Statistica Sinica **7**(1997), 527-544

GENERALIZED ACCUMULATED PREDICTION ERROR AND MODEL SELECTION FOR CATEGORICAL PANEL DATA

Ping Zhang

University of Pennsylvania

Abstract: A unique feature of panel data is that temporal and cross-sectional variations are often confounded with one another. Numerous models have been proposed in the literature to describe different aspects of these variations. This article attempts to integrate model selection into categorical panel data analysis. To this end, conventional methods are often inappropriate because most of them are designed to compare submodels that belong to the same parametric class. We introduce the generalized accumulated prediction error (GAPE) for panel data and propose to use it as a model selection criterion. Theoretical properties of GAPE will be discussed. The results are applied to a set of scanner data drawn from marketing research.

Key words and phrases: AIC, BIC, consumer behavior, heterogeneity, marketing research, multivariate logit model, scanner data, Simpson's Paradox.

1. Introduction

Panel data are longitudinal records taken from a group of individuals selected randomly from a population. In economics and many other social sciences, researchers use longitudinal survey data to monitor and to predict aggregate social trends (Hsiao (1986)). In medical research, cohorts of subjects are usually followed for a long period of time in order to determine or to establish evidence for causal factors (Diggle, Liang and Zeger (1994)). In marketing research, the use of consumer panel and the collection of household purchasing records has been a common practice for decades (Guadagni and Little (1983)). In general, panel data contain substantially more information than traditional cross-sectional survey data or aggregate time series data. By looking at individual traces, researchers can study patterns of temporal change in much greater detail. That aggregate and individual data do not necessarily tell the same story is well documented in the literature. Table 1 shows an example, known as Simpson's Paradox, in which the observed trend at the aggregate level contradicts trends observed at a disaggregate level (Cohen (1986), Samuels (1993), Benjamini and Krieger (1992)). See Haccou and Meelis (1994) and Massy, Montgomery and Morrison (1970) for more applications.

Table 1. An example of Simpson's Paradox

	Period I			Period II		
	N	n	n/N	N	n	n/N
Group I	40	10	10/40	90	25	25/90
Group II	60	20	20/60	10	4	4/10
Combined	100	30	30/100	100	29	29/100

From time period I to time period II, the market share of brand A increases within each of the two consumer groups. When the two groups are combined, the aggregate market share of brand A decreases (N=number of people buying; n =number of people buying brand A).

Technically, we can write panel data as $(\mathbf{y}_{it}, \mathbf{x}_i, \mathbf{z}_t)$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where i labels individuals and t labels time. Here $\mathbf{y}_{it}$ is a response variable of main interest; $\mathbf{x}_i$ is a vector of covariates that vary with i (e.g., demographic attributes that do not change over time); and $\mathbf{z}_t$ is a vector of covariates that depend only on t (e.g., environmental variables that have impact on all individuals). Occasionally, we also observe covariates that depend on both i and t. The basic goal of panel data modeling is to describe and, hopefully, to understand the cross-sectional and cross-time variation of the response variable $\mathbf{y}_{it}$. This can be accomplished, for instance, by building regression models that relate $\mathbf{y}_{it}$ to the observed covariates $\mathbf{x}_i$ and $\mathbf{z}_t$. Unfortunately, for panel data, the two sources of variation are often confounded with one another. This creates interpretation problems since models that seem entirely different can fit the data equally well.

To illustrate, suppose, in the absence of any covariate, that $\mathbf{y}_{it}, t = 1, \ldots, T$, are i.i.d. random variables following the normal distribution $N(\theta_i, \sigma^2)$. In other words, we assume that the observations from a given individual are not correlated over time. The population heterogeneity is reflected through variations in θ_i. In the spirit of the empirical Bayes method, let us assume that $\theta_i, i = 1, \ldots, N$, are i.i.d. random variables following a normal distribution $N(\mu, \tau^2)$. It is then straightforward to verify that the unconditional joint distribution of $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ is multivariate normal $N(\mu J, \sigma^2 I + \tau^2 JJ')$, where I is the identity matrix and $J = (1, \ldots, 1)'$ is a T-dimensional column vector. Thus unconditionally, the $\mathbf{y}_{it}$'s are correlated over time. What appears to be a temporal correlation at the aggregate level is actually the side effect of population heterogeneity. Likewise for categorical panel data, the order of dependence in a Markov chain would appear to be higher when cross-sectional heterogeneity is present (Massy, Montgomery and Morrison (1970), Haccou and Meelis (1994)).

Ideally, a model for panel data should address both temporal and cross-sectional variations. However, a survey of the current literature quickly reveals that the modeling of cross-sectional variation receives much more attention than the modeling of temporal variation. Moreover, models that deal with non-observable heterogeneity (e.g., random-effects and state-space models) dominate the field. Most of these latent variable models ignore temporal correlation by assuming that variables observed at different points of time are independent of each other. Technical convenience is clearly not the main reason for such seemingly unrealistic assumptions. In practice, cross-sectional variation is often a more dominant data feature than temporal variation, especially when the panel record is "short". Even for "long" panel records, it is not clear whether one can improve a model in any way simply by relaxing assumptions of independence (e.g., mixtures of independence models). The confounding problem described earlier plays an important role in this dilemma. In other words, it is intrinsically impossible to separate temporal and cross-sectional variations by means of modeling techniques.

Of considerable practical importance therefore is the problem of model selection, the goal being to assess the merits of different models objectively. What sets our problem apart from conventional model selection is that models of different types are often involved in panel data analysis. Usually, the number of parameters ceases to be an effective measure of model complexity, rendering standard model selection methods such as AIC (Akaike (1973)) and BIC (Schwarz (1978)) completely inappropriate. In this paper, we shall adopt a predictive approach, i.e., select the model that minimizes the prediction error. To be specific, suppose that $\mathbf{g}_t = \mathbf{g}(\mathbf{y}_{1t}, \ldots, \mathbf{y}_{Nt})$ is a cross-sectional statistic that we wish to monitor. The performance of a model can be measured by its ability to predict future values of $\mathbf{g}_t$. We define the generalized accumulated prediction error (GAPE) as

$$\text{GAPE} = \sum_{t=t_0}^{T} D(\mathbf{g}_t, \hat{\mathbf{g}}_t), \tag{1}$$

where $\hat{\mathbf{g}}_t$ is a model based predictor of $\mathbf{g}_t$, $D(\cdot,\cdot)$ is a distance function, and t_0 is an initial value usually chosen to be large enough so that $\hat{\mathbf{g}}_{t_0}$ is reasonably accurate. Suppose that a set of models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ is under consideration. The models could assume similar mathematical forms or they could differ in arbitrary ways. Model selection here amounts to choosing the model $\mathcal{M}_k$ that yields the minimum GAPE. An important special case of (1) is when the target of prediction equals to the vector of individual data points, i.e., $\mathbf{g}_t = (\mathbf{y}_{1t}, \ldots, \mathbf{y}_{Nt})$. The corresponding GAPE is reduced to the ordinary accumulated prediction error (APE) criterion, originally proposed by Rissanen (1986a, 1986b, 1987). A more thorough treatment of APE can be found in Wei (1992). See also Dawid (1984). By and large, the asymptotic behavior of the ordinary APE has been shown to be very similar to that of BIC, which penalizes model complexity by a factor that is proportional to the log of sample size (Schwarz (1978)).

The existing literature on panel data analysis focuses largely on continuous data and linear models (Hsiao (1986)). The current paper treats the case when the observed response $\mathbf{y}_{it}$ is a categorical variable. In Section 2, we discuss issues that arise when modeling such data. In particular, we introduce a special case of the GAPE criterion, referred to as GAPE_{KL}, for categorical panel data. In Section 3, we derive some asymptotic properties for the newly defined GAPE_{KL}. It is well known, for linear models, that the ordinary APE is asymptotically equivalent to BIC. By contrast, we show in Section 4 that GAPE_{KL} behaves rather differently. In addition to penalizing model complexity, our GAPE_{KL} criterion also represents a trade-off between different levels of heterogeneity. In Section 5, we apply the GAPE_{KL} criterion to an actual set of scanner panel data drawn from marketing research and show that modeling heterogeneity yield more accurate prediction than modeling temporal dependency. Some concluding remarks are given in the final section.

2. Categorical Panel Data

This paper is concerned with categorical panel data of the following form:

$$\mathbf{y}_{it} = \{y_{it}(0), y_{it}(1), \ldots, y_{it}(K)\},\ i = 1, \ldots, N,\ t = 1, \ldots, T,$$

where $y_{it}(k)$ are 0-1 variables satisfying $\sum_k y_{it}(k) = 1$. Thus at any given point of time, a panelist can choose, either voluntarily or passively, among K possible outcomes. The panelist could also decline all the available choices. When this happens, we observe $y_{it}(0) = 1$. For fixed t and i, we have a categorical variable with $K + 1$ categories. For given t, it is natural to assume that observations obtained from different individuals are independent. When i is fixed, $\mathbf{y}_{it}, t = 1, \ldots, T$, can be viewed as a categorical time series. Figure 1 shows some examples of such time series records in the marketing context. For later expositions, we need the following

Assumption I. The response variables $\mathbf{y}_{it}, i = 1, \ldots, N; t = 1, \ldots, T$, are independent across both i and t.

Technically, we could treat the event $\{y_{it}(0) = 1\}$ as one of the available choices and model the $K+1$ probabilities $\Pr\{y_{it}(k) = 0\}, k = 0, 1, \ldots, K$, simultaneously. In practice, depending on the type of data available, it is often more realistic to separate the event $\{y_{it}(0) = 1\}$ from the rest of the choices. In the context of marketing research, for example, one might want to study the impact of advertising on consumer behavior (e.g., brand choice). This would be impossible if none of the panelists made any purchases. Thus models for categorical panel data often consist of two parts. First, at any given time, one needs to separate people who are active in making choices and people who are idle. In other words, we first need to model the probability

$$\pi_{it}(\alpha) = \Pr\{y_{it}(0) = 1\}.$$

Then, choice decisions can take place only if the person is active. A choice model is therefore a model for the conditional probability

$$p_{it}(k,\theta) = \Pr\{y_{it}(k) = 1 | y_{it}(0) = 0\},\ k = 1,\ldots,K.$$

Most measured covariates have impact only on $p_{it}(k,\theta)$. One has little information as far as the modeling of $\pi_{it}(\alpha)$ is concerned. For binary data, we have $K = 1$ and there is no need to address the conditional event. In choice models, however, the conditional probability is the focus.

Throughout this paper, to emphasize to purpose of model selection, we shall not specify the mathematical forms of $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$, with the understanding that different specifications would lead to different models. For example, logistic regression models can be used to link both $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$ to covariates $\mathbf{x}_i$ and $\mathbf{z}_t$. Notice, however, that our description excludes random effects models. When Assumption I holds, models $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$ give an adequate account of the data. Otherwise, we would need to model the evolution of conditional probabilities over time (e.g., Markov chains). See Sections 1 and 5 for more discussion on the feasibility of Assumption I. Throughout this paper, we use the terms cross-sectional and cross-time heterogeneity to describe deterministic (as opposed to random) variations in $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$.

The main goal of the current article is to explore the properties of GAPE as a model selection criterion. To accomplish this, we need to specify the three components in the definition of (1). First, the performance of GAPE will obviously depend on the prediction target $\mathbf{g}_t$. Next, we need to construct a predictor $\hat{\mathbf{g}}_t$ based on models $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$. Different constructions will lead to different criteria. Finally, one should also be specific about the choice of the distance function $D(\cdot,\cdot)$. We illustrate the three steps as follows:

The Prediction Target. For categorical data, statistical quantities of practical interests are various sample proportions. In our case, since the goal is to model choice behavior, a natural target of prediction is the vector of conditional frequencies

$$s_t(k) = \left\{\sum_{i=1}^{N} y_{it}(k)\right\} \Bigg/ \left\{N - \sum_{i=1}^{N} y_{it}(0)\right\},\ k = 1,\ldots,K. \tag{2}$$

In the context of marketing research, $s_t(k)$ represents the market share of brand k. Throughout this paper, we shall assess the merits of models $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$

according to their ability to predict the market share vector $\mathbf{s}_t=\{s_t(1),\ldots,s_K(t)\}$. In other words, we consider only the special case when $\mathbf{g}_t = \mathbf{s}_t$ in (1).

Construction of Predictor. We shall derive an intuitive predictor of $\mathbf{s}_t$ by means of asymptotic approximation. As $N \to \infty$, it is easy to verify that

$$N^{-1}\sum_{i=1}^{N} y_{it}(0) = N^{-1}\sum_{i=1}^{N} \pi_{it}(\alpha) + o_p(1).$$

Likewise, we have

$$N^{-1}\sum_{i=1}^{N} y_{it}(k) = N^{-1}\sum_{i=1}^{N}\{1-\pi_{it}(\alpha)\}p_{it}(k,\theta) + o_p(1),\ \ k=1,\ldots,K.$$

Hence an asymptotic approximation of $s_t(k)$ is

$$\tilde{p}_t(k,\alpha,\theta) = \sum_{i=1}^{N} r_{Ni}(\alpha)p_{it}(k,\theta), \tag{3}$$

where $r_{Ni}(\alpha) = \{1-\pi_{it}(\alpha)\}/\sum_{i=1}^{N}\{1-\pi_{it}(\alpha)\}$. In other words, $s_t(k) = \tilde{p}_t(k,\alpha,\theta)$ $+o_p(1)$. It is interesting to point out that the average of conditional probabilities, i.e., $N^{-1}\sum_{i=1}^{N} p_{it}(k,\theta)$, is not the correct approximation of $s_t(k)$, due to possible cross-sectional heterogeneity in $\pi_{it}(\alpha)$. We propose to predict $s_t(k)$ by

$$\hat{s}_t(k) = \tilde{p}_t(k,\hat{\alpha}_{t-1},\hat{\beta}_{t-1}) = \sum_{i=1}^{N} r_{Ni}(\hat{\alpha}_{t-1})p_{it}(k,\hat{\theta}_{t-1}), \tag{4}$$

where $\hat{\alpha}_{t-1}$ and $\hat{\theta}_{t-1}$ are parameter estimates (e.g., mle) based on data observed up to time $t-1$. Thus in equation (1), we have $\hat{\mathbf{g}}_t = \hat{\mathbf{s}}_t = \{\hat{s}_t(1),\ldots,\hat{s}_t(K)\}$.

The Distance Function. Since both $\mathbf{s}_t$ and $\hat{\mathbf{s}}_t$ are probability vectors, i.e., vectors with non-negative components that sum up to one, we should choose a metric that measures the distance between density functions. Throughout this paper, for technical convenience, we use the Kullback-Leibler distance

$$D_{KL}(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{K} p(k)\log\{\,p(k)/q(k)\},$$

where $\mathbf{p}$ and $\mathbf{q}$ are two arbitrary probability vectors with components $p(k)$ and $q(k)$ respectively. Other common choices include the chi-square distance

$$D_2(\mathbf{p},\mathbf{q}) = 2^{-1}\sum_{k=1}^{K}\{\,p(k)-q(k)\}^2/p(k)$$

and the Hellinger distance

$$D_H(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{K} \left\{ \sqrt{p(k)} - \sqrt{q(k)} \right\}^2 .$$

To summarize, the problem that we shall treat in this paper is that of predicting $\mathbf{s}_t$ using $\hat{\mathbf{s}}_t$ under the Kullback-Leibler distance. The corresponding GAPE criterion shall be denoted as

$$\mathrm{GAPE}_{KL} = \sum_{t=t_0}^{T} D_{KL}(\mathbf{s}_t, \hat{\mathbf{s}}_t), \tag{5}$$

whose asymptotic properties will be discussed in the next section.

3. Asymptotic Properties of GAPE$_{KL}$

There are two basic concerns in a typical model selection problem: goodness-of-fit and model complexity. One wants to fit the data well with models that are as simple as possible. The goal of model selection is to find a compromise between the two contrasting goals. It is therefore of crucial importance to understand the trade-off mechanism of a proposed model selection criterion. Currently known results on the ordinary APE all indicate that it behaves similarly to the well known BIC criterion. Why should we expect anything different from GAPE$_{KL}$? First, the short answer is that the definitions of APE and GAPE$_{KL}$ are different. More specifically, as we discussed in the previous section, properties of $\hat{\mathbf{g}}_t$, hence the properties of GAPE$_{KL}$, depend not only on the dimensions of α and θ, as BIC does, but also on the cross-sectional heterogeneity of $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$, which BIC does not address.

To derive a meaningful decomposition of (5), let us first define

$$\tilde{\mathbf{p}}_t = \{\, \tilde{p}_t(1,\alpha,\theta), \ldots, \tilde{p}_t(K,\alpha,\theta)\},$$

where $\tilde{p}_t(k,\alpha,\theta)$ is as defined in (3). We can decompose GAPE$_{KL}$ as

$$\mathrm{GAPE}_{KL} = \sum_{t=t_0}^{T} D_{KL}(\mathbf{s}_t, \tilde{\mathbf{p}}_t) + \xi_{NT} + \eta_{NT}, \tag{6}$$

where

$$\xi_{NT} = \sum_{t=t_0}^{T} \sum_{k=1}^{K} s_t(k) \log\{\, \tilde{p}_t(k,\alpha,\theta)/\tilde{p}_t(k,\alpha,\hat{\theta}_{t-1})\}$$

and

$$\eta_{NT} = \sum_{t=t_0}^{T} \sum_{k=1}^{K} s_t(k) \log\{\, \tilde{p}_t(k,\alpha,\hat{\theta}_{t-1})/\tilde{p}_t(k,\hat{\alpha}_{t-1},\hat{\theta}_{t-1})\}.$$

The three terms on the right hand side of (6) can be roughly described as follows: If the underlying models are correct, we should have $\mathbf{s}_t = \tilde{\mathbf{p}}_t + o_p(1)$. Hence the first term is a measure of goodness-of-fit. The second term, i.e., ξ_{NT}, is small when $\hat{\theta}_{t-1}$ is close to θ. Hence it is related to the accuracy of $\hat{\theta}_{t-1}$, which in most cases is proportional to the dimension of θ. Likewise, the third term, i.e., η_{NT}, is related to the dimension of α. To further understand the decomposition in (6), let us define matrices

$$\mathbf{V}_T = \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K} \frac{\{1-\pi_{it}(\alpha)\}}{p_{it}(k,\theta)} \left\{\frac{\partial p_{it}(k,\theta)}{\partial\theta}\right\} \left\{\frac{\partial p_{it}(k,\theta)}{\partial\theta}\right\}'$$

and

$$\mathbf{W}_T = \sum_{t=1}^{T}\sum_{i=1}^{N} \frac{1}{\pi_{it}(\alpha)\{1-\pi_{it}(\alpha)\}} \left\{\frac{\partial \pi_{it}(\alpha)}{\partial\alpha}\right\} \left\{\frac{\partial \pi_{it}(\alpha)}{\partial\alpha}\right\}'.$$

Note that $\mathbf{V}_T$ and $\mathbf{W}_T$ are the Fisher information matrices for θ and α respectively.

Define random variables A_{NT} and B_{NT} such that $\xi_{NT} = A_{NT}N^{-1}\log\det(\mathbf{W}_T)$ and $\eta_{NT} = B_{NT}N^{-1}\log\det(\mathbf{V}_T)$. Thus (6) becomes

$$\mathrm{GAPE}_{KL} = \sum_{t=t_0}^{T} D_{KL}(\mathbf{s}_t, \tilde{\mathbf{p}}_t) + A_{NT}N^{-1}\log\det(\mathbf{W}_T) + B_{NT}N^{-1}\log\det(\mathbf{V}_T). \tag{7}$$

The main result of this paper is the following theorem, which shows that A_{NT} and B_{NT} in the above equation are asymptotically non-negative and bounded.

Theorem 1. *Suppose that the maximum likelihood estimates of θ and α are asymptotically consistent and efficient. Furthermore, suppose that there exist positive constants ρ and γ such that $\pi_{it}(\alpha) \le \rho$ and $N^{-1}\sum_{i=1}^{N}\{\tilde{p}_t(k,\alpha,\theta) - p_{it}(k,\theta)\}^2 \le \gamma$. Then as $N \to \infty$ and $T \to \infty$, the random variables A_{NT} and B_{NT} in* (7) *satisfy*

$$0 \le E(A_{NT}) \le 2^{-1}(1-\rho)^{-1}$$

and

$$0 \le E(B_{NT}) \le 8^{-1}(1-\rho)^{-2}\gamma.$$

Proof. See the Appendix.

A number of remarks are in order here. By comparing (6) with (7), it is clear that ξ_{NT} and η_{NT} serve as penalties for the complexity of models $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$ respectively. The decomposition in (7) is quite similar to standard results for linear models (Wei (1992)). The difference is that the two terms A_{NT}

and B_{NT}, both equal to a constant in ordinary APE, are now dependent on the underlying models. Moreover, what is not shown in Theorem 1 is the fact that the two penalty terms are related. In fact, we shall argue in the next section that A_{NT} and B_{NT} tend to move in opposite directions. The same conditions that make one term large would make the other term small. Furthermore, we will also argue that this negative correlation is driven by the level of heterogeneity in the underlying models. How large the penalty terms are depend on how heterogeneous the models are. Thus the GAPE$_{KL}$ criterion not only penalizes model complexity, but also determines the level of heterogeneity that we should build into our models.

4. The Difference Between GAPE$_{KL}$ and BIC

4.1. The ambiguity of BIC

The general expression of the BIC criterion is

$$-2\log(\text{maximum likelihood}) + (\text{model dimension}) \times \log(\text{sample size}).$$

Accordingly, the BIC criterion in our case can be written as

$$\text{BIC} = -2\log L(\hat{\theta}_T, \hat{\alpha}_T) + \{\dim(\theta) + \dim(\alpha)\}\log(NT) = \text{BIC}_\pi + \text{BIC}_p, \qquad (8)$$

where

$$\begin{aligned}\text{BIC}_\pi = -2\sum_i\sum_t & [y_{it}(0)\log\{\pi_{it}(\hat{\alpha}_T)\} + \{1-y_{it}(0)\}\log\{1-\pi_{it}(\hat{\alpha}_T)\}] \\ & + \dim(\alpha)\log(NT)\end{aligned}$$

and

$$\text{BIC}_p = -2\sum_i\sum_t\sum_k y_{it}(k)\log\{p_{it}(k,\hat{\theta}_T)\} + \dim(\theta)\log(NT)$$

correspond to models $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$ respectively. Alternatively, one can treat the modeling of the initial decision $y_{it}(0)$ and the modeling of the choice activities $\{y_{it}(1),\ldots,y_{it}(K)\}$ as separate problems. By doing so, one can then choose models $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$ separately. The problem with this latter approach, often used in marketing research, is that the meaning of sample size becomes ambiguous. For example, the BIC criterion for $\pi_{it}(\alpha)$ shall be given by the BIC$_\pi$ component in (8). However, the BIC criterion for $p_{it}(k,\theta)$, when viewed as conditional models, is different from the BIC$_p$ component in (8), the reason being that for any fixed t, the effective sample size in the conditional model $p_{it}(k,\theta)$ is not N, but $N - \sum_i y_{it}(0)$ instead. For example, in the marketing research literature, one often selects brand choice models by minimizing

$$-2\sum_i\sum_t\sum_k y_{it}(k)\log\{p_{it}(k,\hat{\theta}_T)\} + \dim(\theta)\log(N_c),$$

where $N_c = NT - \sum_i \sum_t y_{it}(0)$ is the number of "purchase occasions". The upshot, of course, is that the results of model selection via BIC would be very different depending on whether one treats the modeling of $\pi_{it}(\alpha)$ and $p_{it}(k, \theta)$ as separate problems or as the integral parts of a single modeling problem. This controversy does not arise when using GAPE$_{KL}$ because the construction of the prediction rule, hence the performance of GAPE$_{KL}$, depends on both $\pi_{it}(\alpha)$ and $p_{it}(k, \theta)$. In other words, one is not allowed to treat the two parts as separate problems.

4.2. GAPE$_{KL}$ and heterogeneity

A simple comparison between GAPE$_{KL}$ and BIC can be made by comparing the asymptotic orders of the corresponding goodness-of-fit and penalty terms. To this end, let us assume that $(NT)^{-1}\mathbf{V}_T \to \mathbf{V}$ and $(NT)^{-1}\mathbf{W}_T \to \mathbf{W}$ for some positive definite matrices $\mathbf{V}$ and $\mathbf{W}$. Under these assumptions, we have

$$\log\det(\mathbf{W}_T) \approx \dim(\alpha)\log(NT) + \log\det(\mathbf{W})$$

and

$$\log\det(\mathbf{V}_T) \approx \dim(\beta)\log(NT) + \log\det(\mathbf{V}).$$

Note also that $D_{KL}(\mathbf{s}_t, \tilde{\mathbf{p}}_t) = O(N^{-1})$. Table 2 shows the results of a side-by-side comparison between GAPE$_{KL}$ and BIC.

Table 2. Comparison of asymptotic orders

	Goodness-of-fit	Penalty for $\pi_{it}(\alpha)$	Penalty for $p_{it}(k,\theta)$
GAPE$_{KL}$	$O(TN^{-1})$	$A_{NT}\dim(\alpha)N^{-1}\log(NT)$	$B_{NT}\dim(\beta)N^{-1}\log(NT)$
BIC	$O(TN)$	$\dim(\alpha)\log(NT)$	$\dim(\beta)\log(NT)$

The three terms in GAPE$_{KL}$ are as in (7). The three terms in BIC are as in (8)

A number of interesting observations can be drawn from Table 2. First, the goodness-of-fit terms in GAPE$_{KL}$ and BIC are not directly comparable. One needs to multiply GAPE$_{KL}$ by N^2 in order for it to be comparable with BIC. After the modification, the penalty terms in GAPE$_{KL}$ become $A_{NT}\dim(\alpha)N\log(NT)$ and $B_{NT}\dim(\theta)N\log(NT)$ respectively. Thus, if A_{NT} and B_{NT} were bounded away from zero, GAPE$_{KL}$ will penalize model complexity more heavily than BIC does, especially when N is large. In other words, models selected by GAPE$_{KL}$ will be simpler than models selected by BIC. However, since the zero lower bounds in Theorem 1 are both reachable, there is more to the comparison of GAPE$_{KL}$ and BIC than Table 2 indicates. In fact, the following result holds:

Theorem 2. *Let $\tilde{p}_t(k,\alpha,\theta)$ be the asymptotic approximation of $s_t(k)$ as defined in* (3). *Let A_{NT} and B_{NT} be as in Theorem* 1. *Then* (a) *A sufficient condition for $E(A_{NT}) \to 0$ is $\partial\tilde{p}_t(k,\alpha,\theta)/\partial\theta \to 0$;* (b) *$E(B_{NT})$ reaches its lower bound when either $\pi_{it}(k,\theta)$ or $p_{it}(k,\theta)$ does not depend on i; and* (c) *The upper bound for $E(A_{NT})$ is reached when $p_{it}(k,\theta)$ does not depend on i.*

Proof. See the Appendix.

In addition to compromising between goodness-of-fit and model complexity, as most model selection criteria do, Theorem 2 shows that the GAPE$_{KL}$ criterion also determines the level of heterogeneity allowed in the models $\pi_{it}(\alpha)$ and $p_{it}(k,\theta)$. The ξ_{NT} term in (6) penalizes homogeneous models, whereas the η_{NT} term in (6) penalizes heterogeneous models.

5. Application

A rapidly growing area of empirical research where categorical panel data are frequently encountered is marketing research. The increasing availability of electronic scanner data have enabled researchers to study in greater detail a consumer's reaction to marketing stimulations. Consider the widely used multivariate logit model:

$$p_{it}(k,\theta) = \frac{\exp\{\mathbf{x}'_{it}(k)\beta(k)\}}{\sum_{j=1}^{K}\exp\{\mathbf{x}'_{it}(j)\beta(j)\}}, \tag{9}$$

where $\mathbf{x}_{it}(k)$ is a vector of covariates, $\beta(k)$ is a vector of unknown parameters, and $\theta = \{\beta'(1),\ldots,\beta'(K)\}'$. The implicit assumption in (9) is that all variations can be described through observable variables. The chief attraction of such "fixed effects" models is their straightforward interpretations and easy computational implementation. However, much of the variation, as is often argued, cannot be observed directly. Random effects models are often used in the latter case to handle the problem of hidden heterogeneity. Unfortunately, the literature on random effects models is largely confined to linear regression models due to formidable technical difficulties encountered in general non-linear models. As a compromise, discrete mixture models are often used as simple approximations to general random effects models (Kamakura and Russel (1989), Gupta and Chintagunta (1994)). Efforts have also been made to model or to test the existence of temporal correlations. Using the technique of Gibbs sampling, Allenby and Lenk (1994) incorporate an autocorrelation structure in the random effects logit model. For other related models of serial correlation, see Gilula and Haberman (1994), Diggle, Liang and Zeger (1994), Cosslett and Lee (1985) and Gottschau (1994).

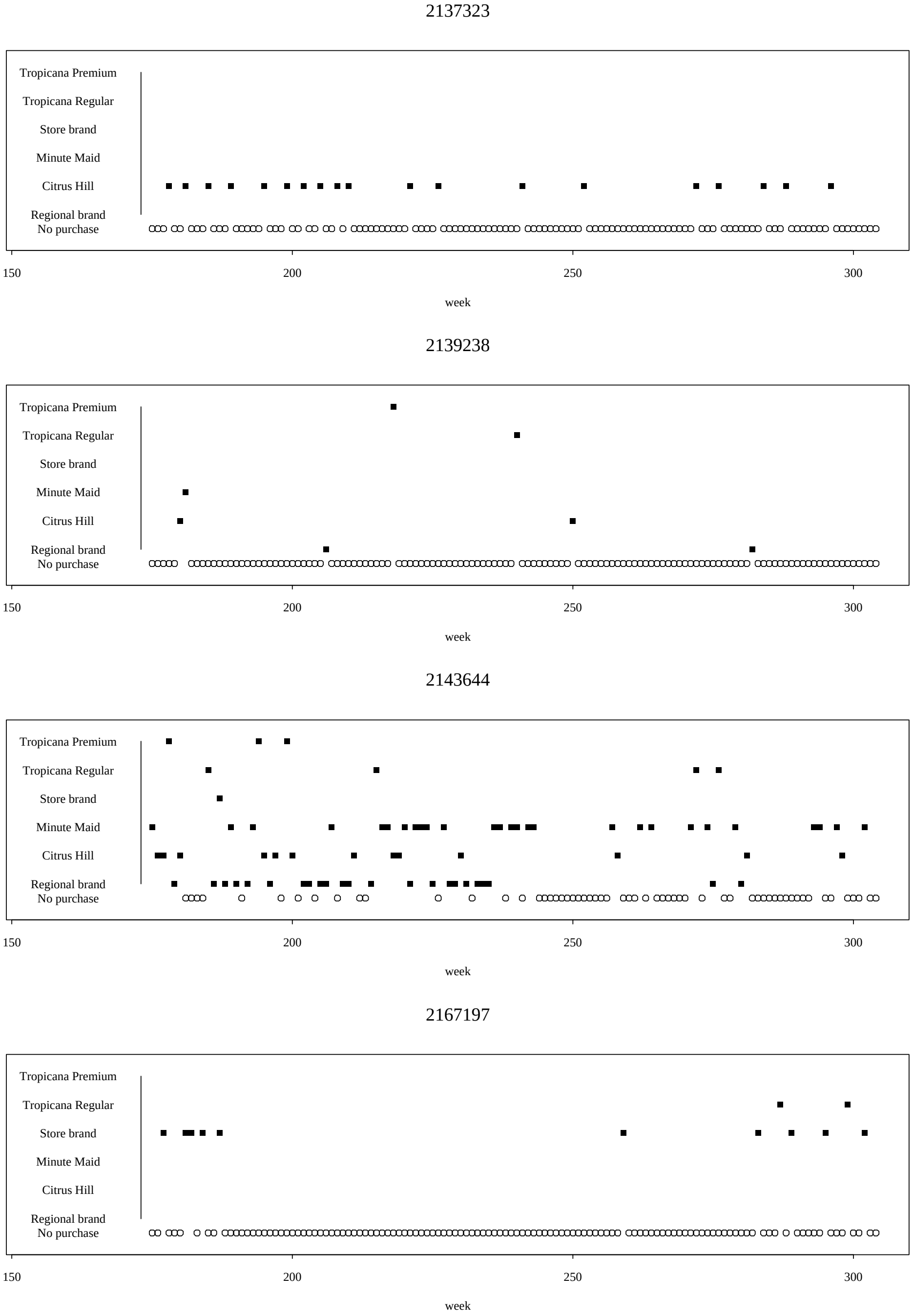

Figure 1. Purchasing records of four households

The solid boxes represent purchase occasions. The empty circles at the bottom correspond to no purchase weeks. Week 175 is the beginning of 1983.

One approach to assess the merits of different models is through model selection. An objective comparison of different models would not only enhance our understanding of the data structure, but also point to the right directions for further research.

The data used in the following application, obtained from Information Resources, Inc., consist of weekly purchasing records for 200 households in Marion, Indiana, from the beginning of 1983 through the middle of 1985 (Fader and Lattin (1993)). Six popular brands of orange juice are included as the product class for this study. Figure 1 depicts the purchasing patterns of some typical panel households. Clearly, there is substantial amount of heterogeneity across consumers. Some make frequent purchases. Others buy occasionally. Some stay loyal to a certain brand while others switch among different brands in an apparently random fashion. In addition to the purchasing records, we also have data on store environment like price and feature advertisement.

Guadagni and Little (1983) introduce the variable

$$\mathrm{LOY}_{it}(k) = \lambda \mathrm{LOY}_{i,t-1}(k) + (1-\lambda) y_{i,t-1}(k),$$

as a measure of consumer i's loyalty towards brand k, where $\lambda \in [0,1]$ is an unknown constant. The so called loyalty model is a logit regression model with $\mathrm{LOY}_{it}(k)$ as one of the covariates. When $\lambda = 1$, $\mathrm{LOY}_{it}(k)$ becomes a measure of heterogeneity since it does not depend on t. When $\lambda = 0$, $\mathrm{LOY}_{it}(k) = y_{i,t-1}(k)$ and the corresponding loyalty model reduces to a simple Markov regression model (Diggle, Liang and Zeger (1994)). By varying the value of λ, the loyalty model can be used to gauge various combinations of temporal and cross-sectional variations. We consider the choice of λ as a model selection problem. Incidentally, both AIC and BIC give the maximum likelihood estimate of λ because the dimension of the parameter space does not depend on λ.

For the data described above, we have $N = 200$, $T = 130$ and $K = 6$. The two environmental variables that we derive from scanner data are (i) $\mathrm{PRICE}_{it}(k)$, which equals the acting price of brand k at time t in the store where consumer i made the purchase; and (ii) $\mathrm{ADVER}_{it}(k)$, which equals 0,1,2 respectively when none, one, or both of feature advertisement and price discount is present for brand k at time t in the store where consumer i made the purchase. We use the 1983 data for initialization ($t_0 = 52$) and calculate the one week ahead prediction error for each of the subsequent weeks until the end of the study period ($T = 130$). Figure 2 depicts the GAPE as a function of λ under different distance functions. Figure 3 shows the maximum log-likelihood as a function of λ. Regardless of the choice of distance functions in (1), the value of λ that minimizes GAPE_{KL} is substantially larger than the maximum likelihood estimate of λ, indicating that

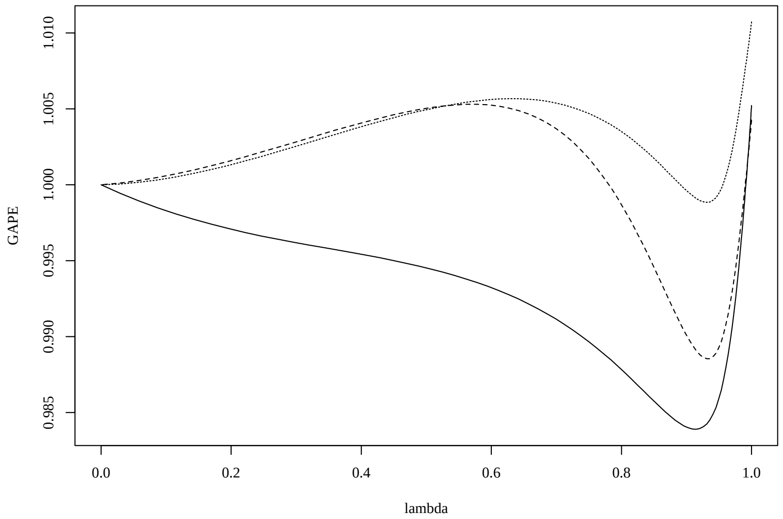

Figure 2. GAPE as a function of λ

The solid line corresponds to GAPE_{KL}. The dotted and the dashed lines correspond to GAPE under the χ^2-distance, and the Hellinger-distance respectively. The values of GAPE are rescaled so that $\text{GAPE}(0) = 1$. The optimal values of λ for the three distance functions are 0.9148, 0.9311, 0.9311 respectively.

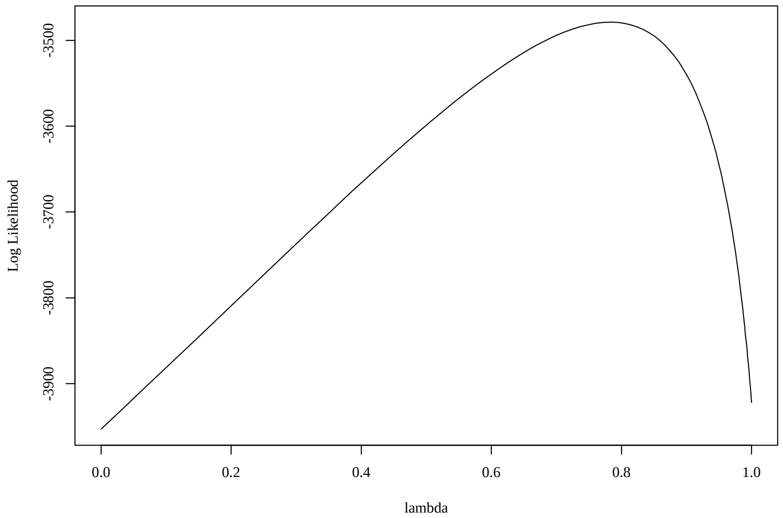

Figure 3. Log likelihood as a function of λ

The maximum likelihood estimate of λ is $\hat{\lambda} = 0.7846$, which coincides with the value chosen by AIC and BIC.

modeling cross-sectional heterogeneity (i.e., large λ) would yield better market share predictions than modeling temporal correlation (i.e., small λ).

6. Final Remarks

In the previous section, we did not specify the form of $\pi_{it}(\alpha)$. To calculate market share predictions, we simply estimated the r_{Ni} terms in (3) by the corresponding sample proportions. Technically, this corresponds essentially to a nonparametric model for $\pi_{it}(\alpha)$, or alternatively, a parametric model with large $dim(\alpha)$. Thus in our application, the first penalty term in (6) is likely to dominate the second penalty term. According to Theorem 2, it is therefore not surprising that the GAPE_{KL} criterion favors heterogeneous models.

Modeling categorical panel data is a difficult problem. Much of the difficulty, nevertheless, lies in the unique structure of panel data rather than deficiencies in methodology. A good understanding of individual behavior does not necessarily lead to accurate predictions of aggregate outcomes, no matter how good the models are otherwise. For instance, it is counter-intuitive to assume that individual decisions are independent across time. Yet decades of empirical research have shown that cross-sectional heterogeneity plays a more important role in the modeling of panel data. We reached the same conclusion in our application. The irony here is that a "wrong" model might wind up doing better than a "true" model. It is therefore not the true model that we should be looking for in model selection, but models that would help us accomplish specific goals.

Finally, we point out that aggregate prediction is not always the natural goal of modeling. In many cases, cross-sectional prediction is of more interest. A typical example is the classification problem. In the context of marketing research, mixture models are used mainly to address the problem of market segmentation. In other words, the goal here is to determine whether there are different types of consumer behavior and how to classify individuals into one of a few segments. Still, it is fair to say that models that fail to yield accurate market share predictions are not serving their ultimate goals.

Acknowledgement

This research is partially supported by NSA Grant MDA904-93-H-3014 and the Research Foundation of the University of Pennsylvania. The author wishes to thank Peter Fader for providing the scanner panel data and helpful conversations.

Appendix

The proofs

The main step in the proof of Theorem 1 is the following

Lemma 1. *Let* $\mathbf{B}_t = \sum_{j=1}^{t} \mathbf{A}_j$*, where* $\mathbf{A}_t$*,* $t = 1, \ldots, T$*, are positive definite* $d \times d$ *matrices satisfying* $\|\mathbf{B}_{t-1}^{-1}\mathbf{A}_t\| = o(t)$ *and* $\|\mathbf{B}_t\| \to \infty$*. Then there exists* t_0 *large enough such that*

$$\sum_{t=t_0}^{T} \mathrm{tr}(\mathbf{B}_{t-1}^{-1}\mathbf{A}_t) = \log\det(\mathbf{B}_T)\{1 + o(T)\}.$$

Proof. Let $\lambda_i(t)$, $i = 1, \ldots, d$, be the eigenvalues of $\mathbf{B}_{t-1}^{-1}\mathbf{A}_t$. Then by assumption, we have $\max_i \lambda_i(t) \to 0$, as $t \to \infty$, which further implies that

$$\log\det(\mathbf{I} + \mathbf{B}_{t-1}^{-1}\mathbf{A}_t) = \log\prod_i\{1 + \lambda_i(t)\} \approx \sum_i \lambda_i(t) = \mathrm{tr}(\mathbf{B}_{t-1}^{-1}\mathbf{A}_t).$$

The conclusion follows since

$$\log\det(\mathbf{B}_t) - \log\det(\mathbf{B}_{t-1}) = \log\det(\mathbf{I} + \mathbf{B}_{t-1}^{-1}\mathbf{A}_t).$$

Proof of Theorem 1. Since $\tilde{p}_t(k, \alpha, \hat{\theta}_{t-1})$ depends only on the past data, we have from the definition of ξ_{NT} that

$$E(\xi_{NT}) = \sum_{t=t_0}^{T}\sum_{k=1}^{K} E\{s_t(k)\} \cdot E\log\{\,\tilde{p}_t(k, \alpha, \theta)/\tilde{p}_t(k, \alpha, \hat{\theta}_{t-1})\}.$$

Next, from Section 2, we have $E\{s_t(k)\} \to \tilde{p}_t(k, \alpha, \theta)$ as $N \to \infty$. Hence asymptotically, $E(\xi_{NT}) \approx \sum_{t=t_0}^{T} E\{D_{KL}(\tilde{\mathbf{p}}_t, \tilde{\mathbf{s}}_t)\}$, where $\tilde{\mathbf{p}}_t = \{\,\tilde{p}_t(1, \alpha, \theta), \ldots, \tilde{p}_t(K, \alpha, \theta)\}$ and $\tilde{\mathbf{s}}_t = \{\tilde{p}_t(1, \alpha, \hat{\theta}_{t-1}), \ldots, \tilde{p}_t(K, \alpha, \hat{\theta}_{t-1})\}$. Under the assumptions, $\hat{\theta}_t$ is consistent and efficient. Hence $E\{(\hat{\theta}_t - \theta)(\hat{\theta}_t - \theta)'\} \approx \mathbf{V}_t^{-1}$. Following the Taylor expansion $\tilde{p}_t(k, \alpha, \hat{\theta}_{t-1}) - \tilde{p}_t(k, \alpha, \theta) \approx \{\partial\tilde{p}_t(k, \alpha, \theta)/\partial\theta\}'(\hat{\theta}_{t-1} - \theta)$, we get

$$\begin{aligned} E(\xi_{NT}) &\approx \frac{1}{2}\sum_{t=t_0}^{T}\sum_{k=1}^{K} \frac{1}{\tilde{p}_t(k, \alpha, \theta)} E\{\,\tilde{p}_t(k, \alpha, \hat{\theta}_{t-1}) - \tilde{p}_t(k, \alpha, \theta)\}^2 \\ &\approx 2^{-1}\sum_{t=t_0}^{T} \mathrm{tr}\{\mathbf{V}_{t-1}^{-1}\Delta\mathbf{U}_t\}, \end{aligned}$$

where

$$\mathbf{U}_t = \sum_{j=1}^{t}\sum_{k=1}^{K} \frac{1}{\tilde{p}_j(k, \alpha, \theta)} \left\{\frac{\partial\tilde{p}_j(k, \alpha, \theta)}{\partial\theta}\right\}\left\{\frac{\partial\tilde{p}_j(k, \alpha, \theta)}{\partial\theta}\right\}'$$

and $\Delta\mathbf{U}_t = \mathbf{U}_t - \mathbf{U}_{t-1}$. Next, the Cauchy-Schwarz inequality implies that

$$\Delta\mathbf{U}_t \le \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{r_{Ni}}{p_{it}(k, \theta)} \left\{\frac{\partial p_{it}(k, \theta)}{\partial\theta}\right\}\left\{\frac{\partial p_{it}(k, \theta)}{\partial\theta}\right\}' \le (1 - \rho)^{-1}N^{-1}\Delta\mathbf{V}_t,$$

where for two matrices A and B, $A \leq B$ means that $B - A$ is non-negative definite. The second inequality in the above expression holds since

$$r_{Ni} = \{1 - \pi_{it}(\alpha)\} \Big/ \sum_{i=1}^{N} \{1 - \pi_{it}(\alpha)\} \leq (1-\rho)^{-1} N^{-1} \{1 - \pi_{it}(\alpha)\}.$$

Consequently, we get

$$\sum_{t=t_0}^{T} E\{D_{KL}(\tilde{\mathbf{p}}_t, \tilde{\mathbf{s}}_t)\} \leq 2^{-1}(1-\rho)^{-1} N^{-1} \sum_{t=t_0}^{T} \text{tr}\{\mathbf{V}_{t-1}^{-1} \Delta \mathbf{V}_t\}.$$

It follows from Lemma 1 of the Appendix that

$$\sum_{t=t_0}^{T} E\{D_{KL}(\tilde{\mathbf{p}}_t, \tilde{\mathbf{s}}_t)\} \leq 2^{-1}(1-\rho)^{-1} N^{-1} \log\det(\mathbf{V}_T).$$

Hence the first inequality in Theorem 1 holds.

To prove the second inequality, let $\psi_{Nt}(k) = \partial \tilde{p}_t(k, \alpha, \theta)/\partial \alpha = \sum_{i=1}^{N} p_{it}(k, \theta)$ $\partial r_{Ni}/\partial \alpha$. Then it is easy to verify that

$$\psi_{Nt}(k) = (1 - \bar{\pi}_t)^{-1} N^{-1} \sum_{i=1}^{N} \{\tilde{p}_t(k, \alpha, \theta) - p_{it}(k, \theta)\} \frac{\partial \pi_{it}(\alpha)}{\partial \alpha},$$

where $\bar{\pi}_t = N^{-1} \sum_{i=1}^{N} \pi_{it}(\alpha)$. Next, define $\tilde{\mathbf{U}}_t = \sum_{j=1}^{t} \sum_{k=1}^{K} \psi_{Nj}(k) \psi'_{Nj}(k)$. Then the Cauchy-Schwarz inequality implies that

$$\psi_{Nt}(k) \psi'_{Nt}(k) \leq 4^{-1}(1-\rho)^{-2} N^{-1} \gamma \Delta \mathbf{W}_t.$$

Notice that $\Delta \tilde{\mathbf{U}}_t = \psi_{Nt}(k) \psi'_{Nt}(k)$. The result for η_{NT} follows from the same argument that we used to treat ξ_{NT}.

Proof of Theorem 2. This is nearly a corollary of Theorem 1. The following brief statements will be adequate: (a) When $\partial \tilde{p}_{it}(k, \alpha, \theta)/\partial \theta \to 0$, we have $(NT)^{-1} \mathbf{W}_T \to 0$. Hence the conclusion; (b) By definition, $\gamma = 0$ when the model is homogeneous. The conclusion follows from the second inequality in Theorem 1; (c) Homogeneity is the condition for equality in the Cauchy-Schwarz inequality, which is used to establish the upper bounds in Theorem 1.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2*nd International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csáki), 267-281. Akadémiai Kiadó, Budapest.

Allenby, G. and Lenk, P. (1994). Modeling household purchase behavior with logistic normal regression. *J. Amer. Statist. Assoc.* **89**, 1218-1231.

Benjamini, Y. and Krieger, A. M. (1992). Market share paradox and heterogeneous chains. *Ann. Appl. Probab.* **2**, 1019-1023.

Cohen, J. E. (1986). An uncertainty principle in demography and the unisex issue. *Amer. Statist.* **40**, 32-39.

Cosslett, S. R. and Lee, L. F. (1985). Serial correlation in latent discrete variable models. *J. Econometrics* **27**, 79-97.

Dawid, A. P. (1984). Statistical theory, the prequential approach. *J. Roy. Statist. Soc. Ser. A* **147**, 278-292.

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Clarendon Press, Oxford.

Fader, P. and Lattin, J. M. (1993). Accounting for heterogeneity and nonstationarity in a cross-sectional model of consumer purchase behavior. *Marketing Science* **12**, 304-317.

Gilula, Z. and Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *J. Amer. Statist. Assoc.* **89**, 645-656.

Gottschau, A. (1994). Markov chain models for multivariate binary panel data. *Scand. J. Statist.* **21**, 57-72.

Guadagni, P. M. and Little, J. D. C. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science* **2**, 203-238.

Gupta, S. and Chintagunta, P. K. (1994). On using demographic variables to determine segment membership in logit mixture models. *J. Marketing Research* **31**, 128-136.

Haccou, P. and Meelis, E. (1994). *Statistical Analysis of Behavioural Data.* Oxford University Press, New York.

Hsiao, C. (1986). *Analysis of Panel Data.* Cambridge University Press, New York.

Kamakura, W. A. and Russell, G. J. (1989). A probabilistic choice model for market segmentation and elasticity structure. *J. Marketing Research* **26**, 379-390.

Massy, W. F., Montgomery, D. B. and Morrison. D. G. (1970). *Stochastic Models of Buying Behavior.* The M.I.T. Press, Cambridge, Mass.

Rissanen, J. (1986a). A predictive least square principle. *IMA J. Math. Control. Inform.* **3**, 211-222.

Rissanen, J. (1986b). Order estimation by accumulated prediction errors. *J. Appl. Probab.* **23**, 55-61.

Rissanen, J. (1987). Stochastic complexity and modeling. *J. Roy. Statist. Soc. Ser. B* **49**, 223-239.

Samuels, M. L. (1993). Simpson's paradox and related phenomena. *J. Amer. Statist. Assoc.* **88**, 81-88.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Wei. C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20**, 1-42.

Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6302, U.S.A.

(Received November 1994; accepted September 1996)