Statistica Neerlandica (1999) Vol. 53, nr. 1, pp. 21–35

Description of purchase incidence by multivariate heterogeneous Poisson processes

A. W. Hoogendoorn*

Economics Institute Tilburg, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

D. Sikkel

CentERdata, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

In this paper the application of bivariate Poisson heterogeneous models to budget data is studied. This study was motivated from inconsistencies that we encountered when univariate Poisson based models were applied to cumulative data sets. Application of a multivariate Poisson based model is a possible solution to this problem. In this paper we will study the feasibility of estimators based on these models.

Key Words & Phrases: budget survey, consumer models, product penetration, bivariate gamma distribution, latent variable models, latent class models.

1 Introduction

In research on consumer behaviour one is interested in key indices that describe the purchasing of a product. One such index is the *product period penetration*, that denotes the fraction of households that buys at least one product item in a given period (see e.g. EHRENBERG, 1988). In this paper we are interested in estimating this quantity for a large set of products in the field of meat, poultry and eggs for periods of a month, quarter and their cumulative periods (half year, three quarters, full year).

We have budget data that were collected for the product boards for livestock, meat and eggs. The data were collected by use of electronic diaries that were developed as a data collection system for a Completely Automated System for Information Processing (CASIP, see SARIS et al., 1992). The gathered information includes product, volume, price and the place of purchase (outlet). To estimate product period penetrations we use numbers of purchases in the time reported. If all households reported over the full period, we could estimate the period penetration by a direct estimator, i.e. the sample fraction of households that bought at least one product in that period. If, however, this is not the case we need model assumptions to estimate penetrations.

* A.W.Hoogendoorn@kub.nl

Table 1. Four models and their parameters

Model	Parameters
Poisson	λ (intensity)
Poisson Gamma	b (scale), k (shape)
Poisson Spike	λ, p_0 (probability of being a non-buyer)
Poisson Gamma Spike	p_0, b, k

Models that are based on a Poisson assumption for the underlying process are most popular to fit to purchase incidence data, because of their computational ease (see e.g. MORRISON and SCHMITTLEIN, 1988). We applied four Poisson based models to these data that differ in the way heterogeneity is treated (see SIKKEL and HOOGENDOORN, 1995). These models and their parameters are listed in Table 1.

If we are interested to estimate the model parameters of *two subsequent periods*, we can do this either *separately*, where we find a different set of parameters for each period, or *combined*, where we add the data of the two periods, and estimate on the combined data one set of new parameters. These parameters can be used to describe the distribution of the number of purchases M_1, M_2 and M for the first, second and combined quarters respectively. In the course of the process we first estimate the model for the first quarter. In the case of a Poisson Gamma Spike model we have estimates for p_{01}, b_1 and k_1, that may be obtained by ML-estimation (see SIKKEL and HOOGENDOORN, 1995). From these estimates the penetration for the first quarter is derived and published. The next quarter we estimate p_{02}, b_2 and k_2 from which we derive the penetration for the second quarter. In order to estimate a half year penetration we can apply the same estimation procedure to the combined data of the complete period. Then the parameters that reflect the complete period are p_0, b and k. This, however, can create a problem. When the parameter set p_{01}, b_1 and k_1 is different from p_{02}, b_2 and k_2 this is 'interpreted' by the estimation procedure as extra heterogeneity within households, since the coefficient of variation of the purchases in time is fixed. The more the heterogeneity, the higher the variance of the distribution of the number of purchases. So this may lead to the inconsistency that the estimate of $P\{M=0\}$ is larger than the estimate of $P\{M_1=0\}$ or $P\{M_2=0\}$ (in half a year there are more non-buyers than in one or each of the quarters). The penetrations, which are the complements of the zero-probabilities, are also inconsistent. Practically, it is impossible to avoid this problem, since an estimated penetration has to be published immediately after a quarter. Theoretically, it is a drawback of the combined model; with the exceptions of the Poisson process and the Poisson Gamma process with equal shape parameters, it is simply not true that a succession of two processes with different parameters leads to the same process with 'average' parameters. As a consequence, we can not use the combined model and have to look for an alternative. Therefore, we will treat the quarters as two separate dimensions. Although we separate these dimensions, it does not mean that they are independent, so that we must take the dependency between these dimensions into account.

In this paper we will generalise the existing models to a *multivariate model*. We will formulate and analyse bivariate models for both the Poisson Gamma model and the Poisson Spike model under the assumption that the model parameters of two subsequent quarters were estimated separately by application of univariate estimators to the marginal data, and that the corresponding estimates for penetrations were published. Our goal is to find an estimator for the penetration of the *combined* period, that will not contradict earlier published figures. Therefore, we will study the dependency between two marginal processes by estimation of the extra parameters from the combined data, where we consider the marginal processes given. To gain insight we will link this problem to *classical measurement theory*, where attenuation of dependencies plays a crucial role.

The outline of this paper is as follows. In section 2 we will discuss the bivariate Poisson Gamma and the bivariate Poisson Spike model. In section 3 we will discuss a heuristic estimation method for these bivariate models based on the method of moments. In section 4 we will discuss ML-estimation for these models. In section 5 we will present some results, that we discuss in section 6. Section 7 concludes.

2 Two bivariate Poisson processes

In this section we will discuss two bivariate Poisson processes: the bivariate Poisson Gamma process and the bivariate Poisson Spike process. We will describe both models in terms of latent variables for the intensity of the Poisson process. In order to formulate two estimation procedures, we will derive expressions for both the bivariate distribution and for the correlation between two such variables.

2.1 *The bivariate Poisson Gamma process*

The Poisson Gamma process is a latent variable model. It assumes the existence of a latent variable Λ that indicates the intensity of the individual Poisson processes. We assume that Λ has a $gamma(b, k)$ distribution, where b is the scale parameter and k is the shape parameter of the gamma distribution. In the bivariate case we have variables Λ_1 and Λ_2 for the first and the second quarter and with $gamma(b_1, k_1)$ and $gamma(b_2, k_2)$ distributions respectively. We will discuss a multivariate version of the gamma distribution that describes the dispersion of the Poisson parameter over the population in each of the periods as given in Johnson and Kotz (1972). Therefore, we will restrict ourselves to the bivariate case, and suppose that Λ_1 and Λ_2 may have different shape parameters k_1 and k_2, but have both a scale parameter 1. The idea behind the formulation of a bivariate gamma distribution is that the two jointly distributed random variables Λ_1 and Λ_2 have a common part X_0 besides independent parts X_1 and X_2, so that we can write

$$\Lambda_1 = X_0 + X_1$$

and

$$\Lambda_2 = X_0 + X_2$$

Here X_0, X_1 and X_2 all have the same scale parameter 1, but have different shape parameters h_0, h_1 and h_2 respectively, with $k_1 = h_0 + h_1$ and $k_2 = h_0 + h_2$. Note that by definition $h_0 \leqslant \min(k_1, k_2)$. Since the joint density of X_0, X_1 and X_2 is defined as

$$f(x_0, x_1, x_2) = \frac{1}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} x_0^{h_0-1} x_1^{h_1-1} x_2^{h_2-1} \, \mathrm{e}^{-(x_0+x_1+x_2)}$$

the joint density of X_0, Λ_1 and Λ_2 is

$$f(x_0, \lambda_1, \lambda_2) = \frac{1}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} x_0^{h_0-1} (\lambda_1 - x_0)^{h_1-1} (\lambda_2 - x_0)^{h_2-1} \, \mathrm{e}^{x_0-\lambda_1-\lambda_2} \quad (1)$$

so that we can find the joint density of Λ_1 and Λ_2 by integrating formula (1) with respect to x_0 from 0 to $\tilde{\lambda}$, where $\tilde{\lambda}$ is the minimum of λ_1 and λ_2:

$$f(\lambda_1, \lambda_2) = \frac{\mathrm{e}^{-\lambda_1-\lambda_2}}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} \int_0^{\tilde{\lambda}} x_0^{h_0-1} (\lambda_1 - x_0)^{h_1-1} (\lambda_2 - x_0)^{h_2-1} \, \mathrm{e}^{x_0} \, \mathrm{d}x_0 \quad (2)$$

The correlation ρ^* between Λ_1 and Λ_2 is

$$\rho^* = \frac{h_0}{\sqrt{(h_0 + h_1)(h_0 + h_2)}} = \frac{h_0}{\sqrt{k_1 k_2}} \quad (3)$$

In this approach the model has three free parameters, due to the fact that we fixed the scale parameters all to unity. However, there will be no loss of generality, since time can be rescaled in such a way that the scale parameters are equal to one.

2.2 *The bivariate Poisson Spike process*

The Poisson Spike process is essentially a latent class model. It assumes the existence of a latent variable C that indicates that an individual is a buyer. When $C = 0$ the individual is a non-buyer, when $C = 1$ the individual buys according to a Poisson process. The model contains two parameters: p_0 the probability of being a non-buyer and μ the intensity of the Poisson process of the buyers. In the bivariate case we have variables C_1 and C_2 for the first and the second quarter respectively with the same meaning as C. The variables C_1 and C_2, however, are not independent and not identically distributed. We can describe the distribution of the latent classes by a transition matrix Q, where the elements q_{ij} are the transition probabilities to move from class i to class j ($i, j = 0, 1$). Let the initial distribution of the latent Markov process be p_{01} and $1 - p_{01}$, then

$$p_{02} = p_{01} q_{00} + (1 - p_{01}) q_{10} \quad (4)$$

The bivariate Poisson Spike model then has five free parameters. The model can be specified in such a way that four parameters (μ_1, p_{01}, μ_2 and p_{02}) determine the marginal distributions, and that one remaining parameter (e.g. q_{00}) determines the dependency in the model. Given the marginals there are some restrictions for

the values of q_{00}. Apart from the obvious bounds of 0 and 1, q_{00} should also be smaller than p_{02}/p_{01} and should be larger than $(p_{01}+p_{02}-1)/p_{01}$.

We can regard the Poisson Spike process also as a latent variable model. Therefore, we will say that a non-buyer 'buys' according to a Poisson process with intensity 0. The latent variable Λ that indicates the intensity of the individual Poisson process is then *discrete* and takes two values: zero (with probability p_0) and μ (with probability $1-p_0$). In the bivariate case we have two latent variables Λ_1 and Λ_2, representing the intensities of the first and second period respectively. This bivariate latent variable has a discrete distribution in the four points (0, 0), (μ_1, 0), (0, μ_2) and (μ_1, μ_2) with probabilities π_{00}, π_{10}, π_{01} and π_{11} respectively. Relationships between the p_{0i}s and the q_{ij}s on the one hand and the π_{ij}s on the other can easily be derived. The correlation ρ^* between Λ_1 and Λ_2 is

$$\begin{aligned}\rho^* &= \frac{\pi_{11}-(\pi_{10}+\pi_{11})(\pi_{01}+\pi_{11})}{\sqrt{(\pi_{10}+\pi_{11})(\pi_{00}+\pi_{01})(\pi_{00}+\pi_{10})(\pi_{01}+\pi_{11})}} \\ &= \frac{1-p_{01}-p_{02}+p_{01}q_{00}-(1-p_{01})(1-p_{02})}{\sqrt{p_{01}(1-p_{01})p_{02}(1-p_{02})}}\end{aligned} \tag{5}$$

3 Estimation based on correlation in a two way table

The data on which we base the estimation of the parameters consist of numbers of purchases of respondents in the two periods. Let $M_1(t_1)$ be the number of purchases of a respondent who reported over a time t_1 in the first period and $M_2(t_2)$ be the number of purchases of a respondent who reported over a time t_2 in the second period. If all respondents reported over the same time period t_1 and t_2, then the data can be written in the form of a two-way table $\{m_{ij}\}$, where m_{ij} is the number of respondents that reported i purchases of a product in the first quarter and j purchases in the second quarter (see Table 2).

In this section we will describe how we can estimate the bivariate model using a direct estimator of the correlation in the table $\{m_{ij}\}$ for both the bivariate Poisson

Table 2. Two way table showing fictive numbers of households that made M_1 purchases in t_1 and M_2 purchases in t_2

Number of purchases in t_1	Number of purchases in t_2							
	$M_2=0$	1	2	3	4	5	6	...
$M_1=0$	92	44	12	7	2	0	1	...
1	52	46	28	14	5	4	3	...
2	19	22	41	18	10	8	4	...
3	12	15	17	29	12	8	5	...
4	9	13	17	16	21	9	7	...
5	6	5	8	9	10	14	6	...
6	1	2	0	7	9	9	12	...
...	...	...	...	...	...	...	...	...

Gamma model and the bivariate Poisson Spike model. The estimation method is a generalisation of the method of moments, and uses the sample correlation coefficient in the two-way table $\{m_{ij}\}$. In order to formulate this estimator we will first discuss the notion of attenuation.

3.1 *Attenuation*

In classical test theory (see e.g. LORD and NOVICK, 1968) the notion of attenuation is used to describe the effect of measurement error on correlations between observed variables. If we have a classical model for tests the observed score X is split into two parts: a true score T and a measurement error E

$$X = T + E$$

We assume that the measurement error E has zero expectation and is uncorrelated with the latent true score T. The square of the correlation $\rho(X, T)$ between observed and true score is called the *reliability* of the measurement. It is equal to the quotient σ_T^2/σ_X^2 of variances of T and X and takes values between 0 and 1. If we observe more than one variable we also have correlations between observed scores and correlations between true scores. Because the true scores are measured with some error, the correlation between the observed scores will be lower than the correlation between the true scores. Suppose that we have two measurement instruments X_1 and X_2 measuring true scores T_1 and T_2, then the correlation $\rho(T_1, T_2)$ between the true scores T_1 and T_2 is given by

$$\rho(T_1, T_2) = \frac{\rho(X_1, X_2)}{\sqrt{\rho(X_1, T_1)\rho(X_2, T_2)}}$$

The idea is that the correlation between observed scores is less than the correlation between corresponding true scores, because the former correlation is *attenuated* by the unreliability of the measurements. We also write

$$\rho(T_1, T_2) = \frac{\rho(X_1, X_2)}{c_1 c_2} \tag{6}$$

where we call c_i the attenuation factor of measurement i, being $c_i = \rho(X_i, T_i)^{1/2}$, $(i = 1, 2)$. The notion of attenuation has also been applied to heterogeneous renewal processes (see SIKKEL and JELIERSE, 1987). Here the role of the true score is played by f, the frequency of the individual renewal process. In general f will be a function of a latent variable Λ, that models the heterogeneity of the individuals. Note that f itself can also be considered to be a latent variable. The role of the observed score is played by $M(t)$, the number of renewals in an observation time period t. Although we assume that $M(t)$ is observed without measurement error, we find lower correlations between observed scores than between latent scores. The attenuation comes from observing the latent variable f as if we have some measurement error through the

randomness of the renewal data. For a heterogeneous renewal process the analogue of equation (6) is

$$\rho(f_1, f_2) = \frac{\rho(M_1(t_1), M_2(t_2))}{c_1(t_1)c_2(t_2)} \tag{7}$$

where the attenuation factor $c_i(t_i)$ is now a function of t_i, the length of the observation period. The attenuation factors for the renewal process are then

$$\begin{aligned} c_i(t_i) &= \sqrt{\rho(M_i(t_i), f_i)} \\ &= \sqrt{\frac{t_i^2 \sigma_{f_i}^2}{E_{\Lambda_i}[\sigma^2(M_i(t_i) \mid \Lambda_i)] + t_i^2 \sigma_{f_i}^2}} \end{aligned} \tag{8}$$

It follows that $c_i(t_i) \downarrow 0$ for $t_i \downarrow 0$, and that $c_i(t_i) \to 1$ for $t_i \to \infty$ (see **Sikkel** and **Jelierse**, 1987).

3.2 *The bivariate Poisson Gamma model*

In order to formulate an estimator based on the correlation in a two way table, we will first find an expression for the attenuation for the univariate Poisson Gamma model in terms of the estimated parameters. In the case of a univariate Poisson Gamma model the heterogeneity is modelled in such a way that the frequency f_i of the individual Poisson process in period i has a *gamma*(b_i, k_i) distribution. Note that in this case we have no distinction between f_i and Λ_i. Conditional on Λ_i then $M_i(t_i)$ is a Poisson process, so that $E_{\Lambda_i}[\sigma^2(M_i(t_i) \mid \Lambda_i)] = E_{\Lambda_i}[\Lambda_i t_i] = b_i k_i t_i$. Since $\sigma_{f_i}^2 = b_i^2 k_i$ it follows from (8) the attenuation factor $c_i(t_i)$ for the Poisson Gamma model is

$$c_i(t_i) = \sqrt{\frac{b_i t_t}{1 + b_i t_i}} \tag{9}$$

Now we can formulate a procedure to find an estimate for h_0. Using the table $\{m_{ij}\}$ we first apply standard methods to find maximum likelihood estimates for the four parameters b_1, k_1, b_2 and k_2 of the univariate Poisson Gamma models using only the marginals $\{m_{i+}\}$ and $\{m_{+j}\}$ of the table. Based on the table $\{m_{ij}\}$ we can also compute a correlation coefficient $\hat{\rho}$ between $M_1(t_1)$ and $M_2(t_2)$. We can relate this to ρ^*, the correlation between Λ_1 and Λ_2 by (7) using the attenuation factors of (9), so that ρ^* can be estimated by

$$\hat{\rho}^* = \sqrt{\frac{1 + \hat{b}_1 t_1}{\hat{b}_1 t_1} \frac{1 + \hat{b}_2 t_2}{\hat{b}_2 t_2}} \hat{\rho} \tag{10}$$

The next step is to transform the marginal distributions such that they become standard gamma distributions ($b_1 = b_2 = 1$). Without loss of generality, we change the unit of time by a factor b_1 and b_2, respectively. By defining $u_1 = \hat{b_1} t_1$ and $u_2 = \hat{b_2} t_2$ the

table $\{m_{ij}\}$ can be interpreted as if it refers to the time interval u_1 and u_2. The corresponding random variables have a gamma distribution with scale parameters 1. From (3) we find that the value of h_0 can be estimated by

$$\hat{h}_0 = \hat{\rho}^* \sqrt{\hat{k}_1 \hat{k}_2} \tag{11}$$

3.3 *The bivariate Poisson Spike model*

In the case of a univariate Poisson Gamma model the heterogeneity is such that the frequency f_i is either 0, with probability p_{0i}, or μ_i. Note that there is no distinction between f_i and Λ_i again. We find here that $E_{\Lambda_i}[\sigma^2(M_i(t_i) \mid \Lambda_i)] = E_{\Lambda_i}[\Lambda_i t_i] = \mu_i p_{0i} t_i$ and $\sigma^2_{\Lambda_i} = \mu_i^2 p_{0i}(1 - p_{0i})$. The attenuation factor $c_i(t_i)$ for the Poisson Spike model is then

$$c_i(t_i) = \sqrt{\frac{p_{0i}\mu_i t_i}{1 + p_{0i}\mu_i t_i}} \tag{12}$$

From here we can define in the same way as for the Poisson Gamma model an estimator for the fifth parameter of the bivariate Poisson Spike model. We start with univariate ML-estimates for the four parameters μ_1, p_{01}, μ_2 and p_{02}. Next we find an estimate $\hat{\rho}$ for the correlation in the table $\{m_{ij}\}$ that we can use to estimate ρ^*, the correlation between the latent intensities by

$$\hat{\rho}^* = \sqrt{\frac{1 + \hat{p}_{01}\hat{\mu}_1 t_1}{\hat{p}_{01}\hat{\mu}_1 t_1} \frac{1 + \hat{p}_{02}\hat{\mu}_2 t_2}{\hat{p}_{02}\hat{\mu}_2 t_2}}\, \hat{\rho} \tag{13}$$

From equation (5) it then follows we estimate q_{00} by

$$\hat{q}_{00} = \frac{\sqrt{p_{01}(1 - p_{01})p_{02}(1 - p_{02})}\hat{\rho}^* + p_{01}p_{02}}{p_{01}} \tag{14}$$

4 Maximum likelihood estimation

In this section we will derive ML-estimators based on the individual observations of $M_1(t_1)$ and $M_2(t_2)$ for both the bivariate Poisson Gamma and the bivariate Poisson Spike model. Again we will consider the situation where we already estimated the marginal distributions, so that we maximise the conditional likelihood given the marginals.

4.1 *The bivariate Poisson Gamma model*

In order to find the maximum likelihood estimator for h_0 we need an expression for $P\{M_1(t_1) = m_1,\ M_2(t_2) = m_2\}$. Since we consider the estimates of b_1, k_1, b_2 and k_2 to be given we can transform the time by changing t_1 into $u_1 = \hat{b}_1 t_1$ and $u_2 = \hat{b}_2 t_2$. Now in the transformed time the marginal Gamma distributions have scale parameter 1. Let

$L(h_0; m_1, m_2, u_1, u_2)$ be the likelihood $P\{M_1(u_1) = m_1,\ M_2(u_2) = m_2\}$ for h_0. According to the Poisson Gamma process, by formula (1) we have

$$
\begin{aligned}
&L(h_0; m_1, m_2, u_1, u_2)\\
&\quad = \int_{x_0=0}^{\infty}\int_{\lambda_1=x_0}^{\infty}\int_{\lambda_2=x_0}^{\infty} \frac{\mathrm{e}^{-(\lambda_1 u_1+\lambda_2 u_2)}}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} \frac{(\lambda_1 u_1)^{m_1}(\lambda_2 u_2)^{m_2}}{m_1!m_2!} x_0^{h_0-1}(\lambda_1 - x_0)^{h_1-1}\\
&\qquad \times (\lambda_2 - x_0)^{h_2-1}\ \mathrm{e}^{-(x_0+\lambda_1-x_0+\lambda_2-x_0)}\ \mathrm{d}x_0\ \mathrm{d}\lambda_1\ \mathrm{d}\lambda_2\\
&\quad = \int_{x_0=0}^{\infty}\int_{x_1=0}^{\infty}\int_{x_2=0}^{\infty} \frac{\mathrm{e}^{-[(x_0+x_1)u_1+(x_0+x_2)u_2]}}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} \frac{[(x_0+x_1)u_1]^{m_1}[(x_0+x_2)u_2]^{m_2}}{m_1!m_2!}\\
&\qquad \times x_0^{h_0-1}x_1^{h_1-1}x_2^{h_2-1}\ \mathrm{e}^{-(x_0+x_1+x_2)}\ \mathrm{d}x_0\ \mathrm{d}x_1\ \mathrm{d}x_2\\
&\quad = \frac{u_1^{m_1}u_2^{m_2}}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)m_1!m_2!} \sum_{j_1=1}^{m_1}\sum_{j_2=1}^{m_2}\int_{x_0=0}^{\infty}\int_{x_1=0}^{\infty}\int_{x_2=0}^{\infty} \mathrm{e}^{-x_0(u_1+u_2+1)}x_0^{k_0+j_1+j_2-1}\\
&\qquad \times \binom{m_1}{j_1}\binom{m_2}{j_2}\mathrm{e}^{-x_1(u_1+1)}x_1^{h_1+m_1-j_1-1}\ \mathrm{e}^{-x_2(u_2+1)}x_2^{h_2+m_2-j_2-1}\ \mathrm{d}x_0\ \mathrm{d}x_1\ \mathrm{d}x_2\\
&\quad = \frac{u_1^{m_1}u_2^{m_2}/(m_1!m_2!)}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} \sum_{j_1=1}^{m_1}\sum_{j_2=1}^{m_2}\\
&\qquad \times \frac{\binom{m_1}{j_1}\binom{m_2}{j_2}\Gamma(h_0+j_1+j_2)\Gamma(h_1+m_1-j_1)\Gamma(h_2+m_2-j_2)}{(u_1+u_2+1)^{h_0+j_1+j_2}(u_1+1)^{h_1+m_1-j_1}(u_2+1)^{h_2+m_2-j_2}}\\
&\quad = \sum_{j_1=1}^{m_1}\sum_{j_2=1}^{m_2} \frac{u_1^{j_1}u_2^{j_2}h_0^{[j_1+j_2]}}{j_1!j_2!(u_1+u_2+1)^{h_0+j_1+j_2}}\\
&\qquad \times \frac{u_1^{m_1-j_1}h_1^{[m_1-j_1]}}{(m_1-j_1)!(u_1+1)^{h_1+m_1-j_1}} \frac{u_2^{m_2-j_2}h_2^{[m_2-j_2]}}{(m_2-j_2)!(u_2+1)^{h_2+m_2-j_2}}
\end{aligned}
\tag{15}
$$

where we write $h^{[j]} = h(h+1)\ldots(h+j-1)$ for $j > 0$, and $h^{[0]} = 1$. Note that since the marginal parameter estimates are given, we have $h_1 = \hat{k_1} - h_0$ and $h_2 = \hat{k_2} - h_0$. Then the log-likelihood of a given sample is equal to

$$
l(h_0) = \sum_{m_1}\sum_{m_2}\sum_{u_1}\sum_{u_2} n(m_1, m_2, u_1, u_2)\log(L(h_0; m_1, m_2, u_1, u_2))
$$

where $n(m_1, m_2, u_1, u_2)$ is the number of respondents who reported m_1 and m_2 purchases over transformed periods of lengths u_1 and u_2 respectively. An ML-estimate is obtained by numerically maximising this one-dimensional likelihood function.

4.2 *The bivariate Poisson Spike model*

Our situation is that the parameters p_{01}, μ_1, p_{02} and μ_2 of the marginal distributions of the bivariate Poisson Spike distribution are given. In order to find the ML-estimator for the fifth parameter q_{00} we need expressions for the probability distribution of the outcomes. These can easily be written down, e.g.

$$P\{M_1(t_1) = 0, M_2(t_2) = 0\} = p_{01}(q_{00} + (1 - q_{00})\exp\{-\mu_2 t_2\}) + (1 - p_{01})\exp\{-\mu_1 t_1\}(q_{10} + (1 - q_{10})\exp\{-\mu_2 t_2\}) \quad (16)$$

Note that q_{10} can be written as a function of p_{01}, p_{02} and q_{00} by formula (4). Then from this and other expressions for the probabilities we can derive the likelihood function for the bivariate distribution of $M_1(t_1)$ and $M_2(t_2)$.

5 Results

We applied the bivariate Poisson models and the described estimation methods to a set of budget data. We used the purchase data of 450 households on three meat products (pork, beef and horse) of the first two quarters of 1994. The results for the Poisson Gamma and the Poisson Spike models are shown in Tables 2 and 3 respectively, but before we present these results we will discuss the estimation scheme of the two methods for the bivariate Poisson Gamma model (see Fig. 1a and Fig. 1b).

Both methods start with estimation of the marginal distributions. From these marginals the attenuation factors are estimated. In the case of the moment estimator the next step is to compute the sample correlation coefficient from the two way table

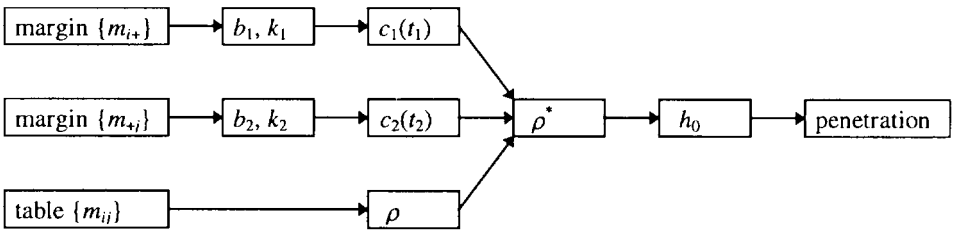

Fig. 1a. Estimation scheme for bivariate moment estimators for the Poisson Gamma model.

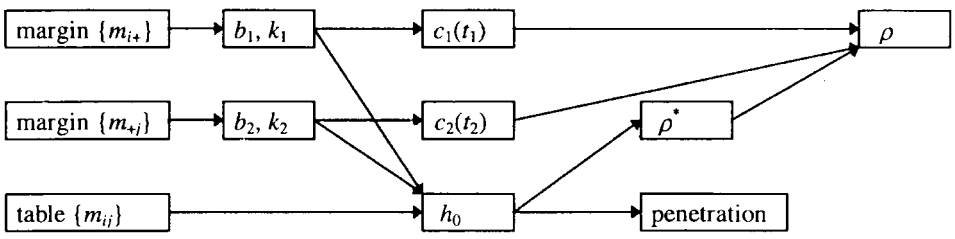

Fig. 1b. Estimation schemes for bivariate ML-estimators for the Poisson Gamma model.

$\{m_{ij}\}$. Together with the estimated attenuation it is possible to give an estimate for the correlation at the latent level ρ^*. This helps us to an estimate for h_0, and finally to an estimate for the penetration. In the case of ML-estimation the scheme is more simple. Based on the two way table $\{m_{ij}\}$ an estimate for h_0 is obtained by maximising the one dimensional conditional likelihood given the estimates of the marginal parameters b_1, k_2, b_2 and k_2. From the parameter estimates we can find an estimate for the penetration. As a by-product we may then derive estimates for the correlations ρ^* and ρ.

In Table 3 the rows labelled *Quarter 1* and *Quarter 2* show parameter ML-estimates of b and k obtained from the univariate Poisson Gamma distribution, and the quarter penetration estimates derived from these parameter estimates. The rows labelled *Half year combined* show the results of combining the data, and estimate a new set of ML-estimates b and k. From this new set of estimates the half year penetration of the products is estimated. Although it does not occur in these examples, it may happen that the estimated half year penetration is lower than one of the quarter penetrations. The rows labelled *Half year bivariate mom.* and *Half year bivariate ML* show results for the bivariate estimators explained in sections 3 and 4 respectively. For the product *pork* an estimate for the penetration by way of a moment estimator based on the correlation in the two way table, has the value 0.9410 (column labelled *penetr.*). This estimate is based on the table correlation ρ that was estimated at 0.8048 (column labelled ρ). Together with the 'given' estimates for b_1 and b_2 the attenuation of the

Table 3. Estimates of parameters, attenuation, penetrations and correlations for three meat products using the univariate *Poisson Gamma* model for Quarter 1, Quarter 2 and the combined Half year, and using two estimation methods for the bivariate Poisson Gamma model for the Half year

	b	k	Attenuat.	Penetr.	h_0	ρ	ρ^*
pork							
Quarter 1	0.6897	0.9465	0.9485	0.8865			
Quarter 2	0.5757	0.9471	0.9392	0.8680			
Half year combined	0.6536	0.9062		0.9271			
Half year bivariate mom.				0.9410	0.8554	0.8048	0.9035
Half year bivariate ML				0.9370	0.9043	0.8509	0.9551
beef							
Quarter 1	0.5212	0.8683	0.9335	0.8315			
Quarter 2	0.4864	0.8070	0.9292	0.7995			
Half year combined	0.5359	0.7791		0.8783			
Half year bivariate mom.				0.8933	0.8205**	0.8503**	0.9802**
Half year bivariate ML				0.8949	0.8070*	0.8362	0.9641
horse							
Quarter 1	0.2848	0.0539	0.8873	0.0801			
Quarter 2	0.1021	0.0949	0.7552	0.0771			
Half year combined	0.1857	0.0664		0.1104			
Half year bivariate mom.				0.1063	0.0728**	0.6826**	1.0186**
Half year bivariate ML				0.1028	0.0539*	0.5050	0.7535

* Estimate of h_0 at upperbound. ** Estimate of h_0 outside domain.

Poisson Gamma process could be estimated, leading to an estimate of ρ^* of 0.9035 using formula (8). As a next step h_0 was estimated at 0.8554 using formula (9), leading finally to an estimate of the half year penetration of 0.9410 using

$$P\{M_1(t_1) = 0, M_2(t_2) = 0\} = \frac{1}{(1 + b_1 t_1 + b_2 t_2)^{h_0}} \frac{1}{(1 + b_1 t_1)^{k_1 - h_0}} \frac{1}{(1 + b_2 t_2)^{k_2 - h_0}} \tag{17}$$

The rows labelled *Half year bivariate ML* show estimates of the half year penetration based on an ML-estimate for h_0, as described in section 4. For pork h_0 was estimated at 0.9043. From h_0 estimates for the penetration and for the correlations ρ^* and ρ were derived.

Table 4 has basically the same structure as Table 3. The parameters to be estimated for the univariate Poisson Spike model are μ and p_0. For the bivariate case both estimation methods discussed in section 3 and 4 are applied to estimate q_{00} and the penetration. The row labelled *Half year bivariate mom.* shows an estimate of the penetration based on an estimate of ρ from the observed table $\{m_{ij}\}$. Note that these estimates are identical to those in Table 3, since these are direct estimates, and do not depend on the model. From ρ we can estimate ρ^*, this time using formula (12). From ρ^* we can derive an estimate for q_{00}, and finally estimate the penetration. The row labelled *Half year bivariate ML* shows the ML-estimate of q_{00}. From this estimate the penetration, ρ and ρ^* are derived.

Table 4. Estimates of parameters, attenuation, penetrations and correlations for three meat products using the univariate *Poisson Spike* model for Quarter 1, Quarter 2 and the combined Half year, and using two estimation methods for the bivariate Poisson Spike model for the Half year

	μ	p_0	Attenuat.	Penetr.	q_{00}	ρ	ρ^*
pork							
Quarter 1	0.7980	0.1821	0.8086	0.8179			
Quarter 2	0.6737	0.1906	0.7908	0.8093			
Half year combined	0.6870	0.1378		0.8622			
Half year bivariate mom.				0.7745	1.2383**	0.8048**	1.2586**
Half year bivariate ML				0.8690	0.7189	0.4058	0.6346
beef							
Quarter 1	0.5741	0.2119	0.7827	0.7876			
Quarter 2	0.5178	0.2419	0.7871	0.7572			
Half year combined	0.5051	0.1733		0.8267			
Half year bivariate mom.				0.7073	1.3816**	0.8503**	1.3801**
Half year bivariate ML				0.8366	0.7704	0.3942	0.6399
horse							
Quarter 1	0.1807	0.9150	0.8261	0.0769			
Quarter 2	0.0862	0.8876	0.7063	0.0758			
Half year combined	0.1098	0.8877		0.1058			
Half year bivariate mom.				0.0840	1.0002**	0.6826**	1.1699**
Half year bivariate ML				0.1087	0.9559	0.4138	0.7092

* Estimate of q_{00} at upperbound. ** Estimate of q_{00} outside domain.

6 Discussion of the results

Tables 3 and 4 show some interesting features that we like to discuss here. A first thing we see is that the estimation method based on the correlation in a two way table may lead to an estimated correlation ρ^* that is larger than 1. This happened in the Poisson Gamma model for the product *horse*, and in the Poisson Spike case for all three products. In these cases estimates of h_0 and q_{00} are out of their domain. Secondly we noticed that although the log-likelihood function for the bivariate Poisson Gamma model appeared to be a concave function, we found for two products (*beef* and *horse*) that the ML-estimate of h_0 is at the upper bound of its domain (being the minimum of the two estimates k_1 and k_2). This does not suggest a good fit for the models. Thirdly we found a large difference in attenuation between the Poisson Gamma and the Poisson Spike process by comparison of Tables 3 and 4 shows. For example, the total attenuation for the product *pork* (estimated by the quotient of ρ/ρ^*) in the PG-model is 0.89, whereas for the PS-model it is 0.64. We expected the attenuation for the two models to be equal, since that would have been the case if moment-estimators were used for the marginal parameters. Apparently using ML-estimates tends to give quite different results.

Where does all this leave us? We may question the Poisson assumption for the purchasing process. For both Poisson models we encountered problems from unmanageable estimates for the attenuation. In the literature Poisson based models are criticised for the fact that they imply the highly irregular exponential distribution for the interpurchase times (see MORRISON and SCHMITTLEIN, 1988). Several other distributions for the interpurchase times were studied that provide more regularity. Examples are the Erlang-2 distribution (CHATFIELD and GOODHARDT, 1973) and the inverse Gaussian distribution (BANERJEE and BHATTACHARYYA, 1976). These approaches, however, still fix the variation coefficient of the process at the individual level to a value that does not necessarily reflect reality. Let us generalise the Poisson Gamma model to the Gamma Gamma model, where we assume that the interpurchase times A have a *gamma* distribution with unknown shape parameter α, and that the heterogeneity variable Λ has a *gamma*(b, k) distribution. The α parameter reflects the variation coefficient of the process at the individual. Although it is numerically extremely hard to solve this model, we may gain some insight from it. The interpurchase times A for a certain individual with $\Lambda = \lambda$ have a *gamma*$(1/\lambda, \alpha)$ distribution, and the purchasing frequency is $f = 1/E[A] = \lambda/\alpha$. Note that this time there is a distinction between the latent variables Λ and f. In a process with given Λ the asymptotical variance for $t \to \infty$ is given by

$$\sigma^2(M(t) \mid \Lambda) = \frac{\sigma^2(A \mid \Lambda)t}{E[A \mid \Lambda]^3} = \frac{\Lambda t}{\alpha^2}$$

(see Cox, 1962) so that

$$E_\Lambda[\sigma^2(M(t) \mid \Lambda)] = \frac{bkt}{\alpha^2}$$

Since

$$\sigma_f^2 = \sigma^2(\Lambda/\alpha) = \frac{b^2 k}{\alpha^2}$$

we find by substitution of these results into formula (8) that the attenuation for the Gamma Gamma model is

$$c(t) = \sqrt{\frac{bt}{1+bt}}$$

This attenuation formula for the Gamma Gamma model is exactly the same as for the Poisson Gamma model in formula (9). However, this result is misleading. If we fix the average purchasing frequency bk/α to 1, we find

$$c(t) = \sqrt{\frac{\alpha t}{k+\alpha t}}$$

This result has significant consequences. It implies that without careful estimation of the regularity parameter α, the time scale that is used for the attenuation $c(t)$ is arbitrary. When α increases, ρ^* decreases. As a result the estimation method can only be used for products for which the Poisson assumption is reasonable.

7 Conclusions

The results of the application of bivariate Poisson models to budget data are not too good. The estimation method based on the correlation of the observed table $\{m_{ij}\}$ leads to estimates that get out of their domain. The maximum likelihood estimation method does not suffer this problem, but in the case of the Poisson Gamma model we may end up with estimates at the edge of the domain. In the case of the Poisson Spike model we find a severe gap between the estimated correlation from the model compared with the observed correlation in the table $\{m_{ij}\}$.

The argument in the previous section, on generalising the Poisson process to a process with Gamma distributed interpurchase times, shows that the Poisson assumption is essential and that the method described here can only be used for processes where the Poisson assumption is reasonable. If one is not sure, it is necessary to take the regularity parameter α into account. Variance component models (see e.g. ZWINDERMAN, VAN HOUWELINGEN and SCHWEITZER, 1995) allow such an approach, and their use to the budget data is an object of study.

In our approach we neglect the effect of covariates and unobserved heterogeneity. This may be an alternative explanation of the bad fit of the Poisson based models. Recently GUPTA (1991) extended Poisson and Erlang-2 models by inclusion of time varying covariates. These variables may improve the fit of these models significantly. Other recent work shows the use of a hazard rate model that includes *marketing mix variables* (time varying covariates) and *unobserved heterogeneity* (JAIN and VILCASSIM,

1991). They use a Box-Cox formulation for the baseline hazard, for which the exponential, Weibull and Erlang-2 distributions for interpurchase times are special cases. They conclude that the flexible Box-Cox formulation is preferred over the exponential, Weibull and Erlang-2 distributions for interpurchase times. Further they found that there is significant unobserved heterogeneity, and that certain covariates influence the interpurchase times significantly. However, such model generalisations come with a certain cost (see MORRISON and SCHMITTLEIN, 1988, and GUPTA, 1991). Estimation of these models becomes very time consuming, the estimates may not be very robust, and prescribes registration of marketing mix variables.

References

BANERJEE, A. K. and G. K. BHATTACHARYYA (1976), A purchase incidence model with Inverse Gaussian interpurchase times, *Journal of the American Statistical Association* **71**, 823–829.

CHATFIELD, C. and G. J. GOODHARDT (1973), A consumer purchasing model with Erlang interpurchase times, *Journal of the American Statistical Association* **68**, 828–835.

COX, D. R. (1962), *Renewal theory*, Methuen, London.

EHRENBERG, A. S. C. (1988), *Repeat-buying: facts, theory and data*, 2nd ed. Oxford University Press, New York.

GUPTA, S. (1991), Stochastic models of interpurchase time with time-dependent covariates, *Journal of Marketing Research* **28**, 1–15.

JAIN, D. C. and N. J. VILCASSIM (1991), Investigating household purchase timing decisions: a conditional hazard function approach, *Marketing Science* **10**, 1–23.

JOHNSON, N. L. and S. KOTZ (1972), *Distributions in statistics: continuous multivariate distributions*, Wiley, New York.

LORD, F. M. and M. R. NOVICK (1967), *Statistical theories of mental test theories*, Addison-Wesley, Reading, Massachusetts.

MORRISON, D. G. and D. C. SCHMITTLEIN (1988), Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort?, *Journal of Business and Economic Statistics* **6**, 145–166.

SARIS, W. E., P. PRASTACOS and C. A. JACOBS (1989), CASIP: A complete automated system for information processing in family budget research, in: *New technologies and techniques for statistics*, proceedings of the conference, Bonn, February 1992, 80–87.

SIKKEL, D. and G. JELIERSE (1987), Retrospective questions and correlations, *Psychometrika* **52**, 251–261.

SIKKEL, D. and A. W. HOOGENDOORN (1995), Models for monthly penetrations with incomplete panel data, *Statistica Neerlandica* **49**, 378–391.

ZWINDERMAN, A. H., J. C. VAN HOUWELINGEN and D. SCHWEITZER (1995), Prediction of the next hypercalcemia free period: application of random effect models with selection on first event, *Statistica Neerlandica* **49**, 310–323.

Received: May 1997. Revised: October 1997.