A simulated data example in “Power Link Functions in Modeling Dependent Ordinal Data”

Dan Li, Xia Wang and Dipak K. Dey

April 13, 2018

Contents

1 Background

We simulate an ordinal data example with 1-dimensional correlation (ordinal time series data) based on the splogit model framework: $z_i = \beta_0 + \beta_1 x_i + f_i + \varepsilon_i$, and $y_i = j$, if $\gamma_{j-1} < z_i \leq \gamma_j$, where $\mathbf{f} = (f_1, \ldots, f_n)' \sim \mathcal{N}\left(0, \sigma^2 \Sigma(\rho)\right)$, the random error term $\varepsilon_i \sim F_{Splogit}(r)$, $i = 1, \ldots, n$ and $j = 1, \ldots, J$.

We fit both splogit and logit models to the data. Two parallel Markov chains are run with dispersed initial values for each model fit. For each single chain, we run 1,000 iterations and discard the first 500 as a warm-up phase, yielding a total of 1,000 posterior samples which are used to calculate posterior summaries of parameters and model comparison measurements. We calculate DIC and LPML for model comparison.

The simulated data example has $n = 400$, $J = 3$, $r = 0.5$, $\gamma_0 = -\infty$, $\gamma_1 = 0$, $\gamma_3 = +\infty$, and x_i generated from $\mathcal{N}(0, 1)$. The true values of the model parameters are: $\beta_0 = 2$, $\beta_1 = 2$, $r = 0.5$, $\gamma_2 = 3$, $\sigma^2 = 2$ and $\rho = 0.2$.

2 Simulated data

Load R packages and related R functions:
Simulate an ordinal time series data set under the splogit model with $r = 0.5$:

```
source('../R/ordinal_splogit.R')
rng_seed <- 20180306
set.seed(rng_seed)

n <- 400
## parameters shared by all models
beta <- c(2, 2)       # regression parameters
eta <- c(0, 3)        # cut points
x1 <- rnorm(n, 0, 1) # covariate
r <- 0.5

## time series ##
t <- seq(-1, 1, length = n)
## GP hyperparameters
sig_sq <- 2            # magnitude parameter
rho_sq <- 0.2          # length-scale parameter
jitter <- 0.0001       # jitter

subjects <- data.frame(id = 1:n, x1 = x1, t = t)

dd <- subjects
dd$Y <- NA
setup <- setup.ordinal.timeseries(Y ~ 1 + x1 | t, data = dd)
#                                   X*beta | location
dd2 = simulate.ordinal.timeseries(setup,
                                  beta = beta,
                                  eta = eta,
                                  a = sig_sq,
                                  b = rho_sq,
                                  r = r,
```

```
                                                   jitter = jitter,
                                                   seed = rng_seed)

dd$Y <- dd2$y
write.table(dd, file = "data_example.txt", row.names = F)
```

Show the first 20 lines of the data set:

```
   id          x1          t Y
1   1 -0.62338386 -1.0000000 2
2   2  0.30089121 -0.9949875 2
3   3  0.16246220 -0.9899749 2
4   4 -0.42849552 -0.9849624 2
5   5 -1.14209539 -0.9799499 2
6   6  0.44252310 -0.9749373 2
7   7  0.46554700 -0.9699248 2
8   8  0.42176921 -0.9649123 2
9   9  0.61079337 -0.9598997 2
10 10 -1.77304212 -0.9548872 1
11 11  1.24612062 -0.9498747 3
12 12 -0.09155203 -0.9448622 2
13 13  0.54770287 -0.9398496 1
14 14  0.79113478 -0.9348371 2
15 15  0.15111209 -0.9298246 2
16 16 -0.12422253 -0.9248120 1
17 17 -1.62659638 -0.9197995 1
18 18 -1.11504624 -0.9147870 2
19 19  0.45256499 -0.9097744 2
20 20 -1.46054365 -0.9047619 1
```

The table below summarizes the number of observations in each category:

```
  1   2   3
139 168  93
```

3 Fit model with Stan

3.1 Fit splogit model

The posterior mean, 95% confidence interval and the potential scale reduction statistic $\hat{R}$ of the key model parameters:

```
             mean       2.5%     97.5%      Rhat
beta[1] 1.6853382 -1.4564231 4.6484117 1.0021604
beta[2] 1.8431947  1.3038494 2.5400788 0.9998491
cuts[1] 0.0000000  0.0000000 0.0000000       NaN
cuts[2] 2.9690858  2.1252158 4.0787068 1.0002965
```

```
eta_sq   4.5507466  0.5933430 18.7566185 1.0014146
rho_sq   0.8392007  0.1564435  2.1092701 1.0005173
r        0.5171032  0.2430674  0.9554493 1.0024284
```

3.2 Fit logit model

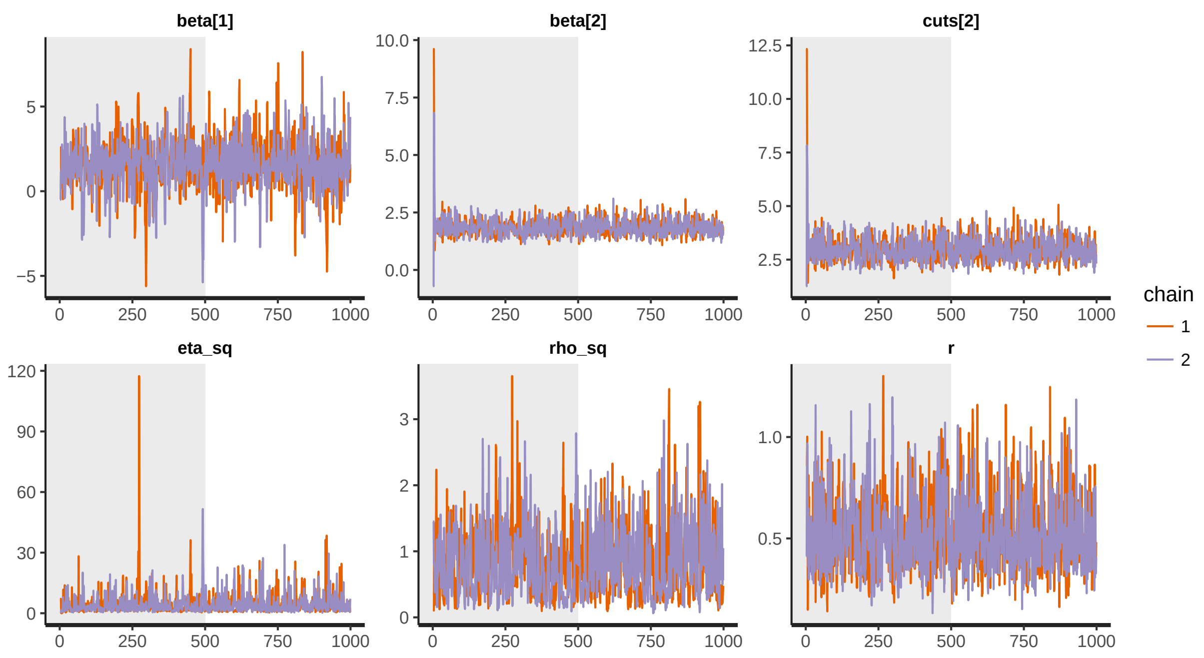
The posterior mean, 95% confidence interval and the potential scale reduction statistic $\hat{R}$ of the key model parameters:

```
             mean       2.5%     97.5%      Rhat
beta[1] 1.2320401 -1.8643585  4.383007 1.0064763
beta[2] 2.6218609  2.2534731  3.014763 0.9985367
cuts[1] 0.0000000  0.0000000  0.000000       NaN
cuts[2] 4.2210959  3.6231779  4.839127 1.0047667
eta_sq  8.5380285  1.0712264 65.826786 1.0794162
rho_sq  0.7063771  0.1190431  1.580570 1.0172482
```

4 Diagnostic plots

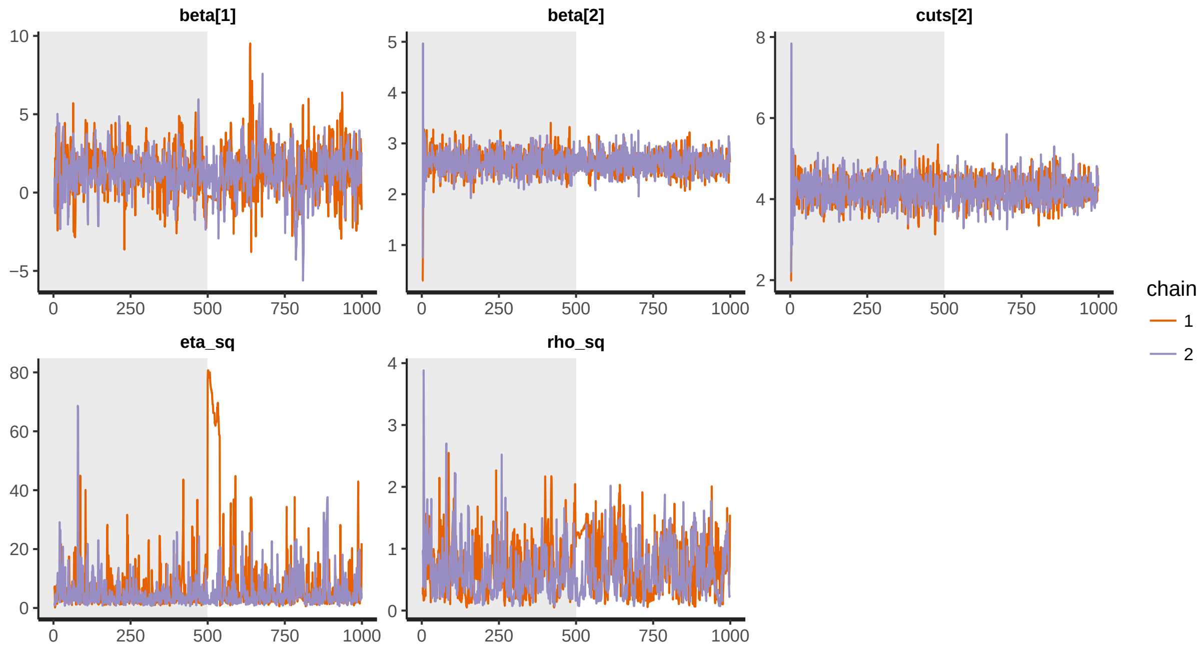
4.1 Trace plots for parameters of splogit model

```
#### trace plot ####
traceplot(fit_splogit,
          pars = c('beta[1]', 'beta[2]', 'cuts[2]', 'eta_sq', 'rho_sq', 'r'),
          inc_warmup = TRUE)
```



4.2 Trace plots for parameters of logit model

```
#### trace plot ####
traceplot(fit_logit,
          pars = c('beta[1]', 'beta[2]', 'cuts[2]', 'eta_sq', 'rho_sq'),
          inc_warmup = TRUE)
```



5 Model comparision criteria

```
criteria.splogit <- comparison.splogit(fit_splogit, dd, dependent = TRUE)
criteria.logit <- comparison.logit(fit_logit, dd, dependent = TRUE)
tab <- data.frame(DIC = rep(NA, 2), LPML = rep(NA, 2))
rownames(tab) <- c("splogit", "logit")
tab[1,1] <- criteria.splogit$DIC
tab[1,2] <- criteria.splogit$LPML
tab[2,1] <- criteria.logit$DIC
tab[2,2] <- criteria.logit$LPML
```

The minimum DIC and maximum LPML indicate the best fit. Splogit model (true model) fits better than logit model.

```
kable(tab, format = "latex", booktabs = T)
```

	DIC	LPML
splogit	463.9696	-232.2134
logit	468.2883	-234.1064