Measuring more or less: Estimating product period penetrations from incomplete panel data
Hoogendoorn, A.W.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
4. Multivariate Heterogeneous Poisson Processes

4.1 Introduction

In chapter 3 we described four Poisson based models that were used for budget data. The models differ in the way heterogeneity is treated. The first and most complex of these models distinguishes a class of never buyers and allows Gamma heterogeneity for the class of buyers (Poisson Gamma Spike). The other models are all simplifications. The second model allows Gamma heterogeneity but no class of never buyers (Poisson Gamma), the third provides a class of never buyers but no heterogeneity within the class of buyers (Poisson Spike) and the fourth allows no heterogeneity at all (Poisson). The models and their parameters are listed in Table 4.1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>( \lambda ) (intensity)</td>
</tr>
<tr>
<td>Poisson Gamma</td>
<td>( b ) (scale), ( k ) (shape)</td>
</tr>
<tr>
<td>Poisson Spike</td>
<td>( \lambda, \ p_0 ) (probability of being a never buyer)</td>
</tr>
<tr>
<td>Poisson Gamma Spike</td>
<td>( p_0, b, k )</td>
</tr>
</tbody>
</table>

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 strategy, 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 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

\footnote{An adapted version of this chapter will appear in Statistica Neerlandica, 1998, Vol. 52}
year there are more non-buyers than in one or each of the quarters). The penetrations, which are the compliments of the zero-probabilities then, 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. In fact we will find correlated data as a result from unobserved heterogeneity. We must take the dependency between these dimensions into account.

In this chapter we will generalise the existing models to multivariate models. 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 estimating 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.

4.2 Bivariate Poisson processes

In this section we will discuss two bivariate Poisson processes: the bivariate Poisson Gamma process and the bivariate Poisson Spike process. For each model we will formulate two estimation procedures: moment estimation and maximum likelihood estimation. We will describe the models in terms of latent variables for the intensity of the Poisson process.

4.2.1 The bivariate Poisson Gamma process

The Poisson Gamma process is a latent variable model. It assumes the existence of a latent variable $\Lambda$ that indicates the intensity of the individual Poisson processes. The distribution of $\Lambda$ is assumed to be 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 $\Lambda_1$ and $\Lambda_2$ for the first and the second period and with gamma$(b_1,k_1)$ and gamma$(b_2,k_2)$ distributions respectively. We will discuss a bivariate version of the gamma distribution as given in JOHNSON and KOTZ (1972). Therefore we suppose that $\Lambda_1$ and $\Lambda_2$ 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 $\Lambda_1$ and $\Lambda_2$ 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 $h_0 \leq \min(k_1, k_2)$. Since the joint density of $X_0$, $X_1$ and $X_2$ is defined as

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

the joint density of $X_0$, $\Lambda_1$ and $\Lambda_2$ is

$$f(x_0, \Lambda_1, \Lambda_2) = \frac{1}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} x_0^{h_0 - 1} (\Lambda_1 - x_0)^{h_1 - 1} (\Lambda_2 - x_0)^{h_2 - 1} e^{h_0\Lambda_1 + h_2\Lambda_2}$$  \hspace{1cm} (4.1)

By integrating formula (4.1) with respect to $x_0$ from 0 to $\tilde{\lambda}$, where $\tilde{\lambda}$ is the minimum of
$\Lambda_1$ and $\Lambda_2$, we can find the joint density function of $\Lambda_1$ and $\Lambda_2$

$$f(\Lambda_1, \Lambda_2) = \frac{e^{\lambda_1 - \lambda_2}}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} \int_0^{\tilde{\lambda}} x_0^{h_0 - 1} (\Lambda_1 - x_0)^{h_1 - 1} (\Lambda_2 - x_0)^{h_2 - 1} e^{x_0} dx_0$$  \hspace{1cm} (4.2)

The correlation $\rho^*$ between $\Lambda_1$ and $\Lambda_2$ can easily be derived from the definition

$$\rho^* = \frac{h_0}{\sqrt{(h_0 + h_1)(h_0 + h_2)}} = \frac{h_0}{\sqrt{k_1 k_2}}$$  \hspace{1cm} (4.3)

Since we fixed the scale parameters all to unity, the number of free parameters in this
approach is three. However there will be no loss of generality: we can re-scale time in such
a way that the scale parameters are equal to one.
The bivariate Poisson Spike process

The Poisson Spike process is essentially a mixture model at the zero observations. 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 $\mu$ 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 $p_{00}$ and $1 - p_{00}$, then

$$p_{02} = p_{00}q_{00} + (1 - p_{00})q_{10}$$ \hspace{1cm} (4.4)

The number of free parameters in the bivariate Poisson Spike model then has five free parameters. The model can be specified in such a way that four parameters ($\mu_1$, $p_{01}$, $\mu_2$ and $p_{02}$) determine the marginal distributions, and that one remaining parameter (e.g. $q_{00}$) determines the dependency in the model. The marginal distributions put restrictions to the values of $q_{00}$. Apart from the obvious bounds of 0 and 1 the value of $q_{00}$ should also be smaller than $p_{02}/p_{01}$ and should be larger than $(p_{01} + p_{02} - 1)/p_{01}$.

The Poisson Spike process can also be formulated in terms of a latent variable model. In that case we will say that a non-buyer ‘buys’ according to a Poisson process with intensity 0, so that the latent variable $\Lambda$ indicating the intensity of the individual Poisson process is discrete and takes two values: zero (with probability $p_0$) and $\mu$ with (with probability $1 - p_0$). In the bivariate case we have two latent variables $\Lambda_1$ and $\Lambda_2$, representing the intensities of the first and second period respectively. The bivariate latent variable has a discrete distribution in the four points (0,0), $(\mu, 0)$, $(0, \mu_2)$ and $(\mu_1, \mu_2)$ with probabilities $\pi_{00}$, $\pi_{10}$, $\pi_{01}$ and $\pi_{11}$ respectively. Relationships between the probabilities $p_{0i}$ and $q_{ij}$ on the one hand and the probabilities $\pi_{ij}$ on the other can easily be derived.

The correlation $\rho^*$ between $\Lambda_1$ and $\Lambda_2$ is

$$\rho^* = \frac{\pi_{11} - (\pi_{01} + \pi_{11})(\pi_{01} + \pi_{11})}{\sqrt{(\pi_{01} + \pi_{11})(\pi_{00} + \pi_{01})(\pi_{00} + \pi_{11})(\pi_{01} + \pi_{11})}}$$

$$= \frac{1 - p_{01} - p_{02} + p_{00}q_{00} - (1 - p_{01})(1 - p_{02})}{\sqrt{p_{01}(1 - p_{01})p_{02}(1 - p_{02})}}$$ \hspace{1cm} (4.5)
4.3 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 quarter and \( M_2(t_2) \) be the number of purchases of a respondent who reported over a time \( t_2 \) in the second quarter. 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. 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 Gamma model and the bivariate Poisson Spike model. The estimation procedure is based on equating the first two marginal moments and the correlation coefficient of \( M_1(t_1) \) and \( M_2(t_2) \) in the table \( \{m_{ij}\} \) to the theoretical population values. The method can be seen a generalisation of the method of moments (see e.g. HOEL, PORT and STONE, 1971). In order to formulate this estimator we will first discuss the notion of attenuation. The attenuation relates the correlation of the latent variable that indicates the frequency of the individual renewal process to the observable variable \( M(t) \).

4.3.1 Attenuation

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

\[
X = T + E
\]

We assume that the measurement error \( E \) has zero expectation and is uncorrelated with respect to the true score \( T \). The correlation \( \rho(X, T) \) between observed and true score is called the reliability of the measurement, and is equal to the quotient \( \sigma_T^2 / \sigma_X^2 \) of variances of \( T \) and \( X \). As a consequence the reliability takes values between 0 and 1. Suppose that we have two measurement instruments \( X_1 \) and \( X_2 \) that measure the true scores \( T_1 \) and \( T_2 \). 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. 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 correlations between observed scores are less than the correlations between corresponding true scores, because the former correlations are attenuated by the unreliability of the measurements. We can also write

\[ \rho(T_i, T_j) = \frac{\rho(X_i, X_j)}{c_i c_j} \]  

where we call \( c_i \) the attenuation factor of measurement \( i \):

\[ c_i = \sqrt{\rho(X_i, T_i)}, \quad (i = 1, 2). \]

The notion of attenuation can also be 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 \( \Lambda \), 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 (4.6) is

\[ \rho(f_1, f_2) = \frac{\rho(M_1(t_1), M_2(t_2))}{c_1(t_1) c_2(t_2)} \]  

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

\[ c_i(t_i) = \sqrt{\rho(M_i(t_i), f_i)} = \sqrt{\frac{t_i^2 \sigma^2_f}{\text{E}_\Lambda [\sigma^2(M_i(t_i)\Lambda_i)] + t_i^2 \sigma^2_i}} \]  

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).

### 4.3.2 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 $\text{gamma}(b_i,k_i)$ distribution. Note that in this case we have no distinction between $f_i$ and $\Lambda_i$. Conditional on $\Lambda_i$, then $M_i(t_i)$ is a Poisson process, so that $E_{\Lambda_i}[\sigma^2(M_i(t_i)|\Lambda_i)] = E_{\Lambda_i}[\Lambda_i t_i] = b_i k_i t_i$. Since $\sigma^2_{\Lambda_i} = b_i^2 k_i$, it follows from equation (4.8) the attenuation factor $c_i(t_i)$ for the Poisson Gamma model is

$$c_i(t_i) = \frac{b_i t_i}{\sqrt{1 + b_i t_i}} \quad (4.9)$$

Now we can formulate a procedure to find an estimate for $h_0$. Using the table $\{m_e\}$ 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 marginal totals $\{m_e\}$ and $\{m_f\}$ of the table. Based on the table $\{m_e\}$ we can also compute a correlation coefficient $\hat{\rho}$ between $M_1(t_1)$ and $M_2(t_2)$. We can relate this to $\rho^*$, the correlation between $\Lambda_1$ and $\Lambda_2$ by (4.7) using the attenuation factors of equation (4.9), so that $\rho^*$ can be estimated by

$$\hat{\rho}^* = \frac{1 + \hat{b}_1 t_1}{\hat{b}_1 t_1} \frac{1 + \hat{b}_2 t_2}{\hat{b}_2 t_2} \hat{\rho} \quad (4.10)$$

The next step is to transform the marginal distributions such that they become standard gamma distributions ($b_1 = b_2 = 1$). This can be achieved by changing 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 $\{m_e\}$ 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 equation (4.3) we find that the value of $h_0$ can be estimated by

$$\hat{h}_0 = \hat{\rho}^* \sqrt{k_1 k_2} \quad (4.11)$$

4.3.3 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 $/u_i$. Note that there is no distinction between $f_i$ and $\Lambda_i$ again. We find here that $E_{\Lambda_i}[\sigma^2(M_i(t_i)|\Lambda_i)] = E_{\Lambda_i}[\Lambda_i t_i] = \mu_i p_{0i} t_i$ 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
\[ c_i(t^*_i) = \sqrt{\frac{p_{0i} \mu_i t_i}{1 + p_{0i} \mu_i t_i}} \]  

(4.12)

Now 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 \( \mu_1, p_{01}, \mu_2, \) and \( p_{02} \). Next we find an estimate \( \hat{\rho} \) for the correlation in the table \( \{m_{ij}\} \) that we can use to estimate \( \rho^* \), the correlation between the latent intensities by

\[ \hat{\rho}^* = \sqrt{\frac{1 + \hat{p}_{01} \hat{\mu}_1 t_1}{\hat{p}_{02} \hat{\mu}_2 t_2}} \]  

(4.13)

From equation (4.5) it then follows we estimate \( q_{00} \) by

\[ \hat{q}_{00} = \frac{\sqrt{p_{01}(1 - p_{01})p_{02}(1 - p_{02})}}{p_{01} + p_{01}p_{02}} \]  

(4.14)

### 4.4 Maximum likelihood estimation

In this section we will derive maximum likelihood 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 marginal distributions.

#### 4.4.1 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 \( t_i \) into \( u_i = b_i t_i \) and \( u_2 = b_2 t_2 \). 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 equation (4.1) we have

\[ L(h_0; m_1, m_2, u_1, u_2) = \]
\[
= \int_{x_0}^{\infty} \int_{x_1}^{\infty} \int_{x_2}^{\infty} e^{-(\lambda_1 + \lambda_2 + \lambda_3)} \frac{(\lambda_1 u_1)^{m_1}(\lambda_2 u_2)^{m_2}}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} \frac{x_0^{h_0-1}(\lambda_1 - x_0)^{h_1-1}(\lambda_2 - x_0)^{h_2-1}}{m_1! m_2!} \times e^{-(x_0 + x_1 + x_2)} dx_0 dx_1 dx_2
\]

\[
= \frac{u_1^{m_1} u_2^{m_2}}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} \sum_{j=1}^{m_1} \sum_{j=2}^{m_2} \int_{x_0=0}^{\infty} \int_{x_1=0}^{\infty} \int_{x_2=0}^{\infty} e^{-(x_0 + x_1 + x_2)} \frac{x_0^{h_0-1} x_1^{h_1-1} x_2^{h_2-1}}{m_1! m_2!} dx_0 dx_1 dx_2
\]

\[
= \frac{u_1^{m_1} u_2^{m_2}}{\Gamma(h_0)\Gamma(h_1)\Gamma(h_2)} \sum_{j=1}^{m_1} \sum_{j=2}^{m_2} \left( \begin{array}{c} m_1 \\ j \end{array} \right) \left( \begin{array}{c} m_2 \\ j \end{array} \right) \Gamma(h_0 + j_1 + j_2) \Gamma(h_1 + m_1 - j_1) \Gamma(h_2 + m_2 - j_2)
\]

\[
= \sum_{j=1}^{m_1} \sum_{j=2}^{m_2} \frac{u_1^{m_1-j_1} u_2^{m_2-j_2}}{(m_1 - j_1)(u_1 + u_2 + 1)^{h_1 + j_1} (m_2 - j_2)(u_1 + u_2 + 1)^{h_2 + m_2 - j_2}}
\]

where we write \( h^{j+1} = h(h+1) \ldots (h+j-1) \) for \( j > 0 \), and \( h^{10} = 1 \). 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

\[
l(h_0) = \sum_{m_1} \sum_{m_2} \sum_{u_1} \sum_{u_2} n(m_1, m_2, u_1, u_2) \log(L(h_1; m_1, m_2, u_1, u_2))
\]

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.
4.4.2 The bivariate Poisson Spike model

Our situation is that the parameters $\mu_1$, $p_{01}$, $\mu_2$ and $p_{02}$ of the marginal distributions of the bivariate Poisson Spike distribution are given. In order to find the ML estimator for the fifth parameter $q_{10}$ we need expressions for the probability distribution of the outcomes. These can easily be written down, e.g.

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

$$+ (1 - p_{01})\exp(-\mu_1 t_1)(q_{10} + (1-q_{10})\exp(-\mu_2 t_2))$$

(4.16)

Note that $q_{10}$ can be written as a function of $p_{01}$, $p_{02}$ and $q_{00}$ by equation (4.4). 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 $M_2(t_2)$.

4.5 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 and Table 4.2 respectively, but before we present these results we will discuss the estimation scheme of the two methods for the bivariate Poisson Gamma model (see Figure 4.1 and Figure 4.2).

Figure 4.1 Estimation scheme for bivariate moment estimators for the Poisson Gamma model.

Both methods start with estimation of the marginal distributions. From these distributions 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 $\{m_{ij}\}$. Together with the estimated attenuation it is possible to give an estimate for the correlation at the latent level.
This helps us to an estimate for \( h_0 \), and finally to an estimate for the penetration. In the ML case the estimation scheme is simpler. Based on the two-way table \( \{ m_{ij} \} \) an estimate for \( h_0 \) is obtained by maximising the conditional likelihood given the estimates of the marginal parameters \( h_1, k_1, h_2 \) and \( k_2 \). This leads directly to an estimate for the penetration. As a by-product we may then derive estimates for the correlations \( \rho^* \) and \( \rho \).

**Figure 4.2. Estimation schemes for bivariate ML estimators for the Poisson Gamma model.**

**Table 4.1 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.**

<table>
<thead>
<tr>
<th></th>
<th>( b )</th>
<th>( k )</th>
<th>attenuation</th>
<th>penetration</th>
<th>( h_0 )</th>
<th>( \rho )</th>
<th>( \rho^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pork</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 1</td>
<td>0.6897</td>
<td>0.9465</td>
<td>0.9485</td>
<td>0.8865</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 2</td>
<td>0.5757</td>
<td>0.9471</td>
<td>0.9392</td>
<td>0.8680</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year combined</td>
<td>0.6536</td>
<td>0.9062</td>
<td></td>
<td>0.9271</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year bivariate mom.</td>
<td></td>
<td></td>
<td></td>
<td>0.9410</td>
<td>0.8554</td>
<td>0.8048</td>
<td>0.9035</td>
</tr>
<tr>
<td>Half-year bivariate ML</td>
<td></td>
<td></td>
<td></td>
<td>0.9370</td>
<td>0.9043</td>
<td>0.8509</td>
<td>0.9551</td>
</tr>
<tr>
<td><strong>beef</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 1</td>
<td>0.5212</td>
<td>0.8683</td>
<td>0.9335</td>
<td>0.8315</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 2</td>
<td>0.4864</td>
<td>0.8070</td>
<td>0.9292</td>
<td>0.7995</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year combined</td>
<td>0.5359</td>
<td>0.7791</td>
<td></td>
<td>0.8783</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year bivariate mom.</td>
<td></td>
<td></td>
<td></td>
<td>0.8933</td>
<td>0.8205**</td>
<td>0.8503**</td>
<td>0.9802**</td>
</tr>
<tr>
<td>Half-year bivariate ML</td>
<td></td>
<td></td>
<td></td>
<td>0.8949</td>
<td>0.8070*</td>
<td>0.8362</td>
<td>0.9641</td>
</tr>
<tr>
<td><strong>horse</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 1</td>
<td>0.2848</td>
<td>0.0539</td>
<td>0.8873</td>
<td>0.0801</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 2</td>
<td>0.1021</td>
<td>0.0949</td>
<td>0.7552</td>
<td>0.0771</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year combined</td>
<td>0.1857</td>
<td>0.0664</td>
<td></td>
<td>0.1104</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year bivariate mom.</td>
<td></td>
<td></td>
<td></td>
<td>0.1063</td>
<td>0.0728**</td>
<td>0.6826**</td>
<td>1.0186**</td>
</tr>
<tr>
<td>Half-year bivariate ML</td>
<td></td>
<td></td>
<td></td>
<td>0.1028</td>
<td>0.0539*</td>
<td>0.5050</td>
<td>0.7535</td>
</tr>
</tbody>
</table>

*: estimate of \( h_0 \) at upper bound
**: estimate of \( h_0 \) outside domain

In Table 4.1 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 attenuation). This estimate is based on the table correlation \( \rho \) that was estimated at 0.8048 (column labelled penetration). Together with the 'given' estimates for \( b_1 \) and \( b_2 \), the attenuation of the Poisson Gamma process could be estimated, leading to an estimate of \( \rho^* \) of 0.9035 using equation (4.8). As a next step \( h_0 \) was estimated at 0.8554 using equation (4.9), leading finally to an estimate of the half-year penetration of 0.9410 using

\[
P[M_1(t_1) = 0, M_2(t_2) = 0] = 1 \frac{1}{(1 + b_1 t_1 + b_2 t_2)^{h_0}} \frac{1}{(1 + b_1 t_1)^{1-h_0}} \frac{1}{(1 + b_2 t_2)^{1-h_0}}
\]

The rows labelled Half-year bivariate ML show estimates of the half-year penetration based on a ML-estimate for \( h_0 \), as described in section 4.4. For pork \( h_0 \) was estimated at 0.9043. From \( h_0 \) estimates for the penetration and for the correlations \( \rho^* \) and \( \rho \) were derived.

Table 4.2. 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.

<table>
<thead>
<tr>
<th></th>
<th>( \mu )</th>
<th>( \rho_0 )</th>
<th>attenuation</th>
<th>penetration</th>
<th>( q_{00} )</th>
<th>( \rho )</th>
<th>( \rho^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>pork</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 1</td>
<td>0.7980</td>
<td>0.1821</td>
<td>0.8086</td>
<td>0.8179</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 2</td>
<td>0.6737</td>
<td>0.1906</td>
<td>0.7908</td>
<td>0.8093</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year combined</td>
<td>0.6870</td>
<td>0.1378</td>
<td></td>
<td></td>
<td>0.8622</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year bivariate mom.</td>
<td></td>
<td></td>
<td></td>
<td>0.7745</td>
<td>1.2383**</td>
<td>0.8048**</td>
<td>1.2586**</td>
</tr>
<tr>
<td>Half-year bivariate ML</td>
<td></td>
<td></td>
<td></td>
<td>0.8690</td>
<td>0.7189</td>
<td>0.4058</td>
<td>0.6346</td>
</tr>
<tr>
<td>beef</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 1</td>
<td>0.5741</td>
<td>0.2119</td>
<td>0.7827</td>
<td>0.7876</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 2</td>
<td>0.5178</td>
<td>0.2419</td>
<td>0.7871</td>
<td>0.7572</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year combined</td>
<td>0.5051</td>
<td>0.1733</td>
<td></td>
<td>0.8267</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year bivariate mom.</td>
<td></td>
<td></td>
<td></td>
<td>0.7073</td>
<td>1.3816**</td>
<td>0.8503**</td>
<td>1.3801**</td>
</tr>
<tr>
<td>Half-year bivariate ML</td>
<td></td>
<td></td>
<td></td>
<td>0.8366</td>
<td>0.7704</td>
<td>0.3942</td>
<td>0.6399</td>
</tr>
<tr>
<td>horse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 1</td>
<td>0.1807</td>
<td>0.9150</td>
<td>0.8261</td>
<td>0.0769</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarter 2</td>
<td>0.0862</td>
<td>0.8876</td>
<td>0.7063</td>
<td>0.0758</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year combined</td>
<td>0.1098</td>
<td>0.8877</td>
<td></td>
<td>0.1058</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Half-year bivariate mom.</td>
<td></td>
<td></td>
<td></td>
<td>0.0840</td>
<td>1.0002**</td>
<td>0.6826**</td>
<td>1.1699**</td>
</tr>
<tr>
<td>Half-year bivariate ML</td>
<td></td>
<td></td>
<td></td>
<td>0.1087</td>
<td>0.9559</td>
<td>0.4138</td>
<td>0.7092</td>
</tr>
</tbody>
</table>

**: estimate of \( q_{00} \) outside domain
Table 4.2 has basically the same structure as Table 4.1. The parameters to be estimated for the univariate Poisson Spike model are $\mu$ and $\beta_0$. 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 \textit{Half-year bivariate mom.} shows an estimate of the penetration based on an estimate of $\rho$ from the observed table $\{m_{ij}\}$. Note that these estimates are identical to those in Table 4.1, since these are direct estimates, and do not depend on the model. From $\rho$ we can estimate $\rho^*$, this time using equation (4.12). From $\rho^*$ we can derive an estimate for $q_{00}$, and finally estimate the penetration. The row labelled \textit{Half-year bivariate ML} shows the ML-estimate of $q_{00}$. From this estimate the penetration, $\rho^*$ and $\rho$ are derived.

4.6 Discussion of the results

Table 4.1 and Table 4.2 show some interesting features that we like to discuss here. A first thing that occurred, is that the estimation method based on the correlation in a two-way table, may lead to an estimated correlation $\rho^*$ that is larger than 1. This happened in the Poisson Gamma model for the product \textit{horse}, and in the Poisson Spike case for all three products. In that case estimates of $h_0$ and $q_{00}$ are out of their domain and the estimates of penetrations should not be taken seriously. Secondly we noticed that for the bivariate Poisson Gamma model there are two products (\textit{beef} and \textit{horse}) where the ML-estimate of $h_0$ is at the maximum 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, comparison of Table 4.1 and Table 4.2 shows a large difference in attenuation between the Poisson Gamma and the Poisson Spike process. For example: the total attenuation for the product \textit{pork} (estimated by the quotient of $\rho/\rho^*$ ) 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 or moment estimators gives quite different results. Where does this all leave us? We may question the Poisson assumption of the underlying purchasing process. In the case of the Poisson Spike model we encountered problems related to a badly fitting attenuation, leading to overestimation of $\rho^*$ based on an estimate of $\rho$ - the correlation in the observed table $\{m_{ij}\}$ - and to an underestimation of $\rho$ based on ML-estimate of $\rho^*$. But also in the case of the Poisson Gamma model the attenuation may not be reflected by the data. This becomes obvious if we generalise the interpurchase times in the Poisson Gamma model from an exponential distribution to a \textit{gamma} distribution with shape parameter $\alpha$. Since in that case both the interpurchase times $A$ and the heterogeneity variable $\Lambda$ have a \textit{gamma} distribution, this model is called the \textit{gamma-gamma} model. Let us assume that the heterogeneity parameter $\Lambda$ has a \textit{gamma}(b,k) distribution. Then the interpurchase times $A$ for a certain individual with $\Lambda = \lambda$ have a \textit{gamma}(1/$\lambda$, $\alpha$) distribution and the purchasing frequency is $f = 1/E[A] = \lambda/\alpha$. Note that here there is a distinction between the variables.
If we take $\alpha = 1$ we have a Poisson Gamma process or the NBD model. Setting $\alpha = 2$, brings us to the situation of Erlang-2 distributed interpurchase times, i.e. the CNBD model (see section 4.1). In a process with given intensity $\Lambda$ the asymptotic variance for $t \to \infty$ is given by

$$\sigma^2(M(t) | \Lambda) = \frac{\sigma^2(\Lambda | \lambda) t}{E[\lambda | \Lambda]} = \frac{\Lambda t}{\alpha^2}$$

(see Cox, 1962) so that

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

Since

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

we find by substitution of these results into equation (4.8) that the attenuation for the gamma-gamma model

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

Note that the attenuation formula for the gamma-gamma model is exactly the same as for the Poisson Gamma model in equation (4.9). This result, however, is misleading. If we fix the average purchasing frequency $bk/\alpha$ to $1$, then we find

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

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

4.7 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 are 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 use for processes where the Poisson assumption is reasonable. If one is not sure, it is necessary to take the regularity parameter \( \alpha \), the variance within time, 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 the object of study of chapters 5, 6 and 7.