Bayes methods for categorical data

April 25, 2017
Motivation for joint probability models

- Increasing interest in high-dimensional data in broad applications
- Focus may be on prediction, variable selection, inference on dependence, etc
- Most literature focuses on $y_i = (y_{i1}, \ldots, y_{ip})^T \in \mathbb{R}^p$
- Today’s focus: general class of flexible joint probability models for high-dimensional categorical data
Motivation for joint probability models

- Flexible joint probability model for $y_i$ can be used directly to predict a subset of the elements of $y_i$ given the other values
- Univariate & multivariate classification problems dealt with automatically
- Accommodates higher order interactions automatically without explicitly parameterizing these interactions
- Joint modeling of responses & predictors makes it easy to handle missing data
- Adapted easily for joint nonparametric modeling for general data types (functions, images, text, etc) by using the model for latent class indices
Motivating application

- Modeling dependence of nucleotides within the p53 transcription factor binding motif.
- p53 tumor-suppressor = short DNA sequence, regulates the expression of genes involved in variety of cellular functions.
- A, C, G, T nucleotides at 20 positions for 574 sequences (Wei et al. 2006).
- Flexibly characterize the dependence structure and test for positional dependencies.
- Models of nucleotide sequences useful for finding gene regulatory regions & for other uses
Recap: Modeling multivariate ordinal data

Suppose we have $y_i \in \{1, \ldots, C\}$, with the ordering in the levels important.

For example, $y_i$ may measure severity of response, with $y_i = 1$ mild, $y_i = 2$ moderate, $y_i = 3$ severe.

Likelihood of data is multinomial:

$$\prod_{i=1}^{n} \prod_{j=1}^{C} \pi_{ij}^{I(y_{ij} = j)}$$

where $\pi_{ij} = Pr(y_i = j \mid x_i)$—how to model??
A typical approach is to let

\[ Pr(y_i \leq j \mid x_i) = F(\alpha_j - x_i^t \beta), \]

where \( F(\cdot) \) is a cdf.

Here, \(-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_{C-1} < \alpha_C = \infty\) characterize the baseline distribution of the categorical response.

For example, if we choose \( F(z) = \Phi(z) \), then we obtain a generalized probit model.

If we choose \( F(z) = 1/\{1 + \exp(-z)\} \), then we obtain a generalized logit model.

These models represent direct extensions of probit and logistic regression models for binary response data.
Recap: Modeling multivariate nominal data

- $y_i = (y_{i1}, \ldots, y_{ip})^T$, with $y_{ij} \in \{1, \ldots, d_j\}$.
- Generalized latent trait models (GTLM) accommodate different data types (continuous, count, binary, ordinal).
- Define glm for each outcome with shared normal latent traits in these models (Sammel et al., 1997; Moustaki & Knott, 2000; Dunson, 2000, 2003).
- Motivated by the nucleotide application, Barash et al. (2003) used Bayes networks (BN) to explore models with varying degrees of complexity.
- Even with very efficient model search algorithms, only feasible to visit a tiny subset of the model space for moderate $p$.
- Difficult to define an appropriate penalty for model complexity, overfitting tends to occur in practical examples.
Link each $y_{ij}$ to an underlying continuous variable $z_{ij}$, with $y_{ij}$ assumed to arise via thresholding $z_{ij}$.

When $y_{ij} \in \{0, 1\}$, a MVN on $z_i = (z_{i1}, \ldots, z_{ip})^T$ induces the widely used multivariate probit model (Ashford and Sowden, 1970; Chib and Greenberg, 1998).

Can accommodate nominal data with $d_j > 2$ by introducing a vector of variables $z_{ij} = (z_{ij1}, \ldots, z_{ijd_j})^T$ underlying $y_{ij}$ with $y_{ij} = l$ if $z_{ijl} = \max_z z_{ij}$: multivariate multinomial probit model.

Model $z_i$ as $\sum_{j=1}^p d_j$ dimensional Gaussian with covariance matrix $\Sigma$. 
Recap: Multivariate probit models

- A Gaussian latent variable needed for each level of the response.
- The relationship between the dependence in the latent variables and dependence in the observed categorical variables is complex and difficult to interpret.
- Need to constrain at least \( p \) diagonal elements of \( \Sigma \) for identifiability.
- Complicates sampling from the full conditional posterior of \( \Sigma \).
- Zhang et al. (2006, 2008) used parameter-expanded MH for posterior computation in multivariate multinomial probit models.
When $y_i \in \mathbb{R}^p$, factor models useful for dimension reduction (West 03; Carvalho et al. 08; Bhattacharya & Dunson 10) explain dependence among high dimensional observations through $k << p$ underlying factors.

The Gaussian linear factor model is most commonly used,

$$y_i = \mu + \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Sigma), \quad i = 1, \ldots, n,$$

$\Lambda$ is a $p \times k$ factor loadings matrix, $\eta_i \sim N_k(0, I_k)$ are latent factors. Marginally, $y_i \sim N_p(0, \Omega)$ with $\Omega = \Lambda \Lambda^T + \Sigma$.

Easily adapted to accommodate binary & ordered categorical $y'_{ij}$s through use of underlying variables
Motivation

- Aim to explain dependence among the high-dimensional nominal variables in terms of relatively few latent factors.
- Similar to Gaussian factor models, but factors on simplex more natural here.
- Joint distribution of $y_i$ induced by our model corresponds to a PARAFAC decomposition (De Lathauwer et al., 2000) of probability tensors.
- Related to mixed membership models, such as latent Dirichlet allocation (Blei et al. 2003) for topic modeling, also Pritchard et al. (2000, 2003).
Focus on $p = 2$, so that data for subject $i$ consist of a pair of categorical variables, $x_i = (x_{i1}, x_{i2})'$.

Results in a $d_1 \times d_2$ contingency table with cell one can let $(c_1, c_2)$ containing the count $\sum_{i=1}^{n} 1(x_{i1} = c_1, x_{i2} = c_2)$, for $c_1 = 1, \ldots, d_1$ and $c_2 = 1, \ldots, d_2$.

Our focus is on parsimonious modeling of the cell probabilities, $\pi = \{\pi_{c_1c_2}\}$, with $\pi_{c_1c_2} = Pr(x_{i1} = c_1, x_{i2} = c_2)$.

Reduce $d_1d_2 - 1$ free parameters.

Let $\psi^{(1)}, \psi^{(2)} \in S_{d_1-1} \times S_{d_2-1}$.

One simple way is to have $Pr(x_{i1} = c_1) = \psi_{c_1}^{(1)}$ and $Pr(x_{i2} = c_2) = \psi_{c_2}^{(2)}$ with $x_{i1}$ and $x_{i2}$ independent.

In this case, we obtain $\pi_{c_1c_2} = \psi_{c_1}^{(1)} \psi_{c_2}^{(2)}$.

Highly parsimonious $d_1 + d_2 - 2$ free parameters.
Overly restrictive

Latent structure analysis (Lazarsfeld and Henry 1968; Goodman 1974)

Relies on the finite mixture specification

\[ Pr(x_i^1 = c_1, x_i^2 = c_2) = \pi_{c_1c_2} = \sum_{h=1}^{k} \nu_h \psi_{hc_1}^{(1)} \psi_{hc_2}^{(2)} \]

where \( \nu = (\nu_1, \ldots, \nu_k)' \) is a vector of mixture probabilities,

\( z_i \in \{1, \ldots, k\} \) denotes a latent class index,

\( Pr(x_i^1 = c_1 | z_i = h) = \psi_{hc_1}^{(1)} \) is the probability of \( x_i^1 = c_1 \) in class \( h \),

\( Pr(x_i^2 = c_2 | z_i = h) = \psi_{hc_1}^{(2)} \) is the probability of \( x_i^2 = c_2 \) in class \( h \)

\( x_i^1 \) and \( x_i^2 \) are conditionally independent given \( z_i \).
Let $\Pi_{d_1...d_p}$ = set of probability tensors, with $\pi \in \Pi_{d_1...d_p} \rightarrow$

$$\pi = \{\pi_{c_1...c_p} \geq 0, c_j = 1, \ldots, d_j, j = 1, \ldots, p : \sum_{c_1=1}^{d_1} \ldots \sum_{c_p=1}^{d_p} \pi_{c_1...c_p} = 1\}$$

A decomposed tensor (Kolda, 2001) $D = u^{(1)} \otimes u^{(2)} \ldots \otimes u^{(p)}$, or elementwise, $D_{c_1...c_p} = u^{(1)}_{c_1} \ u^{(2)}_{c_2} \ldots \ u^{(p)}_{c_p}$.

PARAFAC rank (Harshman, 1970) – minimal $r$ such that $D$ is a sum of $r$ decomposed tensors.
Dunson & Xing (2009) decompose probability tensor $\pi$ as

$$\pi_{c_1 \ldots c_p} = \sum_{h=1}^{k} \nu_h \psi^{(1)}_{hc_1} \cdots \psi^{(p)}_{hc_p}$$  \hspace{1cm} (1)

where $\nu_h = \text{pr}(z_i = h)$, and $\psi^{(j)}_h \in S_{d_j-1}$.

(1) is a form of non-negative PARAFAC decomposition.
Although any multivariate categorical data distribution can be expressed as above for a sufficiently large $k$, a number of practical issues arise in the implementation.

Firstly, it is not straightforward to obtain a well-justified approach for estimation of $k$.

Because the data are often very sparse with most of the cells in the $d_1 \cdots d_p$ contingency table being empty, a unique maximum likelihood estimate of the parameters often does not exist even when a modest $k$ is chosen.

Such problems may lead one to choose a very small $k$, which may be insufficient.

Follow a Bayesian nonparametric approach.
We propose to induce a prior, \( \pi \sim P \) through the following specification

\[
\pi = \sum_{h=1}^{\infty} \nu_h \Psi_h, \quad \Psi_h = \psi_h^{(1)} \otimes \ldots \otimes \psi_h^{(p)}
\]

\( \psi_h^{(j)} \sim P_{0j} \), independently for \( j = 1, \ldots, p; h = 1, \ldots, \infty \)

\( \nu \sim Q \).

\( P_{0j} \) is a probability measure on \( S_{d_j - 1} \).

\( Q \) is a probability measure on the countably infinite probability simplex, \( S_\infty \).
Choice of prior for $P_{0j}$ and $Q$

- $P_{0j}$ may correspond to a Dirichlet measure with

$$\psi_h^{(j)} \sim \text{Diri}(a_{j1}, \ldots, a_{jc_j})$$

- $Q$ corresponds to a Dirichlet process $\sum_h \pi_h \delta_h$ where

$$\pi_h = V_h \prod_{l<h} (1 - V_l)$$

with $V_h \sim \text{beta}(1, \alpha)$ independently for $h = 1, \ldots, \infty$ where $\alpha > 0$ is a precision parameter characterizing $Q$. 
Interest to test for independence of the elements of $x_i = (x_{i1}, \ldots, x_{ip})'$.

In the motif application, considerable debate on the appropriateness of the independence assumption.

Under our proposed formulation, the null hypothesis of independence is nested within a nonparametric alternative that accommodates a sequence of models of increasing complexity including the saturated model.

In particular, the independence model corresponds to $H_0 : \nu_1 = 1$.

As noted in Berger and Sellke (1987), interval null hypotheses are often preferred to point null hypotheses.
Motivated by this reasoning and by computational considerations, we focus instead on the interval null

\[ H_0 : \nu_* > 1 - \epsilon, \quad \nu_* = \max\{\nu_h, h = 1, \ldots, k_*\} \]

Fix \( \epsilon > 0 \) (usually 0.05)
Measures of association for nominal data

- Infer dependence structure from pairwise dependencies between $y_{ij}$ and $y_{ij'}$ for $j \neq j' \in \{1, \ldots, p\}$
- Pairwise Cramer’s V association matrix $\rho = (\rho_{jj'})$

\[
\rho_{jj'}^2 = \frac{1}{\min\{d_j, d_{j'}\} - 1} \sum_{c_j=1}^{d_j} \sum_{c_{j'}=1}^{d_{j'}} \frac{(\pi_{c_j c_{j'}} - \bar{\psi}_{c_j} \bar{\psi}_{c_{j'}})^2}{\bar{\psi}_{c_j} \bar{\psi}_{c_{j'}}}
\]

with $\bar{\psi}_l^{(j)} = \sum_{h=1}^{k^*} \nu_h \psi_{hl}^{(j)}$.

- $\rho_{jj'}$ ranges from 0 to 1, with $\rho_{jj'} \approx 0$ when $x_{ij}$ and $x_{ij'}$ are independent.
Posterior distribution of $\rho_{jj'}$ for all $(j, j')$ pairs based on the output of the Gibbs sampler.

Construct recommend reporting a $p \times p$ association matrix, with the elements corresponding to posterior means for each $\rho_{jj'}$.

In addition, we can calculate posterior probabilities and Bayes factors for local null hypotheses, $H_{1,jj'} : \rho_{jj'} > \epsilon$ from the Gibbs sampler output.
Simulation studies

- Simulated data consisted of A, C, G, T nucleotides \( (d_j = d = 4) \) at \( p = 20 \) positions for \( n = 100 \) sequences.
- 2 settings: generate the nucleotides (1) independently, and (2) assuming dependence in locations 2, 4, 12, and 14.
Figure 1. Histograms of estimated posterior probabilities of $H_1$ in each of the 100 simulations under (a) case 1 (no positional dependence—$H_0$ is true) and (b) case 2 (positional dependence—$H_1$ is true).
Simulation studies

Figure 2. Results of simulation case 2—percentages of simulations for which (a) $\Pr(H_{1,j'| X}) > 0.95$, and (b) the Xie and Geng (2008) method estimated an association between positions $j, j'$. The true model has dependence in positions 2, 4, 12, and 14.