The Gamma-Gamma Model of Monetary Value

Peter S. Fader
www.petefader.com

Bruce G. S. Hardie[†]
www.brucehardie.com

February 2013

1 Introduction

This note presents a detailed derivation of the gamma-gamma "spend" model presented in Fader et al. (2005), along with details of how to estimate the model parameters and use them to predict likely spend per transaction in the future at the customer level.

Our model of spend per transaction is based on the following three general assumptions:

- The monetary value (e.g., \$, £, €) of a customer's given transaction varies randomly around their average transaction value.
- Average transaction values vary across customers but do not vary over time for any given individual.
- The distribution of average transaction values across customers is independent of the transaction process.

For a customer with x transactions, let $z_1, z_2, \ldots, z_x$ denote the value of each transaction. The customer's observed average transaction value by

$$\bar{z} = \sum_{i=1}^{x} z_i / x \, .$$

$\bar{z}$ is an imperfect estimate of their (unobserved) mean transaction value ζ. Our goal is to make inferences about ζ given $\bar{z}$, which we denote as $E(Z \mid \bar{z}, x)$. As a first step, we need to derive the distribution of $\bar{z}$ given x transactions.

2 Model Development

Schmittlein and Peterson (1994) proposed that the z_i be distributed normal. Two problems with this choice of distribution are that i) it is not bounded from below by 0, and ii) it results in a symmetric spend distribution. Since spend data tend to be right skewed, we follow Colombo and Jiang (1999) and use the gamma distribution. More formally,

i) we assume that $z_i \sim \text{gamma}(p, \nu)$, with $E(Z_i \,|\, p, \nu) = \zeta = p/\nu$.
   - given the convolution properties of the gamma, it follows that total spend across x transactions is distributed gamma(px, ν).
   - given the scaling property of the gamma distribution, it follows that $\bar{z} \sim \text{gamma}(px, \nu x)$.

ii) we assume $\nu \sim \text{gamma}(q, \gamma)$.

This results in what we call the gamma-gamma (GG) model of monetary value (or spend per transaction), which is also known as the beta of the second kind (B2).

Deriving $f(\bar{z} \,|\, x)$

Given these assumptions, the distribution of $\bar{z}$ given x is

$$\begin{aligned} f(\bar{z} \,|\, p, q, \gamma; x) &= \int_0^\infty \frac{(\nu x)^{px} \bar{z}^{px-1} e^{-\nu x \bar{z}}}{\Gamma(px)} \frac{\gamma^q \nu^{q-1} e^{-\gamma\nu}}{\Gamma(q)} d\nu \\ &= \frac{\bar{z}^{px-1} x^{px} \gamma^q}{\Gamma(px)\Gamma(q)} \int_0^\infty \nu^{px+q-1} e^{-(\gamma + x\bar{z})\nu} d\nu \\ &= \frac{\Gamma(px+q)}{\Gamma(px)\Gamma(q)} \frac{\bar{z}^{px-1} x^{px} \gamma^q}{(\gamma + x\bar{z})^{px+q}} \qquad (1a) \\ &= \frac{1}{\bar{z} B(px, q)} \left(\frac{\gamma}{\gamma + x\bar{z}}\right)^q \left(\frac{x\bar{z}}{\gamma + x\bar{z}}\right)^{px} . \qquad (1b) \end{aligned}$$

Deriving $f(\zeta)$

We denote a customer's (unobserved) mean transaction value by ζ. Conditional on p and ν, $\zeta = p/\nu$. However, since ν varies across customers according to a gamma(q, γ) distribution, we view the (unobserved) mean transaction value as a random variable Z with realization ζ. We derive its distribution using the change of variables method. For $\zeta = h(\nu)$, we have

$$f_\zeta(\zeta) = \left| \frac{d}{d\zeta} h^{-1}(\zeta) \right| f_\nu(h^{-1}(\zeta)) \,;$$

$\zeta = p/\nu \Rightarrow \nu = p/\zeta$ which in turn implies $d\nu/d\zeta = -p/\zeta^2$. Therefore,

$$f(\zeta \mid p, q, \gamma) = \frac{p}{\zeta^2} \frac{\gamma^q \left(\frac{p}{\zeta}\right)^{q-1} e^{-\gamma \frac{p}{\zeta}}}{\Gamma(q)}$$
$$= \frac{(p\gamma)^q \zeta^{-q-1} e^{-\frac{p\gamma}{\zeta}}}{\Gamma(q)}, \tag{2}$$

which is an inverse-gamma distribution with shape parameter q and scale parameter $p\gamma$. The mean and variance of this distribution are

$$E(Z \mid p, q, \gamma) = \frac{p\gamma}{q-1}, \text{ and} \tag{3}$$

$$\text{var}(Z \mid p, q, \gamma) = \frac{p^2\gamma^2}{(q-1)^2(q-2)}. \tag{4}$$

Deriving $E(Z \mid \bar{z}, x)$

Finally, we wish to make inferences about an individual customer's ζ given $\bar{z}$, which we denote as $E(Z \mid \bar{z}, x)$. By Bayes' theorem,

$$g(\nu \mid p, q, \gamma; \bar{z}, x) = \frac{f(\bar{z} \mid p, \nu; x) g(\nu \mid q, \gamma)}{f(\bar{z} \mid p, q, \gamma; x)}$$
$$= \frac{(\nu x)^{px} \bar{z}^{px-1} e^{-\nu x \bar{z}}}{\Gamma(px)} \frac{\gamma^q \nu^{q-1} e^{-\gamma\nu}}{\Gamma(q)} \Bigg/ \frac{\Gamma(px+q)}{\Gamma(px)\Gamma(q)} \frac{\bar{z}^{px-1} x^{px} \gamma^q}{(\gamma + x\bar{z})^{px+q}}$$
$$= \frac{(\gamma + x\bar{z})^{px+q} \nu^{px+q-1} e^{-\nu(\gamma+x\bar{z})}}{\Gamma(px+q)}.$$

In other words, the posterior distribution of ν is gamma with shape parameter $px + q$ and scale parameter $\gamma + x\bar{z}$. It follows that

$$E(Z \mid p, q, \gamma; \bar{z}, x) = \frac{p(\gamma + x\bar{z})}{px + q - 1}$$
$$= \left(\frac{q-1}{px+q-1}\right) \frac{p\gamma}{q-1} + \left(\frac{px}{px+q-1}\right) \bar{z}. \tag{5}$$

We call this the conditional expectation of Z. We note that this is the weighted average of the population mean, $E(Z)$, and the observed average transaction value, $\bar{z}$. As the number of observations (x) used to compute $\bar{z}$ increases, less weight is placed on the population mean and more weight is placed on the customer's observed average.

3 Empirical Analysis

We now repeat the analysis of the CDNOW data presented in Fader et al. (2005). We assume that an RFM summary of the data has been created — see Fader and Hardie (2008) — and that it is contained in a worksheet titled `RFM summary` in the Excel workbook `gamma-gamma_calibration.xlsx`.

Summarizing the Data

Of the 2357 individuals in the dataset, 946 made at least one repeat purchase in the calibration period. We compute the average spend per transaction (i.e., $\bar{z}$) for each of the 946 individuals and report the basic descriptive statistics in Table 1. We note that the distribution of observed individual means is highly skewed to the right.

	$
Minimum	2.99
25th percentile	15.75
Median	27.50
75th percentile	41.80
Maximum	299.63
Mean	35.08
Std. deviation	30.28
Mode	14.96

Table 1: Summary of average transaction value per customer.

The nonparametric density plot of the observed individual means given in Figure 1 is created in MATLAB using the following commands:

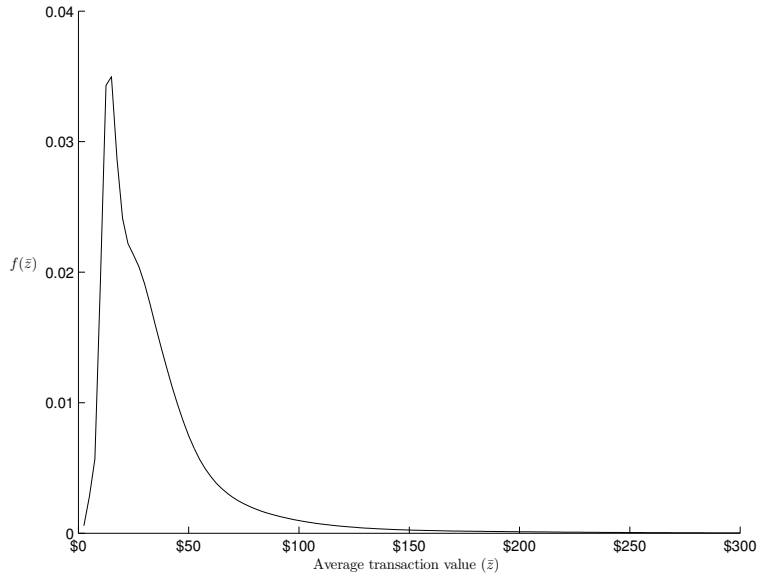

Figure 1: Observed distribution of average transaction values across customers.

i) We read the data into MATLAB,

```
RFM = xlsread('d:\gamma-gamma_calibration','RFM summary');
```

and extract the "frequency" and "monetary value" data (x and $\bar{z}$) for those individuals who made a repeat purchase in the calibration period:

```
use = find(RFM(:,2)>0);
x = RFM(use,2);
zbar = RFM(use,5);
```

ii) We create the probability density estimate of the sample,

```
m = cumsum(2.5*ones(1,120));
f = ksdensity(zbar,m,'support','positive');
```

and the associated plot:

```
plot(m,f,'k-');
xlabel('Average transaction value ($\bar{z}$)','Interpreter','LaTeX');
ylabel('$f(\bar{z})\qquad $','Interpreter','LaTeX','rot',0);
axis([0,300,0,.04]);
set(gca,'XTick',[0 50 100 150 200 250 300]);
set(gca,'XTickLabel',{'$0','$50','$100','$150','$200','$250','$300'});
set(gca,'YTick',[0 0.01 0.02 0.03 0.04]);
set(gca,'YTickLabel',{'0','0.01','0.02','0.03','0.04'});
```

Parameter Estimation

Given the frequency (x_i) and monetary value ($\bar{z}_i$) data for each individual ($i = 1, \ldots, I$), the sample log-likelihood function is simply

$$LL(p, q, \gamma \,|\, \text{data}) = \sum_{i=1}^{I} \ln \left[f(\bar{z}_i \,|\, p, q, \gamma; x_i) \right], \tag{6}$$

the maximum of which can be found using standard numerical optimization methods.

It is a simple exercise to "code up" (6) in Excel. To illustrate this, consider the worksheet `Calibration - zbar`:

- We first note that this worksheet contains a copy of the data presented in the worksheet `RFM summary`, and that the records associated with those customers who made no repeat purchases in the calibration period (i.e., $x_i = 0$) have not been deleted.

- Given the parameter values located in cells `B1:B3`, and the data in cells `B7:C7`, we compute in cell `D7` the log of $f(\bar{z} \,|\, p, q, \gamma; x)$ using the expression given in (1a). (For those customers who made no repeat purchases in the calibration period, we return the value of 0.) This is copied down to the end of the dataset, row 2363.

- The sum of cells `D7:D2363` is located in cell `B4` and is the value of the log-likelihood function for the parameter values located in cells `B1:B3`.

We use the Excel Solver add-in to find the values of p, q, and γ (cells `B1:B3`) that maximize the value of the log-likelihood function (cell `B4`), subject to the constraint that cells `B1:B3` are $\geq$ a small positive number (e.g., 0.0001). We find that the maximum likelihood estimates of the model parameters are $\hat{p} = 6.25$, $\hat{q} = 3.74$, and $\hat{\gamma} = 15.44$.

Estimated Distribution of Average Spend per Transaction

Given the parameter estimates and (2), we can plot the distribution of the (unobserved) mean transaction value ζ across individuals — see the worksheet `f(zeta)` and Figure 2.

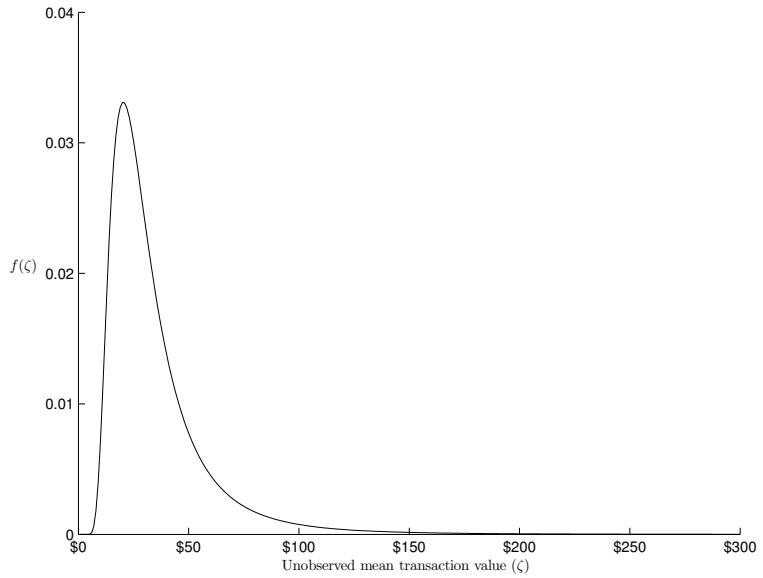

Figure 2: Distribution of the (unobserved) mean transaction value (ζ).

While knowledge of this distribution of the underlying mean spend per transaction is interesting, Figure 2 cannot be be compared to Figure 1 so as to get a sense of how well the model fits the data. Figure 1 is the distribution of average spend per (repeat) transaction across the 946 individuals who

made a repeat transaction in the calibration period. Each customer's average is computed across a (typically very) small number of transactions. Figure 2, on the other hand, is effectively the distribution where the means have been computed across $x \to \infty$ transactions.

In order to get a sense of how well the model fits the data, we need to compare the model-based distribution of $\bar{z}$ with the empirical distribution (Figure 1). Since the distribution of $\bar{z}$ is conditional on x, we compute a weighted average of (1),

$$\sum_{x=1}^{\max(x)} f(\bar{z} \,|\, p, q, \gamma; x) f_x / I \,,$$

where, for a sample of size I, f_x is the number of customers who made x repeat transactions in the model calibration period.

We first determine the distribution of repeat transactions in the calibration period (`Pivot Table`) and copy the data into cells `D1:Y2` of the worksheet `f(zbar)`. In column `D` we compute $f(\bar{z})$, $\bar{z} = 1, 2, \ldots, 300$ for the case of $x = 1$. We copy this across to column `Y` to evaluate this function for all observed values of x.[1] For each value of $\bar{z}$, we compute in column `B` the weighted average of the $f(\bar{z} \,|\, x)$ contained in columns `D-Y`.

We compare in Figure 3 this model-based distribution of $\bar{z}$ with the distribution of the (observed) average spend per transaction (as contained in Figure 1). As Fader et al. (2005, p. 421) note, "[t]he fit is reasonable. However, the theoretical mode of \$19 is greater than the observed mode of \$15, which corresponds to the typical price of a CD at the time the data were collected. The model is not designed to recognize the existence of threshold price points (e.g., prices ending in .99), so this mismatch is not surprising."

This plot is created in MATLAB using the following code. We first create the predicted distribution,

```
% *** compute the density of average transaction value ***
p=6.25;
q=3.74;
gam=15.44;

% how many people with each level of x?
nx = [];
for i=1:max(x),
   nx(i)=length(find(x==i));
end
```

[1] Note that while we use (1a) for model calibration (`Calibration - zbar`), we use (1b) for this calculation. When evaluating (1a) for large values of $\bar{z}$ and x, the result exceeds the largest allowed positive number; this is not a problem when using (1b) as the base of the exponentiation is less than one. We can use (1a) for model calibration as we are evaluating its logarithm.

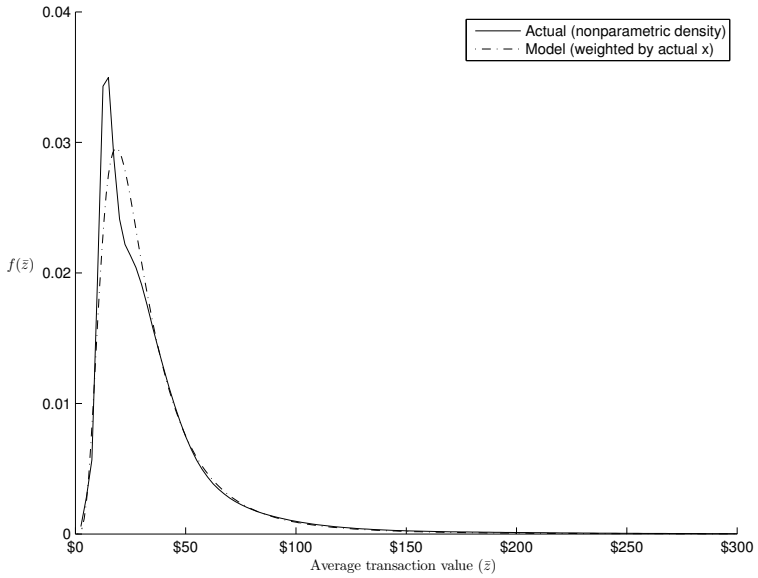

Figure 3: Observed versus theoretical distribution of average transaction value across customers.

```
% compute the density of zbar for each x
y = cumsum(ones(1,300))';
for i=1:max(x),
   a1 = gammaln(p*i+q)-gammaln(p*i)-gammaln(q);
   a2 = q*log(gam);
   a3 = (p*i-1)*log(y);
   a4 = (p*i)*log(i);
   a5 = (p*i+q)*log(gam+y*i);
   g1(:,i) = exp(a1+a2+a3+a4-a5);
end

% compute the weighted average
g = nx*g1'./sum(nx);
```

and then overlay the two plots:

```
hold on
plot(m,f,'k-');
plot(y,g,'k-.');
legend('Actual (nonparametric density)','Model (weighted by actual x)');
hold off
xlabel('Average transaction value ($\bar{z}$)','Interpreter','LaTeX');
ylabel('$f(\bar{z})\qquad $','Interpreter','LaTeX','rot',0);
axis([0,300,0,.04]);
set(gca,'XTick',[0 50 100 150 200 250 300]);
set(gca,'XTickLabel',{'$0','$50','$100','$150','$200','$250','$300'});
set(gca,'YTick',[0 0.01 0.02 0.03 0.04]);
set(gca,'YTickLabel',{'0','0.01','0.02','0.03','0.04'});
```

Computing Conditional Expectations

Our primary reason for utilizing a model of monetary value is to enable us to make inferences about a customer's (unobserved) mean transaction value ζ given $\bar{z}$, which we denote as $E(Z \mid \bar{z}, x)$. As noted in our comment on (5), this can be written as a weighted average of the population mean, $E(Z)$, and the observed average transaction value, $\bar{z}$.

We compute this conditional expectation in the worksheet `Conditional expectation` by first computing the weight (cells `D6:D2362`) placed on the population mean (cell `D1`). Given this weight, the conditional expectation is computed in column `E`.

References

Colombo, Richard and Weina Jiang (1999), "A Stochastic RFM Model," *Journal of Interactive Marketing*, **13** (Summer), 2–12.

Fader, Peter S. and Bruce G. S. Hardie (2008), "Creating an RFM Summary Using Excel." <`http://brucehardie.com/notes/022/`>

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005), "RFM and CLV: Using Iso-value Curves for Customer Base Analysis," *Journal of Marketing Research*, **42** (November), 415–430.

Schmittlein, David C. and Robert A. Peterson (1994), "Customer Base Analysis: An Industrial Purchase Process Application," *Marketing Science*, **13** (Winter), 41–67.