



**Web Supplement to
“Recommendation Systems with Purchase Data,”
Journal of Marketing Research, February 2008**

Anand V. Bodapati
UCLA Anderson Graduate School of Management
bodapati@ucla.edu
phone: 310-206-8624

We elaborate here on some details on several statistical issues with respect to estimation of the dual latent class model.

1 Specifications of the full conditional densities

Drawing values of π_{user} and π_{item} :

Considering drawing for the posterior of π_{user} . This vector has C_{user} elements. Let us index the vector by s which takes values from 1 to C_{user} . Let π_s denote the value in π in the s th position. The parameters of the Dirichlet are the class sample sizes, which we denote by $\{n_s\}_{s=1}^{C_{\text{user}}}$. The Dirichlet distribution has the function form:

$$\text{Dirichlet}(\pi) = \frac{\Gamma(\sum_{s=1}^{C_{\text{user}}} n_s)}{\prod_{s=1}^{C_{\text{user}}} \Gamma(n_s)} \prod_{s=1}^{C_{\text{user}}} \pi_s^{n_s-1}.$$

Let n_0 be a scalar used to denote the “prior” sample size for each of the latent classes. In our empirical work, we took the value of n_0 to be 1. Let m_s be used to denote the number of users that fall into latent class s in any particular iteration. Then the conditional posterior for π_{user} is the Dirichlet density with class sample sizes given by $n_s = m_s + n_0$.

The conditional posterior density for π_{user} is also Dirichlet, constructed analogously.

Drawing values of Z_u^{user} and Z_i^{item} :

Each Z_u^{user} takes values from 1 to C_{user} . The posterior conditional distribution of Z_u^{user} is Multinomial with sample size 1 and the probability for the s th latent class being

$$\frac{\pi_{\text{user}}(s) L_{u,s}^{\text{user}}}{\sum_{s'} \pi_{\text{user}}(s') L_{u,s'}^{\text{user}}}.$$

To draw from this multinomial density we use a sequence of binomial draws as explained on page 583 of Gelman *et al* (2003).

The posterior conditional distribution of each Z_i^{item} is constructed analogously.

Drawing values of $\{\beta_s\}_{s=1}^{C_{\text{user}}}$:

There is no simple mechanism to draw from the posterior density of β_s . One approach is described in the technical appendix. It might initially appear that we would be forced to use the Metropolis-Hastings algorithm, with the accompanying difficult task of constructing a suitable proposal density mechanism. However, we do not have to do this. Because the conditional posterior of β_s turns out to be globally log-concave, we can instead use the Gilks-Wild Adaptive Rejection Sampling algorithm embedded in Gibbs sampling iterations. This is discussed in Gilks and Wild (1992). In our application, we find this to be more computationally efficient than using Metropolis-Hastings.

Drawing values of $\{\alpha_s\}_{s=1}^{C_{\text{user}}}$: These values are drawn using ideas very close to what was discussed for the β vectors. Again, because the conditional posterior of α_s is globally log-concave we can use Adaptive Rejection Sampling in Gibbs iterations instead of Metropolis-Hastings.

Drawing values of $\{x_t\}_{t=1}^{C_{\text{item}}}$:

The conditional posterior of x_t is not globally log-concave and we have to use the Metropolis-Hastings algorithm. We use the random-walk version, with the proposal density being a Gaussian density centered at the current iteration's value of x_t . The dispersion of the proposal density is adapted over the iterations following ideas in Chen, Shao and Ibrahim (2000) and Rossi, Allenby and McCulloch (2006). While the conditional posterior of x_t is not globally log-concave, it is *locally* log-concave near the mode of the posterior. This suggests the following strategy which we follow to increase computational efficiency: Use Metropolis-Hastings initially until the chain stabilizes. Then switch over to the Adaptive Rejection Sampling framework. To guard against unexpected departures from log-concavity in the vicinity of the mode, we use a variant called "Adaptive Rejection Metropolis Sampling," which is discussed in Gilks *et al* (1995,1997).

2 Comments on Identification

There are two distinct approaches to address identification issues described on the paper. The first approach is to accept the fact that the likelihood does not identify the parameters. Recall that the priors on these parameters are zero-centered Gaussians. Therefore, it turns out that the posterior density is proper and the MCMC iterations would (at least *in theory*) converge to a stable distribution despite the likelihood function not identifying. The posterior draws of \mathbf{A} , \mathbf{B} , \mathbf{X} would of course vary widely over the iterations because the only stability is coming via the prior which is generally chosen to be weak. Despite this variability, it may turn out that the posterior draws of $\mathbf{A}^T \mathbf{X}$ and

$\mathbf{B}^T \mathbf{X}$ may not be of high variability. This essentially is the approach taken by McCulloch & Rossi (1994) to address an identification issue that comes up in the multinomial probit model. We do not pursue this approach in this paper because it does not fare well in our application.

The second approach is to reduce the parameter space. The paper presents two different approaches to doing this and gave reasons for favoring the “anchoring” strategy. We now say more about this. Recall that in the anchoring strategy, we fix d rows of the \mathbf{X} matrix to some non-invertible matrix. Recall that the specific anchoring used can affect the mixing level and dispersion of the Markov Chain. We deferred the question of which specific anchoring to use and return to it now. We experimented with a small number of anchoring schemes. One scheme that seems to work fairly well is the following: (a) Anchor the d rows corresponding to the d item-segments with the lowest population sizes, (b) Anchor them to their values in the last iteration of the mode-finding procedure, (c) If the $d \times d$ matrix value so chosen is nearly singular as measured by the matrix’s condition-number, then remove the anchoring for the row contributing most to singularity and instead anchor the row corresponding to the next least populated segment. We repeat step (c) until we obtain a set of d item-segments such that they are of low population and the corresponding \mathbf{X} rows are far from singularity. We do not fully understand why this anchoring works and we have insufficient experience with alternatives to be able to comment properly. Our speculation is that because there is less data on the low population segments, the corresponding \mathbf{X} if not anchored tend to have high variance, and anchoring them reduces the variability. Second, because the matrix \mathbf{X} is of rank d , every row of \mathbf{X} is a linear combination of the d anchored rows. If the anchored rows are nearly singular, then the weights in the linear combination can become very large and this can cause high variance in the Markov Chain and introduce instability.

3 Exact ML estimation for rank-reduced binomial regression via weighted Frobenius norm minimizations

In Appendix B, we presupposed a mechanism for exact ML estimation for rank-reduced binomial regression. This section introduces rank-reduced binomial regression and shows that exact ML estimates for this problem can be obtained by a sequence of weighted Frobenius norm minimizations. We will introduce new notation here which ought not be confused with any of the symbols previously introduced in this paper.

In rank-reduced binomial regression, we are given two matrices \mathbf{S} and \mathbf{T} , each of size $I \times J$. Additionally there is a matrix \mathbf{P} of success probabilities. We are also given two matrices: β of size $I \times d$ and \mathbf{x} of size $d \times J$. Let the i th row of β be denoted by β_i and let the j th column of \mathbf{x} be

denoted by x_j . Finally, the (i, j) th elements of \mathbf{S} , \mathbf{T} and \mathbf{P} are denoted as respectively S_{ij} , T_{ij} and P_{ij} .

If for every (i, j) , the number of “successes” S_{ij} is binomial with success probability P_{ij} and number of trials T_{ij} , and if $\mathbf{P} = \text{logit}(\beta \times \mathbf{x})$ so that

$$P_{ij} = \text{logit}(\beta_i^T x_j),$$

then \mathbf{S} is said to come from a rank-reduced binomial regression with parameters β , \mathbf{x} . It is “reduced-rank” in the sense that the rank of the success probabilities matrix \mathbf{P} is only d which is typically chosen to be considerably smaller than I or J . Consequently, the total number of parameters is $(I + J)d$, a number which is reduced to $(I + J)d - d^2$ because of d^2 constraints that need to be imposed for identification. The number $(I + J)d - d^2$ is typically much smaller than the IJ parameters required for a non-rank-reduced model.

The negative log-likelihood function for the above model is easily written as:

$$NLL(\beta, \mathbf{x}) = - \sum_{i,j} S_{ij} \log P_{ij} + (T_{ij} - S_{ij}) \log(1 - P_{ij}). \quad (1)$$

Consider minimization of the above function by the Newton-Rhapon method. Let the values of the parameters at the k th iteration be denoted by $\beta^{(k)}$, $\mathbf{x}^{(k)}$. Doing a second-order Taylor expansion of NLL about the $\beta^{(k)}$, $\mathbf{x}^{(k)}$, we get

$$NLL(\beta, \mathbf{x}) \approx \sum_{i,j} \left(\frac{(2S_{ij} - 1)\beta^{(k)T}_i x_j^{(k)}}{(2P_{ij}^{(k)} - 1)T_{ij}} - \beta_i^T x_j \right)^2 \frac{(2P_{ij}^{(k)} - 1)T_{ij}}{\beta^{(k)T}_i x_j^{(k)}} \quad (2)$$

where $P_{ij}^{(k)}$ gives the success probability based on parameter values at the k th iteration:

$$P_{ij}^{(k)} = \text{logit}(\beta^{(k)T}_i x_j^{(k)}). \quad (3)$$

The above expression for the Taylor expansion is not an obvious one. Deriving this expression involves considerable algebraic manipulation. For details, see Agresti (2002).

The important implication to take away from the Taylor expansion expression is that because the parameters in the $(k + 1)$ th iteration are chosen to minimize that expression, the estimates of β , \mathbf{x} on the $(k + 1)$ th iteration are obtained by minimizing the weighted Frobenius norm with respect to the matrix of elements $\frac{(2S_{ij}-1)\beta^{(k)T}_i x_j^{(k)}}{(2P_{ij}^{(k)}-1)T_{ij}}$ with the weights being given by $\frac{(2P_{ij}^{(k)}-1)T_{ij}}{\beta^{(k)T}_i x_j^{(k)}}$.

As shown in Srebro & Jaakkola (2003), the minimization of weighted the Frobenius norm can be accomplished by a series of SVD factorizations. (For details, see their paper.) Applying these SVD factorizations therefore produces estimates of $\beta^{(k+1)}$ and $\mathbf{x}^{(k+1)}$ exactly in the way we got

estimates of β and \mathbf{x} in the first stage of the two-stage procedure of the previous sub-section. We can repeat this procedure to produce the $(k + 2)$ th iteration's estimates. Starting from any random initial values and repeating this procedure till convergence, yields the maximum likelihood estimates of β, \mathbf{x} .

4 Some Convergence Issues

Local Maxima: It turns out that, in the prelude procedure, the joint likelihood of the α, β and x parameters is not globally concave. Therefore, there is the potential problem of converging to local maxima. However, in all of our simulation examples, we have yet to see this problem actually occurring. We speculate that this is because the eigenvalue-eigenvector calculation used is a relaxation procedure rather than a gradient-based nonlinear optimizer. On the other hand, when we tried a gradient-based procedure for the mode-finding, that procedure did encounter local maxima, though only in a very few instances. Furthermore, the gradient based procedure took about 4100 seconds to execute. Given that the gradient-based procedure takes much longer and increases the possibility of local maxima, we do not suggest it even though this approach is simpler to program if a non-linear optimization code library is already available.

Data Sparsity and the Prelude Procedure: If data on a user u decrease, this increases the misclassification-rate and entropy of the posterior distribution of Z_u^{user} and this can increase the across-iterations variance of the mode identified by the mode-finding procedure. This could in turn (a) increase the number of iterations for convergence, or (b) increase the number of iterations required to *detect* convergence even if the convergence itself proceeds at the same pace. To get some sense of the effect of sparsity, we halved the amount of data for each individual from our empirical exercise. Rather to our surprise, we found little effect on convergence speed. Two possible explanations: (1) As discussed in Kanaya and Han (1995), misclassification rate decreases exponentially with sample size and not at the Cramer-Rao-type inverse-square-root rate like most other error measures. Therefore, the posterior distribution Z_u^{user} converges quickly and even halving the data may not have much of an effect unless the data were low to begin with. (2) The mode is a function of the across-user distribution of the $\{Z_u^{\text{user}}\}$. This across-user distribution could be very stable with enough users even if each Z_u^{user} individually is not. We expect, however, that if the number of users and the number of observations are small, then data sparsity would have a sizable negative impact on the number of iterations required for the prelude procedure.

References cited in this Appendix

Agresti, A. (2002), *Categorical Data Analysis, 2nd Edition*, Wiley.

Chen, Ming-Hui, Qi-Man Shao and Joseph Ibrahim (2000), *Monte Carlo Methods in Bayesian Computation*, Springer.

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995) “Adaptive rejection Metropolis sampling,” *Applied Statistics*, vol. 44, 455–472.

Gilks, W. R., Neal, R. M., Best, N. G., Tan, K. K. C. (1997), “Corrigendum: Adaptive Rejection Metropolis Sampling,” *Applied Statistics*, vol. 46, 541–542.

Gilks, W. R., Wild, P. (1992), ‘Adaptive rejection sampling for Gibbs sampling’, *Applied Statistics*, vol. 41, 337–348.

Kanaya, F., Han, T. S. (1995), “The Asymptotics of Posterior Entropy and Error Probability for Bayesian Estimation,” *IEEE Transactions on Information Theory*, vol. 41, no. 6, pages 1988–1993.

McCulloch, R., Rossi, P. (1994) “An exact likelihood analysis of the multinomial probit model,” *Journal of Econometrics* 64, 207–240.

Rossi, P., Allenby, G., McCulloch, R. (2006), *Bayesian Statistics and Marketing*, Wiley.

Srebro, N., Jaakkola, T. (2003), “Weighted Low-Rank Approximations”, working paper, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.