Czado:

# Multivariate Probit Analysis of Binary Time Series Data with Missing Responses

Sonderforschungsbereich 386, Paper 23 (1996)

Online unter: http://epub.ub.uni-muenchen.de/

# Multivariate Probit Analysis of Binary Time Series Data with Missing Responses

**Claudia Czado**
**Department of Mathematics and Statistics, York University, 4700 Keele Street, North York, Ontario, M3J 1P3, Canada** [1]

**SUMMARY**

The development of adequate models for binary time series data with covariate adjustment has been an active research area in the last years. In the case, where interest is focused on marginal and association parameters, generalized estimating equations (GEE) (see for example Lipsitz, Laird and Harrington (1991) and Liang, Zeger and Qaqish (1992)) and likelihood (see for example Fitzmaurice and Laird (1993) and Molenberghs and Lesaffre (1994)) based methods have been proposed. The number of parameters required for the full specification of these models grows exponentially with the length of the binary time series. Therefore, the analysis is often focused on marginal and first order parameters. In this case, the multivariate probit model (Ashford and Sowden (1970)) becomes an attractive alternative to the above models. The application of the multivariate probit model has been hampered by the intractability of the maximum likelihood estimator, when the length of the binary time series is large. This paper shows that this difficulty can be overcome by the use of Markov Chain Monte Carlo methods. This analysis also allows for valid point and interval estimates of the parameters in small samples. In addition, the analysis is adopted to handle the case of missing at random responses. The approach is illustrated on data involving binary responses measured at unequally spaced time points. Finally, this data analysis is compared to a GEE analysis given in Fitzmaurice and Lipsitz (1995).

**Keywords:** multivariate binary regression, multivariate probit model, tetrachoric correlation, Bayesian analysis, Markov Chain Monte Carlo methods, missing data.

## 1 Introduction

Binary time series are often observed in biomedical studies. The six cities study (Ware et al. (1984)) investigating the effects of indoor and outdoor air pollution on respiratory health and the asthma studies (Korn and Whittmore (1979)) exploring the effects of air pollution on the occurence of asthma attacks are two well known examples of longitudinal studies with binary outcomes.

Often the basic sampling unit is the individual and measurements of response and covariates are collected on each individual over time. Time constant covariates, such as sex and (initial) age, and time varying covariates, such as daily pollution methods, can occur. In addition, measurement times do not need to be equally spaced.

[1]Parts of this paper were written while C. Czado was visiting the SFB "Statistische Analyse Diskreter Strukturen" at the Maximilians-Universität München, Germany. C. Czado was supported by research grant OGP0089858 of the Natural Sciences and Engineering Research Council of Canada. email: czado@mathstat.yorku.ca

Another commonly occurring problem in longitudinal studies is the problem of missing responses as arising from attrition or by design (Laird (1988)). Fitzmaurice and Lipsitz (1995) present such data from a double blind clinical trial comparing auranofin and placebo therapy for the treatment of rheumatoid arthritis (Bombardier et al. (1986)). In this study, patients self-assess their condition as "poor" or "good" at most five unequally spaced time points during the course of the study with a large proportion of missing responses. The primary question of interest in this study was, whether auranofin increases the probability of a positive self-assessment, while secondary questions concern whether the self-assessment depends on age and sex.

For this purpose, population averaged or marginal approaches are most appropriate in contrast to transitional or cluster specific approaches (for a review see Ashby et al. (1991)). Following a marginal approach, likelihood and non-likelihood based methods have been developed for modelling multivariate regression data with binary or ordinal response. Non-likelihood based methods as general estimation equations (GEE) have been extensively used in this context (see for example Lipsitz, Laird, Harrington (1991), Liang, Zeger and Qaqish (1992), Carey, Zeger and Diggle (1993), Fitzmaurice and Laird (1995) and Lipsitz et al. (1995)).

In general, likelihood based methods are often preferred (see for example the comments to Liang, Zeger and Qaqish (1992)). In the context of handling missing data, likelihood based methods are superior to methods based on GEE's, since they remain valid if observations are missing at random (MAR), while for GEE methods missing values have to be missing completely at random (MCAR). See Little and Rubin (1987) for an introduction to the concepts of MAR and MCAR and for the result for likelihood based methods, and Liang and Zeger (1986) for the corresponding result for GEE methods. However, in a recent paper by Robins et al. (1995) a weighted GEE approach has been proposed to allow the missing responses to be MAR. In the special case of a monotone missing data pattern both likelihood and GEE approaches have been developed by Fitzmaurice, Molenberghs and Lipsitz (1995), while Baker (1995) models a general missing data mechanism. His approach, however, is only feasible in binary time series of length three or less.

Several likelihood based methods for handling regression data with discrete response have been developed. Here methods, which allow the dependence between the responses and the marginals to be independently modelled are only considered. This is in contrast to a model proposed by Prentice (1988). The earliest likelihood based method is the multivariate probit model (Ashford and Sowden (1970), Lesaffre and Molenberghs (1991)) for ordinal response, which models the dependence between binary responses with the help of the well defined correlation structure of underlying quantitative latent variables. This is approach is an extension of tetrachoric correlation (Pearson (1990)) to multivariate regression with discrete outcomes.

More recently, two different likelihood based models have been proposed. Both use odds ratios (see for example Dale (1986)) as measures of association between discrete variables. The one model developed by Molenberghs and Lesaffre (1994) is based on marginal odds ratios using a multivariate extension to the bivariate Plackett distribution (Plackett (1965)) for the construction of the joint likelihood. The other model put forward by Fitzmaurice and Laird (1993) for binary time series is formulated in terms of conditional odds ratios assuming a quadratic exponential model for the joint likelihood (Cox (1972), Zhao and Prentice (1990)). The extension of this approach to the ordinal response has been considered by Heagerty and Zeger (1995) and Heumann (1996). Fitzmaurice, Laird and Lipsitz (1994) use the above models in connection with the EM algorithm for binary time series with missing at random responses.

Even though the interpretation of the association parameters are more straight forward in the approaches based on odds ratios, the number of parameters required for the full specification of the likelihood grows exponentially with the length of the binary time series and in practice often only two way associations are assumed to be nonzero. In this case, the multivariate probit model becomes again an attractive choice.

Cessie and Houwelingen (1994) compared the two approaches for modelling the association, tetrachoric correlation and the odds ratio, for the simplest case of a bivariate binary response and showed that they are approximately equivalent using a first order approximation. They also considered the extension to binary time series of arbitrary length. A full maximum likelihood analysis in this case often proves to be infeasible, since it involves maximizing over multivariate normal probabilities (see also (2.2) in Chapter 2). Therefore, Cessie and Houwelingen (1994) used a pseudo likelihood approach instead. Full likelihood analysis of the multivariate probit model has been restricted to applications to response vectors of dimension at most three (for example see Anderson and Pemberton (1985)).

With the development of Markov Chain Monte Carlo (MCMC) methods, computationally tractable Bayesian analysis for many complex models have become available. For example see Besag, Green, Hidgon and Mengerson (1995) and the many references cited therein. The application of MCMC methods has now become standard practice for a Bayesian analyst (see for example Gelman et al. (1995) and Gelfand and Smith (1995)). An introduction to the Gibbs sampler is given by Casella and George (1992), while the Metropolis-Hastings algorithm is explained in Chib and Greenberg (1994). These analyses often remain feasible when maximum likelihood becomes infeasible and interval estimators are available even in small samples. Recently, in the context of repeated categorical responses a Bayesian analysis based on a cluster specific approach using random effects has been given by Becker and Ten Have (1995).

The goal of this paper is twofold. First, we will provide a computationally tractable analysis of the multivariate probit model for binary time series of arbitrary length and secondly, we adopt the analysis to handle missing at random responses.

The paper is organized as follows. Section 2 introduces the multivariate probit model for binary time series for the complete and missing response case. In Section 3, an algorithm to facilitate inference drawn from the posterior distribution is developed, first for the complete data case and then it is adopted to handle missing at random responses. The arthritis clinical trial will be analysed in Section 4 showing that the missing responses can not be ignored.

# 2 Multivariate Probit Model for Binary Time Series

In this section, the multivariate probit model for binary time series will be formulated. First, it will be assumed that the binary time series is completely observed. Secondly, the extension to binary time series with missing values will be considered.

## 2.1 The Complete Data Case

To formulate a Bayesian approach, we need to specify the joint distribution of the binary response vector. For this, let $Y_i = (y_{i1}, \cdots, y_{iT})^t$ the binary response vector with binary response,

$y_{it} = 1$ or 0, observed at time t and marginal probabilities $\pi_{it} = P(y_{it} = 1)$ for $i = 1, \cdots, n$ and $t = 1, \cdots, T$. We assume, that the response vectors $Y_i$ are independently observed. For each response component $y_{it}$, we have covariate information collected in the vector $(x_{it1}, \cdots, x_{itp})$ available. Some of these covariates might be time stationary. For example, if the jth covariate is time stationary, we have $x_{i1j} = \cdots = x_{iTj}$. We consider now marginal models of the following form

$$\pi_{it} = \Phi(\eta_{it}) \text{ where } \eta_{it}(\beta) = \beta_{0t} + \beta_{1t}x_{it1} + \cdots + \beta_{pt}x_{itp} \tag{2.1}$$

and $\Phi(\cdot)$ denotes the standard normal distribution function. This formulation is the most general, since it allows for both time varying regression parameters $\beta_{jt}$ as well as time varying covariates. Time stationary regression parameters can be achieved by requiring $\beta_{j1} = \cdots = \beta_{jT} = \beta_j$. For the arthritis data set, the model considered will include time varying covariates but only time stationary regression parameters are used.

To give the complete specification of the joint distribution, we introduce independent latent random vectors $Z_i = (Z_{i1}, \cdots, Z_{iT})$ which are jointly normally distributed with mean vector $-\eta_i(\beta) = (-\eta_{i1}(\beta), \cdots, -\eta_{iT}(\beta))^t$ and covariance matrix $\Sigma_i$ with unit diagonal entries. The dependence structure between the binary outcomes $y_{it}$ is modelled indirectly through the dependence structure among the latent variables $Z_{it}$. For this, we assume that

$$y_{it} = 1 \Longleftrightarrow Z_{it} < 0.$$

It is easy to see that this equivalence is consistent with the marginal specification given in (2.1). As in Cessie and Houwelingen (1994), joint probabilities can now be determined by the joint distribution of $Z_i$. For example

$$\begin{aligned} &P(y_{i1} = 1, \cdots, y_{iT} = 1) = P(Z_{i1} < 0, \cdots, Z_{iT} < 0) \\ &= \int_{-\infty}^{0} \cdots \int_{-\infty}^{0} \frac{1}{(2\pi)^{T/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(Z_i + \eta_i(\beta))^t \Sigma_i^{-1}(Z_i + \eta_i(\beta))\} dZ_{i1} \cdots dZ_{iT}. \end{aligned} \tag{2.2}$$

Other joint probabilities can be defined similarly. This represents the multivariate use of tetrachoric correlation in the regression setting. It should be noted, that even though probit margins are used in (2.1), the models for logistic margins or any other margins could have been formulated, by using appropriate transformations of the linear predictor $\eta_{it}(\beta)$ (see Cessie and Houwelingen (1994) for logistic margins).

The specification of the dependence structure $\Sigma_i$ allows for a wide range of association models. We present now some possibilities:

- Covariate independence: $\Sigma_i = \Sigma$
- Serial correlation pattern: $Cor(Z_{is}, Z_{it}) = \rho_i^{|s-t|}$
- Serial correlation pattern with covariate independence: $Cor(Z_{is}, Z_{it}) = \rho^{|s-t|}$.

The last pattern has been used by Fitzmaurice and Lipsitz (1995) for odds ratios. It is appropriate, when the binary responses are measured at unequally spaced time points.

Since the covariance matrix $\Sigma_i$ has unit diagonal entries, $\Sigma_i$ is the correlation matrix of the latent vector $Z_i$, therefore the (s,t)th element of $\Sigma_i$, denoted by $\rho_{ist}$, is restricted to the interval [-1,1]. It is therefore easier to consider a transformation of $\rho_{ist}$ into the real line to incorporate

covariate dependence of the correlation structure. Cessie and Houwelingen (1994) used the following one-to-one transformation

$$\tau_{ist} = \log(\frac{1+\rho_{ist}}{1-\rho_{ist}}).$$

A regression model for $\tau_{ist}$ can now be assumed, for example

$$\tau_{ist} = \alpha_{st0} + \alpha_{st1} W_i, \tag{2.3}$$

where $W_i$ is an appropriate covariate. Additional covariates can be incorporated in the same way. Marginal parameters as defined in (2.1) will be denoted by $\boldsymbol{\beta}$, while the association parameters defined in (2.3) will be denoted by $\boldsymbol{\alpha}$. Since the covariance matrices $\Sigma_i$ depend on $\boldsymbol{\alpha}$, we will denote them with $\Sigma_i(\boldsymbol{\alpha})$.

## 2.2 The Missing Response Case

Following the usual setup for missing data (see Little and Rubin (1987)), we denote by $Y_i$ the ith complete binary response vector, which can be written as $Y_i = (Y_{iobs}, Y_{imiss})$ for $i = 1, \cdots, n$. Here, $Y_{iobs}$ denotes the vector of responses which have been observed, while $Y_{imiss}$ denotes the vector of missing responses. Similarly, we classify also the underlying latent variables $Z_i$ as $Z_i = (Z_{iobs}, Z_{imiss})$. As likelihood methods, Bayesian methods which are based on the observed data only remain valid under the assumption of missing responses, which are MAR (for example see Gelman et. al (1995), Chapter 17).

Since we are interested in inference drawn from the posterior distribution based on a random sample from the posterior, realizations of the latent variables are available. Therefore, the problem of handling missing binary responses can be dealt with by solving the easier problem of handling missing latent variables. This will be done by generating the missing latent values. Since the distribution of the complete latent variables $Z_i$ is multivariate normal with mean $-\eta_i(\boldsymbol{\beta})$ and covariance matrix $\Sigma_i(\boldsymbol{\alpha})$, we have two ways of generating the missing latent variables $Z_{imiss}$.

First, we can simply generate a realization from the distribution of the complete latent variables $Z_i$, or secondly, we can impute the missing latent variables $Z_{imiss}$ by the expected value of the conditional distribution of $Z_{imiss}$ given $Z_{iobs}$, which is given by

$$E(Z_{imiss}|Z_{iobs}) = -\eta_{imiss}(\beta) + \Sigma_{iobsmiss}(\boldsymbol{\alpha})\Sigma_{iobs}(\boldsymbol{\alpha})^{-1}(Z_{iobs} + \eta_{iobs}(\boldsymbol{\beta})), \tag{2.4}$$

where

$$\eta_i(\boldsymbol{\beta}) = \begin{pmatrix} \eta_{iobs}(\beta) \\ \eta_{imiss}(\beta) \end{pmatrix} \text{ and } \Sigma_i(\boldsymbol{\alpha}) = \begin{pmatrix} \Sigma_{iobs}(\boldsymbol{\alpha}) & \Sigma_{iobsmiss}(\boldsymbol{\alpha}) \\ \Sigma_{iobsmiss}(\boldsymbol{\alpha}) & \Sigma_{imiss}(\boldsymbol{\alpha}) \end{pmatrix}.$$

It should be noted, that these ways of generating realizations for the missing variables need to be adopted to handle the conditioning on the observed binary responses $Y_{iobs}$ needed for the application of the MCMC methods. This will be done in Section 3.2.

# 3 Bayesian Inference using Monte Carlo Markov Chain Methods

## 3.1 The Complete Data Case

For the Bayesian analysis, we assume that the response $Y_i$ given the regression parameters $\boldsymbol{\beta}$ and the association parameters $\boldsymbol{\alpha}$ follow the multivariate probit model as specified in (2.1)-(2.3). The prior information about $(\beta, \alpha)$ is summarized in a joint density of the form $\pi(\beta, \boldsymbol{\alpha}) = \pi(\beta) \times \pi(\alpha)$. Noninformative and multivariate normal priors can be used.

MCMC methods allow to draw a sample from the posterior distribution $[\boldsymbol{\beta}, \boldsymbol{\alpha}, Z|Y]$, where $Z = (Z_1, \cdots, Z_n)^t$ and $Y = (Y_1, \cdots, Y_n)^t$. Here, $[u|w]$ denotes the conditional distribution of u given w. A Metropolis within Gibbs approach (Müller (1994)) is now taken, since the conditional distributions $[Z_i|Y_i, \boldsymbol{\beta}, \boldsymbol{\alpha}]$ and $[\boldsymbol{\beta}|\boldsymbol{\alpha}, Z, Y]$ are known, while $[\boldsymbol{\alpha}|\boldsymbol{\beta}, Z, Y]$ is known only up to a normalizing constant, thus requiring a Metropolis-Hastings step.

In particular, $[Z_i|Y_i, \boldsymbol{\beta}, \boldsymbol{\alpha}]$ is a truncated multivariate normal distribution with mean vector $-\eta_i(\boldsymbol{\beta})$ and covariance matrix $\Sigma_i(\boldsymbol{\alpha})$ truncated to the rectangular area given by $[log(1 - y_{i1}), -log(y_{i1})] \times \cdots \times [log(1 - y_{iT}), -log(y_{iT})]$. Note, that $\eta_i(\boldsymbol{\beta})$ and $\Sigma_i(\boldsymbol{\alpha})$ are determined by $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, respectively. For the generation of truncated multivariate random variables, we used a successive generation scheme based on the conditional distribution of the remaining components to be generated given the already generated components. The details of this generation scheme is provided in the Appendix. An alternative to this generation is the algorithm proposed by Geweke (1991). This algorithm can be used when the length of the binary time series is large. In this case, the successive generation scheme is too slow.

We derive now the conditional distribution $[\beta|\boldsymbol{\alpha}, Z, Y]$. The conditional distribution of the latent vector Z given the association parameter $\alpha$ is multivariate normal with mean vector $-X\boldsymbol{\beta}$, where $X$ is a block diagonal matrix with ith block given by

$$X_i = \begin{pmatrix} 1 & x_{i11} & \cdots & x_{i1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{iT1} & \cdots & x_{iTp} \end{pmatrix}$$

and block diagonal covariance matrix $\Sigma(\boldsymbol{\alpha})$ with ith block given by $\Sigma_i(\boldsymbol{\alpha})$. In the case of a multivariate normal prior for $\boldsymbol{\beta}$ with mean vector $\beta_p$ and covariance matrix $\Sigma_p$, it is straight forward to determine that $[\beta|Z, \boldsymbol{\alpha}]$ is again multivariate normal with mean vector

$$-(\Sigma_p^{-1} + X^t\Sigma(\boldsymbol{\alpha})^{-1}X)^{-1}(\Sigma_p^{-1}\beta_p + X^t\Sigma(\boldsymbol{\alpha})^{-1}Z)$$

and covariance matrix

$$(\Sigma_p^{-1} + X^t\Sigma(\boldsymbol{\alpha})^{-1}X)^{-1}.$$

For a flat prior the terms involving the prior parameters $\beta_p$ and $\Sigma_p$ vanish. Finally, we remark that since $Z$ determines $Y$, we have $[\boldsymbol{\beta}|\boldsymbol{\alpha}, Z, Y] = [\boldsymbol{\beta}|\boldsymbol{\alpha}, Z]$.

For updating the association parameters $\alpha$, we require a Metropolis-Hastings update. Here, $[\boldsymbol{\alpha}|\boldsymbol{\beta}, Z, Y]$ is proportional to $[Z|\alpha, \boldsymbol{\beta}]$ considered as function of $\alpha$. A normal proposal density with same mode as $[\boldsymbol{\alpha}|\boldsymbol{\beta}, Z, Y]$ and a user controlled covariance matrix is used for the corresponding Metropolis-Hastings step.

Using the above conditionals, an approximate sample from the posterior can be drawn and point and interval estimates of the parameters can be calculated using this sample. It should be noted that the algorithm can also be used for data with varying cluster sizes. The likelihood approaches based on marginal odds ratios possess also this property, while the ones based on conditional odds ratios do not.

### 3.2 The Missing Response Case

In Bayesian inference based on MCMC methods missing values will be handled as additional parameters (Gelman et al. (1995), Chapter 17). Therefore, as mentioned in Section 2.2 realizations of missing values will be generated. In the context of the multivariate probit model with missing responses, it is enough to generate missing latent variables $Z_{imiss}$. Once they are generated, a missing response $Y_{imiss}$ will be generated as follows:

$$Y_{imiss} = \begin{cases} 1 \text{ if } Z_{imiss} < 0 \\ 0 \text{ otherwise.} \end{cases}$$

To generate missing latent variables, note that the distribution of the complete latent vector $Z_i$ given the observed responses $Y_{iobs}$ and $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is multivariate normal with mean vector $-\eta_i(\boldsymbol{\beta})$ and covariance matrix $\Sigma_i(\boldsymbol{\alpha})$ truncated to the rectangular area $[a_1, b_1] \times \cdots \times [a_T, b_T]$ where

$$a_t = \begin{cases} log(1 - y_{it}) \text{ if } y_{it} \text{ is observed} \\ -\infty \qquad\quad \text{otherwise.} \end{cases},$$

and

$$b_t = \begin{cases} -log(y_{it}) \text{ if } y_{it} \text{ is observed} \\ \infty \qquad\quad \text{otherwise.} \end{cases}.$$

As already mentioned in Section 2.2, there are two different ways to sample the missing latent variables $Z_{imiss}$. First, we can generate the missing latent variables $Z_{imiss}$ together with the observed latent variables $Z_{iobs}$ from $[Z_i|Y_{iobs}, \boldsymbol{\beta}, \boldsymbol{\alpha}]$ specified above, or we can estimate the latent variables $Z_{imiss}$ by the expected value of $[Z_{imiss}|Z_{iobs}, Y_{iobs}, \boldsymbol{\beta}, \alpha]$. It is easy to see that this expectation is the same as the expectation of $[Z_{imiss}|Z_{iobs}, \boldsymbol{\beta}, \boldsymbol{\alpha}]$ since no truncation is involved at the components of $Z_{imiss}$. This last expectation was already specified in (2.4).

Once the missing latent variables are generated, the algorithm described in the previous section proceeds in the same way for the case of updating the regression and association parameters.

Since the missing data mechanism is not specified and therefore ignored, this way of proceeding implicitly requires the responses to be MAR. If the missing responses are MAR, a valid analysis is also obtained by using the MCMC algorithm developed in Section 2.2 for the observed data only. This is a third way to handle MAR responses. Here is important that the algorithm can handle varying cluster sizes. Currently, the extension for missing responses with monotone missing data pattern is studied.

## 4 Example

Fitzmaurice and Lipsitz (1995, p. 57) presented a subset of data on 51 subjects from an arthritis clinical trial (Bombardier et al. 1986). Patient self-assess their condition as "poor" (coded as

0) or "good" (coded as 1) at most five unequally spaced time points. Patients had a base line self-assessment measurement (week 0) and follow-up measurements took place at weeks 1, 5, 9 and 13. Randomization to one of the two treatments, placebo and auranofin, occurred following the second self-assessment. After randomization, patients remained on the assigned treatment for the entire study. Time stationary covariates are sex and age in years at study begin and time varying covariates are treatment and time (in weeks). In this data set 13 of the 51 subjects comprising 25.5 % of the data have some missing responses. The missing response pattern is however not monotone. Following Fitzmaurice and Lipsitz (1995), we investigated a multivariate probit model with margins specified as

$$\Phi(\eta_{it}(\beta)) = \beta_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 treatment_{it} + \beta_4 t \tag{4.1}$$

and the serial correlation pattern specified by

$$\rho_{ist} = \rho^{|s-t|}.$$

Here is the association parameter $\alpha = \tau = log(\frac{1+\rho}{1-\rho})$. Fitzmaurice and Lipsitz (1995) did not find any significant interactions terms in this data set, therefore no interaction was specified in (4.1). Noninformative priors for $(\beta, \tau)$ have been assumed. Results from four different Bayesian analyses will be compared. The first analysis is the naive analysis (NAIVE), where each row which contains at least one missing response is deleted from the analysis. The second one is based on the observed data only (OBS), while the third one assumes the missing responses as additional parameters to be estimated (MRE) and the fourth one imputes the missing responses (MRI). Table 4.1 gives estimated posterior means and their standard errors in parentheses, while Table 4.2 shows the estimated posterior quantiles . All results reported are based on a single run of 1600 iterations for each analysis with the first 400 iterations deleted to account for the burn-in effect.

| Parameter | NAIVE | OBS | MRE | MRI |
|---|---|---|---|---|
| Intercept | .399 (.929) | .704 (.613) | .777 (.670) | .974 (.732) |
| Sex | .468 (.359) | .344 (.249) | .305 (.268) | .253 (.294) |
| Age | -.003 (.015) | -.009 (.011) | -.010 (.012) | -.012 (.013) |
| Treatment | 1.28 (.407) | .779 (.270) | .819 (.280) | .698 (.263) |
| Time | -.044 (.025) | -.021 (.022) | -.030 (.023) | -.037 (.021) |
| $\rho$ | .853 (.089) | .712 (.104) | .798 (.098) | .893 (.050) |
| $\tau$ | 2.700 (.629) | 1.850 (.431) | 2.300 (.564) | 2.960 (.451) |

Table 4.1: Estimated Posterior Means and Estimated Standard Errors in Parentheses for the Arthritis Clinical Trial Data

The Splus library CODA by Best, Cowles and Vines (1995) has been used to assess the convergence of the sampled Markov chains and to produce the output analysis given. Geweke's (1992) convergence diagnostics based on Z-scores, Raftery and Lewis (1992) diagnostic and bounds for the accuracy of the estimated posterior quantiles and Heidelberger and Welsh's (1983) testing method for stationarity of the Markov chains have been used, demonstrating no evidence against the convergence of the sampled chains. See Cowles and Carlin (1995) and Brooks and Roberts (1996) for comparative reviews of MCMC diagnostics. A high autocorrelation between the sampled $\rho$ values was observed indicating that a longer simulation run is needed to reach

convergence. The OBS and MRI analyses produced lower first order autocorrelation among successive $\rho$ values ($\approx .80$) than the naive and the MRI analyses ($\approx .90$). Batching was used to assess the precision of the posterior mean estimates. For all analyses, a run of 1200 cycles was sufficient to achieve a simulation error of less than 2.8 % for the estimated posterior treatment effect and less than 1.7 % for the estimated posterior correlation.

| Parameter | Analysis | Estimated Posterior Quantiles | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | .01% | 2.5% | 25% | 50% | 75% | 975% | 99 % |
| Intercept | NAIVE | -1.8 | -1.5 | -.2 | .4 | 1.0 | 2.2 | 2.5 |
| | OBS | -.74 | -.49 | .31 | .71 | 1.1 | 1.9 | 2.2 |
| | MRE | -.87 | -.58 | .36 | .77 | 1.2 | 2.1 | 2.3 |
| | MRI | -.67 | -.45 | 51 | .96 | 1.4 | 2.5 | 2.7 |
| Sex | NAIVE | -.38 | -.21 | .23 | .47 | .70 | 1.2 | 1.3 |
| | OBS | -.24 | -.15 | .18 | .34 | .50 | .82 | .93 |
| | MRE | -.31 | -.23 | .13 | .30 | .49 | .84 | .92 |
| | MRI | -.47 | -.33 | .06 | .25 | .45 | .84 | .91 |
| Age | NAIVE | -.038 | -.033 | -.014 | -.003 | .008 | .027 | .031 |
| | OBS | -.035 | -.031 | -.017 | -.009 | -.003 | .012 | .016 |
| | MRE | -.037 | -.017 | -.015 | -.010 | -.002 | .015 | .019 |
| | MRI | -.042 | -.038 | -.020 | -.012 | -.004 | .014 | .019 |
| Treatment | NAIVE | .43 | .54 | 1.0 | 1.3 | 1.5 | 2.2 | 2.3 |
| | OBS | .16 | .23 | .61 | .77 | .96 | 1.3 | 1.4 |
| | MRE | .16 | .26 | .64 | .82 | 1.0 | 1.4 | 1.5 |
| | MRI | .14 | .20 | .51 | .68 | .88 | 1.2 | 1.3 |
| Time | NAIVE | -.10 | -.093 | -.062 | -.043 | -.027 | .001 | .011 |
| | OBS | -.072 | -.064 | -.035 | -.022 | -.007 | .027 | .032 |
| | MRE | -.081 | -.076 | -.044 | -.029 | -.015 | .018 | .026 |
| | MRI | -.082 | -.076 | -.050 | -.038 | -.024 | .006 | .016 |
| $\rho$ | NAIVE | .57 | .64 | .81 | .88 | .92 | .96 | .97 |
| | OBS | .41 | .49 | .66 | .72 | .78 | .89 | .91 |
| | MRE | .49 | .58 | .74 | .81 | .87 | .94 | .95 |
| | MRI | .74 | .78 | .87 | .90 | .93 | .95 | .96 |
| $\tau$ | NAIVE | 1.3 | 1.5 | 2.3 | 2.7 | 3.2 | 3.9 | 4.1 |
| | OBS | .87 | 1.1 | 1.6 | 1.8 | 2.1 | 2.8 | 3.0 |
| | MRE | 1.1 | 1.3 | 1.9 | 2.3 | 2.7 | 3.5 | 3.7 |
| | MRI | 1.9 | 2.1 | 2.7 | 3.0 | 3.3 | 3.8 | 3.9 |

Table 4.2: Estimated Posterior Quantiles for the Arthritis Clinical Trial Data

The results presented in Tables 4.1 and 4.2 show, that there is strong evidence for a nonzero correlation parameter $\rho$ from all four analyses. This shows that an analysis based on the independence among the binary responses would be inappropriate for this data set. In particular, only observations taken apart 13 weeks can be considered independent. Since $\rho$ is restricted to the interval [-1,1], it has to be expected, that the posterior distribution is skewed, as can be seen from the density estimate of the posterior distribution for $\rho$ for each analysis given in Figure 4.1. Therefore, the posterior median or mode are more appropriate measures for central tendency than the posterior mean. The estimated posterior modes for $\rho$ corresponding to the

naive and MRI analysis are of the same magnitude, while there are lower for the OBS and MRE analysis. Further, interval estimates should rather be based on posterior quantiles as given above than on interval estimates derived from the asymptotic normal theory.

We observe, that age, sex and time do not influence the marginal probabilities regardless of the analysis performed, while there is evidence for a treatment effect based on the posterior quantile estimates. However, with regard to the size of the treatment effect, the analyses clearly differ. For the naive analysis, the treatment effect is estimated about 1.6 times the size as for the remaining analyses. An explanation for this significant difference is, that the rows with missing responses contain information about the treatment effect, which is discarded in the naive analysis. The information about the treatment effect contained in the rows with the missing responses is conflicting the information about the treatment effect contained in the remaining data as is evidenced by Table 4.3. This explains, why the naive analysis overestimates the treatment effect. Figure 4.2 gives the corresponding posterior density estimates for the treatment parameter. It shows that the analyses based on the complete data (OBS, MRE and MRI) are similar and are more concentrated around the mean compared to the naive analysis, where information is lost due to the removal of rows with missing responses.

| Treatment | Row with missing Responses | Self-assessment | Baseline | Week 1 | Week 5 | Week 9 | Week 13 |
|---|---|---|---|---|---|---|---|
| Auranofin | yes | good | 15 | 14 | 18 | 16 | 17 |
| | | poor | 3 | 4 | 0 | 2 | 1 |
| | no | good | 4 | 6 | 6 | 2 | 0 |
| | | poor | 5 | 3 | 1 | 2 | 2 |
| | | missing | 0 | 0 | 2 | 5 | 7 |
| Placebo | yes | good | 14 | 12 | 9 | 13 | 13 |
| | | poor | 6 | 8 | 11 | 7 | 7 |
| | no | good | 4 | 2 | 2 | 2 | 2 |
| | | poor | 0 | 2 | 1 | 1 | 0 |
| | | missing | 0 | 0 | 1 | 1 | 2 |

Table 4.3: Arthritis Clinical Trial Data Cross-classified according to Treatment, Rows with Missing and Nonmissing responses, Self-assessment and Time Period.

Comparing the performance of the Bayesian analyses which are based on all observed responses, we see that there is not much difference with regard to estimated posterior means. With regard to interval estimates for the regression parameters, the relative interval estimate lengths are similar when adjusted for the magnitude of the posterior mean estimate. For the treatment parameter, the observed relative lengths are about 140% of the estimated posterior mean. For the correlation parameter, however, the MRI analysis produced the smallest relative interval estimate length in this example.

Finally, we compare our results to those obtained by GEE methods in Fitzmaurice and Lipsitz (1995) using a model based on odds. Using the heuristic (see Cessie and Houwelingen, 1994, p. 100) that the coefficients of a logistic model are about 1.7 times larger than the coefficients of the corresponding probit model, we see that the GEE results for the regression parameters are consistent with the Bayesian results based on all observed responses. This consistency indicates that the required assumption of MCAR for the validity of the GEE analyses of Fitzmaurice

and Lipsitz (1995) is tenable for this data set. In contrast, Kenward, Lesaffre and Molenberghs (1994) report on a data set involving missing ordinal responses, where the assumption of MCAR was not tenable and the GEE analysis would give incorrect estimates.

Finally, we compare results for the unrestricted association parameters, i.e. $\tau$ from the Bayesian results and the log odds ratio $\log(\alpha)$ from the GEE results. We see that the estimated relative interval length adjusted for the magnitude of the posterior mean estimate is shorter for all Bayesian analyses compared to the GEE analyses. In the case of the MRI analysis, this estimated relative length is about half as long as for the conditional GEE analysis. This might be an indication that the multivariate probit model provides a better fit in this example than a model based on odds. For the treatment effect, the Bayesian and GEE analyses have about the same relative interval estimate lengths.

In summary, MCMC methods can be used to conduct a computationally feasible Bayesian analysis of high dimensional correlated binary responses with time varying and time constant covariates and complex correlation patterns. Missing responses which are MAR can be accommodated. Three different ways to handle missing responses were investigated. In the data example, we observed a slight preference (lower autocorrelations and shorter interval estimates for the correlation parameter) for imputing the missing latent variables by their expected value (MRI analysis). Extensions to multivariate ordinal response vectors are currently considered.

# References


Anderson, J.A. and Pemberton, J.D. (1985). The grouped continous model for multivariate ordered categorical variables and covariate adjustment, *Biometrics*, **41**, 875-885.

Ashby, M. Neuhaus J.M., Hauck, W.W., Bacchetti P., Heibron, D.C., Jewell, N.P., Segal, M.R. and Fusaro, R.E. (1992) An Annotated Bibliography of Methods for Analyzing Correlated Categorical Data. *Statistics in Medicine*, **11**, 67-99.

Ashford, J.R. and Sowdon, R.R. (1970) Multivariate probit analysis, *Biometrics*, **26**, 535-546.

Baker, S.G. (1995) Marginal Regression for Repeated Binary Data with Outcome subject to Non-ignorable Non-response, *Biometrics*, **51**, 1042-1052.

Becker, M.P. and Ten Have, T.R. (1995). Random effects models for repeated categorical responses, preprint.

Besag, J., Green P., Hidgon D. and Mengersen, K. (1995). Bayesian Computation and Stochastic Systems. *Statistical Science*, **10**, No. 1, 3-66.

Best, N. , Cowles, M.K. and Vines, K. (1995). CODA - Convergence Diagnosis and Output Analysis Software, *MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK, email: bugs@mrc-bsu.cam.ac.uk*

Bombardier, C., Ware, J.H. and Russell, I.J. (1986). Auranofin therapy and quality of in patients with rheumatoid arthritis, *Am. J. Med.*, **81**, 565-578.

Brooks, S. and Roberts, G. (1996) Diagnosing convergence of Markov chain Monte Carlo algorithms, Technical report, Department of Pure Maths and Mathematical Statistics, University of Cambridge.

Carey, V., Zeger, S.L. and Diggle, P.J. (1993) Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517-526.

Casella, G. and George, E.I. (1992) Explaining the Gibbs Sampler, *Amer. Statistician*, **46**, No. 3, 167-174.

le Cessie, S and van Houwelingen, J.C. (1994). Logistic Regression for Correlated Binary Data, *Appl. Statist.*, **43**, No. 1, 95-108.

Chib, S. and Greenberg, E. (1994). Understanding the Metropolis-Hastings Algorithm, preprint.

Cowles, M.K. and Carlin, B.P. (1995) Markov chain Monte Carlo convergence diagnostics: a comparative review, *to appear in J. Am. Statist. Ass.*.

Cox, D.R. (1972). The analysis of multivariate binary data, *Appl. Statist.*, **21**, 113-120.

Dale, J.R. (1986). Global odds-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909-917.

Fitzmaurice, G.M. and Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary responses, *Biometrika*, **80**, 1, 141-151.

Fitzmaurice, G.M., Laird, N.M. and Lipsitz, S.R. (1994) Analysing Incomplete Longitudinal Binary Responses: A likelihood-based Approach, *Biometrics*, **50**, 601-612.

Fitzmaurice, G.M. and Lipsitz, S.R. (1995). A model for binary time series data with serial odds ratio patterns. *Appl. Statist.*, **44**, No. 1, 51-61.

Fitzmaurice, G.M., Molenberghs, G. and Lipsitz, S.R. (1995). Regression Models for Longitudinal Binary Responses with Informative Drop-outs, *J. R. Statist. Soc. B*, **57**, No. 4, 691-704.

Gelfand, A.E. and Smith, A.F.M. (1995). *Bayesian Computation*, New York, Wiley, in preparation.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*, New York, Chapman and Hall.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints, *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface, Seattle, Washington, April 21-24, 1991*, 571-578.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4*, (ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M Smith), Clarendon Press, Oxford, UK.

Heagerty, P.J. and Zeger, S.L. (1995). Marginal regression models for clustered ordinal measurements, Technical report # p790, Department of Biostatistics, Johns Hopkins University.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, **31**, 1109-1144.

Heumann, C. (1996). Marginal regression modeling of correlated multicategorical response: a likelihood approach, Disscusion paper 19, SFB 386, Seminar für Statistics, Ludwig-Maximilians-Universität, München.

Kenward, M.G., Lesaffre, E. and Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*, **50**, 945-953.

Korn, E.L. and Whittemore, A.S. (1979). Methods for analyzing panel studies of acute health effects of air pollution, *Biometrics*. **35**, 795-802.

Laird, N.M. (1988). Missing data in longitudinal studies. *Statist. Med.*, **7**, 305-315.

Lesaffre, E. and Molenberghs, G. (1991). Multivariate Probit Analysis: A neglected procedure in medical statistics, *Statistics in Medicine*, **10**, 1391-1403.

Liang, K.-Y, Zeger,S.L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *J.R. Statist. Soc. B*,**54**, 3-40.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association, *Biometrika*, **78**, 153-160.

Lipsitz, S.R., Fitzmaurice, G.M., Sleeper, L. and Zhao, L.P. (1995). Estimation methods for the joint distribution of repeated binary observations, *Biometrics*, **51**, 562-570.

Little, R.J.A. and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York, John Wiley.

Molenberghs, G. and Lesaffre, E. (1994). Marginal Modeling of Correlated Ordinal Data Using a Multivariate Plackett Distribution. *J. Amer. Statist. Soc.* , **89**, No. 426, 633-644.

Müller, P. (1994) A Generic Approach to Posterior Integration and Gibbs Sampling. to appear in *J. Amer. Stat. Assoc.*

Pearson, K. (1900) Mathematical contribution to the theory of evolution: VII, On the correlation of characters not quantitatively measurable. *Phil. Trans. R. Soc. Lond. A*, **195**, 1-47.

Plackett, R.L. (1965). A class of bivariate distributions, *J. Amer. Statist. Ass.*, **60**, 516-522.

Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation, *Biometrics*, **44**, 1033-1048.

Raftery, A.L. and Lewis, S. (1992). How many iterations in the Gibbs sampler. In *Bayesian Statistics 4*, (ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M Smith), Clarendon Press, Oxford, UK.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Ass.*,**90**, 106-121.

Ware, J.H., Dockery, D.W., Spiro, A. III, Speizer, F.E. and Ferris, B.G., Jr. (1984). Passive smoking, gas cooking and respiratory health in children living in six cities. *Am. Rev. Respir. Dis.*, **129**, 366-374.

Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika*, **77**, 642-648.

# Appendix

**Theorem:** *Sequential Random Generation of a truncated multivariate Random Vector*

Let $N_n(\mu, \Sigma)$ denote a n-dimensional normal distribution with mean vector $\mu = (\mu_1, \cdots, \mu_n)$ and covariance matrix $\Sigma$ with entries $\sigma_{ij}$.

1. Generate $x_n^* \sim N_1(\mu_n, \sigma_{nn})$ truncated to $[a_n, b_n]$
2. Generate $x_{n-1}^* \sim X_{n-1}|X_n = x_n^*$ truncated to $[a_{n-1}, b_{n-1}]$

$\vdots \quad \vdots$

$n$. Generate $x_1^* \sim X_1|X_2 = x_2^*, \cdots, X_n = x_n^*$ truncated to $[a_1, b_1]$

Then $X^* = (x_1^*, \cdots, x_n^*)$ is a realization from $N_n(\mu, \Sigma)$ truncated to $[a_1, b_1] \times \cdots \times [a_n, b_n]$.

**Proof:** For brevity, we consider only the case $n = 2$, the general case can be treated similarly.

Denote with $(X_1^*, X_2^*)$ the random vector resulting from the above scheme and $(X_1, X_2)$ a random vector distributed as $N_2(\mu, \sigma)$ truncated to $[a_1, b_1] \times [a_2, b_2]$. Further, denote with $f^*(\cdot)(f(\cdot))$ the marginal density of $X_2^*(X_2)$ and $F^*(\cdot|X_2^*)$ $(F(\cdot|X_2))$ the conditional distribution function of $X_1^*|X_2^* = x_2^*$ $(X_1|X_2 = x_2)$. The joint distribution of $(X_1^*, X_2^*)$ is therefore given by

$$\begin{aligned} Pr(X_1^* \in [a_1, x_{01}], X_2^* \in [a_2, x_{02}]) &= \int_{a_2}^{x_{02}} [F^*(x_{01}|x_2) - F^*(a_1|x_2)] f^*(x_2) dx_2 \\ &= \int_{a_2}^{x_{02}} \frac{[F(x_{01}|x_2) - F(a_1|x_2)]}{[F(b_1|x_2) - F(a_1|x_2)]} \frac{f(x_2)}{[F(b_2) - F(a_2)]} dx_2. \quad (4.2) \end{aligned}$$

On the other side, we have for $(X_1, X_2)$ truncated to $[a_1, b_1] \times [a_2, b_2]$,

$$Pr(X_1 \in [a_1, x_{01}], X_2 \in [a_2, x_{02}]) = \frac{\int_{a_2}^{x_{02}} [F(x_{01}|x_2) - F(a_1|x_2)] f(x_2) dx_2}{\int_{a_2}^{b_2} [F(b_1|x_2) - F(a_1|x_2)] f(x_2) dx_2}. \quad (4.3)$$

If $F(b_1|x_2) - F(a_1|x_2)$ is independent of $x_2$ it follows that (4.2)=(4.3), which is enough to prove the theorem. Note that the distribution of $X_1|X_2 = x_2$ is normal with mean $\mu_{x_1|x_2} = \mu_1 + \sigma_{12}\sigma_{11}^{-1}(x_2 - \mu_2)$ and variance $\sigma_{x_1|x_2} = \sigma_{11} - \sigma_{12}^2\sigma_{22}^{-1}$. Since $\sigma_{x_1|x_2}$ is independent of $x_2$, it is immediate that $F(b_1|x_2) - F(a_1|x_2)$ is also independent of $x_2$, which proves the theorem.

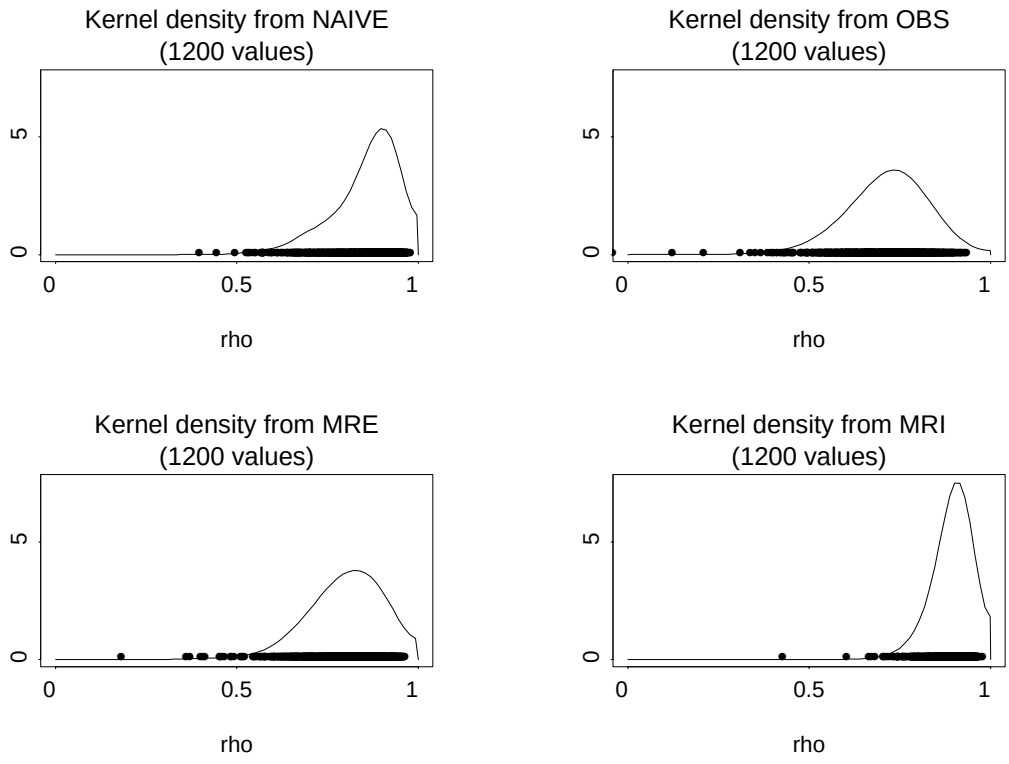


Figure 4.1: Posterior Density Estimates for the Correlation $\rho$ of the Arthritis Clinical Trial Data

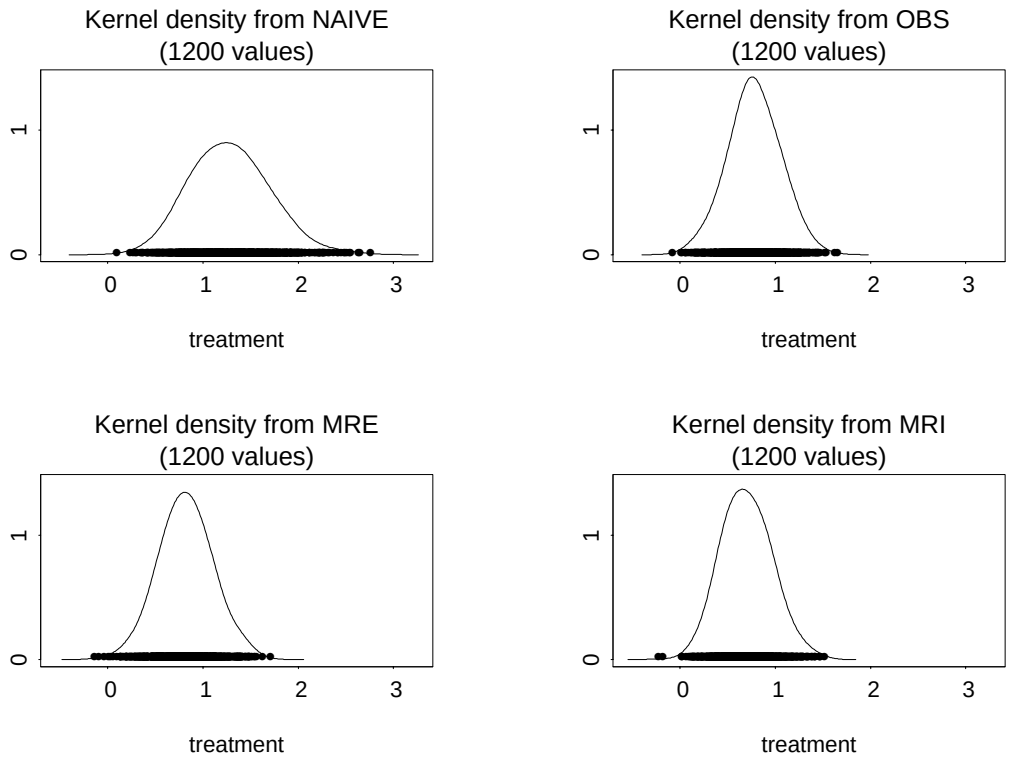


Figure 4.2: Posterior Density Estimates for the Treatment Parameter of the Arthritis Clinical Trial Data