Bayesian Statistics

Latent variable methods for ordinal data

Nan Lin
Department of Mathematics
Washington University in St. Louis

Motivation

- How to model common survey variables such as age, education level and income?
    - Such variables are often binned into ordered categories, the number of which may vary from survey to survey
    - interest often lies not in the scale of each individual variable, but rather in the associations between the variables
- For continuous data, we can use correlation analysis or linear regression, but it is often inappropriate for variables listed above

Example: Education attainment

- Goal: describing the relationship between the educational attainment and number of children of individuals in a population.
- Additionally, we might suspect that an individual's educational attainment may be influenced by their parent's education level.
- DEG_i: the highest degree obtained by individual i
- $CHILD_i$: the number of children of individual i
- $PDEG_i$: the binary indicator of whether or not either parent of individual i obtained a college degree.

Model

- Linear regression

$$\mathrm{DEG}_i = \beta_1 + \beta_2 \times \mathrm{CHILD}_i + \beta_3 \times \mathrm{PDEG}_i + \beta_4 \times \mathrm{CHILD}_i \times \mathrm{PDEG}_i + \epsilon_i,$$

$$\epsilon_1, \ldots, \epsilon_n \sim \text{i.i.d. normal}(0, \sigma^2).$$

- Problems with linear regression
  - The variable DEG takes on only a small set of discrete values → normality assumption will certainly be violated.
  - the regression model imposes a numerical scale to the data that is not really present
    - E.g. A bachelor's degree is not "twice as much" as a high school degree,
- DEG is *ordinal*, and linear regression is inappropriate

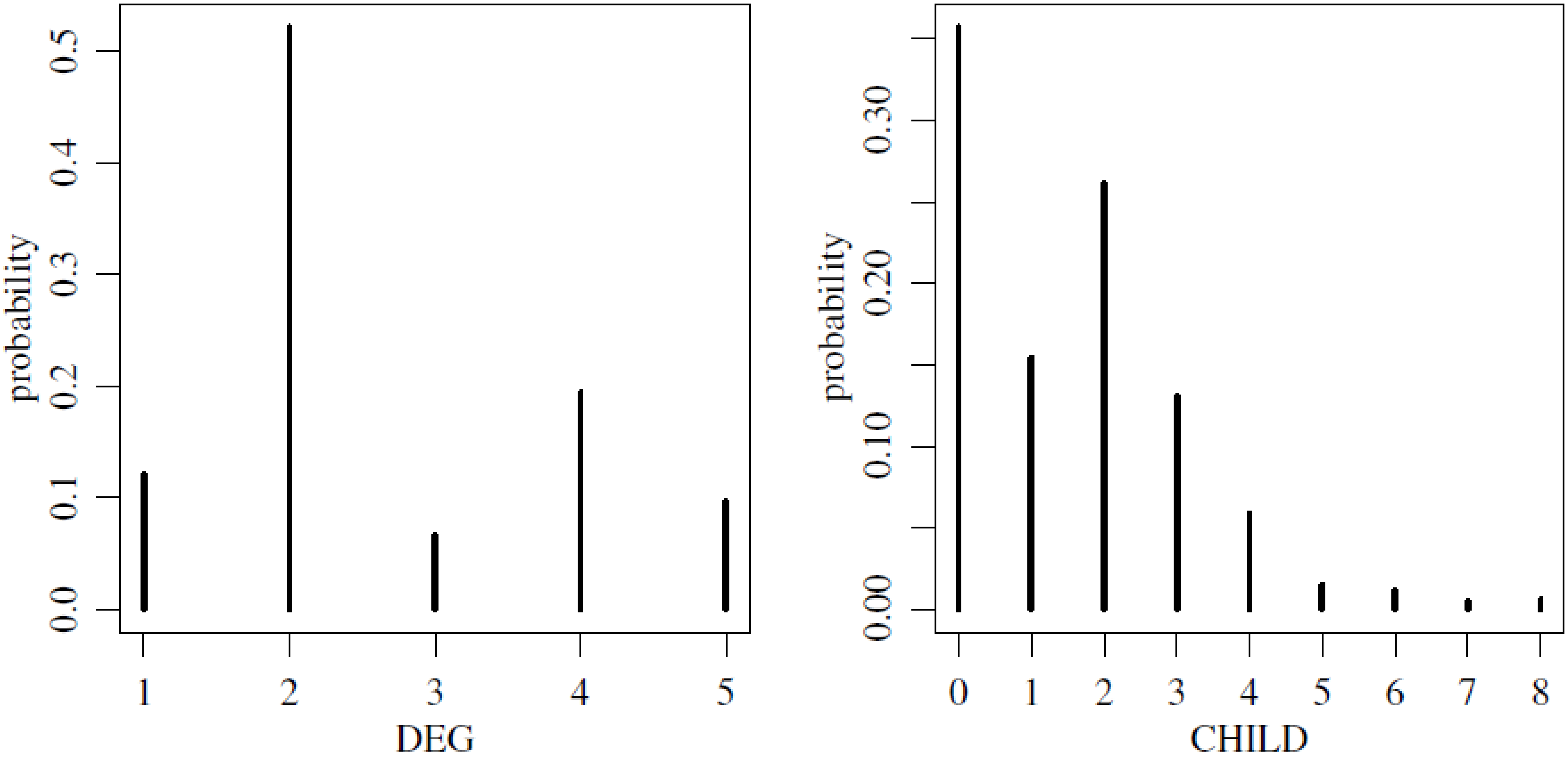


Fig. 12.1. Two ordinal variables having non-normal distributions.

- Ordinal: any variable for which there is a logical ordering of the sample space
- Numeric: variables that have meaningful numerical scales
- Continuous: a variable can have a value that is (roughly) any real number in an interval

- DEG is ordinal but not numeric
- CHILD is ordinal, numeric and discrete
- Variables like height or weight are ordinal, numeric and continuous.

Probit regression

- Assume ordinal, non-numeric variables as arising from some underlying numeric process
  - For example, the severity of a disease might be described "low", "moderate" or "high", although we imagine a patient's actual condition lies within a continuum.
- The amount of effort a person puts into formal education may lie within a continuum, but a survey may only record a rough, categorized version of this variable, such as DEG
- Ordered probit regression

$\epsilon_1, \ldots, \epsilon_n \sim \text{i.i.d. normal}(0, 1)$ ← For identifiability

Latent variable → $Z_i = \beta^T x_i + \epsilon_i$

$Y_i = g(Z_i),$ ← $g()$ is a nondecreasing function

No intercept is needed

Probit regression

- If the sample space for Y takes on K values, say $\{1, \ldots, K\}$, then the function g can be described with only $K - 1$ ordered parameters $\{g_1 < g_2 < \cdots < g_{k-1}\}$

$$
\begin{aligned}
y = g(z) &= 1 \text{ if } & -\infty = g_0 < z < g_1 \\
&= 2 \text{ if } & g_1 < z < g_2 \\
&\vdots & \\
&= K \text{ if } & g_{K-1} < z < g_K = \infty \,.
\end{aligned}
$$

- Unknown parameters
  - “thresholds”: $\boldsymbol{g} = \{g_1 < g_2 < \cdots < g_{k-1}\}$
  - Regression coefficients: β
- To obtain a Gibbs sampling, include the latent variable $\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$ as well.

Gibbs sampling

- Let the observed ordinal response be $\boldsymbol{y} = (y_1, \ldots, y_n)$.
- Prior: $\boldsymbol{\beta} \sim \text{multivariate normal}(\mathbf{0}, n(\mathbf{X}^T\mathbf{X})^{-1})$,
  - Zellner's g-prior
- Full conditional distribution of β: multivariate normal with

$$\mathrm{Var}[\boldsymbol{\beta}|\boldsymbol{z}] = \frac{n}{n+1}(\mathbf{X}^T\mathbf{X})^{-1}, \text{ and}$$

$$\mathrm{E}[\boldsymbol{\beta}|\boldsymbol{z}] = \frac{n}{n+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{z}.$$

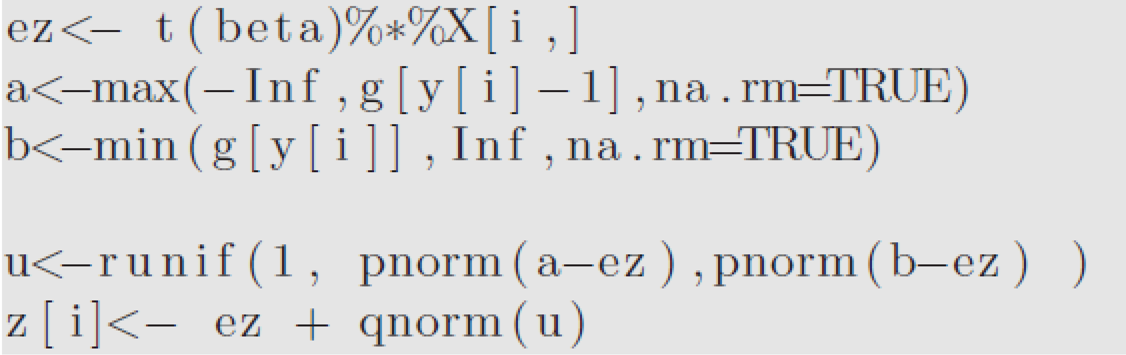
Gibbs sampling

- Full conditional distribution of $\boldsymbol{Z}$ is truncated normal
- Given $Y_i = y_i$, we know that $Z_i \in (g_{y_i-1}, g_{y_i})$
  - Let $(a, b) = (g_{y_i-1}, g_{y_i})$

$$p(z_i|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{g}) \propto \text{dnorm}(z_i, \boldsymbol{\beta}^T \boldsymbol{x}_i, 1) \times \delta_{(a,b)}(z_i).$$

- Sampling from this truncated normal distribution

1. sample $u \sim \text{uniform}(\Phi[(a-\mu)/\sigma], \Phi[(b-\mu)/\sigma])$
2. set $x = \mu + \sigma\Phi^{-1}(u)$



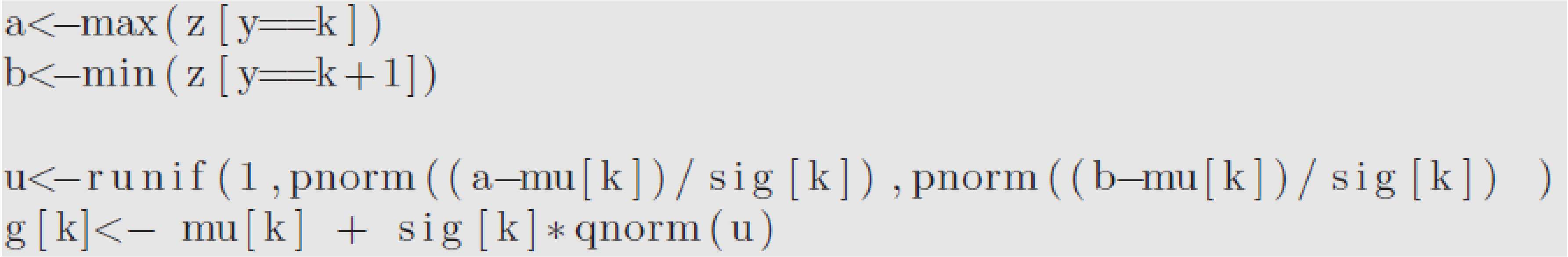
```
ez<- t(beta)%*%X[i,]
a<-max(-Inf,g[y[i]-1],na.rm=TRUE)
b<-min(g[y[i]],Inf,na.rm=TRUE)

u<-runif(1, pnorm(a-ez),pnorm(b-ez) )
z[i]<- ez + qnorm(u)
```

Gibbs sampling

- Given $\boldsymbol{Y} = \boldsymbol{y}$ and $\boldsymbol{Z} = \boldsymbol{z}$, we know from the definition of the thresholds that g_k must be higher than all z_i's for which $y_i = k$ and lower than all z_i's for which $y_i = k + 1.$
- Let $a_k = \max\{z_i : y_i = k\}$ and $b_k = \max\{z_i : y_i = k + 1\}$
- Suppose that the prior of $\boldsymbol{g}$ is $p(\boldsymbol{g})$
- Full conditional distribution of $\boldsymbol{g}$ is then proportional to $p(\boldsymbol{g})$ but constrained to the set $\{g : a_k < g_k < b_k\}$

Gibbs sampling

- For example, if $p(\boldsymbol{g})$ is product of independent normal distributions $\prod_{k=1}^{K-1} \mathrm{dnorm}(g_k, \mu_k, \sigma_k)$ but subject to the constraint $\{g_1 < g_2 < \cdots < g_{k-1}\}$
- Then the full conditional distribution of g_k is a truncated normal distribution with mean and variance (μ_k, σ_k^2) constrained to the interval (a_k, b_k)



```
a<-max(z[y==k])
b<-min(z[y==k+1])

u<-runif(1,pnorm((a-mu[k])/sig[k]),pnorm((b-mu[k])/sig[k])  )
g[k]<-  mu[k]  +  sig[k]*qnorm(u)
```

Example: Education attainment

- Hypothesis: having children reduces opportunities for educational attainment (Moore and Waite, 1977)
- Data: 1994 General Social Survey
  - Variables
    - DEG_i: the highest degree obtained by individual i
    - $CHILD_i$: the number of children of individual i
    - $PDEG_i$: the binary indicator of whether or not either parent of individual i obtained a college degree.
  - 959 of the 1,002 survey respondents we have complete data on the variables DEG, CHILD and PDEG
  -
    ```
    dat<-read.table("http://lib.stat.cmu.edu/aoas/107/data.txt",header=TRUE)
    ```

Example: Education attainment

- $Y_i = DEG_i$ and $\boldsymbol{x}_i = (CHILD_i, PDEG_i, CHILD_i \times PDEG_i)$
- Prior:
  - $\beta \sim N(0, n(X^T X)^{-1})$
  - $p(\boldsymbol{g}) \propto \prod_{k=1}^{K-1} \text{dnorm}(g_k, 0, 100)$ constrained to $g_1 < \cdots < g_{K-1}$.
- Gibbs sampling: 25,000 scans
  - Thining: every 25th value
- Posterior mean regression line
  - For people without a college-educated parent

$$\mathrm{E}[Z|\boldsymbol{y}, x_1, x_2 = 0] = -0.024 \times x_1$$

  - For people with a college-educated parent

$$\mathrm{E}[Z|\boldsymbol{y}, x_1, x_2 = 1] = 0.818 + 0.054 \times x_1.$$

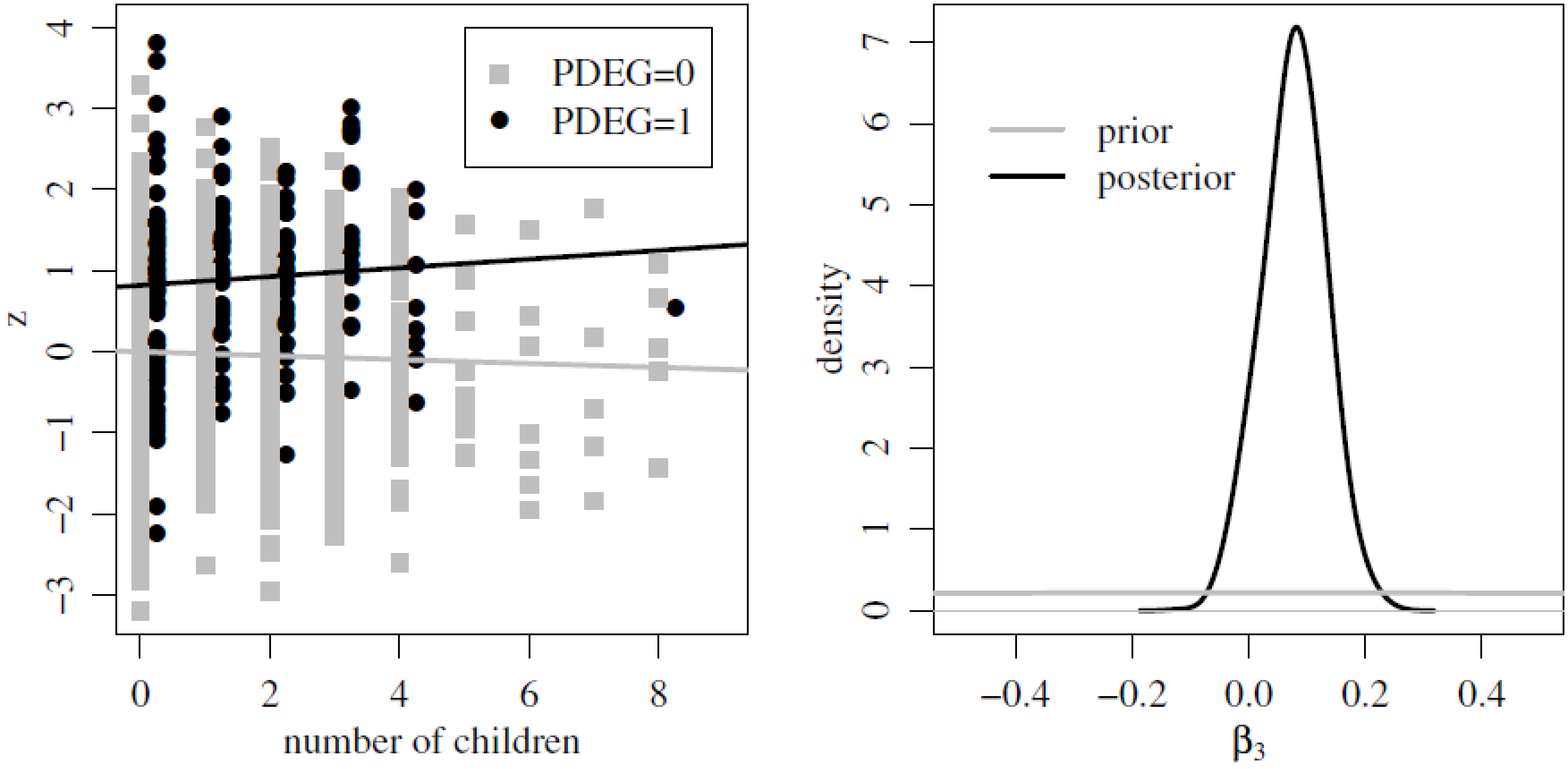


Fig. 12.2. Results from the probit regression analysis.

The lines suggest that for people whose parents did not go to college, the number of children they have is indeed weakly negatively associated with their educational outcome. However, the opposite seems to be true among people whose parents went to college. 95% credible interval of β_3 is (-0.026,0.178), which contains zero, but still shows some evidence of the difference in the slopes.

Transformation models and the rank likelihood

- Specifying the prior for the "thresholds" is not an easy task, especially when the number of categories is large
  - the incomes (INC) of the subjects in the 1994 GSS dataset were each recorded as belonging to one of 21 ordered categories
- An alternative approach using rank likelihood
  - Avoid estimating the function $g(z)$, i.e. the "thresholds"

Transformation models and the rank likelihood

- If Z_i's are observed, the problem is just a linear regression.
- Though not observed, we know something about Z_i's from the observed Y_i's
    - If $y_1 > y_2$, then $z_1 > z_2$
- Given the observed $\boldsymbol{Y} = \boldsymbol{y}$, we know that Z_i's must be in the set $R(\boldsymbol{y}) = \{z \in \mathbb{R}^n : z_{i_1} < z_{i_2} \text{ if } y_{i_1} < y_{i_2}\}$.
- Then we can derive posterior inference by conditioning on $\boldsymbol{Z} \in R(\boldsymbol{y})$. That is,

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{Z} \in R(\boldsymbol{y})) &\propto p(\boldsymbol{\beta}) \times \boxed{\Pr(\boldsymbol{Z} \in R(\boldsymbol{y})|\boldsymbol{\beta})} \quad \longleftarrow \text{Rank likelihood} \\ &= p(\boldsymbol{\beta}) \times \int_{R(\boldsymbol{y})} \prod_{i=1}^{n} \text{dnorm}(z_i, \boldsymbol{\beta}^T \boldsymbol{x}_i, 1)\ dz_i . \end{aligned}$$

Rank likelihood

- It is called a rank likelihood because for continuous data it contains the same information about $\boldsymbol{y}$ as knowing the ranks of $\{y_1, \ldots, y_n\}$, i.e. which one has the highest value, which one has the second highest value, etc
- If Y is discrete then observing $R(\boldsymbol{y})$ is not exactly the same as knowing the ranks, but for simplicity we will still refer to $\Pr(Z \in R(\boldsymbol{y})|\beta)$ as the rank likelihood, whether or not Y is discrete or continuous.
- The key here is that the rank likelihood avoid specifying $g(Z)$.
- But the rank likelihood itself involves complicated integrals

Gibbs sampling

- Gibbs sampling: iteratively sampling from the full conditional distributions of β and Z
- Full conditional of β
  - Given a current value z of Z, the full conditional density $p(\beta|Z = z, Z \in R(\boldsymbol{y}))$ reduces to $p(\beta|Z = z)$ because knowing the value of Z is more informative than knowing just that Z lies in the set $R(\boldsymbol{y})$.
  - Under a multivariate normal prior, the full conditional distribution of β is also multivariate normal

Gibbs sampling

- Full conditional of $Z_i|\beta, \boldsymbol{Z} \in R(\boldsymbol{y}), \boldsymbol{z}_{-i}$
  - Our model assumes that $Z_i|\beta \sim N(\beta^T \boldsymbol{x}_i, 1)$
  - When conditioning also on $(\boldsymbol{Z} \in R(\boldsymbol{y}), \boldsymbol{z}_{-i})$, the distribution will be a truncated normal distribution
- $\boldsymbol{Z} \in R(\boldsymbol{y})$ is obtained from the fact $y_i < y_j$ implies $Z_i < Z_j$, and $y_i > y_j$ implies $Z_i > Z_j$
  - This means that Z_i must lie in the following interval

$$\max\{z_j : y_j < y_i\} < Z_i < \min\{z_j : y_i < y_j\}\,.$$

  - Let a and b be the lower and upper bound of the above interval. Then

$$p(z_i|\boldsymbol{\beta}, \boldsymbol{Z} \in R(\boldsymbol{y}), \boldsymbol{z}_{-i}) \propto \text{dnorm}(z_i, \boldsymbol{\beta}^T \boldsymbol{x}_i, 1) \times \delta_{(a,b)}(z_i).$$

  - Same form as in ordered regression, but the truncation depends directly on other Z_i's but not the "thresholds"

Order probit regression vs rank likelihood

- If the number of categories of the response Y is small and the sample size is large, two methods should perform similarly
- The rank likelihood approach is applicable to a wider array of datasets since with this approach, Y is allowed to be any type of ordinal variable, discrete or continuous.
- The drawback to using the rank likelihood is that it does not provide us with inference about $g(z)$, which describes the relationship between the latent and observed variables.

Example: Education attainment

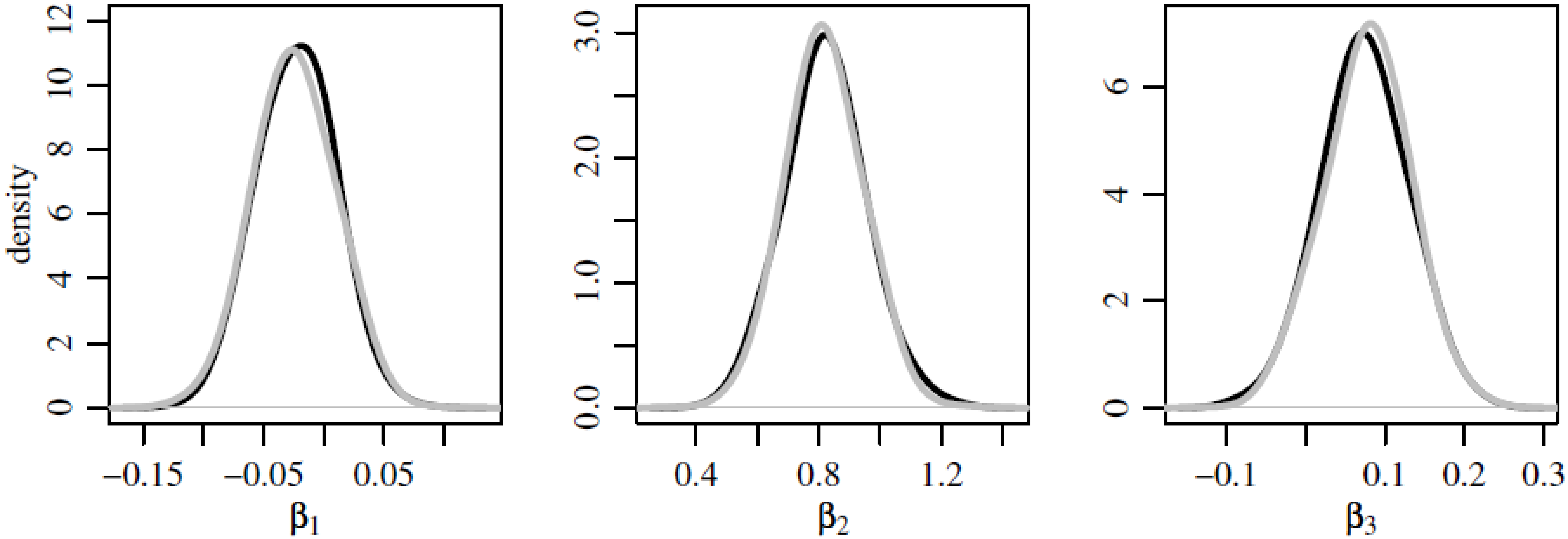


Fig. 12.3. Marginal posterior distributions of $(\beta_1, \beta_2, \beta_3)$, under the ordinal probit regression model (in gray) and the rank likelihood (in black).