -0.004 | 0.388 | 0.370 | 0.112 | 0.900 | 0.920 | 0.900 |
| *η*2 0.409 | 2 0.389 | -0.020 | -0.021 | -0.003 | 0.252 | 0.252 | 0.102 | 0.980 | 0.980 | 0.940 |

1*,*1

1*,*2

1*,*3

2*,*4

2*,*5

2*,*6

2*,*7

1

**Table E2.** Simulation results under prior B (IG(0.5, 0.1)) for our model (“Proposed”), the fixed group model (“Fixed”), and the oracle model (“Oracle”). **Mean estimates for our model are also provided (“Estimate”)**. Reported values for our model are averages over 50 simulated datasets using all retained draws. For identifiability, *α*1*,*1 and *α*2*,*7 are fixed at 1 in our model and in the fixed group model.

Bias Interval Length Coverage

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Truth | **Estimate** | Proposed | Fixed | Oracle | Proposed | Fixed | Oracle | Proposed | Fixed | Oracle |
| *βx*1 0.050 | 0.045 | -0.005 | -0.005 | 0.013 | 0.345 | 0.344 | 0.282 | 0.960 | 0.960 | 0.940 |
| *βx*2 -0.030 | -0.030 | -0.000 | -0.000 | -0.013 | 0.331 | 0.332 | 0.282 | 0.940 | 0.940 | 0.980 |
| *βZy*1 0.150 | 0.152 | 0.002 | 0.002 | 0.003 | 0.172 | 0.172 | 0.172 | 1.000 | 1.000 | 1.000 |
| *βZy*2 -0.100 | -0.098 | 0.002 | 0.002 | 0.003 | 0.172 | 0.172 | 0.172 | 0.960 | 0.960 | 0.940 |
| *βZy*3 0.100 | 0.096 | -0.004 | -0.004 | -0.005 | 0.173 | 0.173 | 0.172 | 0.920 | 0.920 | 0.920 |
| *σ*2 1.000 | 0.952 | -0.048 | -0.046 | -0.044 | 0.239 | 0.240 | 0.240 | 0.980 | 0.980 | 1.000 |
| *α*1*,*1 1.000 | - | - | - | 0.005 | - | - | 0.177 | - | - | 1.000 |
| *α*1*,*2 0.600 | 0.607 | 0.007 | 0.006 | -0.004 | 0.435 | 0.428 | 0.227 | 0.940 | 0.920 | 0.920 |
| *α*1*,*3 0.800 | 0.830 | 0.030 | 0.025 | -0.004 | 0.561 | 0.551 | 0.207 | 0.940 | 0.940 | 1.000 |
| *α*2*,*4 0.500 | 0.516 | 0.016 | 0.017 | -0.002 | 0.348 | 0.345 | 0.234 | 1.000 | 1.000 | 1.000 |
| *α*2*,*5 0.700 | 0.738 | 0.038 | 0.038 | 0.001 | 0.384 | 0.382 | 0.217 | 0.980 | 0.980 | 0.980 |
| *α*2*,*6 0.850 | 0.875 | 0.025 | 0.025 | -0.006 | 0.422 | 0.419 | 0.201 | 0.980 | 0.980 | 0.960 |
| *α*2*,*7 1.000 | - | - | - | 0.003 | - | - | 0.176 | - | - | 1.000 |
| *τ* 2 0.500 | 0.497 | -0.003 | -0.005 | 0.001 | 0.358 | 0.358 | 0.125 | 0.900 | 0.860 | 0.860 |
| *τ* 2 0.820 | 0.827 | 0.007 | 0.008 | 0.005 | 0.242 | 0.241 | 0.206 | 1.000 | 1.000 | 1.000 |
| *τ* 2 0.680 | 0.677 | -0.003 | -0.000 | 0.007 | 0.272 | 0.267 | 0.172 | 0.960 | 0.960 | 1.000 |
| *τ* 2 0.875 | 0.876 | 0.001 | 0.002 | 0.003 | 0.235 | 0.235 | 0.220 | 1.000 | 1.000 | 1.000 |
| *τ* 2 0.755 | 0.747 | -0.008 | -0.006 | 0.001 | 0.225 | 0.224 | 0.189 | 0.980 | 0.980 | 1.000 |
| *τ* 2 0.639 | 0.647 | 0.008 | 0.009 | 0.009 | 0.225 | 0.224 | 0.162 | 1.000 | 1.000 | 0.960 |
| *τ* 2 0.500 | 0.522 | 0.022 | 0.023 | 0.001 | 0.241 | 0.241 | 0.125 | 0.960 | 0.960 | 0.940 |
| *γ*1*,*1 0.071 | 0.064 | -0.006 | -0.006 | -0.003 | 0.151 | 0.152 | 0.117 | 0.940 | 0.940 | 0.960 |
| *γ*1*,*2 0.212 | 0.207 | -0.005 | -0.005 | -0.001 | 0.159 | 0.160 | 0.117 | 0.900 | 0.920 | 0.960 |
| *γ*2*,*1 0.283 | 0.272 | -0.011 | -0.011 | -0.001 | 0.151 | 0.151 | 0.112 | 0.940 | 0.940 | 0.920 |
| *γ*2*,*2 0.106 | 0.104 | -0.002 | -0.002 | 0.000 | 0.141 | 0.141 | 0.112 | 0.940 | 0.940 | 0.960 |
| *η*2 0.450 | 0.451 | 0.001 | 0.005 | -0.004 | 0.370 | 0.370 | 0.112 | 0.920 | 0.920 | 0.900 |
| *η*2 0.409 | 2 0.387 | -0.022 | -0.021 | -0.003 | 0.250 | 0.252 | 0.102 | 0.980 | 0.980 | 0.940 |

1*,*1

1*,*2

1*,*3

2*,*4

2*,*5

2*,*6

2*,*7

1

**Table E3.** Seychelles parameter effective sample sizes across only retained draws corresponding to the top permutation (“Top Draws”) and across all retained draws (“All Draws”) for prior A (IG(0.05, 0.01)) and prior B (IG(0.5, 0.1)).

Prior A Prior B

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Top Draws | All Draws | Top Draws | All Draws |
| *βx*1 | 27298 | 2606 | 23744 | 2089 |
| *βx*2 | 10911 | 2135 | 9833 | 2621 |
| *βZy*1 | 52305 | 79000 | 44748 | 79000 |
| *βZy*2 | 52305 | 79000 | 44231 | 77079 |
| *βZy*3 | 51510 | 79000 | 43714 | 77471 |
| *βZy*4 | 52305 | 79000 | 44748 | 79000 |
| *βZy*5 | 52305 | 67330 | 41737 | 53776 |

*σ*2 52305 79000 44748 78167

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *α*1*,*2 | 4148 | 28 | 4130 | 20 |
| *α*1*,*3 | 2015 | 20 | 1804 | 15 |
| *α*1*,*6 | 1839 | 19 | 1671 | 13 |
| *α*2*,*4 | 462 | 432 | 419 | 396 |
| *α*2*,*5 | 3790 | 1052 | 3428 | 1059 |

2

*τ*

|  |  |  |  |
| --- | --- | --- | --- |
| 28354 | 147 | 23093 | 127 |
| 50623 | 10 | 41539 | 9 |
| 21936 | 7 | 19128 | 8 |
| 19156 | 6 | 16995 | 5 |
| 423 | 81 | 469 | 54 |
| 21902 | 207 | 18672 | 133 |

1*,*1

2

*τ*

1*,*2

2

*τ*

1*,*3

2

*τ*

1*,*6

2

*τ*

2*,*4

2

*τ*

2*,*5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *τ* 2 | 743 | 50 | 691 | 33 |
| *γ*1*,*1 | 31407 | 247 | 25849 | 224 |
| *γ*1*,*2 | 32333 | 2687 | 28168 | 1830 |
| *γ*1*,*3 | 13997 | 15794 | 11461 | 14281 |
| *γ*1*,*4 | 31633 | 2083 | 27325 | 1421 |
| *γ*1*,*5 | 14055 | 323 | 13689 | 263 |
| *γ*1*,*6 | 30182 | 79 | 27523 | 80 |
| *γ*2*,*1 | 27531 | 8114 | 23247 | 5481 |
| *γ*2*,*2 | 26171 | 4250 | 22947 | 3996 |
| *γ*2*,*3 | 5050 | 372 | 5213 | 262 |
| *γ*2*,*4 | 7369 | 464 | 7611 | 310 |
| *γ*2*,*5 | 10570 | 6671 | 11348 | 6088 |
| *γ*2*,*6 | 4135 | 193 | 4249 | 125 |

2*,*7

2

*η*

|  |  |  |  |
| --- | --- | --- | --- |
| 1585 | 482 | 1429 | 365 |
| 2 614 | 61 | 567 | 42 |

1

2

*η*

**Table E4.** Seychelles parameter estimates and 95% posterior intervals under prior A (IG(0.05, 0.01)) and prior B (IG(0.5, 0.1)). Reported values are calculated using all retained draws. For identifiability, *α*1*,*1 and *α*2*,*7 are fixed at 1.

Prior A Prior B

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Posterior Mean | 95% Interval | Posterior Mean | 95% Interval |
| *βx*1 | -0.014 | (-0.161, 0.125) | -0.021 | (-0.172, 0.121) |
| *βx*2 | -0.015 | (-0.156, 0.158) | -0.008 | (-0.153, 0.178) |
| *βZy*1 | 0.142 | (0.093, 0.190) | 0.142 | (0.093, 0.190) |
| *βZy*2 | 0.487 | (0.438, 0.535) | 0.487 | (0.438, 0.535) |
| *βZy*3 | 0.095 | (0.045, 0.144) | 0.094 | (0.045, 0.144) |
| *βZy*4 | 0.071 | (0.021, 0.121) | 0.071 | (0.021, 0.121) |
| *βZy*5 | 0.039 | (-0.010, 0.088) | 0.039 | (-0.010, 0.089) |

*σ*2 0.724 (0.668, 0.784) 0.724 (0.668, 0.784)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *α*1*,*1 | 1.000 | - | 1.000 | - |
| *α*1*,*2 | 1.365 | (0.437, 1.984) | 1.236 | (0.418, 1.969) |
| *α*1*,*3 | 1.037 | (0.242, 1.499) | 0.929 | (0.227, 1.488) |
| *α*1*,*6 | 1.127 | (0.302, 1.604) | 1.015 | (0.288, 1.593) |
| *α*2*,*4 | 1.183 | (0.598, 1.640) | 1.152 | (0.740, 1.596) |
| *α*2*,*5 | 0.492 | (0.253, 0.663) | 0.479 | (0.251, 0.653) |
| *α*2*,*7 | 1.000 | - | 1.000 | - |

2

*τ*

|  |  |  |  |
| --- | --- | --- | --- |
| 0.795 | (0.654, 0.893) | 0.782 | (0.638, 0.891) |
| 0.530 | (0.286, 0.974) | 0.590 | (0.288, 0.980) |
| 0.725 | (0.577, 1.024) | 0.762 | (0.579, 1.031) |
| 0.678 | (0.511, 1.011) | 0.719 | (0.513, 1.019) |
| 0.554 | (0.246, 0.963) | 0.606 | (0.261, 0.968) |
| 0.925 | (0.830, 1.039) | 0.933 | (0.835, 1.046) |

1*,*1

2

*τ*

1*,*2

2

*τ*

1*,*3

2

*τ*

1*,*6

2

*τ*

2*,*4

2

*τ*

2*,*5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *τ* 2 | 0.708 | (0.528, 0.961) | 0.736 | (0.528, 0.967) |
| *γ*1*,*1 | 0.019 | (-0.034, 0.057) | 0.015 | (-0.039, 0.056) |
| *γ*1*,*2 | -0.012 | (-0.047, 0.029) | -0.010 | (-0.046, 0.034) |
| *γ*1*,*3 | 0.082 | (0.043, 0.122) | 0.082 | (0.041, 0.123) |
| *γ*1*,*4 | -0.002 | (-0.040, 0.044) | 0.000 | (-0.039, 0.050) |
| *γ*1*,*5 | -0.063 | (-0.102, -0.012) | -0.059 | (-0.101, -0.006) |
| *γ*1*,*6 | -0.036 | (-0.100, 0.006) | -0.041 | (-0.105, 0.005) |
| *γ*2*,*1 | 0.019 | (-0.023, 0.056) | 0.020 | (-0.022, 0.055) |
| *γ*2*,*2 | -0.024 | (-0.067, 0.016) | -0.023 | (-0.065, 0.012) |
| *γ*2*,*3 | 0.080 | (0.035, 0.136) | 0.076 | (0.034, 0.134) |
| *γ*2*,*4 | -0.032 | (-0.089, 0.011) | -0.028 | (-0.086, 0.011) |
| *γ*2*,*5 | -0.055 | (-0.098, -0.016) | -0.055 | (-0.097, -0.019) |
| *γ*2*,*6 | -0.046 | (-0.101, 0.000) | -0.041 | (-0.098, 0.001) |

2*,*7

2

*η*

|  |  |  |  |
| --- | --- | --- | --- |
| 0.227 | (0.160, 0.334) | 0.235 | (0.161, 0.351) |
| 2 0.288 | (0.067, 0.473) | 0.260 | (0.066, 0.470) |

1

2

*η*

**Table E5.** Summary of Seychelles cytokine permutations for Prior A (IG(0.05, 0.01)) and Prior B (IG(0.5, 0.1)). IL1-*β* (*z*3) and IL-4 (*z*7) were

fixed at 1 and 0 respectively, while classification of IFN-*γ* (*z*1), TNF-*α* (*z*2), IL-2 (*z*4), IL-5 (*z*5) and IL-10 (*z*6) as a Th1 cytokine (*zj* = 1) or a Th2

cytokine (*zj* = 0) was determined from MCMC sampling. The numbers in the “Prior A” and “Prior B” columns are the number of samples (out of 79,000) from the corresponding *zj* permutation.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Perm. Rank | *z*1 | *z*2 | *z*3 | *z*4 | *z*5 | *z*6 | *z*7 | Prior A | Prior B |
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 40209 | 38093 |
| 2 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 32338 | 34276 |
| 3 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 2411 | 2335 |
| 4 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1861 | 1987 |
| 5 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 506 | 499 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 469 | 789 |
| 7 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 348 | 368 |
| 8 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 300 | 5 |
| 9 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 191 | 207 |
| 10 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 56 | 53 |
| 11 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 43 | 46 |
| 12 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 38 | 45 |
| 13 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 37 | 38 |
| 14 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 34 | 57 |
| 15 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 34 | 33 |
| 16 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 28 | 30 |
| 17 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 27 | 27 |
| 18 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 21 | 36 |
| 19 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 14 | 19 |
| 20 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 12 | 14 |
| 21 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 6 | 7 |
| 22 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 6 | 6 |
| 23 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 5 | 4 |
| 24 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 4 | 2 |
| 25 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 2 | 4 |
| 26 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 12 |
| 27 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 8 |

# This table was not included in the original submission

**Table E6.** Summary of Seychelles cytokine permutations for Prior A (IG(0.05, 0.01)) and Prior B (IG(0.5, 0.1)) when only *Zx*1 and *Zx*2 are

included in ***Z****x*. IFN-*γ* (*z*1) and IL-4 (*z*7) were fixed at 1 and 0 respec- tively, while classification of TNF-*α* (*z*2), IL-1*β* (*z*3), IL-2 (*z*4), IL-5 (*z*5) and IL-10 (*z*6) as a Th1 cytokine (*zj* = 1) or a Th2 cytokine (*zj* = 0) was determined from MCMC sampling. The numbers in the “Prior A” and

“Prior B” columns are the number of samples (out of 79,000) from the corresponding *zj* permutation.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Perm. Rank | *z*1 | *z*2 | *z*3 | *z*4 | *z*5 | *z*6 | *z*7 | Prior A | Prior B |
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 54791 | 49432 |
| 2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 19612 | 25473 |
| 3 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 2765 | 2489 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 638 | 251 |
| 5 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 362 | 568 |
| 6 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 322 | 275 |
| 7 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 110 | 210 |
| 8 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 100 | 0 |
| 9 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 47 | 38 |
| 10 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 40 | 50 |
| 11 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 40 | 18 |
| 12 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 31 | 32 |
| 13 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 25 | 50 |
| 14 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 21 | 27 |
| 15 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 17 | 13 |
| 16 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 17 | 24 |
| 17 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 17 | 12 |
| 18 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 15 | 9 |
| 19 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 13 | 14 |
| 20 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 5 | 2 |
| 21 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 4 | 3 |
| 22 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 4 | 4 |
| 23 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 2 | 6 |
| 24 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 25 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |