The ISBA Bulletin

Vol. 14 No. 3 September 2007

The official bulletin of the International Society for Bayesian Analysis

In this issue

APPLICATIONS

Rank likelihood estimation for continuous and discrete data

Peter Hoff
hoff@stat.washington.edu

Consider the semiparametric regression model

$$\begin{aligned} z_i &= \beta^T x_i + \epsilon_i \\ y_i &= g(z_i), \end{aligned}$$

where β is a vector of regression coefficients and g is an unknown non-decreasing function. In many situations, interest lies in the association between y and x (represented by β), but not in the measurement scale of y (represented by g). A rank likelihood is a type of semiparametric marginal likelihood function that is useful for these situations, as it depends on β only and not on the nuisance parameter g. While procedures for obtaining MLEs based on the rank likelihood are complicated, it turns out that the associated Markov chain Monte Carlo procedures for Bayesian inference are extremely simple, often requiring just a few additional lines of R-code beyond those required by ordinary methods. In this short article I motivate the rank likelihood in the context of regression and copula estimation, illustrate the methodology with an example and provide computer code to implement the necessary MCMC algorithm.

Motivation

While the normal model serves us well when describing the variability of the sample mean, many of us find it lacking as a realistic sampling model for much else. I became acutely aware of this the first time I taught estimation for normal populations to a group of social science graduate students. It must have taken me the better part of a day to find an interesting example of a real-life social science dataset that included a variable anywhere close to being normally distributed, and in the end it still needed a log transformation.

The students in my class were used to working with survey data that included variables such as sex, education level, attitudes and income: variables that we may consider binary, ordinal and continuous. Often the scale on which these variables are measured is arbitrary - income, age and attitude variables are often binned into ordered categories, the number of which varies from survey to survey. Furthermore, interest in these variables typically lies not in their univariate marginal distributions, but rather in their multivariate associations: Is the relationship between two variables increasing, decreasing or zero? Is the relationship monotonic or quadratic? What happens if we "account" for a third variable?

Transformation models

Most model-based approaches to answering such questions have one of two undesirable features: either they rely on the observed data being normally distributed, or they require a lot of effort to simultaneously estimate the marginal distributions along with the association parameters. Fortunately, an interesting alternative to these approaches exists: Consider the following regression model for the conditional distribution of $y_1, \ldots, y_n$ given $x_1, \ldots, x_n$:

$$\begin{aligned} \epsilon_1, \ldots, \epsilon_n &\sim \text{i.i.d. normal}(0,1) \\ z_i &= \beta^T x_i + \epsilon_i \\ y_i &= g(z_i) \end{aligned}$$

The unknown parameters in this system are β and g, the latter of which can be assumed to be a nondecreasing function that describes the marginal distribution of y. If y is discrete with a finite number of levels then the above model is an ordered probit model and g is determined by its points of discontinuity. If y is continuous then g is some unknown increasing function. In either case, a full Bayesian analysis would require prior distributions for β and g, even if only β is of interest. However, there is information in the data about β that doesn't depend on the nuisance parameter g: We don't observe the z_i's directly, but since g is monotone we do know the *order* of the z_i's. In particular, we know that z lies in the set

$$R(y) = \{z \in \mathbb{R}^n : z_i < z_j \text{ if } y_i < y_j\}. \qquad (1)$$

Note that since the distribution of z doesn't depend on g, the probability that $z \in R(y)$ for a given y also doesn't depend on the nuisance parameter g:

$$\begin{aligned} p(z \in R(y)|\beta, g) &= \int_{R(y)} \prod_{i=1}^{n} \phi(z_i - \beta^T x_i)\, dz_i \\ &= p(z \in R(y)|\beta) \end{aligned}$$

For continuous data, $p(z \in R(y)|\beta)$ is the same as the probability of the observed ranks. Taken as a function of β, this forms the "rank likelihood," introduced in the regression context by Pettitt [1982]. The rank likelihood is a type of marginal likelihood that depends on the parameter of interest β and not on the nuisance parameter. Doksum [1987] has studied this type of likelihood for general transformation models, which includes the proportional hazards model as a special case, and Bickel and Ritov [1997] study the asymptotic properties of the rank likelihood estimator of β. For discrete data, the information contained in $\{z \in R(y)\}$ is less than that contained in the ranks, because the former does not contain information about ties. However, $p(z \in R(y)|\beta)$ still provides a marginal likelihood for β which doesn't depend on the nuisance parameter g.

Rank likelihood estimation

Given the observed value y_{obs} of y, the rank likelihood estimate of β is obtained by maximizing $p(z \in R(y_{\text{obs}})|\beta)$ as a function of β. The fact that the likelihood involves a complicated integral makes obtaining the MLE very difficult, and existing estimation methods offer only approximate MLEs. This has probably been the greatest obstacle to the widespread adoption of the rank likelihood approach to regression. However, it turns out that Bayesian estimation using the rank likelihood is comparatively straightforward. Taking the event $\{z \in R(y_{\text{obs}}\}$ as our observed information, we can obtain samples of $\{z, \beta\}$ conditional on this information via iterative Gibbs sampling. The relevant full conditional distributions are quite simple:

$p(\beta|z, z \in R(y_{\text{obs}})) = p(\beta|z)$ is a multivariate normal distribution (assuming $p(\beta)$ is multivariate normal).

$p(z_i|\beta, z_{-i}, z \in R(y_{\text{obs}}))$ is a normal density constrained to the interval

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

Example

Let's take a look at how rank likelihood estimation can be implemented in R in the context of an example. The 1996 General Social Survey gathered a wide variety of information on the adult U.S. population, including each survey respondent's sex, their self-reported frequency of religious prayer (on a six-level ordinal scale), and the number of items correct out of 10 on a short vocabulary test. We'll estimate the parameters in a regression model for y_i=prayer as a function of $x_{i,1}$ = sex of respondent (0-1 indicator of being female) and $x_{i,2}$ = vocabulary score. Our model is

$$\begin{aligned} z_i &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_{12} x_{i,1} x_{i,2} + \epsilon_i \\ y_i &= g(z_i) \end{aligned}$$

From these data and this model we hope to learn if the relationship between prayer and vocabulary score is positive, negative or zero, and whether or not the relationship is different for men and women. Letting (`y`, `X`) be R-objects containing the data, R-code for the estimation procedure described above is as follows:

```
##### data setup and starting values
n<-dim(X)[1] ; p<-dim(X)[2]
ranks<-match(y,sort(unique(y)))
uranks<-sort(unique(ranks))
z<-qnorm(rank(y,ties.method="random")/(n+1))
b<-matrix(0,p,1)
#####

for(s in 1:S) {

  ##### update z
  mu<-X%*%b
  for(r in sample(uranks)) {
    ir<-(1:n)[ranks==r]
    lb<-max(z[ranks<r]) ; ub<-min(z[r<ranks])
    z[ir]<-qnorm(
             runif( length(ir),
               pnorm(lb,mu[ir],1),
               pnorm(ub,mu[ir],1)
                   ),
              mu[ir],1
                 )
                           }
  #####

  ##### update b
  V<-solve ( t(X)%*%X +diag(1,nrow=p) )
  E<-V%*%( t(X)%*%z )
  b<-chol(V)%*%rnorm(p) + E
  #####
              }
```

Data and code for this example are available at www.stat.washington.edu/hoff/ISBAexample. In practice, the mixing of the Markov chain is improved if the columns of X are centered to have mean zero.

I ran this algorithm for 25,000 iterations, saving the value of β every 25th iteration leaving 1000 samples with which to estimate the posterior distribution. Some posterior quantiles for the regression parameters are as follows:

	2.5%	50%	97.5%
β_1	0.45	0.88	1.29
β_2	-0.06	-0.02	0.01
β_{12}	-0.10	-0.08	-0.05

These results indicate that the relationship between prayer and vocabulary score differs between men and women: The (2.5,50,97.5)% quantiles for the sex specific slope parameters are (-.13,-.10, -.06) for women and (-.06, -.02, .01) for men, indicating that women's prayer rate decreases more rapidly as a function of vocabulary than does that of the men. This is shown graphically in the figure, which plots the posterior mean regression lines for both sexes, along with a single posterior sample of z (the last sample from the Markov chain).

Copula estimation

In the above example all three variables were sampled. In such situations it may be desirable to estimate the joint dependence among all three variables. This can be accomplished with the Gaussian copula model:

$$\begin{aligned} z_i = (z_{i,1}, \ldots, z_{i,p}) &\sim \text{multivariate normal}(0, \Sigma) \\ y_{i,j} &= g_j(z_{i,j}),\ j \in \{1, \ldots, p\} \end{aligned}$$

As described in Hoff [2007], estimation of Σ using the rank likelihood can be implemented by conditioning on the event

$$R(y) = \{z_1, \ldots, z_n : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}.$$

This estimation procedure does not require modeling the univariate marginal distributions, and is applicable for mixed discrete and continuous data.

Summary

By only using part of the observed information, rank likelihoods allow for estimation of dependence parameters without having to deal with high-dimensional nuisance parameters. Distributions of the dependence parameters, conditional on the partial information, are easily approximated via Gibbs sampling. Besides regression and copula estimate, there are undoubtedly a variety of other semiparametric inference problems that can be addressed by a combination of rank likelihood and Bayesian methodology.

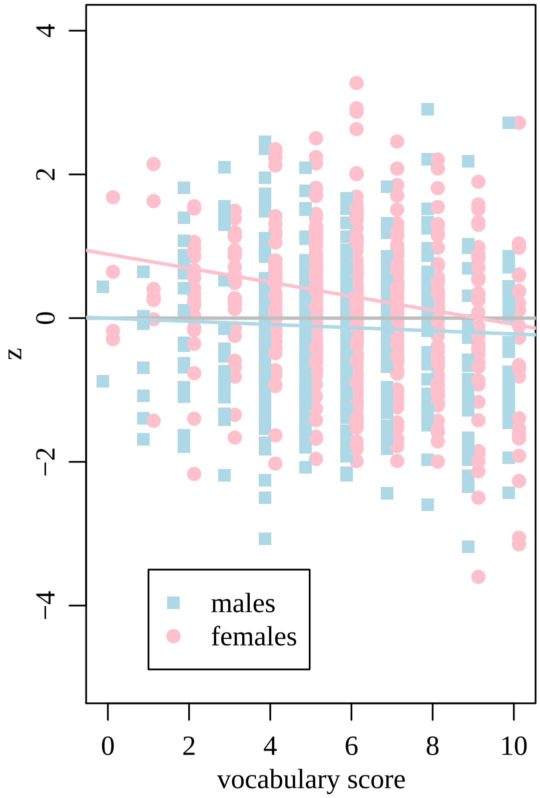
References

P. J. Bickel and Y. Ritov. Local asymptotic normality of ranks and covariates in transformation models. In *Festschrift for Lucien Le Cam*, pages 43–54. Springer, New York, 1997.

Kjell A. Doksum. An extension of partial likelihood methods for proportional hazard models to general transformation models. *Ann. Statist.*, 15(1):325–345, 1987. ISSN 0090-5364.

Peter D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Statist.*, 1(1):265–283, 2007.

A. N. Pettitt. Inference for the linear model using a likelihood based on ranks. *J. Roy. Statist. Soc. Ser. B*, 44(2):234–243, 1982. ISSN 0035-9246.


z
vocabulary score
males
females