8-18-2005

A User-Friendly Introduction to Link-Probit-Normal Models

Brian S. Caffo
The Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, bcaffo@jhsph.edu

Michael Griswold
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, mgriswol@jhsph.edu

A user-friendly introduction to link-probit-normal models

Brian Caffo and Michael Griswold
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

August 18, 2005

Abstract

Probit-normal models have attractive properties compared to logit-normal models. In particular, they allow for easy specification of marginal links of interest while permitting a conditional random effects structure. Moreover, programming fitting algorithms for probit-normal models can be trivial with the use of well developed algorithms for approximating multivariate normal quantiles. In typical settings, the data cannot distinguish between probit and logit conditional link functions. Therefore, if marginal interpretations are desired, the default conditional link should be the most convenient one. We refer to models with a probit conditional link an arbitrary marginal link and a normal random effect distribution as link-probit-normal models. In this manuscript we outline these models and discuss appropriate situations for using multivariate normal approximations. Unlike other manuscripts in this area that focus on very general situations and implement Markov chain or MCEM algorithms, we focus on simpler, random intercept settings and give a collection of user-friendly examples and reproducible code. Marginally, the link-probit-normal model is obtained by a non-linear model on a discretized multivariate normal distribution, and thus can be thought of as a special case of discretizing a multivariate T distribution (as the degrees of freedom go to infinity). We also consider the larger class of multivariate T marginal models and illustrate how these models can be used to closely approximate a logit link.

1 Introduction

In this manuscript we consider mixed effects regression models for binary and ordinal multinomial response variables. In the common setting where the random effects are assumed to be normally distributed, greater consideration should be given to the use of the cumulative probit conditional link functions in lieu of the more popular cumulative logit conditional link. These models are obtained by discretizing a latent normal distribution, a process that has been used extensively in the biometrics and econometrics literature (see Ashford and Sowden, 1970; McFadden, 1989; Hausman and Wise, 1978, for example) and is especially useful for joint modeling of continuous and discrete outcomes (Gueorguieva and Agresti, 2001). The principal benefits of the conditional probit link with normally distributed random effects are twofold. First, these models allow for easy approximations to the log-likelihood that can be more convenient than the usual quadrature or Monte Carlo approximations. Second, these models also allow for easy specification of a marginal

model of interest. That is, these models are both computationally convenient and can be used to obtain coefficients that possess desirable interpretations.

We refer to models with an arbitrary cumulative marginal link, a cumulative probit conditional link, and normally distributed random effects as cumulative link-probit-normal models, or simply link-probit-normal models in the binary case. In this manuscript we outline these models and their computation, focusing on reproducible code for easily understood examples. Notably, we highlight the commonly occurring setting of random intercept models where there are few observations per cluster. We also consider a generalization of these models using the multivariate T distribution. The added flexibility of choosing the degrees of freedom for this distribution allows one to very closely approximate marginal logit links.

2 Notation and likelihood computation

We develop our notation for a general version of the problem, though we focus on random intercept models in the examples. Let Y_{ij} be a response taking ordinal levels $k = 1, \ldots, K$ for cluster/subject $i = 1, \ldots, I$ on measurement/occasion $j = 1, \ldots, J_i$. We assume the hierarchical, conditional model

$$\Pr(Y_{ij} \leq k \mid U_i = u_i) = \Phi(\Delta_{ijk} + z_{ij}^t u_i) \text{ and } U_i \sim \text{Normal}(0, \Sigma),$$

where Φ is the standard normal cdf, Δ_{ijk} is a possibly non-linear, covariate dependent and observation specific fixed predictor, and $U_i = (U_{i1}, \ldots, U_{iP})^t$ are latent effects representing cluster and level specific sources of variation. To obtain an appropriate ordering of the cumulative probabilities it is required that the Δ_{ijk} are increasing in k and that $\Delta_{ij0} = -\infty$ and $\Delta_{ijK} = \infty$. When $K = 2$ (binary data), and Δ_{ijk} is linearized with respect to a covariate vector ($\Delta_{ijk} = x_{ij}^t \beta$), this is a standard probit-normal generalized linear mixed model.

Let W_{ij} be iid standard normal variates, mutually independent of the U_i, and define the convolution $M_{ij} = W_{ij} - z_{ij}^t U_i$, which implies that the M_{ij} are multivariate normal. With this derivation, the contribution of subject i to the likelihood is (see Appendix C)

$$\Pr\left(\cap_{j=1}^{J_i} [\Delta_{ij(y_{ij}-1)} < M_{ij} \leq \Delta_{ijy_{ij}}]\right) \tag{1}$$

(where the y_{ij} are fixed). Therefore, the desired contribution to the likelihood is a multivariate normal probability.

The likelihood, which is the product of (1) across subjects, must be approximated numerically. Bayesian approaches using Markov chain Monte Carlo to simulate from a posterior distribution for both the latent variable process and the fixed parameters are popular (Chib et al., 1998; Imai and van Dyk, 2005; Chib and Greenberg, 1998). Also, the Monte Carlo EM algorithm has been used for maximum likelihood estimation (Chib and Greenberg, 1998; Natarajan et al., 2000). Adaptive quadrature methods have also been widely applied, such as in the SAS procedure `PROC NLMIXED`. Our present discussion differs slightly from these in that we focus on marginal likelihood-based interpretations of parameters (see Section 3) and the common setting where there are few observations per cluster. In these instances well established algorithms for approximating the multivariate normal cdf can be used (see Genz, 1992; Shervish, 1984, for example). In our examples, we use the `mvtnorm` package (Hothorn et al., 2001) in R (Ihaka and Gentleman, 1996).

Because of such packages, we find that coding an approximation to the log-likelihood is often very easy (see Appendix E).

Notice that the dimension over which the multivariate normal probability is calculated increases with the number of observations, J_i, not the dimension of the random effects. Therefore, the direct use of the multivariate normal cdf to calculate the likelihood is well suited for data with large numbers of clusters comprised of few observations, such as in cohort studies where the cluster is a subject measured over few occasions. In contrast it is not well suited for studies with large numbers of observations per cluster. In these settings, we have found that adaptive quadrature approximations work well if the random effects are low dimensional. Monte Carlo based algorithms or Laplace/type approximations appear to be the only applicable options for very high dimensional random effects.

Notice that the marginal variance/covariance matrix of the $M_i = (M_{i1}, \ldots, M_{iJ_i})^t$ is $I + z_i \Sigma z_i^t$ where z_i is the matrix of the z_{ij}^t stacked by rows. If the random effects are not of intrinsic or conceptual interest, one can construct a model by simply specifying this correlation matrix and use (1) to calculate the likelihood; which is the process of discretizing a multivariate normal distribution. In general settings, care must be taken so that the marginal variance of the M_i is identified.

The use of the multivariate normal distribution is primarily due to historical reasons and convenience; the higher order cumulants of the latent variable distribution are not usually identified in most practical settings. Other multivariate distributions, such as the multivariate T and multivariate logistic (Chib et al., 1998; O'Brien and Dunson, 2004), have been suggested. The multivariate T is particularly interesting since it is more flexible than the normal distribution, includes the normal as a limiting case, has widely available accurate approximations to the distribution function (Genz and Bretz, 1999) and it provides a readily available computational approximation to the marginal likelihood (as discussed below).

3 Modeling the Δ_{ijk}

As currently stated, the model is over-specified if the Δ_{ij} are left unstructured. For example, the maxized likelihood always corresponds to a probability of 1 for the observed data. A common, more parsimonious model, linearizes the Δ_{ijk} with respect to some covariates. As discussed in Heagerty (1999) and Heagerty and Zeger (2000), linearizing the Δ_{ijk} on different scales yields coefficients with either marginal or conditional interpretations.

Conditional models

A conditional linear predictor model assumes that

$$\Delta_{ijk} = \alpha_k^c + x_{ij}^t \beta_k^c, \tag{2}$$

where x_{ij} is a vector of covariate values. The required ordering on the Δ_{ijk} is imposed by forcing the α_k^c to be strictly increasing. Under this definition,

$$\Phi^{-1}\{\Pr(Y_{ij} \leq k \mid U_i = u_i)\} = \alpha_k^c + x_{ij}^t \beta^c + z_{ij}^t u_i$$

so that the β^c have a *conditional*, or subject specific interpretation on the probit scale; hence the superscript c. Assuming that neither the fixed slope parameters nor the random effects depend on the ordinal level k is analogous to a proportional odds assumption for cumulative logit models, an assumption we make throughout. This model is a probit-normal instance of multinomial GLMMs (see Hartzel et al., 2001; Liu and Agresti, 2005).

Marginal models

Often marginal, not conditional, interpretations of the coefficients are desired. Let G^{-1} be the marginal cumulative link function of interest with G being the associated distribution function. In this manuscript we are primarily interested in the case where G is the logistic distribution. A linear predictor with marginal parameter interpretations is obtained by assuming that

$$\Pr(Y_{ij} \leq k) = G(\alpha_k^m + x_{ij}^t \beta^m) \tag{3}$$

(see Heagerty and Zeger, 1996, 2000). We demonstrate how this defines the Δ_{ijk} of the conditional model. Because of the conditional probit link and normal random effects, the marginal cumulative probabilities satisfy

$$\Pr(Y_{ij} \leq k) = \Pr(M_{ij} \leq \Delta_{ijk}) = \Phi\{\Delta_{ijk}/(1 + z_{ij}^t \Sigma z_{ij})^{1/2}\}. \tag{4}$$

Therefore setting

$$\Delta_{ijk} = \Phi^{-1}\left\{G(\alpha_k^m + x_{ijk}\beta_k^m)\right\}(1 + z_{ij}^t \Sigma z_{ij})^{1/2} \tag{5}$$

produces a model that satisfies both (3) and (4) (see Griswold, 2005). Notice that if $G = \Phi$ and $z_{ij} = 1$, then the conditional and marginal slopes are rescalings of each other (as in Zeger et al., 1988). This is also nearly true when G is the logistic distribution function, because $\Phi^{-1}G$ is approximately linear.

We reiterate that using (2) produces a linear predictor with subject-specific parameter interpretations while using (5) produces a linear predictor with marginal parameter interpretations. In both cases, the conditional link is the probit link and the random effect distribution is normal. Therefore, we refer to the marginal models as the "cumulative link-probit-normal models". For example, when G is the logistic distribution function, we refer to the model as the cumulative logit-probit-normal model.

4 A multivariate T extension of the link-probit-normal model

As discussed previously, users may want marginalized logit interpretations obtained by discretizing a multivariate distribution other than the normal. In what follows, we discuss the marginal model obtained by discretizing a multivariate T distribution, and illustrate the implied conditional model. Furthermore, we also illustrate how the degrees of freedom of the multivariate T can be chosen so that the marginal distributions are effectively logistic, and hence one can avoid the non-linear definition for the Δ_{ijk}.

Before we introduce the model, we mention some alternative approaches. One class of alternatives, assumes a generalization of the logistic distribution on M_i, so that the M_{ij} all possess (exact) marginal univariate logistic distributions (see O'Brien and Dunson, 2004, for a recent

generalization of the logistic distribution). Another adapts the random effect distribution so that the marginal and conditional links are both logistic (Wang and Louis, 2004, 2003). Keeping with our goal of simple computational methods, we present an approximate method based on the multivariate T distribution that can work quite well. Moreover, the resulting, more general, family of models contains the link-probit-normal models as a limiting case.

Consider the modification of the link-probit-normal model where

$$\Pr(Y_{ij} \leq k \mid U_i = u_i, S_i = s_i) = \Phi\left(\Delta_{ijk}\sqrt{\frac{s_i}{v}} + z_{ij}^t u_i\right)$$

for $U_i \sim \mathrm{Normal}(0, \Sigma)$ and $S_i \sim \chi^2_v$. This model implies that the contribution of subject i to the log likelihood is

$$\Pr\left(\cap_{j=1}^{J_i}\left[\Delta_{ij(y_{ij}-1)} < \frac{M_{ij}}{\sqrt{\frac{S_i}{v}}} \leq \Delta_{ijy_{ij}}\right]\right) \tag{6}$$

(see Appendix D). The $M_i/\sqrt{\frac{s_i}{v}}$ are distributed as a multivariate T with v degrees of freedom and variance matrix $I + z_i\Sigma z_i^t$ (see Genz and Bretz, 1999) and therefore (6) requires evaluating the multivariate T distribution function. Of course, the marginal distributions of the $M_{ij}/\sqrt{\frac{s_i}{v}}$ are scaled univariate T distributions. That is,

$$\Pr(Y_{ij} \leq k) = \Pr\left(\frac{M_{ij}}{\sqrt{\frac{s_i}{v}}} \leq \Delta_{ijk}\right) = T_v\{\Delta_{ijk}/(1 + z_{ij}^t\Sigma z_{ij})\},$$

where T_v is the standard T distribution function with v degrees of freedom. Hence setting

$$\Delta_{ijk} = T_v^{-1}\left\{G(\alpha_k^m + x_{ij}\beta^m)\right\}(1 + z_{ij}^t\Sigma z_{ij})^{1/2} \tag{7}$$

yields a marginal model with link function G^{-1}.

As noted in Chib and Albert (1993), the T distribution function is nearly linearly related to the logistic distribution function. Therefore, when G is the logistic distribution, (7) can be easily approximated with

$$\Delta_{ijk} = c_v(\alpha_k^m + x_{ij}\beta^m)(1 + z_{ij}^t\Sigma z_{ij})^{1/2}.$$

where c_v is a constant value, dependent on the degrees of freedom of the T-distribution used, thereby avoiding the non-linear definition of the Δ_{ijk}.

The constant factor c_v and appropriate degrees of freedom can be chosen to minimize a variety of criteria. For example, minimizing the residual sum of squares of a linear regression approximation of $T_v^{-1}G$ for probabilities between .999 and .001 yields $df = 8.7817$ and $c_{(8.7817)} = 0.6215$ (see the code in Appendix F). Better approximations can be obtained if the range of probabilities is smaller. However, in the examples that follow, these constants performed well. In practice, we would suggest choosing the constant and the degrees of freedom to allow for a reasonable range of probabilities in the given context. Figure 1 plots $T_v^{-1}G$ over the range of logit scale probabilities and gives a useful depiction of how accurate the approximation is. Notice that, for logit-scale probabilities between -6.91 and 6.91 (probabilities of .001 and .999 respectively), the linear approximation is nearly exact. It is only very small or large probabilities that the approximation is breaks down.

5 Examples

Binary matched pairs

Table 1 is a 2×2 contingency table of approval ratings of the British Prime Minister collected at two occasions. Let $Y_{ij} = 1$ correspond to a response of "approval" at sampling occasion j while $Y_{ij} = 2$ corresponds to "disapproval". Under our existing notation, $K = 2$, $J_i = 2$ and $I = 1600$. Consider the model

$$\Phi^{-1}\{\Pr(Y_{ij} = 1 \mid U_i = u_i)\} = \Delta_j + u_i \text{ and } U_i \sim \text{Normal}(0, \sigma^2) \tag{8}$$

where

$$\Delta_1 = \Phi^{-1}\left\{G(\alpha^m)\right\}(1+\sigma^2)^{1/2} \text{ and } \Delta_2 = \Phi^{-1}\left\{G(\alpha^m + \beta^m)\right\}(1+\sigma^2)^{1/2}. \tag{9}$$

Notice that extraneous subscripts have been omitted.

Because MLEs are invariant to monotonic transformations, fitting Model (8) ignoring (9) yields an equivalent fit to when the marginal constraints (9) are employed. In either case, there are three unknowns, either $(\Delta_1, \Delta_2, \sigma)$ or $(\alpha^m, \beta^m, \sigma)$. The log-likelihood is

$$\begin{aligned}
794 \quad & \log\left[\Pr\left\{M_1 \le \Delta_1, M_2 \le \Delta_2\right\}\right] + \\
150 \quad & \log\left[\Pr\left\{M_1 \le \Delta_1, M_2 > \Delta_2\right\}\right] + \\
86 \quad & \log\left[\Pr\left\{M_1 > \Delta_1, M_2 \le \Delta_2\right\}\right] + \\
570 \quad & \log\left[\Pr\left\{M_1 > \Delta_1, M_2 > \Delta_2\right\}\right].
\end{aligned}$$

The fitted values overlaid on a contour plot of the bivariate normal distribution along with the cell counts for the four quadrants are shown in Figure 2.

The saturated model also has three free parameters, and so the fitted counts are identical to the observed. Code for fitting the model is given in Appendix G. The fitted values are $\hat{\alpha}^m = .36$ $(.05)$, $\hat{\beta}^m = -.16$ $(.04)$, $\hat{\sigma} = 2.94$ $(.20)$. Therefore, the fitted model specifies that the marginal log-odds of approval at occasion one is $.36$ while it is $.36 - .16 = .20$ at occasion two. Under this model, $\exp\{\beta^m\} = .85$ is the marginal odds ratio of approval comparing the two sampling occasions, suggesting an estimated 15% decrease in approval.

The estimate of β^m is identical to the marginal ML estimate, $\log(944{\times}720/880{\times}656) = -.16$ (see Agresti, 2002). The fit using the multivariate T approximation yielded $\hat{\beta}^m = -.17$ $(.04)$, differing from the logit-probit-normal results by .006. (The code for the multivariate T approximation is not given, as it is nearly identical to the logit-probit-normal code.) The large estimate for the variance component is due to the large cell counts in the diagonal cells, i.e. a high degree of correlation between the two repeated observations. The likelihood ratio statistic for β^m is 17.58 on 1 degree of freedom, strongly suggesting the lack of marginal homogeneity in the probability approval at the two occasions.

Ordinal matched pairs

We extend the previous example to an ordinal matched pairs data set. The movie rating data given in Table 2 cross-classifies the ratings (1-Con, 2-Mixed, 3-Pro) of Roger Ebert ($j = 1$) to

that of the late Gene Siskel ($j = 2$). Consider a model so that Y_{ij} is the rating of reviewer j on movie i:

$$\Phi^{-1}\{\Pr(Y_{ij} \leq k \mid U_i = u_i)\} = \Delta_{jk} + u_i \text{ and } U_i \sim \text{Normal}(0, \sigma^2),$$

where

$$\Phi^{-1}\{G(\Delta_{1k})\} = \alpha_k^m \text{ and } \Phi^{-1}\{G(\Delta_{2k})\} = \alpha_k^m + \beta^m.$$

Recall that $\alpha_0^m = -\infty < \alpha_1^m < \alpha_2^m < \alpha_3^m = \infty$. It is worth noting that it is easier to fit the sequential differences, $(\alpha_1^m, \alpha_2^m - \alpha_1^m)$ than the α_k^m directly.

Figure 3 displays the fitted bivariate normal distribution with the observed and fitted counts. The resulting Pearson statistic is 7.08 (with 5 df). Appendix H gives R code to perform the fitting. The results of the fit are $\hat{\alpha}_1^m = -1.05$ (.17), $\hat{\alpha}_2^m - \hat{\alpha}_1^m = .84$ (.10), $\hat{\beta}^m = .11$ (.16) and $\hat{\sigma} = 1.25$ (.20). For comparison, the discretized multivariate T approximation yielded $\hat{\alpha}_1^m = -1.08$ (.17), $\hat{\alpha}_2^m - \hat{\alpha}_1^m = .87$ (.10), and $\hat{\beta}^m = .11$ (.16). The fitted link-probit normal model specifies that

$$\begin{aligned}
&\text{logit}\{\Pr(\text{Ebert's rating } \leq 1)\} = -1.05 \\
&\text{logit}\{\Pr(\text{Ebert's rating } \leq 2)\} = -1.05 + .84 \\
&\text{logit}\{\Pr(\text{Siskel's rating } \leq 1)\} = -1.05 + .11 \\
&\text{logit}\{\Pr(\text{Siskel's rating } \leq 2)\} = -1.05 + .84 + .11.
\end{aligned}$$

Notice that $\beta^m = .11$ is the marginal increase in the cumulative odds comparing Siskel to Ebert. An effective way to display the evidence regarding β^m is to plot the profile likelihood. Figure 4 displays the profile likelihood for β^m, which largely overlaps 0.

Crossover data

The data given in Table 3 are from a well studied cross-over experiment. Here the binary response was an indicator of an "abnormal" versus a "normal" reaction to medication when comparing an active drug versus a placebo given in two periods. Models of interest relate the response to the treatment and period while accounting for the correlation incurred by the repeated measures.

Assume that $Y_{ij} = 1$ corresponds to a response of "normal" while $Y_{ij} = 2$ corresponds to a response of "abnormal". Conditional on a random intercept, we assume that the Y_{ij} are binary with success probability

$$\Phi^{-1}\{\Pr(Y_{ij} = 1 \mid U_i = u_i)\} = \Delta_{ijk} + u_i \text{ and } U_i \sim \text{Normal}(0, \sigma^2),$$

where

$$\Delta_{ijk} = \Phi^{-1}\left\{G(\alpha_1^m + x_{ij1}\beta_1^m + x_{ij2}\beta_2^m)\right\}(1 + \sigma^2)^{1/2}.$$

Here x_{ij1} is a binary treatment (1 for the active drug) indicator while x_{ij2} is a binary period indicator (1 for the second period). Appendix I gives code for fitting the model. With G equal to the logistic distribution, the resulting fit is $\hat{\alpha}_1^m = .68$ (.28), $\hat{\beta}_1^m = .59$ (.23), $\hat{\beta}_2^m = -.33$ (.23) and $\hat{\sigma} = 2.80$ (1.06). For comparison, the discretized multivariate T approximation yielded $\alpha_1^m = .71$ (.29), $\beta_1^m = .60$ (.24) and $\hat{\beta}_2^m = -.33$ (.24). The corresponding GEE (Zeger and Liang, 1986) model fit with a logit link and an exchangeable correlation matrix yielded $\alpha_1^m = .67$ (.29), $\beta_1^m = .57$ (.23) and $\hat{\beta}_2^m = -.30$ (.23).

Thus the fitted logit-probit-normal model specifies that

$$\text{logit}\{\Pr(Y_{ij} = 1)\} = .68 + .59x_{ij1} - .33x_{ij2}.$$

Therefore, $\exp(.587) - 1 = 80\%$ estimates the proportional increase in the marginal odds of a normal response for the treatment. Figure 5, displays the profile likelihood for β_1^m with 1/8 and 1/16 reference lines.

Sleep data

Table 4 cross-classifies the categorized time to sleep onset for baseline and followup visits for subjects who received either a hypnotic medication or a placebo. The response, time to sleep onset, was categorized into four ordinal levels. Consider the model

$$\Phi^{-1}\{\Pr(Y_{ij} \leq k \mid U_i = u_i)\} = \Delta_{ijk} + u_i \text{ and } U_i \sim \text{Normal}(0, \sigma^2),$$

where

$$\Delta_{ijk} = \Phi^{-1}\left\{G(\alpha_k^m + x_{ij1}\beta_1^m + x_{ij2}\beta_2^m + x_{ij1}x_{ij2}\beta_3^m)\right\}(1 + \sigma^2)^{1/2}.$$

Here $k = 0, \ldots, 4$ where $\alpha_0^m = -\infty$, $\alpha_4^m = \infty$, G is the logistic distribution, x_{ij1} is a treatment indicator (1 for the active drug) and x_{ij2} is a time indicator (1 for followup).

The resulting fitted values are given in Table 5 with sample R code given in Appendix J. To illustrate the interpretation of the coefficients, consider the fitted marginal models:

$$\begin{aligned}
\text{logit}\{\Pr(Y_{ij} \leq 1)\} &= -2.290 + 1.052x_{i1} + .058x_{i2} + .687x_{i1}x_{i2}, \\
\text{logit}\{\Pr(Y_{ij} \leq 2)\} &= -.983 + 1.052x_{i1} + .058x_{i2} + .687x_{i1}x_{i2}, \\
\text{logit}\{\Pr(Y_{ij} \leq 3)\} &= .332 + 1.052x_{i1} + .058x_{i2} + .687x_{i1}x_{i2},
\end{aligned}$$

Therefore, for example, .058 estimates the marginal increase in the cumulative odds comparing the treatment to the placebo at baseline.

To demonstrate the benefits of a likelihood based approach, consider testing the necessity of the interaction term. The likelihood ratio statistic for testing $H_0 : \beta_3^m = 0$ is estimated to be 7.71 with 1 df. In addition, Figure 6 displays the profile likelihood for β_3^m. (Both methods clearly suggests the need for this term.)

In addition to the cumulative logit-probit-normal results, the discretized multivariate T, the multinomial GEE (see Lipsitz et al., 1994) and marginal-ML (the latter two as reported in Lang and Agresti, 1994) results for the slope coefficients are also given for comparison (as reported in Agresti, 2002). To avoid confusion, we note that both the marginal-ML approach and the cumulative logit-probit-normal model use the method of maximum likelihood with a cumulative logit model on the marginal probabilities. However, the marginal-ML model requires no further model, leaving the multinomial probabilities otherwise unrestricted. In contrast, the cumulative logit-probit-normal model assumes a subject specific model, random effects and a random effect distribution. The results for all three models are similar.

We believe that this example illustrates how the cumulative link-probit-normal model is attractive when compared to these alternatives. When compared to marginal-ML, it can fit a more

parsimonious model. In contrast with GEE results, it completely specifies a likelihood, allowing for likelihood ratio tests and the computation profile likelihoods. Furthermore, we note that currently most software have difficulty fitting marginal models for multinomial data. However, this example illustrates that allowing for a conditional probit link makes model fitting straightforward. There is nothing particular to R, other than well developed packages for non-linear optimization and approximating multivariate normal and multivariate T probabilities.

6 Summary

With regard to model fit, all of the data sets considered exhibited equivalent fit, regardless of the nature of the marginal/conditional link function. For example, Table 6 gives the goodness of fit likelihood ratio statistic comparing models for each of the data sets under consideration to the saturated model. This emphasizes conventional wisdom that the choice of link function should represent preferences in interpretation, as model fit will rarely distinguish between them.

In summary, the likelihood for many binary and ordinal mixed-effect models is a probability from the multivariate distribution function obtained by convolving iid variables from the conditional link distribution and the random effects. For normally distributed random effects and a probit conditional link, this convolution is multivariate normal. This fact makes probit-normal models, in many ways, more convenient than logit-normal models. They can be easy to fit and make the relationship between the marginal and conditional models transparent. Similarly, the introduction of a latent chi-squared variable results in a model with a conditional probit link and marginal T-quantile link function. Moreover, judiciously choosing the degrees of freedom very accurately approximates the logit function.

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, second edition.

Agresti, A. and Winner, L. (1997). Evaluating agreement and disagreement among movie reviewers. *CHANCE*, 10:10–14.

Ashford, J. and Sowden, R. (1970). Multi-variate probit analysis. *Biometrics*, 26:535–546.

Chib, S. and Albert, J. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 422:669–679.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

Chib, S., Greenberg, E., and Chen, Y. (1998). MCMC methods for fitting and comparing multinomial response models. Technical report, University of Washington.

Francom, S. F., Chuang-Stein, C., and Landis, J. R. (1989). A log-linear model for ordinal data to characterize differential change among treatments. *Statistics in Medicine*, 8:571–582.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *The Journal of Computational and Graphical Statistics*, 1:141–149.

Genz, A. and Bretz, F. (1999). Numerical computation of multivariate t-probabilities with applications to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63:361–378.

Griswold, M. (2005). *Complex Distributions, Hmmmm... Hiearchical Mixtures of Marginalized Multilevel Models*. PhD thesis, Johns Hopkins University.

Gueorguieva, R. and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96(455):1102–1112.

Hartzel, J., Agresti, A., and Caffo, B. (2001). Multinomial logit random effect models. *Statistical Modelling*, 1(2):81–102.

Hausman, J. and Wise, D. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogenous preferences. *Econometrica*, 46(2):403–426.

Heagerty, P. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55:688–698.

Heagerty, P. J. and Zeger, S. L. (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, 91:1024–1036.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1):1–26.

Hothorn, A., Bertz, F., and Genz, A. (2001). On multivariate t and gaussian probabilities in r. *R News*, 1(2):27–29.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Imai, K. and van Dyk, D. (2005). A bayesian analysis of the multinomial probit model using data augmentation. *Journal of Econometrics*, 124:311–334.

Jones, B. and Kenward, M. (1987). Modelling binary data from a three ponit cross-over trial. *Statistics in Medicine*, 6:555–564.

Lang, J. B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, 89:625–632.

Lipsitz, S. R., Kim, K., and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, 13:1149–1163.

Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*, 14(1):1–73.

McFadden, D. (1989). A method of simulated moments for estimation of discrete repsonse models without numerical integration. *Econometrica*, 57(2):995–1026.

Natarajan, R., McCulloch, C., and Kiefer, N. (2000). A Monte Carlo EM method for estimating multinomial probit models. *Computational Statistics and Data Analysis*, 34:35–50.

O'Brien, S. and Dunson, D. (2004). Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746.

Shervish, M. (1984). Multivariate normal probabilities with error bound. *Applied Statistics*, 33:81–87.

Wang, Z. and Louis, T. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika*, 90(4):765–775.

Wang, Z. and Louis, T. (2004). Marginalized binary mixed-effects with covariate-dependent random effects and likelihood inference. *Biometrics*, 60(4):884–891.

Zeger, S. and Liang, K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.

Zeger, S., Liang, K., and Albert, P. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44:1049–1060.

A Tables

First	Second Survey	
Survey	Approve	Disapprove
Approve	794	150
Disapprove	86	570

Table 1: Prime minister approval rating. Source Agresti (2002).

	Ebert		
Siskel	Con	Mixed	Pro
Con	24	8	13
Mixed	8	13	11
Pro	10	9	64

Table 2: Movie rating data. Source (Agresti and Winner, 1997).

Response		Treatment sequence	
Period 1	Period 2	Drug-Placebo	Placebo-Drug
Normal	Normal	22	18
Abnormal	Normal	0	4
Normal	Abnormal	6	2
Abnormal	Abnormal	6	9

Table 3: Crossover data, frequency of responses by treatment regimen. Source Jones and Kenward (1987).

		Follow-up			
	Baseline	< 20	$20 - 30$	$30 - 60$	> 60
Active	< 20	7	4	1	0
	$20 - 30$	11	5	2	2
	$30 - 60$	13	23	3	1
	> 60	9	17	13	8
Placebo	< 20	7	4	2	1
	$20 - 30$	14	5	1	0
	$30 - 60$	6	9	18	2
	> 60	4	11	14	22

Table 4: Insomnia data from Francom et al. (1989).

Parameter	logit-probit-normal	discretized T	Marginal-ML	GEE
$\hat{\alpha}_1^m$	-2.290 (.209)	-2.347 (.215)		
$\hat{\alpha}_2^m - \hat{\alpha}_1^m$	1.307 (.113)	1.346 (.116)		
$\hat{\alpha}_3^m - \hat{\alpha}_2^m$	1.315 (.107)	1.362 (.111)		
$\hat{\beta}_1^m$	1.052 (.180)	1.074 (.162)	1.079 (.180)	1.038 (.168)
$\hat{\beta}_2^m$	0.058 (.237)	0.046 (.236)	0.043 (.244)	0.034 (.238)
$\hat{\beta}_3^m$	0.687 (.248)	0.662 (.244)	0.729 (.251)	0.708 (.244)
$\hat{\sigma}$	1.071 (.124)	1.111 (0.133)		

Table 5: Fitted values for the sleep data *estimate (se)*.

	Data set			
	approval	movie	crossover	sleep†
logit-probit-normal	0 (0)	6.38 (4)	2.14 (2)	38.89 (23)
probit-normal		6.39	2.10	41.62
discretized T		5.50	2.22	36.54
logit-normal		5.79	2.10	38.31

Table 6: Goodness-of-fit likelihood ratio statistics (df) testing the various models for each data set versus the saturated model. † - the model for the sleep data set included the interaction term.

B Figures

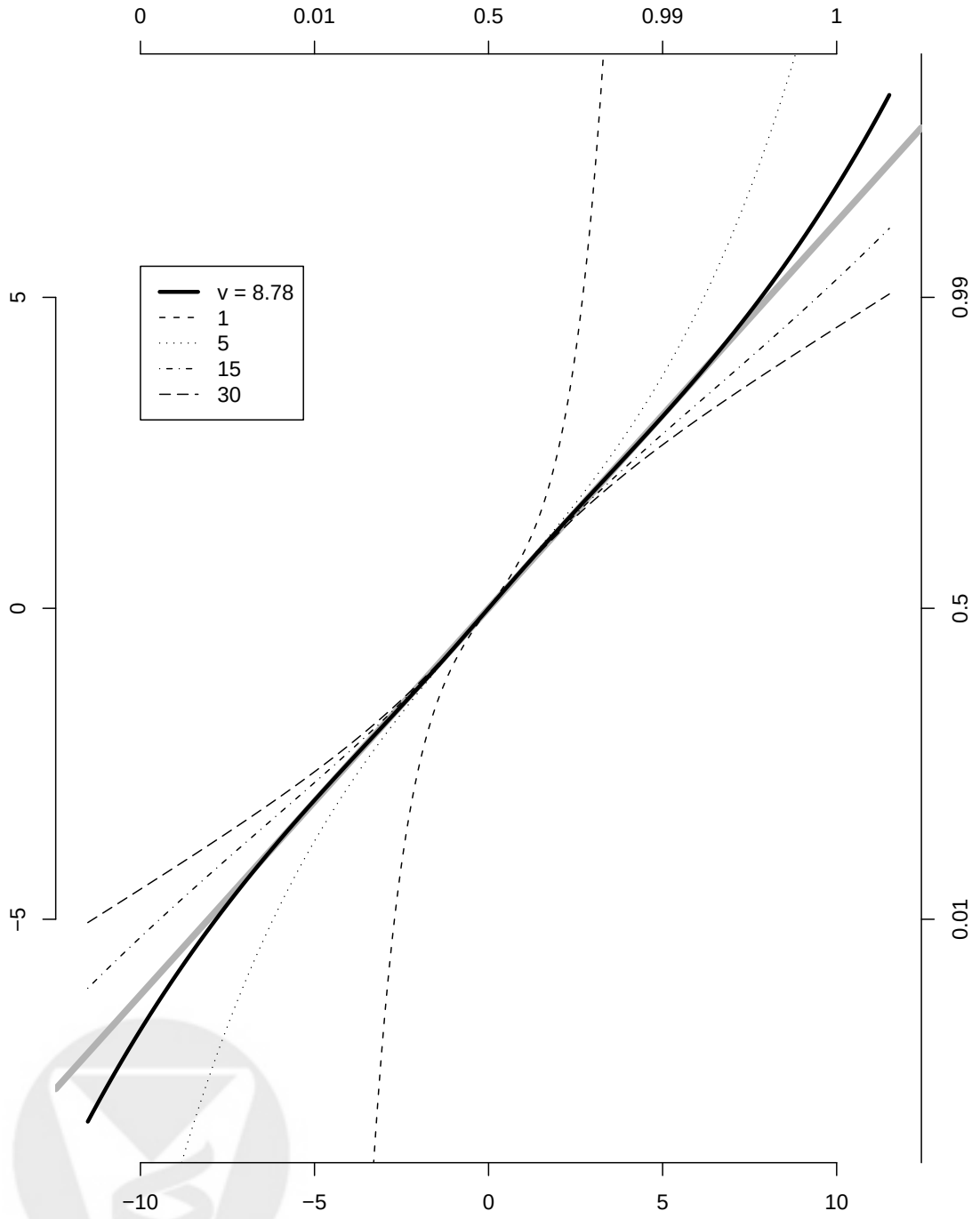

Figure 1: Plots of $T_v^{-1}G$ for various values of v. The dark line is for the suggested 8.7817 degrees of freedom while the grey background line is the linear approximation. The bottom and left axes display the logit scale while the top and right display the corresponding probability scale.

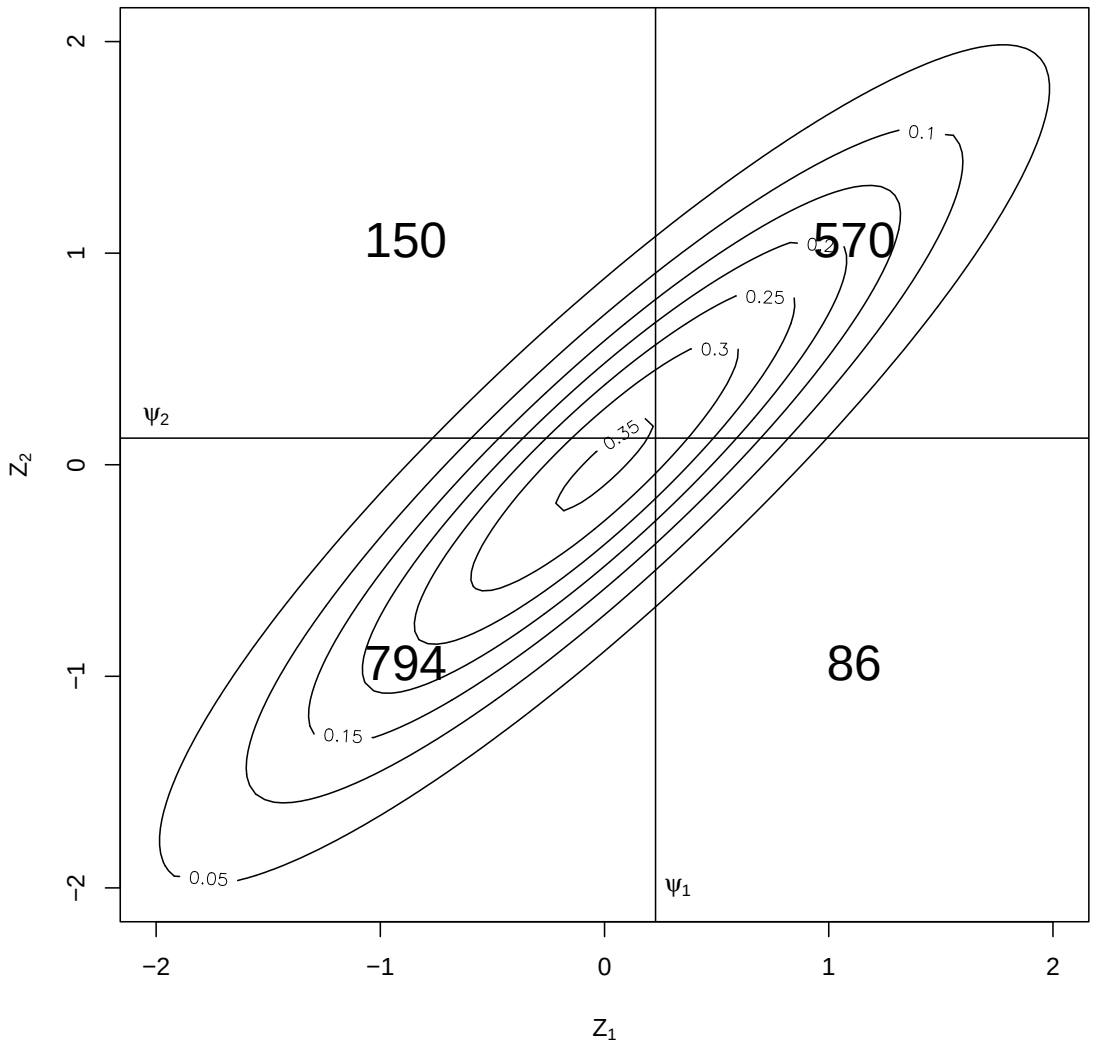

Figure 2: Contour plot of the fitted bivariate normal distribution of the latent variable process, normalized to have unit variances, along with the fitted values of $\psi_j = \Delta_j/(1+\sigma^2)^{1/2}$ and the cell counts for the approval rating data.

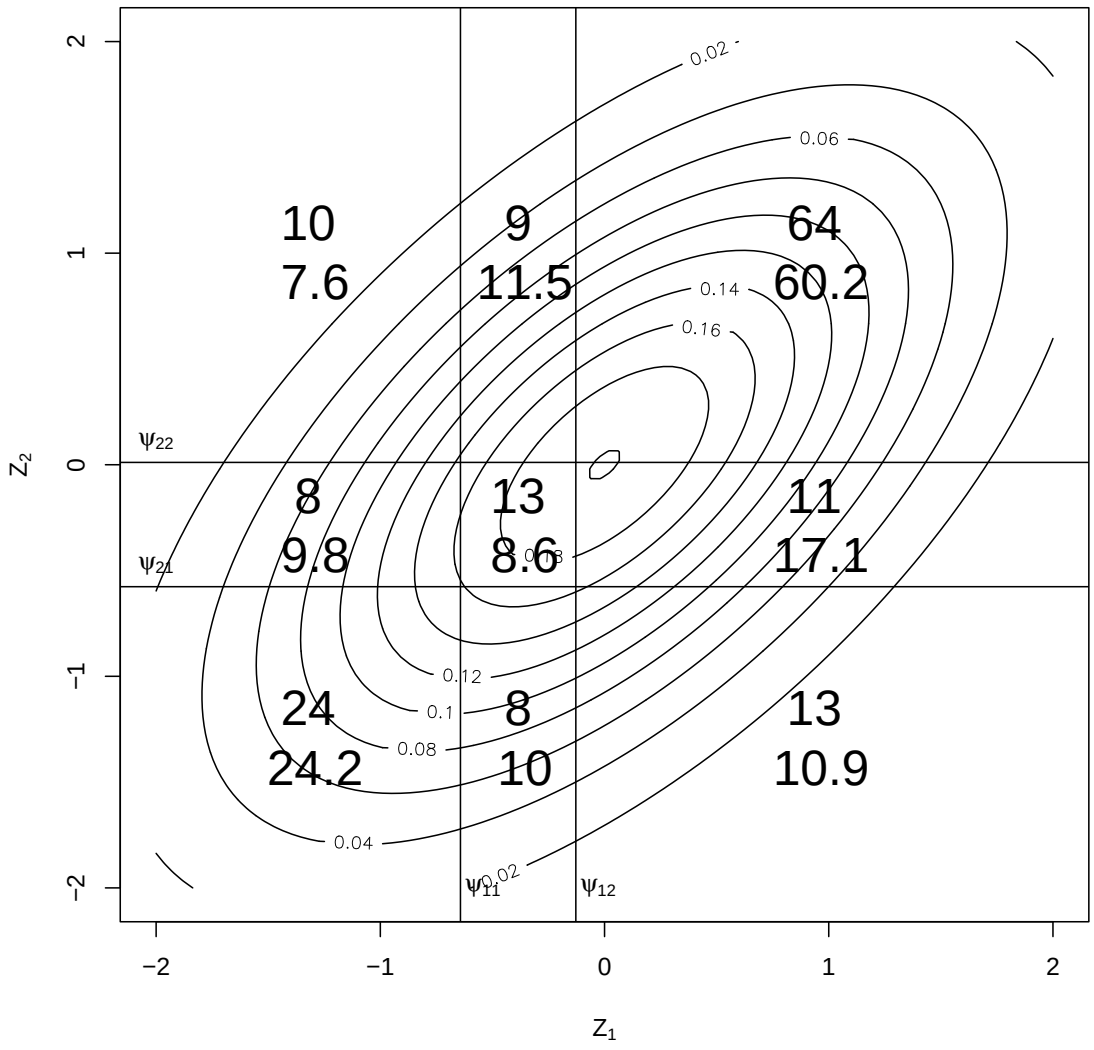

Figure 3: Contour plot of the fitted bivariate distribution, normalized to have unit variances, along with the fitted values of $\psi_{jk} = \Delta_{jk}/(1+\sigma^2)^{1/2}$, the cell counts and the fitted cell counts for the Siskel and Ebert movie data.

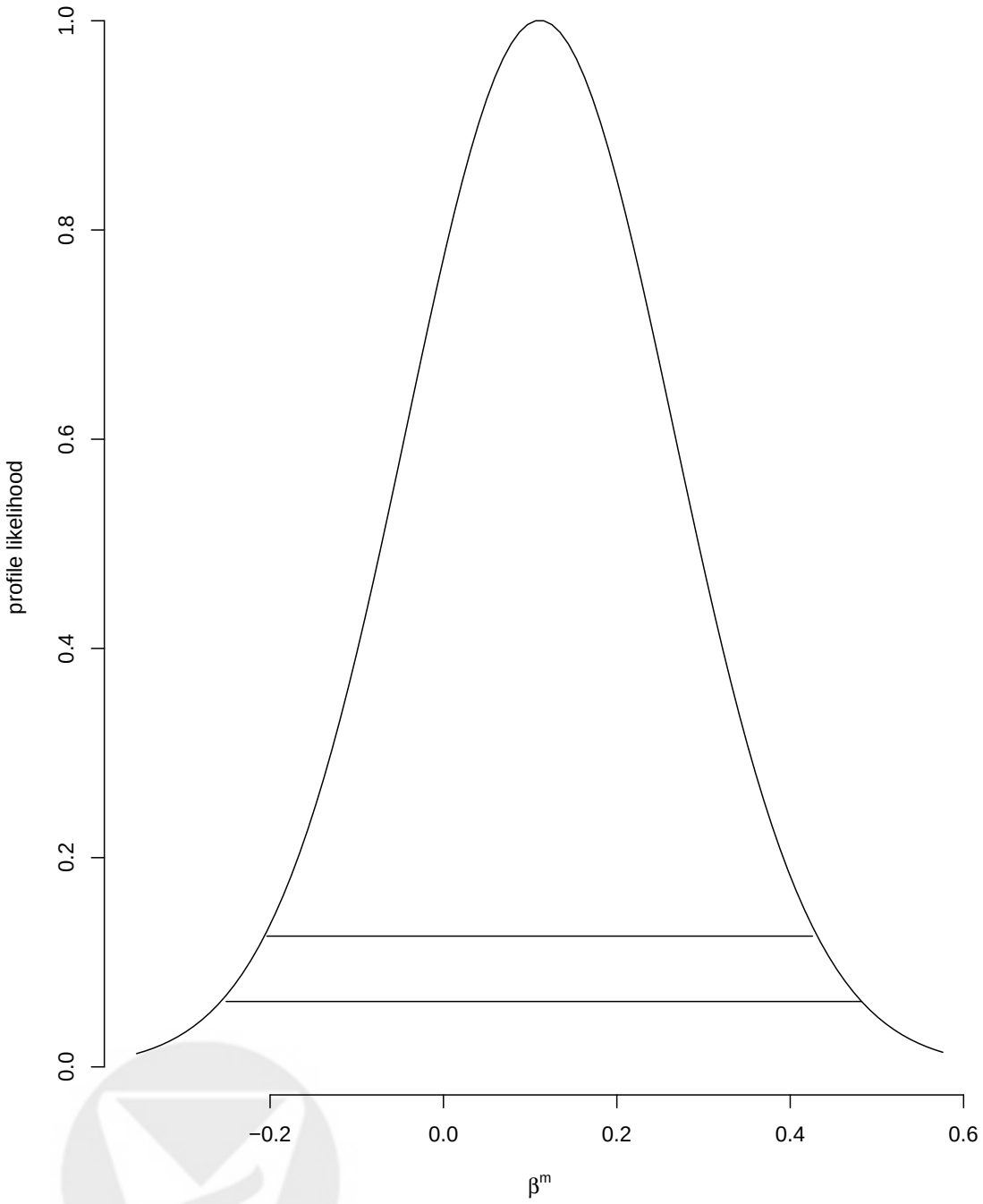

Figure 4: Profile likelihood for the treatment effect, β^m, for movie data with 1/8 and 1/16 reference lines.

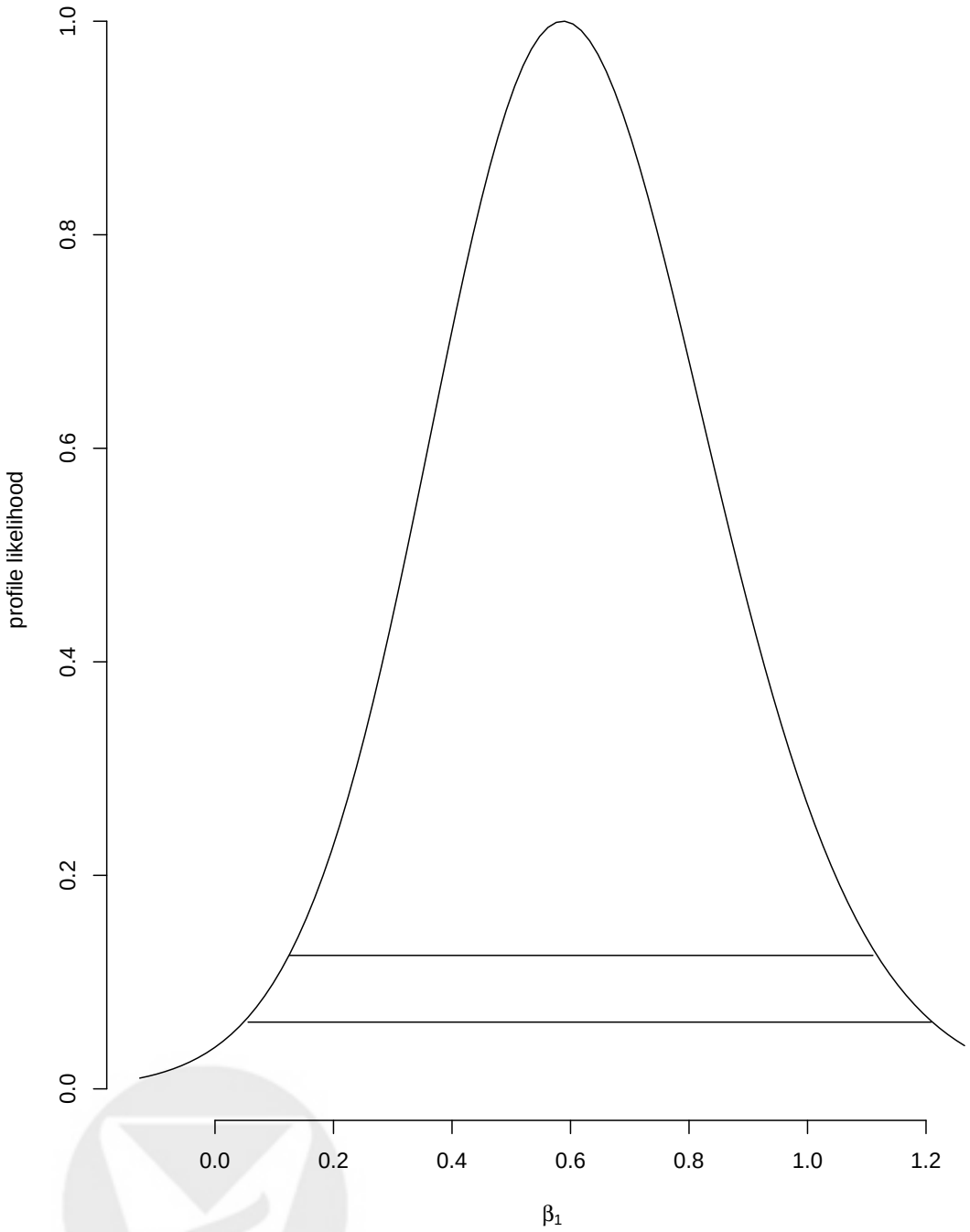

Figure 5: Profile likelihood for the treatment effect, β_1^m, for the crossover data with 1/8 and 1/16 reference lines.

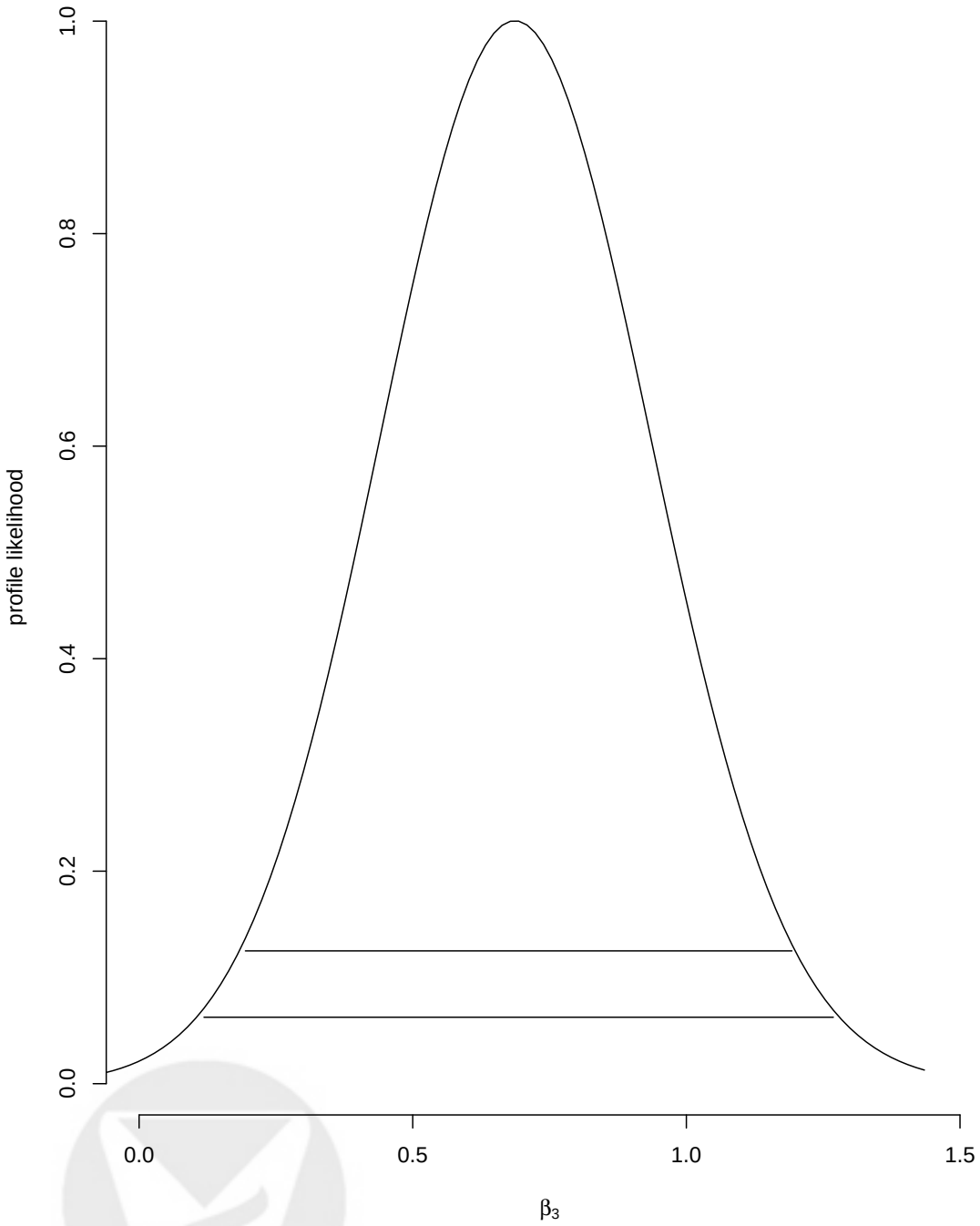

Figure 6: Profile likelihood for the interaction term (β_3^m)for the sleep data with 1/8 and 1/16 reference lines.

C Derivation of the likelihood

Let $\mathcal{L}_i$ be the contribution of subject i to the likelihood. Then, treating the y_{ij} as fixed,

$$\begin{aligned}
\mathcal{L}_i &= \int \prod_{j=1}^{J_i} \prod_{k=1}^{K} \Pr(Y_{ij} = k \mid U_i = u_i)^{I(y_{ij}=k)} dF_u(u_i, \Sigma) \\
&= E\left[\prod_{j=1}^{J_i} \prod_{k=1}^{K} \Pr(\Delta_{ij(k-1)} < W_{ij} - z_{ij}^t u_i \leq \Delta_{ijk} \mid U_i = u_i)^{I(y_{ij}=k)} \right] \\
&= E\left[\prod_{j=1}^{J_i} \Pr(\Delta_{ij(y_{ij}-1)} < W_{ij} - z_{ij}^t u_i \leq \Delta_{ijy_{ij}} \mid U_i = u_i) \right] \\
&= E\left[\Pr\left(\cap_{j=1}^{J_i} [\Delta_{ij(y_{ij}-1)} < M_{ij} \leq \Delta_{ijy_{ij}}] \mid U_i = u_i \right) \right] \\
&= \Pr\left(\cap_{j=1}^{J_i} [\Delta_{ij(y_{ij}-1)} < M_{ij} \leq \Delta_{ijy_{ij}}] \right).
\end{aligned}$$

D Derivation of the multivariate T likelihood

Again let $\mathcal{L}_i$ be the contribution of subject i to the likelihood. Then, again treating the y_{ij} as fixed,

$$\begin{aligned}
\mathcal{L}_i &= \int \prod_{j=1}^{J_i} \prod_{k=1}^{K} \Pr(Y_{ij} = k \mid U_i = u_i, S_i = s_i)^{I(y_{ij}=k)} dF_{us}(u_i, s_i, \Sigma) \\
&= E\left[\prod_{j=1}^{J_i} \prod_{k=1}^{K} \Pr\left(\Delta_{ij(k-1)} \sqrt{\frac{S_i}{v}} < W_{ij} - z_{ij}^t u_i \leq \Delta_{ijk} \sqrt{\frac{S_i}{v}} \mid U_i = u_i, S_i = s_i \right)^{I(y_{ij}=k)} \right] \\
&= \Pr\left(\cap_{j=1}^{J_i} \left[\Delta_{ij(y_{ij}-1)} < \frac{M_{ij}}{\sqrt{\frac{S_i}{v}}} \leq \Delta_{ijy_{ij}} \right] \right).
\end{aligned}$$

E R code for obtaining the log-likelihood

Assume that the data is stored in a list with one element per subject. Each list element is itself a list with components:

`y`	the vector of this subject's responses taking values $1, \ldots, K$
`x`	a design matrix for this subject, *with no intercept*
`z`	the random effect design matrix for this subject
`freq`	a count

In the crossover data, `y` is a vector of ones and twos, `x`, included a treatment indicator and a period indicator; `z` was a vector of ones (a random intercept model.) To describe the variable `freq`, consider the first cell of Table 3. There were 22 subjects who had normal responses on for both the active drug and placebo who received the treatment sequence "Drug-Placebo". It

is wasteful to approximate the associated contributions to the log-likelihood 22 times, instead the contribution log likelihood is appropriately multiplied by 22. For the same reason, the 0 cell should be eliminated.

A function that approximates the log-likelihood is then

```
library(mvtnorm)
linkProbitNormal <- function(alpha, beta, Sigma, dat, G = plogis, ...){
  Li <- sapply(dat,
               function(sub){
                 covMat <- diag(1, length(sub$y)) + sub$z %*% Sigma %*% t(sub$z)
                 Delta <- as.vector(alpha[sub$y + 1] + sub$x %*% beta)
                 DeltaLag1 <- as.vector(alpha[sub$y] + sub$x %*% beta)
                 if (!is.null(G)){
                   Delta <- qnorm(G(Delta)) * sqrt(diag(covMat))
                   DeltaLag1 <-  qnorm(G(DeltaLag1)) * sqrt(diag(covMat))
                 }
                 pHat <- pmvnorm(DeltaLag1, Delta, sigma = covMat, ...)[[1]]
                 return(sub$freq * log(pHat))
               }
               )
  return(sum(Li))
}
```

Here `Alpha` is a vector containing the α_k parameters *including the infinite terms*. By specifying the `G = plogis`, a cumulative logit marginal link is assumed. We can obtain the likelihood for the (subject specific) probit-normal model by specifying `G = NULL`. In contrast `G = pnorm` yields a marginal probit link.

F R code for the logit-T transformation

The following code linearly approximates $T_{df}^{-1}G$ through the origin

```
pLower <- .001
pUpper <- .999
noPoints <- 10 ^ 4
etaVals <- seq(qlogis(pLower), qlogis(pUpper), length = noPoints)
startingValue <- 9
fit <- optim(startingValue,
             function(df)
               mean(resid(lm(I(qt(plogis(etaVals), df = df)) ~ etaVals - 1)) ^ 2),
             method = "BFGS"
             )
df <- fit$par
slope <- coef(lm(I(qt(plogis(etaVals), df = df)) ~ etaVals - 1))
plot(etaVals, qt(plogis(etaVals), df = df), type = "l", col = grey(.7), lwd = 5)
abline(a = 0, b = slope)
```

G R code for the approval rating data

The following code enters the data and gets it into the appropriate format

```
dat <- data.frame(y1 = c(1, 2, 1, 2),
                  y2 = c(1, 1, 2, 2),
                  freq = c(794, 86, 150, 570))

approval <- apply(dat, 1,
                   function(sub){
                     list(y = sub[1 : 2],
                          x = matrix(c(0, 1)),
                          z = matrix(1, 2),
                          freq = sub[3])
                   })
```

The following fits the model

```
startingValues <- c(.2, 0, .5)
fitAlt <- optim(startingValues,
                function(param){
                  alpha <- c(-Inf, param[1], Inf)
                  beta <- param[2]
                  Sigma <- param[3] ^ 2
                  -linkProbitNormal(alpha, beta, Sigma, approval)
                },
                method = "L-BFGS-B",
                lower = c(-Inf, -Inf, 0),
                hessian = TRUE
                )
output <- cbind(fitAlt$par, sqrt(diag(solve(fitAlt$hessian))))
```

H R code for the movie rating data

The following code loads the data set

```
dat <- expand.grid(1 : 3, 1 : 3)
names(dat) <- c("Ebert", "Siskel")
dat$freq <- c(24, 8, 13, 8,13, 11, 10, 9, 64)
movie <- apply(dat, 1,
               function(sub){
                 list(y = sub[1 : 2],
                      x = matrix(0 : 1),
                      z = matrix(1, 2, 1),
                      freq = sub[3])
               }
               )
```

The following code fits the model.

```
startingValues <- c(-0.5827, 0.8673 , 0, .5)
fit <- optim(startingValues,
             function(par){
               alpha <- c(-Inf, cumsum(par[1 : 2]), Inf)
               beta <- par[3]
               Sigma <- par[4] ^ 2
               -linkProbitNormal(alpha, beta, Sigma, movie)
             },
             lower = c(-Inf, 0, -Inf, 0),
             method = "L-BFGS-B",
             hessian = TRUE
             )
output <- cbind(fit$par, sqrt(diag(solve(fit$hessian))))
round(output, 3)
```

I R code for the crossover data

The following enters the data

```
##r1, r2 response at period 1, 2 resp.
##t1, treatment indicator for period 1
shortDat <- expand.grid(1 : 2, 1 : 2, 1 : 0)
names(shortDat) <- c("r1", "r2", "t1")
shortDat$freq <- c(22, 0, 6, 6, 18, 4, 2, 9)
shortDat <- shortDat[shortDat$freq != 0,]
```

The following converts the data to the required format

```
crossover <- apply(shortDat, 1,
                   function(sub)
                     list(y = sub[1 : 2],
                          x = cbind(c(sub[3], 1 - sub[3]), 0 : 1),
                          z = matrix(1, 2, 1),
                          freq = sub[4])
                   )
```

The logit-probit-normal model can be fit with

```
startingValues <- c(.666, .569, -.295, 2)
fit <- optim(startingValues,
             function(param){
               alpha <- c(-Inf, param[1], Inf)
               beta <- param[2 : 3]
               Sigma <- param[4] ^ 2
```

```
              -linkProbitNormal(alpha, beta, Sigma, crossover)
            },
            method = "L-BFGS-B",
            lower = c(-Inf, -Inf, -Inf, 0),
            hessian = TRUE
            )
output <- cbind(fit$par, sqrt(diag(solve(fit$hessian))))
round(output, 3)
```

J R code for the sleep data

The following code enters the data and gets it into the list format

```
dat <- expand.grid(1 : 4, 1 : 4, 1 : 0)
names(dat) <- c("fup", "bl", "trt")
dat$freq <-  c(7,4,1,0,  11,5,2,2,  13,23,3,1,  9,17,13,8,
               7,4,2,1,  14,5,1,0,   6,9,18,2,  4,11,14,22)

sleep <- lapply(1 : nrow(dat),
                function(i)
                list(y = c(dat$bl[i], dat$fup[i]),
                     x = cbind(0 : 1,
                               dat$trt[i],
                               (0 : 1) * dat$trt[i]),
                     z = matrix(1, 2, 1),
                     freq = dat$freq[i]))
```

The following code fits the cumulative logit-probit-normal model

```
fit <- optim(c(1, .1, .1, 1, 1, 1, 1/2),
             function(par){
               alpha <- c(-Inf, cumsum(par[1 : 3]), Inf)
               beta <- par[4 : 6]
               Sigma <- par[7] ^ 2
               -linkProbitNormal(alpha, beta, Sigma, sleep)
             },
             lower = c(-Inf, 0, 0, -Inf, -Inf, -Inf,  0),
             method = "L-BFGS-B",
             hessian = TRUE
             )
output <- cbind(fit$par, sqrt(diag(solve(fit$hessian))))
round(output, 3)
```