

THE APPLICATION OF NEURAL NETWORK BASED  
METHODS TO THE EXTRACTION OF KNOWLEDGE FROM  
ACCOUNTING REPORTS

Duarte Trigueiros

INESC,  
R. Alves Redol 9 2, 1000 Lisbon, Portugal  
E-Mail: dmt@sara.inesc.pt

July the 4<sup>th</sup>, 1991

© Copyright 1991  
by  
Duarte Trigueiros

This report has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that no information derived therefrom may be published without the author's prior written consent.

# Introduction

In this study we develop a new approach to the problem of extracting meaningful information from samples of accounting reports. We show that Neural Network-like algorithms are capable of implementing this approach. Such tools are able to automatically build optimal structures similar to financial ratios.

Some results are presented. They suggest that this approach effectively avoids the search of appropriate ratios by the analyst along with some other major drawbacks of the multivariate statistical modelling techniques used in Accountancy. The organization of the Neural Network models also outlines internal features of accounting data, providing new insights into the relative importance of variables for modelling a particular relation.

**Accounting statistical models:** Accounting reports are an important source of information for managers, investors and financial analysts. Statistical techniques have often been used to extract information from databases where accounting reports and related outcomes are gathered. The goal is to construct models suitable for prediction or for isolating the main features of the firm.

An early model is that of Beaver [5] who used ratios of accounting variables to predict financial distress. Many other researchers followed him, mainly using more sophisticated statistical tools (see [2]). Other examples of accounting models are the prediction of bond ratings [16] and the relationship between market and accounting risk [6].

The procedures used to obtain these models are quite similar. The first stage consists of forming a set of ratios from selected items on an accounting report. This selection is typically made in accordance with the beliefs and expectations of researchers. Next, the normality of these ratio variables is discussed and transformations are applied. Finally some linear modelling technique is used to find optimal parameters in the Least Squares sense. Linear Regressions and Fisher's Multiple Discriminant Analysis are the most popular algorithms. However Logistic Regression can also be found in some studies. Foster [11] offers a review of accounting modelling practice.

All such models use ratios as predictors. The use of ratios as input variables in accounting models seems to be an extrapolation of their normal use in financial analysis. Ratios are supposed to capture in a simple and standard way interesting features of the firm. However, there are difficulties involved in using ratios. As  $M$  meaningful accounting variables can generate up to  $M^2 - M$  ratios, some research seems to get lost in a prolific use of all sorts of combinations of variables. It is easy to find in the accounting literature models with forty and more predictors. Moreover, the cross-sectional distribution of ratios seems to exhibit a non-regular behaviour. Horrigan [15] (1965) is an early work on this subject. He reports positive skewness on ratios, explaining it as a result of effective lower limits of zero for these variables. Other studies followed, [21] considering skewness as an accident and implicitly suggesting the pruning or winsorizing of distributions. Deakin [9] (1976) reacted by showing that the positive skewness shouldn't be ignored and Buijink [8] reported the persistency of this feature over a large period. Barnes [3] (1982) suggested that skewness on ratios could be the result of deviations from strict proportionality between the numerator and the denominator. Frecka [12] (1983) tried to achieve normality by pruning, proposing such procedure as the standard way of dealing with the problem of deviations from normality.

Following the literature on Ratio Analysis, accounting statistical models try to obtain improvements in normality by empirically pruning out tails and imposing transformations which are not always the most appropriate ones. The model after pruning, centering, scaling and rotating becomes difficult to interpret. The entire routine tends to a broad empiricism.

**Contents:** In the first chapter of this study we question the use of ratios as input variables for statistical modelling in Accountancy. We show that financial ratios are a particular case of more general descriptors. In chapter 2 we show that Neural Network optimization is consistent with such descriptors: In section 2.2 we compare our method with the one typical in accounting literature, using a well-known problem of classification. The improvements obtained in the interpretability of the resulting model by means of non-standard training procedures are discussed in section 2.3.

**Acknowledgement:** This study has been carried out in the University of East Anglia, in the U.K. We thank Professor Robert H. Berry for his fruitful cooperation in this research. We also wish to acknowledge the guidance and continued support of Professor Luis B. Almeida at INESC, Portugal.

# Contents

<b>Introduction</b>	<b>iii</b>
<b>1 The Statistical Characterization of Accounting Data</b>	<b>1</b>
1.1 Ratios and Lognormality . . . . .	4
1.1.1 Financial Ratios . . . . .	6
1.1.2 Ratios With More Than Two Items . . . . .	9
1.2 Extending the Notion of Financial Ratio . . . . .	10
1.2.1 Non-Linear Proportionality . . . . .	10
1.2.2 Non-Proportional Ratios . . . . .	12
1.2.3 Other Non-Linear Relations Between Items . . . . .	16
1.3 Discussion and Conclusions . . . . .	17
<b>2 Knowledge Acquisition Using the Multi-Layer Perceptron</b>	<b>18</b>
2.1 Ratios as Internal Representations . . . . .	19
2.2 Learning to Discriminate Industrial Groups . . . . .	22
2.3 Improving Generalisation and Interpretability . . . . .	23
2.3.1 Generalisation: Using the Test Set . . . . .	24
2.3.2 Random Penalization of Small Weights . . . . .	25
2.3.3 Post - Processing of Outputs . . . . .	28
2.4 Discussion and Conclusions . . . . .	30
<b>A Classification Results Using the Multi-Layer Perceptron</b>	<b>32</b>
A.1 Results: The Usual Technique . . . . .	32
A.2 MLP With 8 Factors as Input Variables . . . . .	36
A.3 MLP and MDA With Eight Log Items . . . . .	37
A.4 Using the Devised Set of Ratios With MDA . . . . .	38
A.5 Conclusions . . . . .	38

## Chapter 1

# The Statistical Characterization of Accounting Data

So far ratios have been used as input variables for statistical modelling in Accountancy. In this chapter we question their use. Ratios cannot account for non-proportional and non-linear features. On the other hand, the lognormality observed in items suggests the use of multiplicative or proportional models of which ratios are the simplest example.

In what extent is the lognormal nature of items compatible with non-proportional and non-linear relations between them? The development of new models rely on the ability to answer this question. Therefore, the first task we undertake in this chapter is the answering of the above question as a necessary step towards the building of appropriate tools. We first recall a known mechanism to generate the probability distribution observed in cross-sections. Then we study the extent in which financial ratios and multi-variate relations are consistent with such a mechanism. Finally we introduce on it conditions leading to non-proportional and non-linear relations.

We show that there is no contradiction between proportional mechanisms and a class of non-proportional relations. Financial ratios emerge as a particular case of more general descriptors. They can be extended so as to include non-proportionality.

**Empirical evidence:** Observation suggests that cross-sectional samples of accounting items are lognormal. McLeay [19] observed lognormality in large samples of accounting items which are sums of similar transactions with the same sign like Sales, Creditors or Current Assets. Along with the items already studied by McLeay, we found that lognormality cannot be rejected also for stock variables like Fixed and Total Assets or Net Worth and non-accounting items related to size like the number of employees. Positive values of

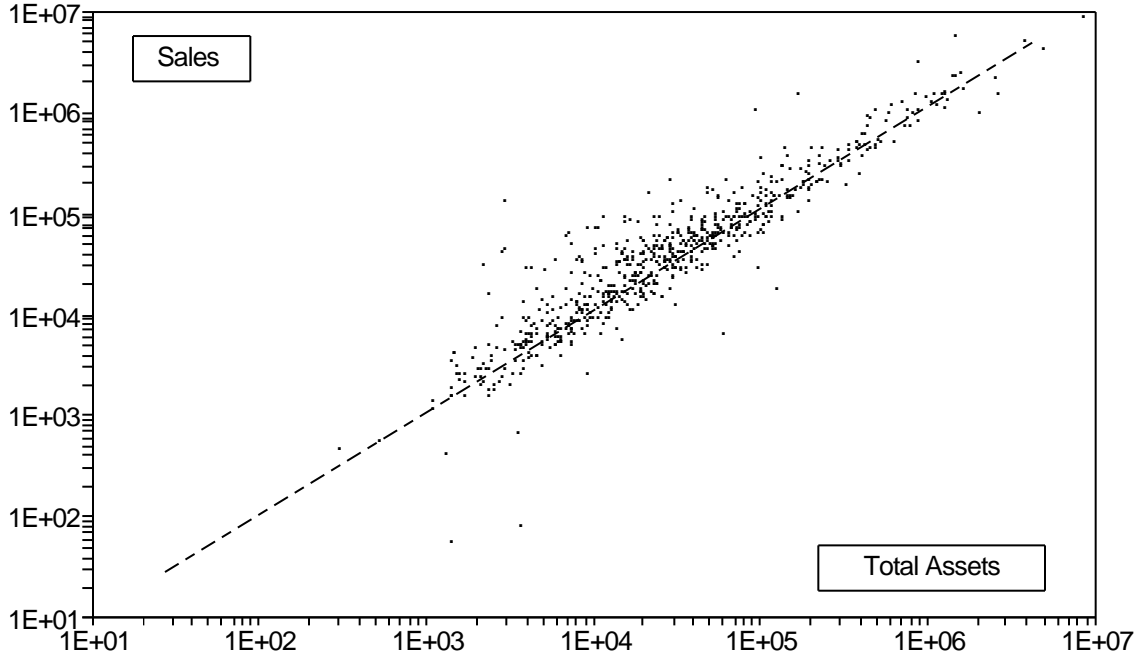


Figure 1: The relation between Sales and Total Assets in log space. All groups together 1984. The dashed line is the axis  $y = x$ .

accounting items having both positive and negative cases, like Working Capital, Earnings, Gross Funds from Operations and the absolute value of the negative cases of these items, are lognormal as well.

We also gathered evidence on the lognormality of small, homogeneous, samples. We examined 18 accounting items for a period of five years (1983-87) belonging to 14 industry groups in the U.K. We concluded that lognormality — either two or three-parametric, see [1] — is a general quality associated with the behaviour of the observed items.

Along with lognormality, it is easy to show that accounting items belonging to the same report share most of its variability. In cross-section, they can be described as a unique, common, process with some particular variability superimposed.

The results referred to here are presented and discussed in [30].

**Empirical models to explain lognormality:** The observed items are lognormal. How far can we go in the building of appropriate models for accounting relations just by considering this empirical finding?

The first consequence of our study is that, once we move into logarithmic space observations can be explained as the mean of the transformed variable plus a deviation from that mean. That is, any lognormal item,  $x$ , is described in log space as an expected value,  $\mu_x$ ,

plus a residual,  $e^x$ :

$$\text{For the } j^{\text{th}} \text{ firm in a sample, } \log x_j = \mu_x + e_j^x$$

When we deflate one item with the median of the same item for that industry the residuals are Gaussian in log space. Lev and Sunder [18] discuss appropriate estimators for the central trend of several possible distributions. Amongst others, the median is also analysed.

Since  $\overline{\log x}$ , the mean value of  $\log x$ , is a good estimator of  $\mu_x$ ,  $\exp(\overline{\log x})$  will be a good estimator of the median of  $x$ . Then,

$$\text{for the } j^{\text{th}} \text{ firm in a sample, the quotient } \frac{x_j}{\exp(\overline{\log x})} = \exp(e_j^x) \quad (1)$$

will reflect the number of times the case  $x_j$  is larger or smaller than the standard for its industry. If  $x_j$  is Sales of firm  $j$ , such quotients would reflect  $j$ 's relative position and relative progress. In general these position quotients seem not to be especially useful in accountancy. They measure size instead of controlling for it.

**Financial ratios:** Since our items are lognormal, financial ratios  $y/x$  can be written in log space as a difference of two position quotients defined in (1):

$$\text{For the } j^{\text{th}} \text{ firm in a sample, } \log(y_j) - \log(x_j) = (\mu_y - \mu_x) + (e^y - e^x)_j \quad (2)$$

This expression is obtained just by subtracting two log items. It is similar to

$$\frac{y_j}{x_j} = R \times f_j \quad \text{with} \quad R = \exp(\mu_y - \mu_x) \quad \text{and} \quad f_j = \exp(e^y - e^x)_j.$$

Here, we arbitrarily used natural logarithms.  $R$  will be the expected proportion in which  $y$  differs from  $x$ . A good estimator for  $R$  is  $\exp(\overline{\log y} - \overline{\log x})$ , the median of the ratio — in log space, the difference between two mean values —.

**The ratio model: Discussion** Ratios are proportions. Lognormal items become homogeneous in a proportional space and their difference is a proportion too. These facts seem to match. But, is the ratio model adequate beyond this apparent matching?

The sole consideration of lognormality on accounting data is not enough to validate the financial ratios themselves as appropriate models. Such an empirical basis cannot prevent the ratio model from being questionable. Ratios are just the simplest relations allowed by the lognormal nature of items. Are ratios able to model all the relations important for financial analysis and knowledge acquisition?



Accounting research seems to give a negative answer to the above question. It is usual to find in the literature a tone of pessimism about the usefulness of ratios. The existence of non-proportional and non-linear relations between items are the main causes of concern. Whittington [32] explains that

... in an empirical relationship between a pair of accounting variables, two of the conditions necessary for proportionality are quite likely to be violated. Firstly, there may be a constant term in a relationship (...). Secondly, the functional form of the relationship may be non-linear.

The potential convenience of more elaborated models like regressions has also been stressed by Barnes [3]. He showed that in any regression  $Y = A + BX$  the distribution of  $Y/X$  will be skewed whenever  $A \neq 0$ . Ratio standards would be likely to misinform since no central trend would exist.

## 1.1 Ratios and Lognormality

In this section we use the well known proportional effect as a basis for explaining ratios. The usual financial ratio emerges as a simple consequence of a strong, common, effect.

The proportional effect has been quoted by McLeay [19] as a mechanism able to explain the existence of lognormality in a few items. He suggests a qualitative distinction between two kinds of items. The first kind would include items reflecting size. The second one, items which cannot “be treated as size measures [19]”. In our study the proportional effect explains size and deviations from size. No attempt is made to specify the particular behaviour of any item. Items having negative cases are considered as a subtraction of two positive ones and explained as such.

**Constant and Proportional Effects:** The Gaussian distribution is often interpreted as the result of many independent elementary perturbations. This approximation entails the strong assumption of a constant effect. For example, the probability of getting odds, when tossing a fair coin, is a constant value of  $1/2$  no matter the number of games or the size of the coin. And the probability of getting particular proportions of odds when tossing a coin in several sequences of games draws a Gaussian distribution. This constant probability of  $1/2$  governing the game referred to is what we call a constant effect. It leads to normality.

If, however, any random change  $dx$  suffered by a variable  $x$  is proportional to the value of  $x$  itself, the effect is no longer constant. It is a proportional effect and a Gaussian generative process will not be able to explain it.

Gaussian variables spread their final realizations around an expected value. They are bounded: It is most unlikely to find cases many orders of magnitude larger or smaller than the expected. This is because the random changes leading to such realizations are expected to be similar — a constant effect. Contrasting with such a mechanism, when the random changes leading to any final realization are expected to be similar only when taken as proportions of the momentary value of the variable, the effect is proportional. The probability distribution of such variables is unbounded. It exhibits strong positive skewness. The observed samples contain cases which are many orders of magnitude larger than the expected ones.

**The Gibrat Law:** The lognormal probability distribution can be viewed as a result of a generative proportional mechanism. This fact is known as the Gibrat law [13]. Let  $x$  be the position of an accounting item. If  $dx$ , the random transactions affecting  $x$ , are expected to be proportional to  $x$  itself,

the quotient  $\frac{dx}{x}$  will be expected to be independent of  $x$ .

So, if we can find a function

$$z = f(x) \quad \text{such that} \quad dz = \frac{dx}{x} \quad (3)$$

then the new variable  $z$  will obey the assumption of a constant effect. In the case of  $dz$  being many, independent, perturbations  $f(x)$  is the logarithmic function. Aitchison and Brown [1] contain a detailed explanation of this reasoning. Singh and Whittington [25] explore the growth of firms as governed by the Gibrat law.

Notice that the logarithmic function emerges as a result of the Central Limit theorem. The normality of the process governing  $dz$  is not required as an assumption, whenever the  $dz$  are many, independent, changes.

**The relative growth:** Any elementary perturbation  $dz$  will produce a small change  $dx$  which is expected to be a proportion of  $x$ . Therefore  $dz$  can be seen as an elementary relative growth and  $z$  as an expected relative growth.

Gaussian final realizations  $z_j = \log x_j$  are explained in the same way. Firstly, by a central trend  $\mu_x$  which is the expected one for the average relative growth affecting all cases in the sample. And secondly, by each particular departure from  $\mu_x$ , the  $e_j^x$ , affecting only firm  $j$ . These  $e_j^x$  are residual average relative growth: When back in multiplicative space, the  $e_j^x$  are the proportion in which the average relative growth of firm  $j$  is above or below the expected.

### 1.1.1 Financial Ratios

Now we study the joint variation of more than one item. The notion of financial ratio as a descriptor stems from the assumption of an effect common to all items for the same firm.

As we saw,  $dz = dx/x$ , the elementary relative changes of  $x$ , have the structure of a relative growth.  $z$  is Gaussian as  $dz$  is commanded by a constant effect.

**Modelling a common effect:** We now assume that in the case of accounting data this Gaussian relative growth is the sum of two components. A common and strong component,  $\sigma_j$ , which accounts for random changes acting over the firm  $j$  as a whole and therefore is the same for all the  $1, \dots, i, \dots, M$  items belonging to the same report. And a weak residual,  $\varepsilon_j^i$ , particular to item  $i$ .

Let  $x$  and  $y$  be the position of two accounting items for firm  $j$ .  $dx_\sigma$  and  $dy_\sigma$  are random changes in  $x$  and  $y$  caused by  $\sigma$ , a disturbance influencing both. Considering the way such common source of variability affects the relative growth which is about to generate  $x$  and  $y$  we can say that

$$\frac{dy_\sigma}{y_\sigma} = \frac{dx_\sigma}{x_\sigma}$$

This basic mechanism yields final realizations of  $x$  and  $y$  obeying general expressions of this kind:

$$\log y_\sigma - C^y = \log x_\sigma - C^x$$

$C$  are constants depending on the initial values of  $x$  and  $y$ . Here, the superscripts are used for identifying corresponding items, not as exponents.

Since we defined  $\varepsilon^x = \log x - \log x_\sigma$  and  $\varepsilon^y = \log y - \log y_\sigma$  as the variability unexplained by  $\sigma$ ,

$$\text{we have: } \log(y) - \varepsilon^y - C^y = \log(x) - \varepsilon^x - C^x$$

If the cross-sectional distribution of the common effect is dictated by the Gibrat law it will be lognormal. In this case, when we consider the whole sample of  $1, \dots, j, \dots, N$  firms, it is easy to see that the statistical model describing the relation between  $y$  and  $x$  for firm  $j$  is

$$\log(y_j) - \log(x_j) = (\mu_y - \mu_x) + (\varepsilon^y - \varepsilon^x)_j \quad (4)$$

a form similar to equation (2), the one based on empirical manipulation.  $\mu_y$  and  $\mu_x$  are the expected values of  $\log y$  and  $\log x$ . Therefore ratios can be viewed as specific models describing the common component of the variability of  $y$  and  $x$  when both  $x$  and  $y$  are supposed to be final realizations of a unique proportional mechanism. The residuals are independent of the common effect.

**Notation:** Equation (4) can be written in the form of a ratio:

$$\frac{y_j}{x_j} = R \times f_j$$

with  $R = \exp(\mu_y - \mu_x)$  and  $f_j = \exp(\varepsilon^y - \varepsilon^x)_j$ .

To express the differences between expected values we use the notation  $\mu_{y/x} = \mu_y - \mu_x$  or, for the ratio standards,  $R_{y/x}$  and so on. We write the differences between residuals as  $\varepsilon^{y/x} = (\varepsilon^y - \varepsilon^x)$ . Superscripts are intended to avoid too many subscripts and should not be taken as exponents. They are used only in the  $C_j, \varepsilon_j, e_j$  and  $f_j$ .

A good estimator for  $\mu_{y/x}$  is  $\overline{\log y} - \overline{\log x}$ , the difference between the mean values of  $y$  and  $x$  in log space. It is, of course, coincident with the median of the ratio expressed in logs. If an homogeneous sample of accounting reports is to be taken as a reference for the building of standards, the value of  $R_{y/x}$ , the ratio standard, should be calculated as

$$R_{y/x} = \exp ( \overline{\log y} - \overline{\log x} )$$

or directly as a median. And if we want to build a new variable from the residuals of the fitted model we can calculate each  $\varepsilon^{y/x}$  as

$$\varepsilon_j^{y/x} = (\log y_j - \overline{\log y}) - (\log x_j - \overline{\log x})$$

Notice that the  $\varepsilon^y$  or the  $\varepsilon^x$  are different from the  $e^y$  or the  $e^x$  in 2, the empirical formulation. However,  $(\varepsilon^y - \varepsilon^x)_j = (e^y - e^x)_j$  for any  $j$ .

**Ratios as functional relations:** As described here, ratios are functional relations. That is, they are not intended to explain one item in terms of the other one. Ratios yield a contrast between two items both affected by errors. Such a contrast measures how big are the discrepancies between the ratio components. Therefore, the above description is intended to assess deviations from standards, not to prediction.

**The weak, particular, effect:**  $\varepsilon^{y/x}$ , the difference between residuals, can be interpreted as the weak effect particular to  $y$  when  $x$  is taken as a proxy for the common effect. Unless we know  $\sigma$ , we cannot determine exactly the real weak effects associated with individual items. We know  $\varepsilon^{y/x}$  but we don't know each  $\varepsilon^x$  and  $\varepsilon^y$ , the components of such a residual difference.

Conversely, it is impossible to determine the value that  $\sigma$ , the common effect, assumes for firm  $j$  unless we know the components of the residual difference. Therefore, both the common effect and the particular one are not directly measurable. Ratios reveal what is different in their components, by concealing what is common in them.

**How to model the common effect: The Case-Average Model.** Is it possible to build a variate reflecting only size and having no particular variability of its own? As we saw, the common effect is not directly accessible. However, there is a way of isolating it by building a model which performs the function inverse of ratios. If we build, for each case in a sample, geometric means (in log space, averages) of items we can ideally self-smooth their particular components so that the common effect emerges.

Considering a group of items  $x_1, \dots, x_i, \dots, x_M$  selected as appropriate, and a common effect,  $s$ , we explain their variability in log space as the result of an effect,  $\sigma = \log s$ , common to them all, plus a residual,  $\varepsilon^i$  particular to each item. In the case of firm  $j$ ,

$$\begin{aligned} \log(x_{1j} + \delta_1) &= \sigma_j + \varepsilon_j^1 \\ \log(x_{2j} + \delta_2) &= \sigma_j + \varepsilon_j^2 \\ &\vdots \\ \log(x_{Mj} + \delta_M) &= \sigma_j + \varepsilon_j^M \end{aligned}$$

The  $\delta_i$  are the base-lines eventually present in  $x_i$  (see section 1.2.2 below). We now average the  $1, \dots, i, \dots, M$  items case by case. For firm  $j$ ,

$$\sigma_j = \frac{1}{M} \sum_{i=1}^M \log(x_{ij} + \delta_i) + \frac{1}{M} (\varepsilon_j^1 + \varepsilon_j^2 + \dots + \varepsilon_j^M)$$

Since an average of independent random deviates tends to zero with  $1/M$ , the number of components, we would have for a large  $M$

$$\sigma_j \approx \frac{1}{M} \sum_{i=1}^M \log(x_{ij} + \delta_i)$$

or the equivalent, in ratio form,

$$s_j \approx \prod_{i=1}^M (x_{ij} + \delta_i)^{\frac{1}{M}} \tag{5}$$

Once obtained,  $\sigma = \log s$  could be used as a proxy for size in statistical models. In [30] we further discuss this topic.

**Assumptions underlying ratios:** An usual topic in accounting literature is to call the attention for the assumption of strict proportionality underlying the use of ratios. Such a statement is descriptive. We could now enumerate the assumptions attached to the ratio model in a generative, rather than in a descriptive way:

1. Accounting items are final realizations of elementary random changes. Such changes, when expressed as proportions of the item they affect, are, in average, the same. This is the Gibrat law.
2. The elementary random changes leading to final realizations of accounting items are, when expressed as proportions of the item they affect, a sum of two components: One which affects in the same way all the items in the same report and another which is particular to each item.

The normality of the process governing  $dz$  is not required as an assumption.

The advantage of using a generative description is that we can now develop models other than simple ratios which are also consistent with this basis. Here, ratios emerge as models obeying to the statistical or expected proportionality of random effects. Proportionality at the end of a growth process is just a particular consequence of a given generative mechanism. Which other models are allowed by such a mechanism?

### 1.1.2 Ratios With More Than Two Items

In section 1.1.1 we noticed that the ratio model emerges when we consider the common variability of two relative growth processes in a generative mechanism. By considering more than two items an obvious extension of the usual ratio emerges.

Let  $x_1, \dots, x_i, \dots, x_M$  be the position of  $M$  items for firm  $j$ .  $dx_{i\sigma}$  are random changes in  $x_i$  caused by  $\sigma$ , a common source of variability. Considering the way such common disturbance affects the relative growth which is about to generate the set of  $x_i$  we can say that

$$\frac{dx_{1\sigma}}{x_{1\sigma}} = \frac{dx_{2\sigma}}{x_{2\sigma}} = \dots = \frac{dx_{M\sigma}}{x_{M\sigma}} \quad (6)$$

For example, we may want to consider two groups of items instead of two simple variables. Given  $y_1, \dots, y_k, \dots, y_K$  and  $x_1, \dots, x_l, \dots, x_L$  and reasoning in the same way as in previous section the mechanism described by 6 leads to the relation

$$\left[ \frac{1}{K} \sum_{k=1}^K \log(y_k) - \frac{1}{L} \sum_{l=1}^L \log(x_l) \right]_j = \frac{1}{K} \sum_{k=1}^K \mu_k - \frac{1}{L} \sum_{l=1}^L \mu_l + \left[ \frac{1}{K} \sum_{k=1}^K \varepsilon^k - \frac{1}{L} \sum_{l=1}^L \varepsilon^l \right]_j$$

Despite its outlook, this model is very simple and can be seen just as an expansion of equation 2. Here, instead of unique variables, expected values and residuals, we have averages of items for every  $j$ .

In ratio form the model would be

$$\frac{\prod_{k=1}^K y_{jk}^{1/K}}{\prod_{l=1}^L x_{jl}^{1/L}} = R \times f_j$$

that is, a ratio of geometric means of variables describes an effect common to them all in the same way simple ratios do.

As the estimators for the  $\mu$  are the mean value of the corresponding log item, the above expression is easily computed just by subtracting to each log item its mean value and then averaging groups of items inside the same firm.

**Degrees of freedom involved:** All the above explanatory models use only one degree of freedom. They are simple translations in log space. One free parameter is enough to account for a unique optimal value. Such an optimum is an estimator of a difference between two central trends. This fact has important implications for the assessment of ratio standards and the interpretation of departures from such standards.

The inclusion of more than one variable in each group will not account for more explained variability. The number of used degrees of freedom remains equal to one. We are still modelling a single parameter. However, more variables, if conveniently selected, can enhance the accuracy of ratios by a self-smoothing process able to make particularities cancel out.

## 1.2 Extending the Notion of Financial Ratio

The ratios introduced in the last section, despite their unusual outlook, are obvious applications to more than two items of the same principle governing the usual ones. In this section we extend the notion of financial ratio in two new directions allowed by the proportional mechanism. First we introduce non-linear proportions consisting of applying the linear model to the log space. Second, we model non-proportionality as part of the Gibrat law.

### 1.2.1 Non-Linear Proportionality

If we wish to model the joint behaviour of  $1, \dots, i, \dots, M$  items after controlling for the common effect we must be able to account for differences amongst them other than the simple position or mean differences the usual ratios account for.

In order to do this we notice that the proportional mechanism is able to yield more complex relations than those developed above. Expression 6 is just the simplest case. Accordingly, we now develop similar models, but able, to some extent, to cope with the variability of individual items.

The introduction of multi-variance in the generating mechanism can be done with different degrees of complexity. The simplest approach consists of using just one parameter,  $b_i$ , individualizing each proportion. This new parameter allows the description, using the

same formulation and without loss of generality, of the two components of the variability of each accounting item. A common effect would have  $b_i = 1$  for all variables.

Similarly to (6), there is a relative growth,  $\rho$ , for which

$$b_1 \times \frac{dx_{1\rho}}{x_{1\rho}} = b_2 \times \frac{dx_{2\rho}}{x_{2\rho}} = \dots = b_M \times \frac{dx_{M\rho}}{x_{M\rho}} = d\rho \quad (7)$$

In this mechanism, the  $b_i$  are gain or attenuation factors expressing different degrees of linear correlation between the generative growth rates leading to these variables. Notice that only  $M - 1$  of these  $b_i$  are independent.

In the simple case of  $b$  being similar across firms the consideration of two items,  $y$  and  $x$ , would yield general expressions like

$$\log y_j - b \times \log x_j = (C^y - C^x)_j + (\varepsilon^y - \varepsilon^x)_j - b \times \log b$$

or similar. Such free-slope ratios would look like this:

$$\frac{y_j}{x_j^b} = \exp(w_0) \times f_j$$

$b$  and  $w_0$  are now the two parameters of the model.  $w_0$  is expressible in terms of  $b$  and the initial values  $C^y$  and  $C^x$ .

And when considering two groups of items instead of two simple items we would have

$$\frac{1}{K} \sum_{k=1}^K b_k \times \log(y_{jk}) - \frac{1}{L} \sum_{l=1}^L b_l \times \log(x_{jl}) = w_0 + \left( \frac{1}{K} \sum_{k=1}^K \varepsilon^k - \frac{1}{L} \sum_{l=1}^L \varepsilon^l \right)_j$$

or similar.  $w_0$  is a parameter expressible in terms of the  $b_i$  and the initial values. In the form of ratio,

$$\frac{\prod_{k=1}^K y_{jk}^{b_k/K}}{\prod_{l=1}^L x_{jl}^{b_l/L}} = \exp(w_0) \times f_j$$

We can simply say that any multi-variate descriptor of this kind has, for  $1, \dots, i, \dots, M$  items, a general form

$$\sum_{i=1}^M w_i \times \log(x_i) = w_0 \quad (8)$$

in which the residual is omitted.  $w_i$  are parameters expressible in terms of the  $b_i$ ,  $M$ , and the initial values. In ratio form,

$$\prod_{i=1}^M x_i^{w_i} = \exp(w_0).$$



We can write (8) as a linear relation in log space

$$\sum_{i=1}^M w_i \times u_i = w_0 \quad \text{where} \quad u_i = \log x_i.$$

that is, in log space a simple inner product can account for linear correlations in the residual behaviour of accounting variables.

**Using free-slope ratios:** There are cases in which the free-slope ratios discussed above could be useful. For example, we might wish to introduce a second free parameter in the model because our goal is the prediction of  $y$  using  $x$  as predictor, not the assessment of a contrast between them. Functional relations describe mechanisms. Mechanisms should be plausible. Free slopes in log space are not plausible to describe mechanisms of the firm since they imply the existence of a unique relative growth for the same item across many firms. Moreover, it would be inadequate to consider non-linear mechanisms as a rule. But when the goal is to predict outcomes using items as predictors, there are no known objections to the use of simple regressions in log space.

### 1.2.2 Non-Proportional Ratios

The relation  $dx/x = dz$  is a simplistic description of generative processes. The Gibrat Law allows a more realistic basis by admitting that the random changes  $dx$  affecting  $x$  are proportional, not to  $x$  itself, but to  $x + x_0$ .

We call this  $x_0$  a base-line. Since the generative process leading to a particular realization of  $x$  starts with a non-zero value for  $x = 0$  the increments  $x$  receives at this point are in average proportional to such base-line. Therefore,

$$\text{instead of } dz = \frac{dx}{x} \quad \text{we should write} \quad dz = \frac{dx}{x + x_0}$$

for describing the generation of a particular item  $x$ .

Such a process leads to a class of ratios which can have many different characteristics according to the magnitude, sign and position of their base-lines. In some cases, but not in all, these base-line ratios will draw non-proportional relations between its components.

Notice that  $x_0$  should not be taken as the initial value of  $x$ , that is, the value of  $x$  at the beginning of the process leading to its final realization. Such initial values — which in our notation are the  $C^x$  — will not induce non-proportionality in the models describing cross-sectional samples. As long as the process is strictly proportional, the outcome is proportional as well. Non-proportionality emerges only when the random changes  $dx$  are proportional to values which are not  $x$ .

Next we briefly describe some of the possible models resulting from base-lines.

**An overall base-line in the denominator:** In the simplest case,  $x_0$  would be a constant value affecting all realizations of  $x_j$  for any  $j$ . That is, for a particular item all firms in the sample were expected to be affected by the same non-zero base-line.

One possible model resulting from a two-variate relation would be

$$\log(y_j) - \log(x_j + x_0) = \mu_{y/x} + \varepsilon_j^{y/x}$$

when the base-line acts on the denominator but not in the numerator. In ratio form,

$$\frac{y_j}{x_j + x_0} = R \times f_j$$

Base-lines occur when any of the ratio components is three-parametric lognormal instead of two-parametric. In the above expression and in all subsequent ones, the item affected by the base-line — in this case it is  $x$  — receives a transformation similar to the one used to achieve three-parametric lognormality (see Aitchison & Brown, 1958 [1]). Ratios of this sort are a non-proportional relation:

$$y_j = x_0 \times R \times f_j + x_j \times R \times f_j$$

The above form is useful just to show that such a model is not a linear regression. The non-proportional term  $x_0 \times R \times f_j$  is not independent. It will introduce displacements proportional to a residual value. Distortions will vary from case to case.

The distortions introduced by this kind of ratio will be small provide  $|\delta_x|$  remains small. The non-proportional term will be significant only for values of  $x_j$  near  $\delta_x$ , that is, whenever the generative process leads to final realizations of items which are near their base-line. Cases far away from their base-lines exhibit proportionality since  $x_j \gg x_0 \times R \times f_j$ .

**An overall base-line in the numerator:** By considering a base-line,  $y_0$  affecting  $y$ , the numerator of the ratio, instead of  $x$ , we get non-proportional terms which can more easily be significant. The expression

$$\log(y_j + y_0) - \log(x_j) = \mu_{y/x} + \varepsilon_j^{y/x}$$

means a ratio

$$\frac{y_j + y_0}{x_j} = R \times f_j$$

which can be written as

$$y_j = x_j \times R \times f_j - y_0$$

In this case the base-line acts as an intercept in a regression. It introduces a displacement affecting all cases in the sample. Notice that this model is still not a regression. The difference, however, is not functional. It stems from the multiplicative nature of the residuals.

**Base-lines both in the numerator and in the denominator:** When considering  $y_0$  and  $x_0$  as both significant, the amount of non-proportionality introduced results from their interaction. A reinforcement of non-proportionality will occur when  $y_0$  and  $x_0$  have different signs. Apart from this, the overall effect will depend on  $R$ , the expected proportion.

In a very particular case,  $y_0 = x_0 \times R$ , both base-lines cancel out. The remaining non-proportionality is residual.

**Multi-variate base-line ratios:** The general multi-variate descriptor, involving free-slopes and base-lines affecting all cases and present in several items would be written as

$$\sum_{i=1}^M w_i \times \log(x_i + x_{0i}) = w_0$$

or, in ratio form,

$$\prod_{i=1}^M (x_i + x_{0i})^{w_i} = \exp(w_0)$$

It is expected that multi-variate models of this sort will eventually generate strong departures from proportionality. The  $x_{0i}$  can easily reinforce their effects creating important joint displacements.

**Proportional base-lines:** The mechanism leading to the above descriptors requires an overall displacement — a base-line acting upon the whole of the sample in the same way —. Overall base-lines suppose the existence of overall costs or income.

We now consider the case of base-lines which are dependent of the size of the firm. Mechanisms internal to the firm are likely to generate base-lines proportional to size. The assumption of such internally generated base-lines being similar for the whole of the cross-section would be difficult to accept.

For  $1, \dots, j, \dots, M$  firms,  $x_{0j}$  is now a particular base-line concerning the generative process of each  $x_j$ . This base-line will act as a new variable, not as a parameter of the model. The model collapses into the no-base-line ones. In fact, if  $x_{0j}$  is proportional to the size of the firm it is similar to any other accounting item. For instance we could write

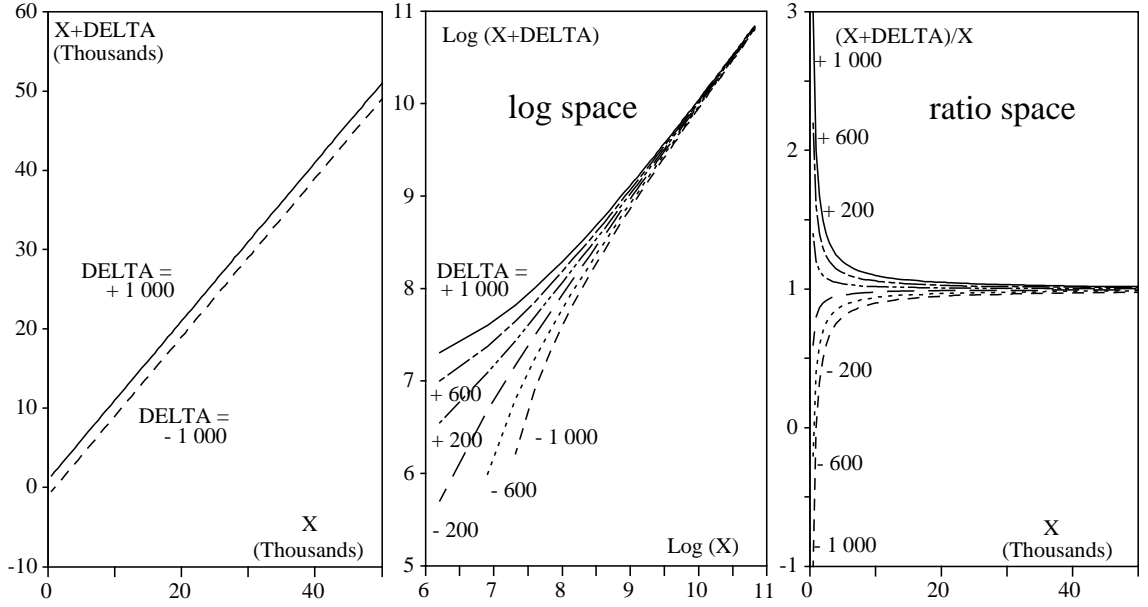


Figure 2: When  $Y = A + X$  is transformed, the fact that  $A \neq 0$  introduces non-linearity in the resulting relation. Such non-linearity affects only values of  $Y$  near  $A$ . On the left several  $Y = A + X$  slopes with very small  $A$ . In the centre the same slopes in log space. On the right, in ratio space.

$x_{0j} = x_j \times R_{0j} \times f_{0j}$  and we would have a relative growth

$$\frac{dx}{x \times (R_{0j} \times f_{0j} + 1)} = dz$$

for the generating process of a particular realization of  $x$ .

And since  $R_{0j}$  and  $f_{0j}$  are not involved in the subsequent growth of  $x$  the resulting model would be a version of the free-slope ratio we explored in section 1.2.1.

Base-lines proportional to the size of the firm will not break proportionality. However, they will induce differences in the way each item is affected by the common variability. In order to account for such differences, mechanisms similar to free slopes are required.

The described model is interesting because it has been often used in the accounting literature as an example of the plausibility of intercept terms in two-variate relations. It was an awkward choice since, as we see, base-lines acting just as another item are not likely to induce overall translations. We now analyze this subject in more detail.

**Assessment of overall departures from proportionality:** In order to assess the significance of overall departures from proportionality it is important to gain insight into the way the introduction of a constant term affects the linearity, in log or ratio space, of an

otherwise proportional relation. By applying log or ratio transformations to both sides of  $Y = A + X$  and observing the distortions resulting from increasing the value of  $A$ , we can acquire a precise idea of the impact of deviations from strict proportionality.

Figure 2 on page 15 shows the results of applying logs (centre) or dividing by  $X$  (right) in both sides of  $Y = A + X$  (left) for small values of  $A$  — thus obtaining relations which are formally similar to the above non-proportional models. Those considered  $A$  are  $\pm 1,000$ ,  $\pm 600$  and  $\pm 200$ . For large  $X$ , the effect of introducing such intercept terms is negligible. The effect of  $A$  becomes significant and visible whenever the order of magnitude of the  $X$  is similar to the order of magnitude of  $A$ .

Accordingly, base-lines must be taken into account only when the final realization of a growth process,  $x$ , is not far away from  $x_0$ . This could happen when the growth is weak (very small relative growth and very few random changes). The examination of two-variate scatter-plots of accounting items in log space can detect departures from strict proportionality when they are significant. As seen above, the log and the ratio transformations produce a trade-off between non-proportionality and non-linearity so that even small departures from proportionality result in departures from linearity.

### 1.2.3 Other Non-Linear Relations Between Items

Apart from the mechanisms described above, non-linearity could emerge in accounting models due to other causes. Two of them seem plausible:

**Higher order effects:** It could occur, for example, when modelling financial risk. Leveraged and non-leveraged firms can behave in opposite directions if they belong to some specific industries. In this case, the significant interaction would emerge owing to the presence of two groupings: Financial structure and industry. A statistical version of the “exclusive-OR” problem, that is, a second-order effect, can arise when modelling an outcome for more than one grouping. Such possibility is important when using linear techniques. Linear tools wouldn’t be able to separate effects other than first-order ones.

**Non-proportional non-linear relations:** Proportional non-linearity is just a particular class of non-linear relations between items. It would require the use of free-slope ratios instead of the usual ones. However, many other kinds of departures from linearity are possible. Whittington [32] reports quadratic relations in profitability ratios. He suggests that this could be explained by saturation effects. Saturation is the kind of distortion free-slope ratios could broadly model since it affects mainly the largest firms in the sample in a way similar to free-slope ratios do for  $b < 1$ .

However, non-linearity affecting, for example, the smallest cases in a sample, wouldn't be modelled by free-slopes. Notice that base-lines represent linear displacements. They only yield non-linearity in log or ratio space. And it is possible that, along with genuine base-lines, other influences affect small firms.

Existing accounting statistical models seem not to be aware of potential sources of non-linearity. Base-line and saturation effects are not very imposing and the second order effect can be avoided by increasing the dimension of the input space, which accounting models implicitly do.

### **1.3 Discussion and Conclusions**

We described a generative mechanism for the probability distribution observed in accounting data. The traditional notion of financial ratio stems from considering a common relative growth impinging over the two items forming it.

Ratios can be extended in several ways consistent with such a mechanism. Firstly, they can have more than two components. The sole requirement for the statistical validity of such ratios is the use of multiplicative residuals. Ratios can also be viewed in log space as a regression. Such free-slope ratios preserve proportionality. They introduce non-linearity in the large firms in the sample.

Finally, the existence of base-lines in the generation of items will eventually introduce non-proportional relations between the components of a ratio.

Distortions in proportionality resulting from overall base-lines depend on several factors. They are maximal for base-lines in the numerator of the ratio or when the signs of the base-lines of the numerator and the denominator are different.

## Chapter 2

# Knowledge Acquisition Using the Multi-Layer Perceptron

The Multi-Layer Perceptron [23], widely known as the MLP, is a supervised learning Neural Network. Topologically it is a layered feed-forward configuration: Nodes are arranged in layers and each node's output is connected to next layer's inputs.

Amongst the algorithms intended to learn a relation input-outcome from a set of examples, the MLP is different in that it approaches relations by stages. During the learning process an MLP creates new sets of variables corresponding to different stages of the modelling of the desired relation. A particular stage uses the variables from the previous one as input. Then, it makes an improvement towards the final modelling of the relation. Finally, it outputs a new set of variables to be used as input for the next stage. The intermediate variables generated by an MLP are often referred to as internal representations.

**Contents:** In this chapter we show that the ability to create internal representations along with other characteristics of the MLP, make it able to automatically extract meaningful knowledge from raw data directly available in accounting reports and the related outcomes thus avoiding the need for searching appropriate ratios. Section 2.1 describes how accounting items can be used as direct inputs for an MLP. Section 2.2 introduces a typical classification problem involving the prediction of the industrial group to which each firm belongs, using accounting data.

Using the above problem as a background example, section 2.3 explains the departures from ordinary techniques we introduced in the training of the MLP. Two contributions are outlined: The post-processing of MLP outputs so that they can be used as scores. The random penalization of small weights for improving generalisation and obtaining meaningful

internal representations.

Appendix A complements this chapter. It is a self-contained study of the performance of the MLP when compared with traditional methods.

## 2.1 Ratios as Internal Representations

Simple ratios have been used for extracting useful experience contained in samples where reports were gathered together with known outcomes. The problem of learning from examples using ratios can be formalized in this way: Let  $x$  and  $y$  be two items forming the ratio  $y_j/x_j = r_j$  in the case of firm  $j$ . For learning we have a sample containing  $1, \dots, j, \dots, N$  examples of these two accounting observations plus  $t$ , the vector of the related outcomes. If we assume the existence of a map  $\mathcal{W}$  such that  $\mathcal{W} : r \mapsto t$  then we learn it by finding a  $\mathcal{W}$  which is optimal in some sense.

**Relating features to outcomes:** The functional relation existing in accounting items — yielding ratios as seen in chapter 1 — is different from the relation between accounting features and outcomes we now study. The last one is the goal of statistical modelling. However, these two relations are not independent. Outcomes are dictated by internal features of the firm which we believe are reflected by appropriate ratios.

In the accounting statistical models used so far, the former relation is embedded in the choice of the input data — ratios. In the framework presented here we let the MLP form both such relations. Appropriate ratios are discovered and used to approach the outcomes, as part of a unique optimization process.

Ratios are size-adjusted variables. Since the size of the firm is generally considered as an important piece of information to model some relations, we also allow our framework to model it — using equation (5) — as an internal representation of the MLP. In short, when modelling a relation we allow ratios to be formed as the output of nodes in the first hidden layer of an MLP, along with a proxy for size.

**Forming ratios in the first hidden layer:** We let the raw data be the input to an MLP. Then, we set it to model the desired relation. As a first stage in this process ratios are formed that approach the outcomes. Other stages follow. At the end, outputs are the final stage. If ratios are the appropriate way of modelling such a relation, the internal representations formed by the MLP in the first hidden layer are extended ratios.

As seen in section 1.1.1 a multi-variate relation able to account for both common and



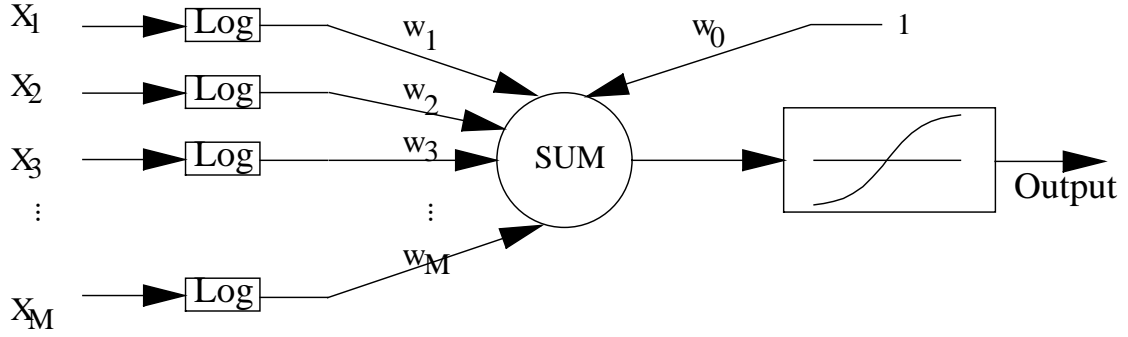


Figure 3: A node able to form a ratio in the first hidden layer of a MLP.

particular components of the variability of accounting data is

$$r = \prod_{i=1}^M x_i^{w_i}$$

containing  $1, \dots, i, \dots, M$  items as input. The residuals are omitted. In logarithmic space,

$$\log r = \sum_{i=1}^M w_i \times \log x_i. \quad (9)$$

Notice that this expression, an inner product, is the same as a Neural Network node's output.

Our approach consists of letting  $w_i$  be the adjustable connections or weights linking the inputs of an MLP with the nodes in the first hidden layer. The inputs are the logs of the accounting items,  $x_i$ , considered as interesting for modelling the desired relation. Thus we create in each node of the first hidden layer an internal representation with the form of an extended ratio. Next layers use such ratios to approach the outcomes.

By using an appropriate training scheme these extended ratios often assume a simple and interpretable form. If the overall model discovered by the MLP is optimal in some sense, it seems reasonable to expect that the discovered ratios represent an optimal choice of combinations of variables as well. Therefore, the best ratios to be used are not imposed by the analyst. Instead, they are discovered by the modelling algorithm. Our approach solves the problem of finding the appropriate set of ratios to model a particular relation. Such problem clearly emerges when reviewing the published literature.

**The transfer function:** Figure 3 is a representation of a node intended to form ratios. The logistic function

$$f(x) = \frac{1}{1 + \exp(-x - \theta)} \quad (10)$$

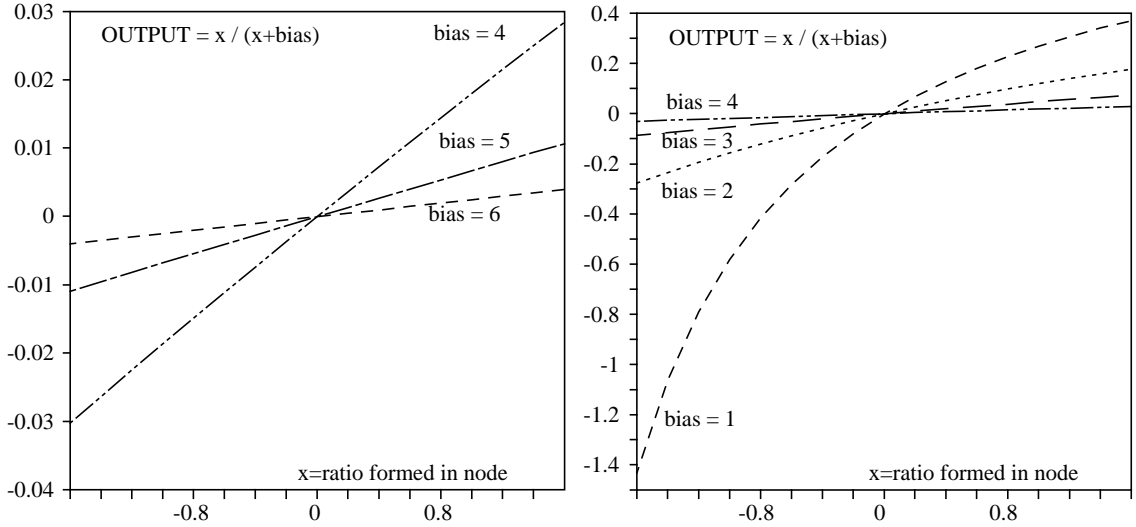


Figure 4: The output of each node in the log MLP will be a concave function approaching linearity for increasing values of the bias. On the left, a magnified view.

which is standard in Multilayer Perceptrons as a transfer function, will bring back the extended ratios from logarithmic space and will also provide a controlled amount of non-linearity for the lower values of  $r$ .

$$f(r) = \frac{1}{1 + \exp(-\log r - \theta)} = \frac{r}{r + \exp(-\theta)}$$

$\theta$  is the bias. Large negative values of  $\theta$  yield a linear relation between  $r$  and the output of the node. Smaller values introduce a concavity affecting small  $r$ .

Figure 4 shows the way  $\theta$  controls the output of its node. For increasing bias the node's response is linear. In general, the training of the bias is directed by the optimization algorithm so that the output is linear. Therefore, the first hidden layer is not apportioning non-linearity to the model. This can be done in next stages. However, there is a class of non-linearity which is accounted for in this layer just by allowing smaller  $\theta$  (a notation also used for  $\theta$  is  $w_0$ ).

**The modelling of base-lines and other non-linearity:** The back-propagation of errors could also be used to discover and account for non-proportionality in individual inputs. Appropriate base-lines could automatically be found for each input just by using the information propagated backwards, in the same way the other free parameters are estimated.

In practice, such a propagation across the log function is not stable. The MLP simply generates  $\delta$  which become more and more negative during the training. Therefore, at least at the present stage of the research, we directly model the non-linearity introduced by

Feature	Ratio	Tr.	Feature	Ratio	Tr.
Operating Scale	$NW/S$	Log Log	Fixed Capital Intensity	$FA/TA$ $S/Av. FA$	Sqrt Log
Labour-Capital Intensity	$W/TA$ $VA/Av. TCE$	Sqrt Sqrt	Short Term Asset Intensity	$D/CA$ $D/I$	None Log
Profitability	$OPP/S$ $EBIT/S$ $OPP/Av. TCE$ $EBIT/Av. TCE$	Sqrt Log Sqrt Sqrt	Asset Turnover	$DD$ $S/Av. TA$ $S/I$	None Log Sqrt
Net Trade Credit	$D/C$	Sqrt	Financial Leverage	$DEBT/NW$ $DEBT/TCE$	Sqrt None

Table 1: Ratios used in the original study and their transformations.

base-lines instead of reproducing its underlying mechanism. This can be done since, as we saw in section 1.2.2 the log space introduces a trade-off between non-proportionality and non-linearity (see figure 2 on page 15) and the MLP can model such a non-linearity.

## 2.2 Learning to Discriminate Industrial Groups

In this section we apply our framework to a known accounting statistical problem, the test of the separability of the components of a particular industry grouping. We compare our procedure with the traditional one and we extract some conclusions.

All companies quoted on the London Stock Exchange are classified into different industry groups according to the Stock Exchange Industrial Classification (SEIC). We selected 14 manufacturing groups according to the SEIC criteria. After discarding some firms (see below) we got accounting information on 500 cases belonging to 1984 reports.

**The data:** The input variables received two different types of processing. The first, usual in finance research, consisted of “forming 18 financial ratios chosen as to reflect a broad range of important characteristics relating to the economic, financial and trade structure of industries (...) [29]” and extracting from them the eight principal components. These new variables were then used as inputs for a Fisher’s Multiple Discriminant Analysis (MDA). A description of these ratios and the modelling procedure can be found in [29]. Table 1 reproduces them along with the transformations applied. DD is the ratio Debtors Days.

Appendix A is a self-contained study of the performance of the MLP compared with traditional methods. There, a description of our reproduction of the method usual in finance research can be found along with the detailed MLP classification results.

The new approach consisted of using eight accounting variables directly, not in the form of ratios. The selected items were Fixed Assets, Inventory, Debtors, Creditors, Long

Term Debt, Net Worth, Wages and Operating Expenses less Wages. All these variables were present in the original 18 ratios, along with others like Earnings, Value Added, Total Capital Employed and Total Assets which we didn't use in the new approach.

**Criteria for selecting the input variables:** The criteria used to select the new variables was threefold. Firstly, they should have been present in the original set in order to allow the comparing of results. No new information was to be introduced in the problem. Secondly, we avoided items representing totals for reasons explained elsewhere [30]. Finally, the input dimension should be eight or less. The number of common factors extracted from ratios in the original study was eight. Eight items or less wouldn't allow a larger flow of information.

The choice of *EX* and Wages instead of Sales and Operating Profit stems from the same reasoning. The discarding of Earnings stems from not being appropriate for the log transformation. The information contained in *EBIT* could be introduced by Sales and *COGS* but for this particular model the residual *EBIT* didn't seem important.

**The selection of cases for the samples:** A major methodological difference between our approach and the usual one was the way firms were selected. In general, one-variate normality criteria is used to prune the original sample of ratios down to an acceptable number of standard deviations. We followed a case-wise method for discarding undesirable firms. It was not based on distributional considerations. Only cases known as distressed firms, non-manufacturing representatives of foreign companies, merged or highly diversified ones were excluded.

Table 2 displays the proportions of cases in the sample. Notice how groups are dissimilar in size, the smallest one having 16 firms and the biggest 80. These proportions entail no prior knowledge of any classification.

## 2.3 Improving Generalisation and Interpretability

In this section we explain the characteristics which make our MLP different from the standard algorithm. They can be summarized as:

- The use of two samples, one to learn and another one to assess the classification performance. This is commented in 2.3.1.
- The random penalization of small weights, explained in 2.3.2.
- The post-processing of outputs, outlined in 2.3.3.

N.	Group Code	Group Name	N. Cases	Proportion
1	14	Building Materials	31	6.2%
2	32	Metallurgy	19	3.8%
3	54	Paper and Pack	46	9.2%
4	68	Chemicals	45	9.0%
5	19	Electrical	34	6.8%
6	22	Industrial Plants	17	3.4%
7	28	Machine Tools	21	4.2%
8	35	Electronics	79	15.7%
9	41	Motor Components	23	4.6%
10	59	Clothing	42	8.4%
11	61	Wool	19	3.8%
12	62	Miscellaneous Textiles	30	6.0%
13	64	Leather	16	3.2%
14	49	Food Manufacturers	80	15.9%

Table 2: Industrial groups and the proportion of each one in our sample.

- Learning rates particular to each weight as described in Silva and Almeida [24].
- Likelihood maximization instead of squared deviations minimization, as explained in 2.3.3.

The first characteristic relates to improvements in the ability to generalise. It is a particular implementation of a known procedure, the Cross-Validation [27] [28]. The random penalization of errors and the post-processing of outputs are specific contributions of this study. They allow the use of the MLP for general-purpose statistical modelling and the interpretability of results.

### 2.3.1 Generalisation: Using the Test Set

In order to obtain an estimate of the generalisation capacity of a model, the original samples were divided randomly into two sub-samples of approximately equal size. All models were constructed twice, first with one half of the sample and a check carried out with the other half, and again reversing the roles of the two half data sets. Results were considered conclusive if both models, when validated with the half-sample not used to build them, produced consistent results.

All classification results reported here concern the test set, not the training set. That is, they were obtained by measuring the rate of correct classification in the half-set not used for learning. The classification performance on the set used for learning depends solely on the number of free parameters and can be increased simply by introducing more nodes on the net. Therefore such results are uninteresting.

The normal approach to test a model, by deleting a single observation and predicting its value with the model estimated on the rest of the data set, and repeating this procedure  $N$  times, is not feasible. This is because the training of a Neural Network is time consuming. The procedure adopted will however, with a large enough data set, produce unbiased estimates [10] [27].

The described procedure, combined with incomplete training, also allows improving the generalisation of the MLP. This is a common practice. Next we describe incomplete training.

**The role of incomplete training:** Since the MLP seeks an optimum iteratively, we can stop its training when an optimum is obtained in the test set rather than in the training set. In doing so we prevent this powerful algorithm from over-fitting the data.

It is generally believed that the Back-Propagation algorithm seeks the modelling of progressively smaller or less important features of the relation during the learning process. Firstly, broad features are accounted for: The mean, a linear trend. Then, more detailed ones are modelled. Hence, the effective degrees of freedom the MLP engages can be viewed as increasing during learning [31].

Assuming that the topology of the net contains plenty of free parameters, the MLP will be able to model, not only the desired features but also the undesirable random uniqueness of a particular sample. We prevent it from doing this by stopping the process before finishing. The appropriate moment for stopping is when the results, as measured by the test set, are optimal.

For a good topology, the fact that the learning stops before a minimum is reached in the learning set clearly enhances the generalisation. The difference between the generalisation performances achieved with analytic tools and the iterative ones stems from this ability to stop. In our example, if we allow the training to proceed, the generalisation obtained with the MLP is worse than the one obtained with analytic tools.

**The role of an appropriate topology:** We found that the generalisation was dependent on the used topology. The number of nodes in a hidden layer seems to determine, not only the dimension of the relation, but also the ability of the MLP to generalise.

### 2.3.2 Random Penalization of Small Weights

Another major goal of this study was to evaluate the power of Neural Networks in knowledge acquisition. Multi-Layer Perceptrons are often considered as not ideal in applications where self-explanatory power is required. However, in the case of accounting variables it seems possible to interpret the way the relation has been modelled by looking into the weights

connecting input variables with the first hidden layer's nodes. These weights are the free slopes of ratios.

In order to enhance interpretability we introduced during training a random penalization of weights with small absolute values. A weight is inhibitory when its absolute value is smaller than the unit. If the input variables were very differently scaled, inhibition values in the input weights could just mean that the MLP was trying to scale down a particular variable. Since the log items used as input to the MLP are mean-adjusted and have very similar spread the only reason for any such weights to remain smaller than the unit throughout the learning is to try to diminish the importance of one variable in the output of the node it belongs to.

In a Neural Network each node acts as a modelling unit with a certain amount of free parameters. The same output can be obtained with very different combinations of weights. Inhibition weights connecting inputs with the first hidden layer appear when the node tries to weaken the contribution of a variable. If we randomly introduce a small penalization of such weights during the training, as the correction of weights is proportional to the input variables, the weights smaller than the unit tend to remain small. In the same way, the large weights tend to have their values strengthen.

The final result is a contrasted set of weights: The first layer now contains only very large or very small weights. The information concerning the modelled relation is concentrated in a few weights instead of distributed by all of them. If the relation to be modelled is consistent with such a contrast, then there is no reason to expect that the described manipulation will damage the performance of the model.

**The algorithm:** The procedure to achieve interpretability involves these steps:

- Let one node in the first hidden layer model the strong common effect and introduce it in subsequent layers. Input variables not convenient for the modelling of size (Debt is an example) have weights connecting to this node set to zero. The others have fixed and equal weights.
- During training, and whenever a new presentation of the entire training set is to begin, one of the remaining nodes of the first layer is randomly selected. Their weights are examined and those with inhibitory weights are penalized by a small factor, typically 0.98.
- Before the end of training, all the weights connecting inputs to the first layer and exhibiting very small values are set to zero and fixed.

This procedure is applied only after discovering the topology yielding the best results. Just by dedicating one node of the first hidden layer to the modelling of size we noticed an improvement in speed of convergence and in the final generalisation. Adding the random penalization of inhibitory weights both speed and generalisation received a further, significant, improvement. When the topology is not the best this procedure can worsen the generalisation.

**Complementary remarks:** The method described here was the one used for this particular experiment. In different cases we found that the performance would not suffer if all the weights below an inhibitory threshold were penalized at the beginning of each new presentation. This threshold typically would begin in 0.1 with the training and then it would be updated to larger values later on. Instead of fixing the weights just before the end of the training we also introduced their fixing during training whenever they would become small enough. By the end of the training the number of free parameters is much reduced.

We never tried this method with the usual, simple, Back-Propagation algorithm. Each one of the weights in our MLP has its own increment, adjusted as described in Silva and Almeida [24]. Other popular methods for pruning the MLP are the “Skeletonization” [20] and “Optimal Brain Damage” [17]. The first one is intended to reduce the number of nodes, not weights. The second one is too general for this task.

**Results:** When the training finishes the number of variables to consider in each node is small and characteristic. Looking at the non-zero weights it is possible to understand, in accounting terms, what the free-slope ratios formed in each node are doing.

Table 3 shows the extended ratios formed in a net with 8 inputs, 6 nodes in one hidden layer and 14 output nodes. The emerging organization reproduces the way an expert in ratio analysis chooses variables. It is usual to build several ratios around one or two variables judged as important to capture a relation. As an example, efficiency is modelled around capital turnover, stock turnover and so on. Analysts put together several points of view around a few significant variables. Extended ratios seem to be trying the same. The item *EX* has been used in all hidden nodes to contrast others. It seems as if it were important for this problem.

The ratios the MLP discovers are not always simple. Ratios like  $(C \times FA)/(W \times EX)$  are not the most familiar ones to accountants. However, in general the combinations of items which emerge as interesting are clearly visible when examining the organization of the hidden nodes.



Variable	Node Number	2	3	4	5	6
Long Term Debt				-6		
Net Worth		8				
Wages		1			-6	
Inventory		8				
Debtors		2				-2
Creditors					3	
Fixed Assets		-9	-4		6	-4
Operating Expenses less Wages		-10	4	8	-2	3

Table 3: Approximate values of weights connecting input variables with nodes in the first hidden layer after training with random penalization.

**Testing the Performance of the Devised Ratios:** Our interpretation of the ratios

formed in the hidden nodes is, according to table 3:  $\left\{ \begin{array}{l} \text{In the 2}^{th} \text{ node: } \frac{NW \times I}{FA \times EX} \\ \text{In the 3}^{th} \text{ node: } \frac{EX}{FA} \\ \text{In the 4}^{th} \text{ node: } \frac{EX}{DB} \\ \text{In the 5}^{th} \text{ node: } \frac{FA \times C}{W \times EX} \\ \text{In the 6}^{th} \text{ node: } \frac{EX}{\sqrt{FA \times D}}. \end{array} \right.$  We

tested the performance of such ratios when used as inputs for linear classifiers in the described problem. The five ratios plus the size effect actually classify the 14 industrial groups with the same accuracy as the original 18 variables.

The gain in performance by using the MLP is, of course, much more visible. Apart from its non-linear modelling capacity — which in this particular problem didn't seem to be very important — such a gain is due to its superior generalisation. Analytic tools cannot control the relative importance of parameters during training nor stop the optimization process before its end, to avoid over-fitting.

### 2.3.3 Post - Processing of Outputs

Discrimination, when overlapping distributions are present, implies a probabilistic interpretation of outputs. In accounting research, Bayesian considerations are in general independent of the proportions observed in the sample. Neural Network application to other sciences can be misleading. There, proportions observed in the sample are generally taken as acceptable prior probabilities.

Following suggestions like those of Baum and Wilczek [4] several authors advocate a direct interpretation of outputs as probabilities [14] [26] and show how the usual squared error criterion can be corrected to achieve likelihood maximization. In such case, the weights are corrected in the gradient direction of the log-likelihood rather than on the gradient of the squared error.

We found that node outputs, when interpreted as probabilities, produce a clear reduction in accuracy. The result is a severe loss of ability to distinguish small groups. Thus, we decided to interpret outputs of the MLP as a multi-dimensional measure of distance to outcomes. If departures from normality are not severe, this interpretation can be carried out by using conventional statistics like Chi-Square, Penrose or Mahalanobis distances. Such measures can be regarded as scores and conditional probabilities can be deduced from them, allowing further Bayesian corrections, independent of proportions observed in the sample. Of course, a Bayesian correction could be done directly over the outputs interpreted as probabilities. However, due to the observed lack of accuracy, a direct correction would lead to a very bold classification.

**An experiment:** Using the MLP with the 1984 data set and implementing learning schemes described by Hopfield [14] and Solla *et al.* [26] we tested the direct interpretation of node outputs as probabilities comparing it with the usual correction of node outputs based on the way linear discriminant analysis, for example, corrects scores. Results are reported in figure 5 when prior probabilities are taken as equal to the size of the group. On the left we can see the result of using post-processing. On the right, the corresponding result derived by directly interpreting node outputs as probabilities.

The post-processing gives detailed classifications. Direct interpretation ignores 9 of the 14 groups, the small ones, but finally achieves a better global performance by classifying the remaining 5 groups, which are the bigger ones, very well. Therefore, although for the sake of efficiency of convergence we adopted the likelihood cost function, node outputs were post-processed as distances. A short description of this post-processing follows.

**MLP outputs as multi-variate distances:** For a training set with  $N$  cases, consider  $o_{jm}$ , the output produced in node  $m$ ,  $m = 1, M$  by case  $j$ ,  $j = 1, N$ . Compute  $K$  square deviations,  $d_{kjm}$ , between the  $m$  node's output and each one of the  $1, \dots, k, \dots, K$  possible outcomes:  $d_{kjm} = (t_{km} - o_{jm})^2$ . The mean sum of squares in node  $m$  for the whole sample will be:  $\sigma_{km}^2 = \sum_{jk} d_{kjm} / (N - 1)$  and the standardized distances between a node's output

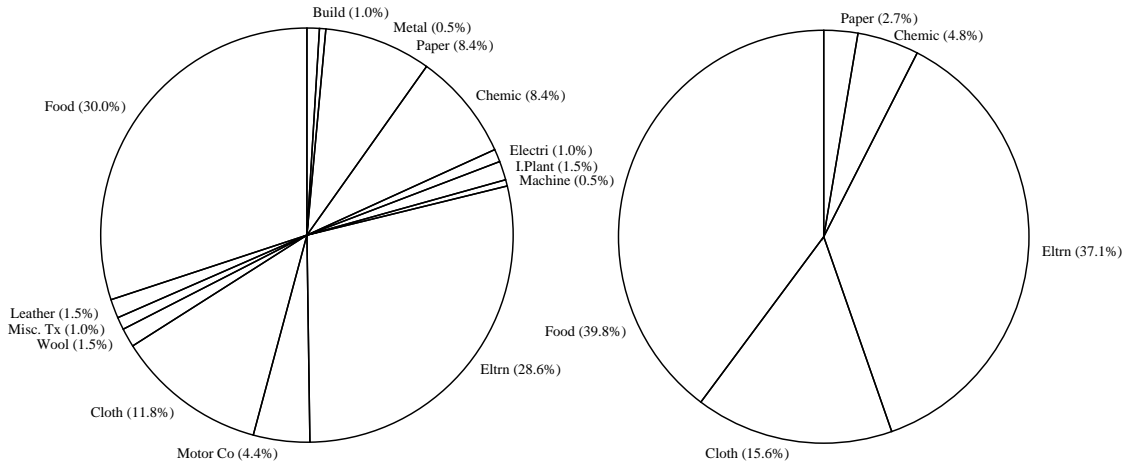


Figure 5: On the left, classification results after post-processing (1984 sample and prior probabilities proportional to the size of the group). On the right, the same with direct interpretation.

and all possible outcomes can now be added over all nodes:

$$D_{kj} = \sum_{m=1}^M \frac{d_{kjm}}{\sigma_{km}^2}$$

The minimum of these distances would identify the outcome predicted by the MLP if no Bayesian corrections were needed — that is, if the assumption of equal prior probabilities is acceptable.

This distance has been compared with a more elaborated measure, the Mahalanobis distance, and it was found that the latter would not achieve a more accurate performance. In order to introduce Bayesian considerations,  $D_{kj}$  ought to be computed as a Chi-Square distance to outcomes. The significance of this distance is the desired conditional probability.

## 2.4 Discussion and Conclusions

So far, expectations about Neural Networks are related to the modelling of difficult relations (pattern recognition) or the mimicking of brain functions. There has been little emphasis in their potential explanatory power. Here we argue that some statistical problems requiring self-explanatory power can take advantage from the existence of meaningful internal representations.

Numerical, continuous-valued observations such as those found in stock returns, or data organized in accounting reports, cannot be efficiently used by actual expert systems as a source of knowledge. Algorithms intended to automatic extraction of rules from examples,

such as the ID3 [22] cannot perform efficiently with non-symbolic, non-hierarchical data. We explore this problem elsewhere [7].

Neural Networks can now be seen as an alternative self-explanatory tool. In our example, hidden units were able to form more appropriate ratios than those commonly used. In other cases the examination of such ratios could shed light in many important issues.

**Self-explanatory models:** The developments of this study are closer to Beaver’s original works than its successors. Beaver tried to discover the most appropriate ratios to model a relation. The goal was not just an efficient modelling. It was mainly the discovering of simple tools for doing the job. After him, statistical modelling focuses on efficiency. The practice of using multi-variate techniques and a large amount of ratios as inputs — along with the trimming, ad-hoc transforming and rotating of inputs — made impossible any interpretation of results. Modelling became a blind automatism.

**Improvements in performance:** The emphasis on interpretation should not hide the other findings of our study. The MLP proved able to outperform the classification performance of a traditional discriminant analysis approach. Neither method came close to adequately classifying the testing sets, but there was a substantial improvement (29% to 38%) when the MLP was used.

The MLP achieved a better performance, with half the number of input variables and within a much simpler framework. Namely, the need for devising appropriate ratios was avoided as well as the blind pruning, and the extraction of a somehow arbitrary number of factors. Several accounting variables used to form the 18 original ratios were not present in our 8 variable set.

**Topology:** The principle of parsimony should also be born in mind. If there are too many hidden nodes the MLP will fail to identify key features and will model the particular randomness in the data set as well. Generalisation will then be lost.

However, Back-Propagation shows a useful ability to take advantage of the topology of the net to improve generalisation. Even with a large number of free parameters, if the number of nodes in a hidden layer is in resonance with some internal feature of the data, high generalisation can arise.

## Appendix A

# Classification Results Using the Multi-Layer Perceptron

In this appendix we gather information concerning the experiment described in section 2.2 about the Multi-Layer Perceptron (MLP) as a modelling tool for accounting relations. But here we examine the MLP as a classifier, intended to be used instead of Multiple Discriminant Analysis (MDA). Therefore, we focus on the classification performance rather than on the acquisition of knowledge.

Each section of this appendix contains the description of a particular test. Firstly, the technique usual in accounting research, involving 18 ratios as input variables, is described. The results of using MDA are compared with those of using MLP. Next we apply the framework developed in the first part of this study instead of the usual one, both with MDA and MLP modelling. It uses eight log items as inputs. Finally, we show the classification obtained with the new ratios devised by the MLP when used as inputs for MDA.

This appendix is intended to show the importance of implementing our framework in a particular, well known, problem. Also, the circumstances leading the MLP to outperform the linear tools can be devised.

### A.1 Results: The Usual Technique

In this section we describe the procedures and results obtained when applying to the classification problem the techniques which are usual in accounting research, that is,

- Input variables are ratios selected so as to reflect desired features.
- Ratios suffer ad-hoc transformations. The goal is to achieve improvements in their

Ratio	Skewness	Kurtosis	Ratio	Skewness	Kurtosis
log NW	0.42	0.01	log S	0.38	0.09
DD	0.59	1.60	FA/TA	0.33	-0.15
S/FA	4.5	28.5	W/TA	1.09	2.17
VA/TCE	2.5	13.8	OPP/S	0.17	6.05
EBIT/S	0.53	5.95	OPP/TCE	1.85	34.8
EBIT/TCE	1.25	24.5	S/TA	2.01	6.85
S/I	2.94	12.1	D/CA	1.45	9.70
D/I	2.41	11.2	D/C	1.78	5.74
DEBT/NW	3.31	18.1	DEBT/TCE	3.31	18.1

Table 4: Skewness and kurtosis of ratios used in the replica of the traditional approach. These values were obtained after applying transformations. DD is the Days Debtors ratio.

normality.

- Factor Analysis is used to extract a few variables from the set of transformed ratios.
- Multiple Discriminant Analysis uses such factors as input variables. In this case, the industrial grouping according to the SEIC is the outcome.

Our study reproduces a reputed one, carried out in 1984 by Sudarsanam and Taffler [29] and quoted by Foster. The ratios used and their transformations are displayed elsewhere (see page 22).

**Normality of transformed ratios:** We obtained a broad set of values for the skewness and kurtosis of the ratios used in the replication of the study referred to. Such values are displayed in table 4.

DEBT has a large number of zero cases corresponding to non-leveraged firms. It will not yield homogeneous distributions with any transformation. The factor extracted from DEBT ratios exhibit a very strong two-modality.

**Extraction of factors from ratios:** After obtaining the transformed ratios we extract the eight largest components of their variability. Next we display the differences between our study and the original one concerning the affinity of input variables with the resulting factors.

1. Operating Scale: We obtained the same groups.
2. Fixed Capital Intensity, the same groups.
3. Labour Capital Intensity, the same groups.

4. Profitability, the same groups.
5. Asset Turnover: This factor was formed with variability from  $S/TA$  and  $S/I$  mainly.
6. Short Term Asset Intensity,  $DD$ ,  $D/C$ ,  $D/CA$  and  $D/I$ .
7. Net Trade Credit,  $DD$ ,  $D/C$ ,  $D/CA$  and  $D/I$ .
8. Leverage, the same groups.

Therefore, our study found differences in the interpretation of the factors related to Short Term. In our data the three factors representing short-term features have their variables mixed up. The co-variance matrix was almost singular. The main correlations were observed between

- $\log S$  and  $\log NW$  (0.995),
- $\sqrt{OPP/S}$  and  $\log(EBIT/S)$  (0.971),
- $\sqrt{OPP/TCE}$  and  $\sqrt{EBIT/TCE}$  (0.980),
- $\sqrt{FA/TA}$  and  $\log(S/TA)$  (0.970),
- $\sqrt{S/TA}$  and  $\log(S/I)$  (0.922) and finally between
- $\sqrt{DEBT/NW}$  and  $DEBT/TCE$  (0.970).

The eigenvalue sequence doesn't exhibit the smallest trace of a break in the rate of decay. It decays smoothly in an exponential way. The factors are, more or less, replicating the original variables. Hence, there is no clear distinction between the selected factors and the rejected ones. There is no real commonality or real uniqueness and each factor contains a good portion of the information others contain. Since the purpose is the reduction in the number of dimensions, not the discovering of features, this is just as well.

A typical eigenvalue sequence have values like these: 20%, 19%, 15%, 15%, 13%, 9%, 8%, 6%. Typically, eight factors account for more than 90% of the variability.

**Transformations:** We must remark that the effect of using different transformations inside the same set of input variables introduces a non-negligible amount of non-linearity in input space. If two linearly related variables are exposed to different transformations, say, one square root and the other logs, the resulting relation between them is no longer linear. Afterwards, when factor analysis is used to extract new variables from these non-linear ones, the clear result will be that most of the variability associated with the extreme

N.	SEIC Code	Group Name	N. Cases	Correct	N. Cases	Correct
1	14	Building Materials	8	3	23	1
2	32	Metallurgy	11	1	8	2
3	54	Paper and Pack	25	5	21	1
4	68	Chemicals	22	4	23	7
5	19	Electrical	16	3	18	4
6	22	Industrial Plants	8	2	9	1
7	28	Machine Tools	11	2	10	1
8	35	Electronics	49	11	35	14
9	41	Motor Components	17	4	6	5
10	59	Clothing	19	10	23	9
11	61	Wool	7	1	12	1
12	62	Miscellaneous Textiles	11	1	19	1
13	64	Leather	8	1	8	3
14	49	Food Manufacturers	43	25	37	23

Table 5: Classification results with MDA and 8 factors.

values — the ones which are most curled by the non-linearity — is flattened away. Factor analysis extract linear patterns. Hence, the final result of this interaction between artificial non-linearity and linear factor analysis is that the extreme values of the distribution will be pushed towards the centre of the distribution.

**Multiple Discriminant Analysis:** A diversion from the original study consisted of dividing the set of examples randomly in two approximately equal sized samples. The MDA model was built with one of the samples but its performance was checked with the other one. In general, the size of each group in one set and in the other are not very similar. This fact introduces a distortion in the classification results since the likelihood of each group in the test set is different from the likelihood in the training set. However, by imposing equal prior probabilities across groups this distortion is minimized.

We are mainly interested in comparing the performance of MDA with that of the Multi-Layer Perceptron. Provide the samples are the same and the prior assumptions coincide, this comparison can be carried out.

Table 5 shows the classification results. *N. Cases* displays the number of cases in a group after split in two random samples. *Correct* shows the number of correct classifications when that group was used to model and the other group was used to test.

The displayed results and all the other results reported were obtained under the supposition of equal prior likelihood of any firm to belong to this group or the other. There is no special reason why a prior knowledge about relative size of groups should be included in this study.

For small groups the classification is very poor. It increases dramatically with the size



N.	SEIC Code	Group Name	N. Cases	Correct	N. Cases	Correct
1	14	Building Materials	8	3	23	0
2	32	Metallurgy	11	0	8	1
3	54	Paper and Packing	25	6	21	1
4	68	Chemicals	22	4	23	7
5	19	Electrical	16	4	18	6
6	22	Industrial Plants	8	2	9	0
7	28	Machine Tools	11	1	10	0
8	35	Electronics	44	12	35	16
9	41	Motor Components	17	3	6	5
10	59	Clothing	19	12	23	10
11	61	Wool	7	0	12	0
12	62	Miscellaneous Textiles	11	2	19	1
13	64	Leather	8	0	8	3
14	49	Food Manufacturers	43	28	37	24

Table 6: Classification results with MLP and 8 factors.

of the group. An overall 29% of success in both cases is attained almost because of very good classification of groups like Food and Electronics.

## A.2 MLP With 8 Factors as Input Variables

The same eight factors which were used as input variables for MDA were also tested as input for an MLP. After several experiments we found that the best results would be achieved with an MLP with one hidden layer and six nodes on it. Table 6 contains the number of correct classifications in the test set, by group.

These results concern an MLP with six nodes in a unique hidden layer and 14 output nodes (one per group). Outputs were post-processed as described in section 2.3.3, on page 28 but no random penalization of weights were applied. The criterion used for convergence was the maximization of the likelihood input-outcomes.

The training was interrupted when the likelihood, measured in the training set, reached a maximum. This procedure is therefore different from the one referred to in section 2.3.1, page 24. It allows a direct comparison with the results of the MDA modelling.

Under the displayed conditions, the MLP shows a performance which is similar to the one of MDA (about 30% of correct classifications), a linear, analytic tool. We believe that the improvements in performance achieved in later experiments stem from the interruption of training before its completion and also from the more robust pre-processing of the input data.

N.	SEIC Code	Group Name	N. Cases	Correct	N. Cases	Correct
1	14	Building Materials	8	2	23	1
2	32	Metallurgy	11	2	8	2
3	54	Paper and Packing	25	5	21	6
4	68	Chemicals	22	4	23	7
5	19	Electrical	16	5	18	4
6	22	Industrial Plants	8	1	9	2
7	28	Machine Tools	11	2	10	2
8	35	Electronics	44	21	35	14
9	41	Motor Components	17	4	6	5
10	59	Clothing	19	10	23	9
11	61	Wool	7	1	12	4
12	62	Miscellaneous Textiles	11	2	19	4
13	64	Leather	8	1	8	1
14	49	Food Manufacturers	43	26	37	23

Table 7: Classification results with MDA and 8 log items.

### A.3 MLP and MDA With Eight Log Items

We now describe our procedure for modelling the relation between accounting information and industry grouping.

We recall that the new approach consisted of using eight accounting items directly, not in the form of ratios. A simple two-parameter log transformation and a mean-adjustment was all the manipulation suffered by these items before being used as input variables for classification. The log basis were the decimal one. Notice that there is a more subtle difference between the MLP and the MDA procedures in what concerns the pre-processing of data. The MDA standardizes the input variables one by one. The MLP uses all the information contained in the differences of spread.

The selected items were Fixed Assets, Inventory, Debtors, Creditors, Long Term Debt, Net Worth, Wages and Sales less Operating Expenses. All these variables were present in the original 18 ratios, along with others like Earnings, Value Added, Total Capital Employed and Total Assets which we didn't use in the new approach.

All the log items were mean-adjusted before being presented as input. The overall mean, not the industry-specific one, was used for this. Therefore, the input variables are not just log items but what we call relative positions (see equation 1 on page 3). No correction for  $\delta$  was introduced.

When using the analytical tool for modelling with these eight positions we obtained about 33-34% of correct classifications in the test set. The detailed results are gathered in table 7. It seems clear that, just by avoiding all the entangling pre-processing of data traditional in accounting research and using the log space instead, some improvements in

N.	SEIC Code	Group Name	N. Cases	Correct	N. Cases	Correct
1	14	Building Materials	8	4	23	10
2	32	Metallurgy	11	1	8	1
3	54	Paper and Packing	25	5	21	2
4	68	Chemicals	22	4	23	9
5	19	Electrical	16	5	18	6
6	22	Industrial Plants	8	2	9	0
7	28	Machine Tools	11	5	10	1
8	35	Electronics	44	17	35	19
9	41	Motor Components	17	5	6	5
10	59	Clothing	19	10	23	11
11	61	Wool	7	2	12	0
12	62	Miscellaneous Textiles	11	2	19	2
13	64	Leather	8	1	8	3
14	49	Food Manufacturers	43	32	37	25

Table 8: The best classification results with MLP and 8 log items.

performance can be observed.

Table 8 shows the best classification results the MLP is able to achieve. The improvement, from 33%-34% to 37%-38%, is due to the interruption of training in the optimum for the test set rather than in the optimum for the training set. It is also a consequence of the better generalisation introduced by forcing a reduction in the number of free parameters in the net.

## A.4 Using the Devised Set of Ratios With MDA

We now show the results obtained when using a devised set of ratios to model with analytic tools. These ratios are a free interpretation of the best topology the MLP builds after learning the relation. Table 9 shows the best generalisation achieved. It is around 29%.

Though the results are not impressive by themselves, we must remember that they approach those obtained with 18 ratios.

## A.5 Conclusions

Generalisation results show that under similar conditions little difference exists between the MDA and the MLP results for the particular problem we studied. Clearly, the relation to be modelled must be near linearity. This is a fortunate circumstance. It allows the strict comparing of these tools in a problem for which the linear, analytical, procedure has not been put in a position of disadvantage.

An interesting achievement is the ability displayed by the MLP to deal with simple,

N.	SEIC Code	Group Name	N. Cases	Correct	N. Cases	Correct
1	14	Building Materials	8	2	23	6
2	32	Metallurgy	11	2	8	2
3	54	Paper and Packing	25	4	21	5
4	68	Chemicals	22	4	23	4
5	19	Electrical	16	5	18	2
6	22	Industrial Plants	8	1	9	0
7	28	Machine Tools	11	0	10	2
8	35	Electronics	44	14	35	14
9	41	Motor Components	17	3	6	1
10	59	Clothing	19	10	23	7
11	61	Wool	7	0	12	3
12	62	Miscellaneous Textiles	11	1	19	3
13	64	Leather	8	1	8	1
14	49	Food Manufacturers	43	25	37	19

Table 9: Classification results with MDA and the five discovered ratios plus size.

linear, relations with no losses in generalisation. Algorithms like polynomial fitting would perform badly if required to model a straight line. The MLP did it easily. Hence, the Multi-layer Perceptron emerges as a general-purpose tool, to which we can trust the task of modelling a broad class of relations, ranging from the simple, linear, ones to the most complex ones.

When using 18 ratios and the procedures typical in accounting research — including the extraction of eight factors — both the MDA and MLP generalisation results range from 29% to 30%. The use of eight log items instead of the eighteen transformed and rotated ratios introduces an expected improvement in the generalisation achieved. Both the MLP and the MDA now range from 33% to 34% of correct classifications in the test set. This clearly shows the disadvantages of such techniques based on standard recipes.

By stopping the learning process in the optimal classification for the test set rather than for the learning one a considerable improvement is added to the experiment with eight items. The generalisation is up to 37% - 38%. Naturally, analytic tools like the MDA cannot replicate this experiment. The classification results are summarized in next table.

INPUT	MDA	MLP
18 ratios	29%	30%
8 variables	34%	38%

Finally, the five ratios inspired by the ones formed inside the MLP plus the estimated size, are able to achieve 28% - 29% of correct classification in the test set, which is similar to the performance of the original 18 ratios.

# Bibliography

- [1] J. Aitchison and J. Brown. *The Lognormal Distribution*. Cambridge University press, 1957.
- [2] E. Altman, R. Haldeman, and P. Narayanan. Zeta analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, pages 29–54, June 1977.
- [3] P. Barnes. Methodological implications of non-normally distributed financial ratios. *Journal of Business, Finance and Accounting*, pages 51–62, Spring 1982.
- [4] E. Baum and F. Wilkzek. Supervised learning of probability distributions by neural networks. In American Institute of Physics, editor, *Neural Information Processing Systems - Natural and Synthetic*, pages 52–61, 1987. Denver, Colorado.
- [5] W. Beaver. Financial ratios as predictors of failure. *Journal of Accounting Research*, pages 71–111, 1966. Supplement.
- [6] W. Beaver, H. Kettler, and M. Sholes. The association between market-determined and accounting-determined risk measures. *The Accounting Review*, pages 654–682, October 1970.
- [7] R. Berry and D. Trigueiros. Using the id3 algorithm to interpret the results of financial models. Technical Report SYS, University of East Anglia — Presented in the British Accounting Association Annual Meeting, April 1990, Dundee, Scotland, 1990.
- [8] W. Buijink and M. Jegers. Cross-sectional distributional properties of financial ratios in belgian manufacturing industries: Some empirical evidence. Technical report, University of Antwerp, Belgium, 1984.
- [9] E. Deakin. Distributions of financial accounting ratios: Some empirical evidence. *The Accounting Review*, pages 90–96, January 1977.

- [10] R. Eubank. *Spline Smoothing and Non-Parametric Regression*. Marcel Dekker, Inc., 1988.
- [11] G. Foster. *Financial Statement Analysis*. Prentice-Hall, 1986.
- [12] T. Frecka and W. Hopwood. The effect of outliers on the cross-sectional distributional properties of financial ratios. *The Accounting Review*, pages 115–128, January 1983.
- [13] R. Gibrat. *Les Inegalites Economiques*. Librairie du Recueil Sirey, 1931.
- [14] J. Hopfield. Learning algorithms and probability distributions in feed-forward and feed-back networks. In *Proceedings of the National Academy of Science USA*, pages 8429–8433. USA Academy of Science, 1987. Volume 84.
- [15] J. Horrigan. Some empirical bases of financial ratio analysis. *The Accounting Review*, pages 558–568, July 1965.
- [16] J. Horrigan. The determination of long-term credit standing with financial ratios. *Journal of Accounting Research, Supplement. Empirical Research in Accounting: Select Studies*, pages 44–68, 1966.
- [17] Y. LeCun. Optimal brain damage. In *Neural Information Processing Systems*, volume 1. Morgan Kaufmann, Denver 1990.
- [18] B. Lev and S. Sunder. Methodological issues in the use of financial ratios. *Journal of Accounting and Economics*, pages 187–210, December 1979.
- [19] S. Mcleay. The ratio of means, the mean of ratios and other benchmarks. *Finance, Journal of the French Finance Society*, 7(1):75–93, 1986.
- [20] M. Mozer and P. Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In *Neural Information Processing Systems*, volume 1. Morgan Kaufmann, Denver 1988.
- [21] M. O’Connor. On the usefulness of financial ratios to the investor in common stock. *The Accounting review*, pages 339–352, April 1973.
- [22] J. Quinlan. Discovering rules from large collections of samples — a case study. In D. Michie, editor, *Expert Systems in the Micro Electronic Age*. Edinburgh University Press, 1979.

- [23] D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing*, volume 1. MIT Press, 1986.
- [24] F. M. Silva and L. B. Almeida. Acceleration techniques for the backpropagation algorithm. In *Neural Networks*. EURASIP Workshop, Sesimbra, Portugal, 1990. L. B. Almeida and C. J. Wellekens (Eds.), Springer-Verlag.
- [25] A. Singh and G. Whittington. *Growth Profitability & Valuation*. Cambridge University Press, 1968.
- [26] S. Solla, E. Levin, and M. Fleisher. Accelerated learning in layered neural networks. *Complex Systems*, 2:625–640, 1988.
- [27] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, B-36, 1974.
- [28] M. Stone. Cross-validation: A review. *Math. Operationsforsch. Statist., Ser. Statistics*, 9(1), 1978.
- [29] P. Sudarsanam and R. Taffler. Industrial classification in u.k. capital markets: A test of economic homogeneity. *Applied Economics*, 17:291–308, 1985.
- [30] D. Trigueiros. *Neural Network Based Methods in the Extraction of Knowledge From Accounting and Financial Data*. PhD thesis, Information Systems, University of East Anglia, 1991.
- [31] A. Weigend, B. Huberman, and Rumelhart D. Predicting the future. a connectionist approach. Technical Report PARC-SSL-90-20, Stanford University, PDP Research Group, 1990.
- [32] G. Whittington. Some basic properties of accounting ratios. *Journal of Business, Finance and Accounting*, 7(2):219–232, 1980.