ZIDM-vignette

Introduction

In this vignette, we provide worked examples on simulated data and the gut microbiome data set analyzed in the main manuscript, `A Bayesian Zero-Inflated Dirichlet-Multinomial Regression Model for Multivariate Compositional Count Data`, to demonstrate how to apply the proposed methods. Specifically, the software contains functionality to implement the proposed zero-inflated Dirichlet-multinomial (ZIDM) model, a standard Dirichlet-multinomial (DM) model, and their extensions to sparsity-induced regression settings, ZIDMbvs and DMbvs (Wadsworth et al. 2017), respectively. Additionally, we provide functionality to implement the approach of Tuyl (2019) for individual-level count probability estimation and a supplementary Monte Carlo sampler for uncertainty estimation.

Data Simualtion

In this section, we first describe how to simulate data investigated in the accompanying manuscript's simulation and sensitivity studies. Having installed and loaded the `ZIDM` package into the R environment (See README for instructions), generate a simulated data set for a zero-inflated Dirichlet-multinomial model using the `simulate_ZIDM()` function.

```
library(ZIDM)
```

```
data_ZIDM <- simulate_ZIDM( rho = 0.3, rho_theta = 0.3)
```

By default 50 subjects (`n_obs = 50`) with 100 taxa (`n_taxa = 100`) and 50 covariates in both levels of the model (`n_vars = 50` and `n_vars_theta = 50`) are simulated. The function requires specification of `rho` (`rho_theta`) or `Sigma` (`Sigma_theta`), where `rho` sets the covariance matrix between simulated covariates as $\Sigma_{ij} = \rho^{|i-j|}$ and `Sigma` is simply a given covariance matrix. Of the 50×100 potential covariate-compositional element associations, 4 of the covariates are associated with 4 of the simulated compositional elements and the zero-inflation indicators, totally 32 active terms across both levels of the model. The total number of counts for each observation is sampled from a Uniform(400,500), where the minimum and maximum values can be adjusted with `n_reads_min` and `n_reads_max`, respectively. Additional arguments are available to control the number of compositional elements with active covariates (`n_relevant_taxa` and `n_relevant_taxa_theta`), the number of active covariates (`n_relevant_vars` and `n_relevant_vars_theta`), minimum (`beta_min` and `beta_min_theta`) and maximum (`beta_max` and `beta_max_theta`) true regression coefficients, signal-to-noise ratio (`signoise`), minimum (`int_zero_min`) and maximum (`int_zero_max`) values for intercept terms $\beta_{\theta 0}$ to control the proportion of at-risk zeros, and the dispersion factor (`theta0`).

In addition to the simulated data, the `simulate_ZIDM()` function also returns the true regression coefficients $\boldsymbol{\beta}_\theta$ (`betas_theta`), $\boldsymbol{\beta}_\gamma$ (`betas`), zero-inflation indicators (`eta`), probabilities used to simulate the zero-inflation indicators (`prob`), probabilities used to simulated the multivariate count data (`beta_0`), and dispersion parameter (`theta0`).

Parameter Estimation

First we demonstrate how to use the proposed ZIDM model for estimation of the population-level zero-inflation probabilities, population-level count probabilities, and individual-level count probabilities. Using the proposed model's notation, these quantities are $\Theta_j = 1/(1+\exp(\beta_{\theta j0}))$ (`post_theta`), $\Gamma_j = \exp(\beta_{\gamma j0})/(\sum_{j=1}^{J} \exp(\beta_{\gamma j0}))$

(post_gamma), and $\psi_{ij} = c_{ij}/T_i$ (post_psi) for all $i = 1, \ldots, N$ and $j = 1, \ldots, J$, respectively. Note that this approach ignores potential covariates' influence in both levels of the model. By default, the method provides 95% credible intervals for each of the estimated parameters (post_theta_lower, post_theta_upper, post_gamma_lower, post_gamma_upper, post_psi_lower, post_psi_upper). By default ZIDM_R runs 10000 MCMC iterations, thins to every 10^{th} iteration, and assumes $\sigma_{\beta_\gamma} = \sigma_{\beta_\theta} = \sqrt{5}$.

```
# Fit the ZIDM model to the data
fit_ZIDM <- ZIDM_R( data_ZIDM$Z )

# Obtain estimates from ZIDM
ZIDM_est <- estimates_ZIDM( zidm_obj = fit_ZIDM, burnin = 500, CI = 0.95  )
```

Compared to the truth, we find the model obtained a Frobenious norm of 1.31, 0.04, and 0.70 for Θ_j, Γ_j, and ψ_{ij}, respectively.

In addition, we provide functionality to estimate $\Gamma_j = \exp(\beta_{\gamma j0})/(\sum_{j=1}^{J} \exp(\beta_{\gamma j0}))$ and $\psi_{ij} = c_{ij}/T_i$ with a Bayesian DM model, as well as ψ_{ij} with Tuyl's approach. Note that we constructed a Monte Carlo sampling algorithm to obtain uncertainty estimates using Tuyl's approach. Estimates from the DM model are obtained with estimates_DM().

```
# Fit the DM model to the data
fit_DM <- DM_R( data_ZIDM$Z )

# Obtain estimates of DM model
DM_est <- estimates_DM( dm_obj = fit_DM, burnin = 500, CI = 0.95  )

# Fit Tuyl's approach to the data and obtain uncertainty estimates via Monte Carlo sampling
fit_tuyl <- tuyl_meaner( data_ZIDM$Z )
uncertainty_tuyl <- tuyl( 40000 , data_ZIDM$Z )
```

Compared to the truth, we find the DM model obtained a Frobenious norm of 0.08 and 0.72 for Γ_j and ψ_{ij}, respectively, and Tuyl's approach obtained a Frobenius norm of 0.71 for ψ_{ij}.

Variable Selection

Next, we demonstrate how to use our approach to identify covariates associated with zero-inflation and compositional counts. The ZIDMbvs_R() function requires a matrix of counts and covariates for both levels of the model (X and X_theta). Note, X and X_theta do not have to be the same, and the function automaticially includes intercept terms. By default the model is run for 10000 iterations, thinning to every 10^{th} iteration. We assume $\sigma_{\beta_\gamma} = \sigma_{\beta_\theta} = \sqrt{5}$ and noninformative prior probabilities of inclusion (i.e., $a_\varphi = b_\varphi = a_\zeta = b_\zeta = 1$). The output of the model contains MCMC samples for φ, β_γ, η, β_θ, ζ, ω, c, and the acceptance probability of η (eta_accept).

```
fit_ZIDMbvs <-  ZIDMbvs_R( Z = data_ZIDM$Z, X = data_ZIDM$X[,-1], X_theta = data_ZIDM$X_theta[,-1]  )
```

Inclusion is determined with the marginal posterior probability of inclusion (MPPI) for each compositional element-by-covariate inclusion indicator. By default, the MPPI threshold for significant terms is set to 0.50. To obtain MPPIs for both levels of the model from the ZIDMbvs_R output, run

```
MPPI_zeta <- apply( fit_ZIDMbvs$zeta[ ,, 501:1000 ], c(1,2), mean )
MPPI_varphi <- apply( fit_ZIDMbvs$varphi[ ,, 501:1000 ], c(1,2), mean )
```

The selection performance of the model is evaluated using the select_perf() function which calculates the sensitivity, specificity, Matthew's correlation coefficient (MCC), as well as the F1 score (as defined in the main manuscript). To obtain these results, simply supply the selected terms and true active terms. For example in this analysis, we obtained a 0.75 sensitivity, 0.96 specificity, 0.20 MCC, and 0.11 F1 for the associations between the compositional count and covariates.

```
# Selection performance for covariates associated with zero-inflation indicators using ZIDMbvs
select_perf_zeta <- select_perf( selected = (MPPI_zeta[ , -1 ] > 0.5)*1,
                                 truth =  (data_ZIDM$betas_theta[ , -1 ] != 0 )*1 )
# Selection performance for covariates associated with compositional counts using ZIDMbvs
select_perf_varphi <- select_perf( selected = (MPPI_varphi[ , -1 ] > 0.5)*1,
                                 truth =  (data_ZIDM$betas[ , -1 ] != 0 )*1 )

select_perf_varphi$sens
#> [1] 0.75
select_perf_varphi$spec
#> [1] 0.961878
select_perf_varphi$mcc
#> [1] 0.2041967
select_perf_varphi$f1
#> [1] 0.1100917
```

Additionally, users can implement the DMbvs method of Wadsworth et al. (2017) using the following code:

```
fit_DMbvs <-  DMbvs_R( Z = data_ZIDM$Z, X = data_ZIDM$X[,-1])
```

The model has similar default settings at `ZIDMbvs_R`, where applicable. The selection performance of the `DMbvs` model is calculated as

```
# Selection performance for covariates associated with compositional counts using DMbvs
MPPI_varphi_DM <- apply( fit_DMbvs$varphi[ ,, 501:1000 ], c(1,2), mean )
select_perf_varphi_DM <- select_perf( selected = (MPPI_varphi_DM[ , -1 ] > 0.5)*1,
                                      truth =  (data_ZIDM$betas[ , -1 ] != 0 )*1 )

select_perf_varphi_DM$sens
#> [1] 0.375
select_perf_varphi_DM$spec
#> [1] 0.9773274
select_perf_varphi_DM$mcc
#> [1] 0.1305472
select_perf_varphi_DM$f1
#> [1] 0.08888889
```

Here, we see how the ZIDMbvs greatly outperforms the DMbvs approach in the presence of zero-inflation.

Application Study

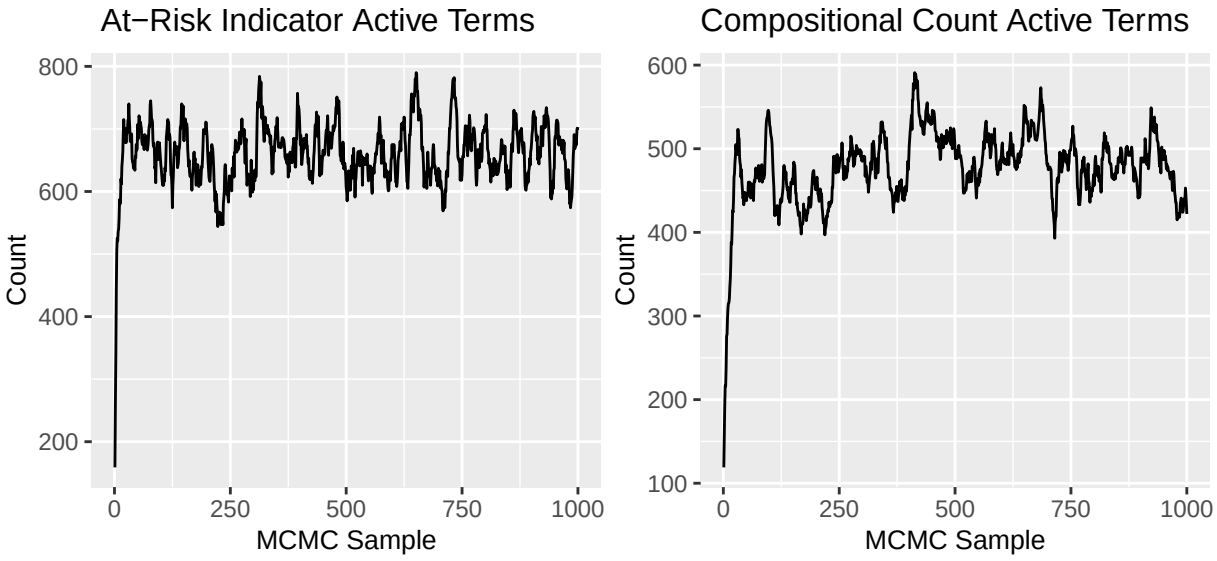
In this section, we demonstrate how to apply the proposed ZIDMbvs method to the data investigated in the application study in the main manuscript. First, load the data into the working environment

```
data("Gut_dietary")
data("Gut_micro")
```

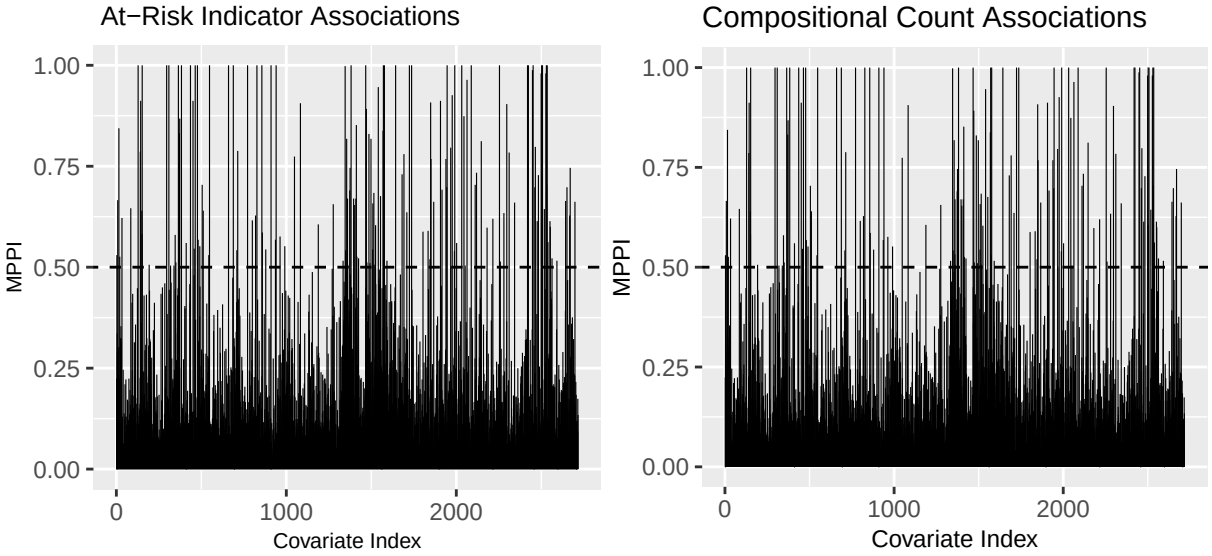
The `Gut_micro` data set contains 28 taxa count reads for 98 participants and the `Gut_dietary` data set contains their corresponding 97 dietary covariates. To investigate these data, simply run:

```
fit_gut <-  ZIDMbvs_R( Z = Gut_micro, X = Gut_dietary, X_theta = Gut_dietary )
```

To demonstrate the convergence of the algorithm, we plot the number of active terms in the model over MCMC iterations



The plots of the corresponding MPPIs for both levels of the model are presented below. The horizontal dotted line indicates the selection threshold. Covariates with corresponding MPPIs above 0.50 are considered active in the model.



References

Tuyl, Frank. 2019. "A Method to Handle Zero Counts in the Multinomial Model." *The American Statistician.*

Wadsworth, Duncan, Raffaele Argiento, Michele Guindani, Jessica Galloway-Pena, Samuel A Shelburne, and Marina Vannucci. "An Integrative Bayesian Dirichlet-Multinomial Regression Model for the Analysis of Taxonomic Abundances in Microbiome Data." *BMC Bioinformatics* 18 (1): 94.