Using a copula-based model of GST data to visualise the New Zealand economy

R. T. R. Vale
Modelling Team
Inland Revenue Department
New Zealand
richard.vale@ird.govt.nz

Abstract

Each New Zealand business reports its expenses and sales income when it files a GST (Goods and Services Tax) return. Economic forces cause the joint distribution of expenses and sales to change over time. The purpose of this work is to visualise these changes.

A six-parameter model of the joint distribution of GST expenses and sales is constructed using a copula. We explain this model and an associated visualisation. This visualisation allows us to track changes in the economy over the last twelve years and identify key events such as the Global Financial Crisis. It also enables us to see a clear business cycle, with the distribution of expenses and sales returning to an earlier state after some time.

This analysis shows that we can get new insights by considering the shape of the joint distribution of expenses and sales, rather than just using the aggregate amount of expenses and sales themselves.

1 Introduction

Goods and Services Tax (GST) is a tax on most goods and services in New Zealand. All New Zealand businesses with an annual turnover (sales and income) exceeding $60,000 are required to register for GST and to file a GST return. Many smaller businesses also choose to be registered for GST. The GST return must be filed on either a monthly, two-monthly, or six-monthly basis. Usually only small businesses are allowed to file on a six-monthly basis. Every business must declare its sales and income and its purchases and expenses on its GST returns. These returns are therefore a source of information about how much money is being made by New Zealand businesses and are one of the sources of information used by Statistics New Zealand to calculate the Gross Domestic Product [6].

In this paper, we refer to sales and income as *sales* and purchases and expenses as *expenses*. Since each business must declare both of these on its GST return, we can consider the joint distribution of sales and expenses for all GST filers in a particular month. The aim of this paper is to perform an exploratory analysis of this joint distribution and visualise how it varies from month to month.

2 The Joint Distribution of Sales and Expenses

Figure 1 shows two scatterplots of log(sales) versus log(expenses) for March 2011, where logs have been taken to base 10. Businesses with sales or expenses under $1 have been omitted from the plot. The bulk of the data form a teardrop shape with pronounced upper tail dependency.

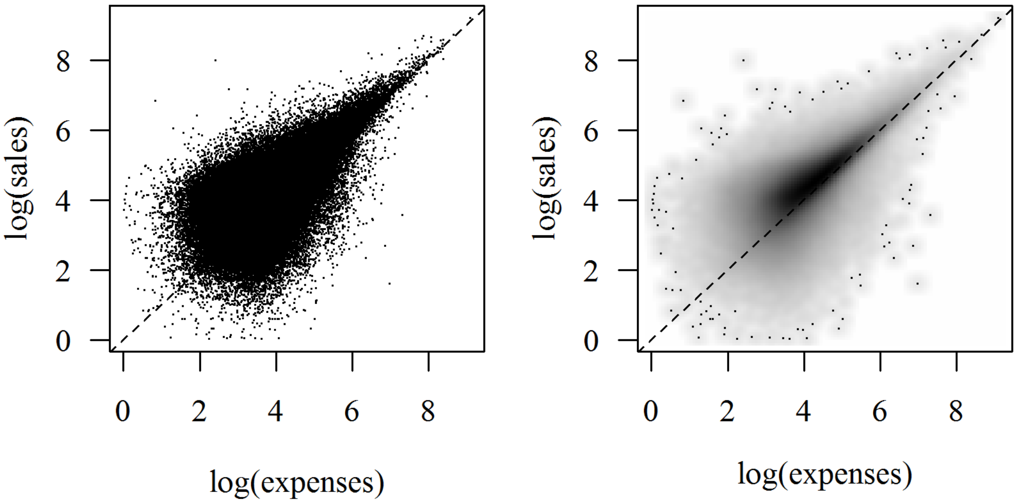

Figure 1: Scatterplot (left) and smoothed scatterplot (right) of log(sales) and log(expenses) for March 2011 GST returns. The dotted line is $y = x$.

The smoothed scatterplot in the right-hand half of Figure 1, produced with the `smoothScatter` function in the statistical package R [3], reveals that the bulk of the probability mass lies above the line $y = x$, which is to be expected since most businesses make a profit.

The goal of modelling the joint distribution is dimension reduction rather than prediction. We wish to find a distribution with a small number of parameters which provides an adequate fit to the shape of Figure 1 and then to examine the time series of the fitted parameters.

2.1 Copulas

Copulas are a standard tool for modelling joint distributions. A bivariate copula is simply a distribution function on the unit square whose marginals are uniform. *Sklar's Theorem* [5] states that every joint distribution function $F(x, y)$ can be written as $F(x, y) = C(F_X(x), F_Y(y))$ where F_X and F_Y are the marginal distribution functions of F and C is a copula. Consequently, a common way of fitting a joint distribution to a random vector (X, Y) is to first fit marginal distributions $\widehat{F}_X$ and $\widehat{F}_Y$ and then to fit a copula to $(\widehat{F}_X(X), \widehat{F}_Y(Y))$.

There is not always much to be gained by using copulas, but many families of copulas can be fitted by software, making them a convenient choice for data analysis.

Figure 2 shows kernel density estimates for the marginal distributions of the data in Figure 1. These look roughly normal, although they are leptokurtic (meaning that they have higher peaks and heavier tails than the normal.) We choose to model them as normal, so that sales and expenses follow a lognormal distribution, and we use a copula to describe their dependence structure.

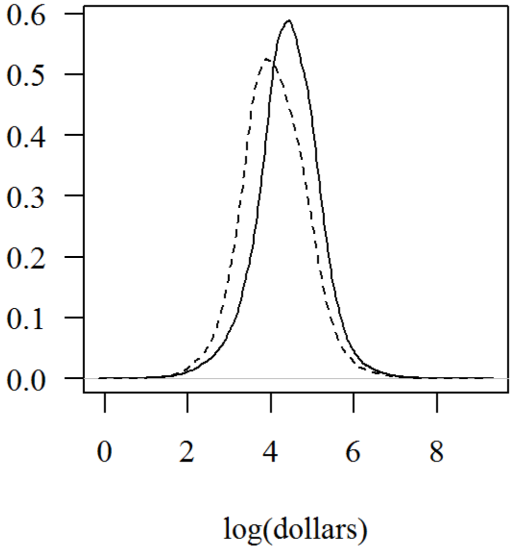

Figure 2: Marginal distributions of log(sales) (solid) and log(expenses) (dotted) for March 2011 GST returns.

A suitable copula is the *Joe copula* of H. Joe [2] fitted using the `VineCopula` R package [4]. The Joe copula depends on a parameter $\theta \geq 1$ and has the distribution function

$$C(x, y) = \varphi^{-1}(\varphi(x) + \varphi(y)), \qquad 0 < x, y < 1$$

where

$$\varphi(t) = -\log(1 - (1 - t)^\theta). \tag{1}$$

For our purposes, this copula has to be flipped through 180°.

The Joe copula is symmetric about the line $y = x$, but we have seen in Figure 1 that the distribution of sales and expenses is not symmetric. The model therefore includes an additional parameter π which is the proportion of filers whose sales exceed their expenses. We now describe the model in full.

2.2 The Model

The model has six parameters.

- μ_e, μ_s, the mean log-expenses and log-sales respectively.

- σ_e, σ_s, the standard deviation of the log-expenses and log-sales.

- θ, the parameter of the fitted Joe copula.
- π, the proportion of filers whose sales exceed their expenses.

The model is fitted in two stages. First, the parameters $\mu_e, \sigma_e, \mu_s, \sigma_s$ and π are estimated from the data. Then the fitted marginal distribution functions are used to transform the data to the unit square, as described in Section 2.1, and a Joe copula is fitted to the result using maximum likelihood estimation.

2.3 Quality of the Fit

Figure 3 shows smoothed scatterplots of the March 2011 data together with an equal number of simulated points from the fitted model. The fitted distribution is the right shape and has the bulk of its probability mass in the right place, but it fails to capture the variability seen in the real data. Various methods exist for evaluating the quality of the fit, including two-dimensional histograms and plots of the lambda function [1, Section 3]. The performance of the fitting procedure itself can be assessed by simulating from the model and then fitting the model to the simulated data, to see whether the parameters used for simulation can be recovered. This was done 1000 times for the model with the following parameters (which are typical values for the GST expenses/sales data)

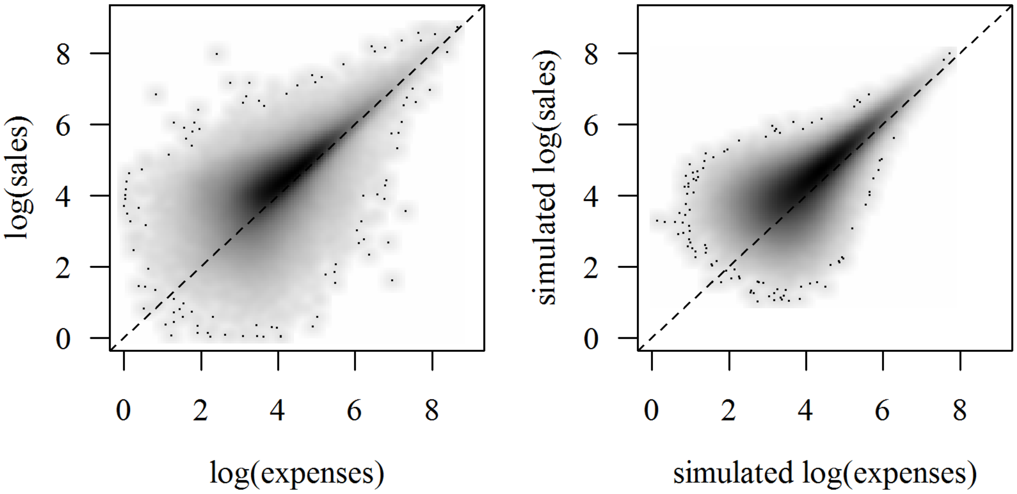

Figure 3: Smoothed scatterplot of March 2011 GST data (left) and 380610 points simulated from the fitted model (right). The dotted line is $y = x$.

$$(\mu_e, \sigma_e, \mu_s, \sigma_s, \theta, \pi) = (4.2, 0.87, 4.5, 0.85, 3.9, 0.78) \qquad (2)$$

and the resulting relative errors are plotted in Figure 4. The most severe bias is in the estimate of θ. Experiments with different values in (2) indicate that the estimate of θ tends to be biased upwards by about +3%. This is not surprising as there is no reason why the fitting procedure of Section 2.2 would yield an unbiased estimator. Although the issues with the model fit mean that this model is unlikely to be useful for prediction, they do not imply that the model is of no use as a dimension reduction technique.

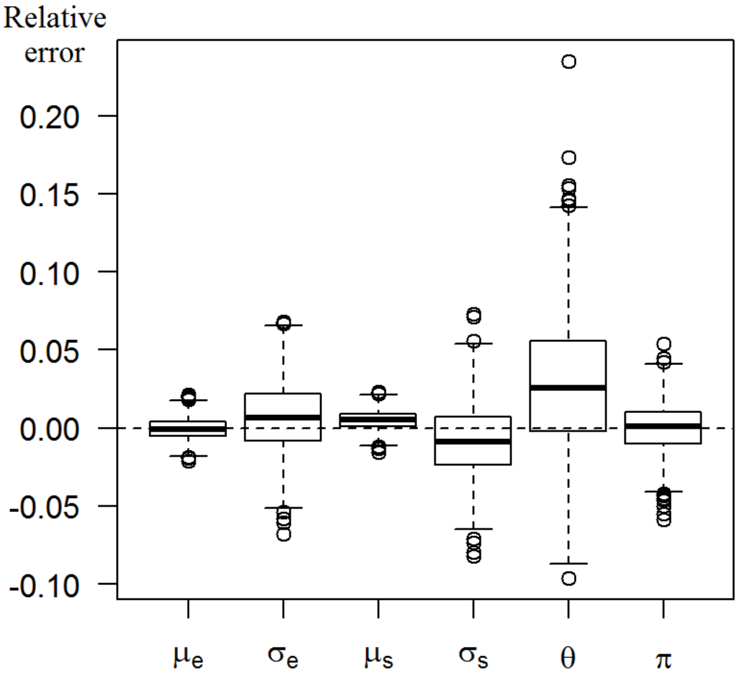

Figure 4: Boxplots of relative errors in 1000 estimates of the model parameters using data simulated from the model with parameters given by (2).

2.4 Interpretation of θ

It is useful to consider the meaning of the parameter θ as it is the only parameter whose interpretation is not obvious. Recall that $\theta \geq 1$. When $\theta = 1$, the Joe copula (1) reduces to the *independence copula* $C(x, y) = xy$, and there is no relationship between expenses and sales. As $\theta \to \infty$, the distribution becomes more and more concentrated along the line $y = x$. We can therefore view θ as a generalised measure of correlation between sales and expenses. Indeed, by [1, (3)], there is a one-to-one correspondence between θ and *Kendall's tau*, which is a measure of correlation based on the ranks of the observations in two data sets.

When communicating with stakeholders, we refer to θ as the *shape* of the distribution.

3 Visualisation

The model was fitted to the GST data from each month from January 2000 to July 2013, giving time series for the six parameters. Since μ_e and μ_s increase over time, these were adjusted for inflation by subtracting the log of the Consumer Price Index. The six time series exhibit very strong seasonality due to the various filing frequencies, as explained in Section 1. This seasonality was removed by taking a twelve-point moving average.

In order to visualise the resulting six-dimensional time series, principal component analysis was used to choose a suitable three-dimensional projection. The first two principal components can be plotted, with shading used to indicate the third principal component. Together, these components account for 96% of the variance in the fitted parameters. The first two principal components account for 83% of the variance.

The resulting visualisation is shown in Figure 5. The most obvious feature of the plot is the vertical movement of the line between mid-2008 and mid-2009, which roughly corresponds to the recession caused by the Global Financial Crisis of 2007-2008. It is appealing that the line "turns the corner" in mid-2009. Notice that the line crosses over itself in early 2012 (although the third principal component has changed) which might be interpreted as the beginning of a new business cycle.

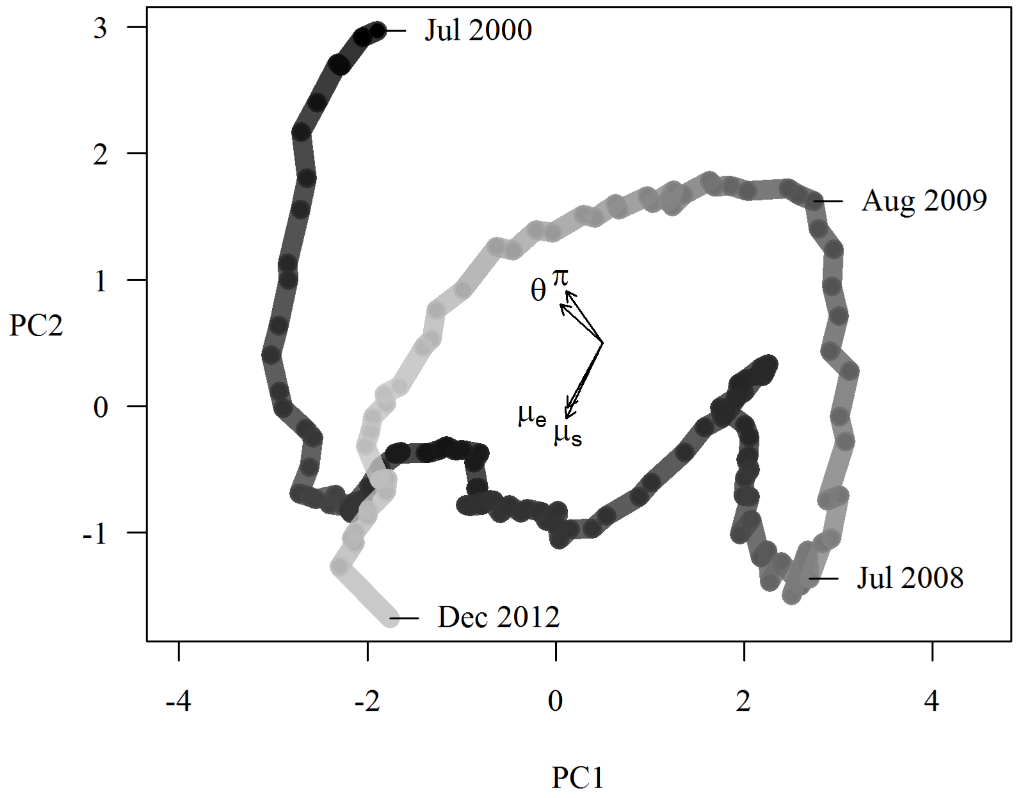

Figure 5: Principal components of the time series of the fitted model parameters, found using the R `prcomp` command. Some dates are indicated on the plot. The darkness of the line represents the third principal component. The arrows indicate directions in which μ_s, μ_e, θ and π are *increasing.*

To present the model to stakeholders, an animated and coloured version of Figure 5 was created using the html5 `<canvas>` element. JavaScript controls enable the viewer to pause and reverse the animation and to change the speed.

4 Conclusion

Even using a simple model, an interesting picture of the New Zealand economy can be obtained. The key idea is to introduce a statistic θ which describes the shape of the expenses/sales distribution. Identifying such statistics can add to our understanding and allow us to visualise existing data sources in new ways.

Acknowledgments

The author thanks Dr. John Holt for suggesting the problem and for reading a draft of the paper, and Dr. Michael Duggan for valuable comments.

References

[1] C. Genest, L.-P. Rivest, 1993. Statistical inference procedures for bivariate Archimedean copulas. J. Amer. Statist. Assoc. 88, 10341043.

[2] H. Joe, *Multivariate Models and Dependence Concepts*, London: Chapman and Hall, 1997.

[3] R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL www.R-project.org.

[4] U. Schepsmeier, J. Stoeber, and E. C. Brechmann, (2013). VineCopula: Statistical inference of vine copulas. R package version 1.1-1. http://CRAN.R-project.org/package=VineCopula

[5] A. Sklar, (1959), Fonctions de repartition à n dimensions et leurs marges, *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229-231

[6] Statistics New Zealand (2013) Quarterly Gross Domestic Product: Series and Methods (third edition) Available from www.stats.govt.nz. Retrieved 31 October 2013.