Estimating Directional Intra-Group Loan Volumes from Group Firm's Unconsolidated Financial Statements[⋆]

Matthias Eckerle

University of Augsburg
Chair of Business Taxation

Mark Trede

University of Münster
Institute for Econometrics and Economics Statistics

Robert Ullmann

University of Augsburg
Chair of Business Taxation

—

Working Paper
Early Draft
Please do not cite without the authors' written permission

—

November 4, 2019

Abstract

This is a methodical note. We develop a technique to estimate joint frequencies from marginal frequencies. Since any setting in this regard implies an underdetermined system of equations, we use simulation methods. We then apply the technique to estimate the most likely distribution of joint frequencies for directional intra-group loan volume between firm group entities when only marginal frequencies of overall directional intra-group loan volume per entity are known.

Keywords: Marginal Frequencies, Joint Frequencies, Simulation, Research Design

JEL: C81, C82, G32

[⋆] The authors thank the participants of the 2019 VHB Annual Meeting, the 2019 EAA Annual Congress, the 2019 ADFM Doctoral Seminar on Business Taxation between the Universities of Augsburg, Düsseldorf, Freiburg, and Münster, the 2018 EAA Annual Congress, and the 2018 Doctoral Seminar on Business Taxation between the Universities of Augsburg and Eichstätt-Ingolstadt for their helpful comments.

Email addresses: `matthias.eckerle@wiwi.uni-augsburg.de` (Matthias Eckerle), `mark.trede@uni-muenster.de` (Mark Trede), `robert.ullmann@wiwi.uni-augsburg.de` (Robert Ullmann)

1. Introduction

It is common in economics research that groups of observational units interact among each other only bilaterally. Examples of these settings include business transactions between different enterprises, trade and capital flows between economic regions, and flows of laborers or refugees between countries. Nonetheless, frequencies of bilateral interactions between two observational units (joint frequencies) is often only observable as an aggregated number for each observational unit (marginal frequencies). Unavailability of joint frequencies in such bilateral settings naturally imposes limitations on research.

We develop a technique that allows the estimation of joint frequencies when only marginal frequencies are available in bilateral interactions between $n > 3$ observational units. Such an estimate must in principle rely on the entire equation system derived from the marginal frequencies. Any such equation system is underdetermined, and hence, we must investigate distributional parameters of the solution space of the equation system.

The technique can be applied to a broad range of actual frequency values (e. g. number of people migrating) as well as to monetary values (e. g. trade and capital flows, loan volume flows).

We provide a baseline modification and then two different alterations that can also be combined. In the baseline modification, only the marginal frequencies are used to estimate the joint frequencies. In the first alteration, panel data techniques are applied to increase precision of the estimation, if such panel data is available. In the second alteration, in the case that additional data with impact on the data under observance can be included in the technique to better specify the search process.

The paper proceeds as follows. Chapter 2 defines the applicable research settings. Chapter 3 provides an overview over relevant literature. Chapter 4 describes the estimation technique of the baseline model and both alterations. Chapter 6 shows the demonstration of the three specifications to the practical case of intra-group loan volumes within a simulated set of corporate groups. Chapter 7 summarizes.

2. Relevant Research Settings

Observational units (i) with different but simultaneously occuring attributes (A_i and B_i) are commonly displayed in a contingency table. Such a contingency table visualizes information about the distribution of the marginal frequencies (for each attribute individually) and joint frequencies (between two attributes) taking into account the observational units' attributes A_i and B_i as rows and columns as shown in table 1.

	A_1	A_2	A_3	Sum
B_1	x_{11}	x_{21}	x_{31}	b_1
B_2	x_{12}	x_{22}	x_{32}	b_2
B_3	x_{13}	x_{23}	x_{33}	b_3
Sum	a_1	a_2	a_3	

Table 1: Illustration of a contingency table of three observational units i with each two attributes A_i and B_i

In the case of a contingency table of two observational units with two attributes and respective known amounts of a_i and b_j for each attribute (i. e. creating a $2x2$ matrix within the contingency table), the frequencies can be calculated easily or probabilities estimated quite precisely by using the Theorem of Bayes (1763). Also in the case of the observational units with two attributes, the system remains solvable. However, most data used in research contains more than three observational units. In this case there is no precise mathematical method to estimate the joint frequencies within the contingency table.

Our technique is able to analyze given vectors of attributes' marginal frequencies (a_i and b_j in table 1) of $n > 3$ observational units. It can handle natural numbers and zero values as marginal frequencies. The joint frequencies are also limited to this number space.

We argue that an extension to the positive rational number space $\mathbb{Q}_0^+$ is not necessary, as positive rational numbers (e. g. amounts of money) can be rounded or extended to natural numbers without relevant loss of information. An extension to the integer number space $\mathbb{Z}_0$ (containing also negative numbers) is not done here as we only allow frequencies for which the direction is known and which are bilateral in nature.

Furthermore, we limit the possible interactions between observational units in a way that interactions with itself is not possible. Therefor, the diagonal items of the contingency table are all equal to zero.

The baseline model of our technique provides an estimated solution of the contingency table by distributing the marginal frequency into the joint frequencies in a one-period setting. We therefor use a search process which is described in detail in chapter 4.2. Our technique searches a given amount of possible solutions for the contingency table and thereof calculates the estimated solution.

To also handle panel data of marginal frequencies, we extend our technique for a multi-period setting. The technique optimizes the estimated solution considering all periods. This is described in detail in chapter 4.3. As second alteration, we allow additional data to be included in the search process. This data shall have a verified or anecdotal deduced impact on the data under observance and influences the search process. This alteration is described in chapter 4.4.

3. Literature

The first study to our knowledge which aims at estimating joint frequencies within a contingency table was made by Deming and Stephan (1940). They deal with data where the joint frequencies for a small sample of the whole observed population are available. Their approach extrapolates the known sample's joint frequencies to the population out of the population's marginal frequencies. In the more recent literature similar approaches of additional information about the data were mainly used in marketing related research.

Most closely related, Putler et al. (1996) use a Bayesian approach estimating the target market potential having only limited geodemographic information. They use variables' correlation of the Public Use Microdata Sample (PUMS) data, representing 5 % of the total data, as additional information to estimate the joint frequencies of the whole census data to optimize marketing-related budget optimization.

Romeo (2005) develops an approach for estimating joint frequencies of demographic characteristics corresponding to market areas for individual retail stores. He uses marginal frequencies available for a bigger geographic area to estimate joint frequencies for the retail store's market area.

The mentioned literature has in common that specific knowledge about the distributions and/or the variables is available and that only one period in time is observed. Hence, we contribute to the literature by creating a new estimation technique that is applicable in situations when only marginal frequencies are available for a one-period or a multiple-period setting. In addition and if useful, the preciseness of our technique can be specified by adding external data without a direct relation to the marginal frequencies is available.

4. Estimation Technique

4.1. Model Setup

Estimating joint frequencies from marginal frequencies implies an underlying equations system that consists of a matrix A, with rows i and columns j. Each i and j, respectively, represents an observational unit while $a_{i,j}$ gives the joint frequency of the bilateral interaction between i and j. Matrix A is quadratic with rows and columns representing the same observational units, as we only consider within group bilateral interactions. The equation system also has two limitation vectors $\vec{x}$ with entries x_i and $\vec{z}$ with entries z_j. Limitation vectors show the amount of marginal frequencies of each i and j, respectively, and thus represent the system boundaries.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & a_{24} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & a_{34} & \dots & a_{3n} \\ a_{41} & a_{42} & a_{43} & a_{44} & \dots & a_{4n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & a_{n3} & a_{n4} & \dots & a_{n,n} \end{pmatrix}; \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{pmatrix}; \vec{z} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ \vdots \\ z_n \end{pmatrix} \tag{1}$$

The equation system can also be written in contingency table notation:

a_{11}	a_{12}	a_{13}	a_{14}	$\dots$	a_{1n}	x_1
a_{21}	a_{22}	a_{23}	a_{24}	$\dots$	a_{2n}	x_2
a_{31}	a_{32}	a_{33}	a_{34}	$\dots$	a_{3n}	x_3
a_{41}	a_{42}	a_{43}	a_{44}	$\dots$	a_{4n}	x_4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		$\vdots$
a_{n1}	a_{n2}	a_{n3}	a_{n4}	$\dots$	$a_{n,n}$	x_n
z_1	z_2	z_3	z_4	$\dots$	z_n	

Table 2: Basic equation system

Each row i and each column j results in one equation in the equation system. The sum of all items $a_{i\cdot}$ for a given column i in the matrix A is equal to the x_i and the sum of all items $a_{\cdot j}$ in the matrix A has to be equal with the respective amount of z_j. Hence:

$$A \cdot \mathbb{1}_n = \vec{x}|_{\vec{z}^T} \tag{2}$$

$$A^T \cdot \mathbb{1}_n = \vec{z}|_{\vec{x}^T} \tag{3}$$

Furthermore, the sum of all items x_i and z_j, respectively, within the two limitation vectors $\vec{x}$ and $\vec{z}$ has to be equal. The sum of all items in each limitation vector also represents the amount of joint frequencies to be estimated in total.

$$\sum_{k=1}^{l} x_k = \sum_{k=1}^{l} z_k \tag{4}$$

Note that both the marginal frequencies x_i and z_j present maximum thresholds for the joint frequency between the two observational units i and j, hence for $a_{i,j}$. This can be written as follows:

$$a_{i,j} \in [0, min(x_i, z_j)] \tag{5}$$

The specific setting under investigation allows two restrictions for the estimation of joint frequencies of matrix A as already mentioned in chapter 2. First, given that we investigate frequencies, the equation system can contain only natural numbers including zero.

$$a_{i,j} \in \mathbb{N}_0 \tag{6}$$

Second, observational units are not able to interact with themselves. Thus, all items $a_{i,j}$ with $i = j$ are equal to zero.

$$a_{i,j} = 0|_{i=j} \tag{7}$$

An determined equation system would contain of the same amount of equations and unknown variables. This equation system is obviously underdetermined, containing $(i * j) - i$ variables $a_{i,j}$ and only $i + j$ equations. An underdetermined equation system can either have zero or infinitely many solutions (Datta, 2010). However, by including the restrictions in the equation system that are given by the specific setting under investigation, we are able to limit the equation system's solution space to a finite number of solutions. [1]

[1] The restriction to only natural numbers and zero $\mathbb{N}_0$ creates a subset out of the infinite set of all numbers. By including a maximum value for each item $a_{i,j}$ to be within $[0, min(x_{i\cdot}, z_{\cdot j})]$ we limit the subset of possible solutions to a finite number.

4.2. Baseline Modification

4.2.1. Drawing the Estimators

We naturally assume that marginal frequencies are available and joint frequencies are not. Consequently, we begin by using only the information given by the two limitation vectors $\vec{x}$ and $\vec{z}$. Generally speaking, the simulation uses a specific random search process to identify estimators of matrix A, estimators $\hat{A}_m$ with m being the amount of estimates required. Out of the estimators $\hat{A}_m$, the search process calculates the estimated solution. Since each estimator $\hat{A}_m$ contains a valid solution to the equation system as described by equations (1) to (5) and taking the restrictions of equations (6) and (7) into account, the relative probability of the randomly found solution in each matrix $\hat{A}_m$ has an impact on the estimated solution to be calculated. We argue that the higher the probability of an estimator $\hat{A}_m$ the more often it is found. The respective shape of estimator $\hat{A}_m$ influences the estimated solution to a higher extent than more unlikely estimators.

The random search process for each estimator $\hat{A}_m$ runs row by row, using a randomly picked order within the rows i. Each frequency of the respective marginal frequency x_i for each row i is allocated in a single allocation sequence. To clarify this approach, if row i contains a marginal frequency of $x_i = 250$ the search process will perform 250 allocation sequences for this row. In each allocation sequence one frequency will be allocated to the respective items $a_{i\cdot}$.

Each allocation sequence uses a multinomial distribution to allocate the individual frequencies. The multinomial distribution uses the compilation of marginal frequencies of the input vector $\vec{z}$ (i. e. the relative distribution of marginal frequencies over the observational units j) to distribute the relative probability over the observational units j. Hence, items $a_{i,j}$ with greater respective limit vector items z_j are more likely to be drawn within an allocation sequence. The multinomial distribution is shown in equation (8).

$$p_i = \left(\frac{a_{i+1}}{\sum_{k=i+1}^{n} a_k}, \ \ldots, \ \frac{a_n}{\sum_{k=i+1}^{n} a_k} \right); \|p_i\| = 1 \tag{8}$$

After each allocation sequence for one frequency, the input vectors $\vec{x}$ and $\vec{z}$ are updated. Input vector item x_i is decreased by 1. Also, the respective input vector item z_j of the receiving item $a_{i,j}$ is decreased by 1. Then, the multinomial distribution for the following allocation sequence is recalculated with the new input vector $\vec{z}$. The allocation sequences continue until the last frequency of input vector item x_i is allocated to the items $a_{i\cdot}$. After one row i is finished, the next row $i+1$ is treated simultaneously.

If the treated items $a_{i\cdot}$ of row i violate one of the mentioned restrictions, the respective items $a_{i\cdot}$, and the input vectors $\vec{x}$ and $\vec{z}$ are reset and drawn again. Already drawn rows i are not affected by this reset.

The last remaining row i is self-releasing. Either the items $a_{i,j}$ fit the respective remaining input vector items z_j without violating any restrictions, or not. If the items fit to each other, the drawing of the items $\hat{a}_{i,j}$ is saved as one respective estimator $\hat{A}_m$. If they do not fit, the drawing of this respective estimator $\hat{A}_m$ is reset in total. Already drawn estimators are not affected by this reset.

The estimators $\hat{A}_m$ are drawn and saved until the required amount m of such estimators $\hat{A}_m$ is found.

4.2.2. Calculating the Estimated Solution $\hat{S}$

After the required amount of m estimators $\hat{A}_m$ is found, the search process calculates the estimated solution $\hat{S}$. First, the respective median values for all items $\hat{a}_{i,j,m}$ are computed. The median values are then saved as median items $\hat{med}_{i,j}$ in matrix $\hat{Med}$.

Second, the search process compares all median items $\hat{med}_{i,j}$ in matrix $\hat{Med}$ with the respective item $\hat{a}_{i,j}$ for each estimator $\hat{A}_m$. For each estimator $\hat{A}_m$, a sum of distances between the median items $\hat{med}_{i,j}$ and its items $\hat{a}_{i,j}$ is computed.

Third, the estimator $\hat{A}_m$ with the least sum of distances to $\hat{Med}$ is denoted as the estimated solution $\hat{S}$.[2]

4.3. Alteration 1: Panel Data Analysis

Assume now that we have panel data available. Since the majority of items $a_{i,j}$ should not change materially over periods, panel data reduces the relevant solution space of the system under observation.

For panel data analysis, the search process first treats each period y individually as described in chapter 4.2. It creates m estimators $\hat{A}_{M,y}$ and the respecting estimated solution $\hat{S}_y$ for each individual period y.

For combining multiple periods and to find results, the search process cuts out 30 % of the

[2] The used characteristics of the estimated solution $\hat{S}$ are still work in progress. We will examine other sums of distances and other location parameters besides $\hat{Med}$.

m estimators $\hat{A}_{m,y}$[3] around the calculated estimated solution $\hat{S}_y$ for each period. Within these estimators $\hat{A}_{m,y}$, the search process is looking for an estimated panel solution (hereafter: estimated panel solution $\hat{P}$) with the least sum of distances to the combination of respective single-period estimated solutions $\hat{S}_y$. The best fitting estimated panel solution $\hat{P}$ is taken as the estimation result which fulfill all restrictions considering all periods for the multi-period system.

4.4. Alteration 2: Correction Data Analysis

This alteration is still work in progress. However, we want to give an overview about its modification and our expectations.

As can be seen in chapter 3, previous attempts to estimate joint frequencies from marginal frequencies use additional data sets to precise their estimation approaches. These additional data sets include specific knowledge about the (expected) distribution of the joint frequencies under observation. Our technique also provides the possibility to include additional data but without limiting its origin and/or connection to the frequencies under observation.[4]

Since the additional data has a corrective effect on our search process, we call it correction data.

Out of the correction data, the correction distribution cf_i is determined. It is another drawing probability beside p_i reflecting the correction data's relative proportions. Standardization of the correction data by hand is unnecessary and performed within the search process to the correction distribution cf_i. The correction distribution items $b_{i\cdot}$ are automatically calculated that their sum is equal to one.

$$cf_i = \left(b_{i,i+1},\ \ldots,\ b_{i,n}\right); \|cf_i\| = 1 \tag{9}$$

After the automatic standardization, the correction distribution cf_i is then combined with the general multinomial distribution of p_i from equation (8) with a weighting variable k.

$$c_i = p_i + k \cdot cf_i \tag{10}$$

To use this combination of multinomial distribution and correction factor distribution c_i (hereafter: overall probability c_i), it is necessary to normalize c_i as shown in equation (11) so that the

[3] The percentage rate of 30 % is chosen intuitively. It still has to be checked for fitting.

[4] Possible data would be, for example, macroeconomic information, but also already known information about connections in the data set itself.

sum of probabilities is again equal to one. This is mandatory to meet the axioms of Kolmogorov (1933) regarding probability distributions:

$$c_i = \frac{\left(\frac{a_{i+1}}{\sum_{h=i+1}^{n} a_h} + k \cdot b_{i,i+1},\ \ldots,\ \frac{a_n}{\sum_{h=i+1}^{n} a_h} + k \cdot b_{i,n}\right)}{\|\left(\frac{a_{i+1}}{\sum_{h=i+1}^{n} a_h} + k \cdot b_{i,i+1},\ \ldots,\ \frac{a_n}{\sum_{h=i+1}^{n} a_h} + k \cdot b_{i,n}\right)\|}; \qquad (11)$$

$$\|c_i\| = 1; k \in \mathbb{R}_0^+$$

With the weighting variable k, the intensity of the correction distribution cf_i on the overall probability of c_i can be modified. This leads to the two extreme cases of $\lim_{k \to 0} c_i = p_i$ and $\lim_{k \to \infty} c_i = cf_i$. It becomes clear that an excessive use of the correction distribution cf_i by using a high value of k suppresses the effect of the multinomial distribution p_i and only mirrors the correction distribution cf_i - which cannot be the intention at all.

As already mentioned, this alteration is still work in progress. We have to examine a feasible value for the weighting variable k as well as necessary characteristics for possible correction data.

5. Treatment of Errors

In theory, we assume that the marginal frequencies to be analyzed are free of errors. Nevertheless, we are aware that several sources of errors can occur in real data. Such sources can be unsystematic errors like typos or transmission errors, or systematic errors like over- or underreporting.

Analyzing these sources of errors is still work in progress.

6. Demonstration of the Technique

6.1. Context

To show the functionality of our technique, we use it on sets of simulated corporate groups. We examine the hypothetical intra-group loan volumes from group firm's unconsolidated financial statements. Each corporate group contains of several group entities, i. e. the group firms. The entities are able to lend money to each other. We can observe the loan volumes in the firm's individual financial statements as liabilities and receivables to/from related parties. However, we are only able to see the frequencies on a firm level. The joint frequencies within the group stay unobservable. We argue that an estimation of such joint frequencies of intra-group loan volumes is of interest to researchers, e. g. to investigate corporate debt shifting.

6.2. Literature

The findings of Modigliani and Miller (1958) that the financing structure of a company has no influence on its value only holds in a world where, among other things, no taxes exist. Verschueren and Deloof (2006) find evidence that intra-group financing indeed affects the company's leverage. Group entities that borrow internally seem to have higher investment, leverage, and return on equity (ROE) than other firms (Buchuk et al., 2014). This may be because corporate headquarters decide which projects are financed and thus form a group-optimized financial structure for value creation (Stein, 1997).

From a tax point of view, the Organisation for Economic Co-operation and Development (OECD, 2013) claims debt shifting one of the two main vehicles of tax avoidance, besides transfer pricing. Using a German data set, Buettner and Wamser (2013) confirm that internal debt is used more by multinationals with affiliates in low-tax countries and increases with the spread between the host-country tax rate and the lowest tax rate among all affiliates. According to Hanlon and Heitzman (2010), the existence of tax havens have an impact on debt location. Huizinga et al. (2008) find evidence that a foreign subsidiary's capital structure reflects local corporate tax rates as well as tax rate differences. Hopland et al. (2018) investigate the flexibility of income shifting under losses but cannot find empirical evidence for internal debt.

However, none of the mentioned literature uses complete bilateral data sets containing joint frequencies. We argue that the availability of joint frequencies for intra-group loan volumes would have been useful for observing financing effects. E. g. Hopland et al. (2018) can only observe the Norwegian company and the respective foreign related party. The observation of the total group is not possible. Furthermore, in current research, both Amberger et al. (2019) and De Simone et al. (2018) investigate bilateral intra-group topics related with repatriation taxes. We argue that all three mentioned papers could benefit by the use of our technique.

6.3. Hypothetical Setting

We create sets of 100 simulated corporate groups by using a simulation algorithm.[5] Hence, we are able to observe both the marginal and joint frequencies of each group.

[5] The operation method of the simulation algorithm is illustrated in Appendix A.

Each group contains randomly of 4 to 10 firms. The firms are able to lend money to each other firm but not to itself. Each bilateral connection between two firms can have a maximum amount of 250. It is reasonable that the intra-group loans structure is designed efficiently by the group's headquarter (Stein, 1997). So, only few great amounts of loans exist within one group and many entities do not have any loans between each other. Therefor, the simulation algorithm uses a specific distribution to select one specific amount of loan: the probability for no loan (i. e. 0) is 40 %. For the amount of 1 to 250, the simulation algorithm creates five equally sized buckets, each containing 50 values. Each bucket has an individual probability to be chosen. Within the buckets, all values are equally distributed. The probabilities for the five buckets are 20 %, 5 %, 5 %, 10 %, and 20 %.

The direction of the loan is definite (liabilities for the lenders, receivables for the borrowers), hence all values are positive. We also allow that several entities give loans reciprocally to one another at the same time. With these settings we ensure that all restrictions made in chapter 4.1 are met.

In the following we first examine the baseline model by analyzing one year under observation. Second, we analyze a three-year panel data set. Third, since the second alteration of correction data is still work in progress, we only give a short example of one group under observation.

6.4. Testing a Baseline Modification Analysis

We use our search process and draw $m = 100$ estimators $\hat{A}_m$ for each simulated group. Each simulated solution $\hat{S}$ is then compared with it hypothetical solution, created in the simulation process. We calculate the correlations of Pearson's r, Kendall's τ, and Spearman's ρ for each group. The correlations are shown in boxplots in figure 1.

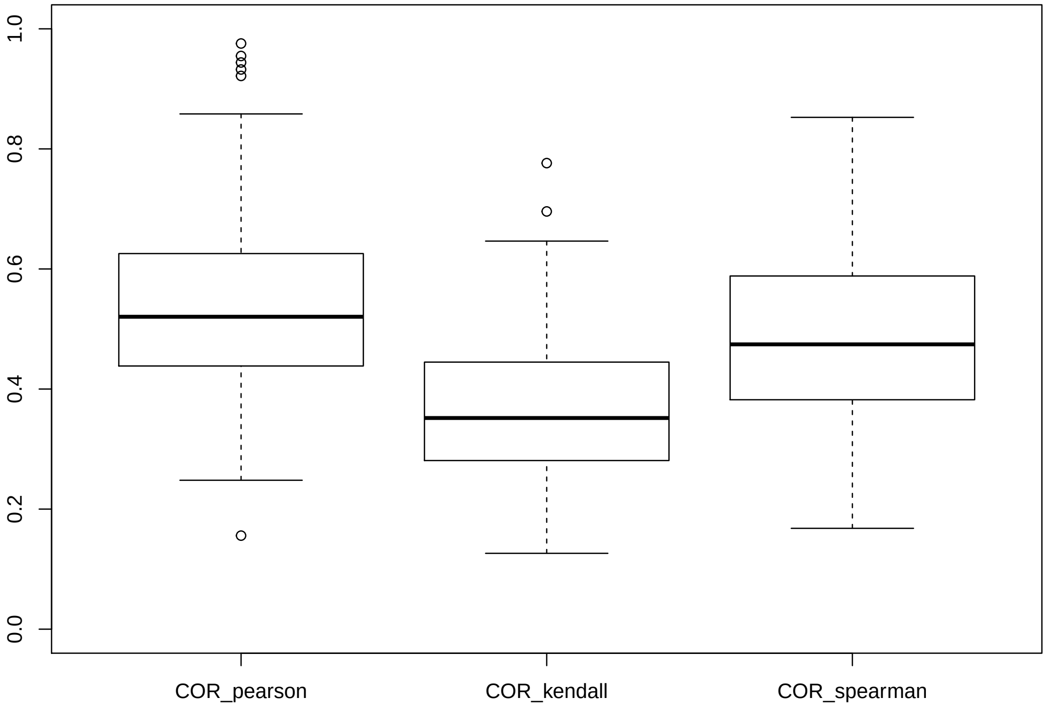

Figure 1: Boxplots of 100 correlations between estimated solution $\hat{S}$ with $m = 100$ and its respective hypothetical solution

Table 3 shows the correlation analysis in more detail.

	Min	25%Q	Median	Mean	75%Q	Max
Pearson	0.1558	0.4384	0.5204	0.5442	0.6244	0.9756
Kendall	0.1263	0.2813	0.3517	0.3693	0.4436	0.7763
Spearman	0.1680	0.3828	0.4744	0.4885	0.5871	0.8525

Table 3: Descriptive statistics of the correlation analysis with $m = 100$

Cohen (1988) suggests that, as a rule of thumb, a Pearson's $r > |0.1|$ shows a small, $r > |0.3|$ a medium, and $r > |0.5|$ a large effect size. Regarding the correlations at hand, all 100 estimated solutions $\hat{S}$ have at least a small effect size and over 50 % a large effect size. Nevertheless, two correlations between estimated solutions $\hat{S}$ and their hypothetical solutions is statistical insignificant with $p > 0.1$. Regarding Kendall's τ (Spearman's ρ), we find a similar minimum but four (seven) statistical insignicant correlations with $p > 0.1$.

We repeat the analysis, now searching $m = 500$ estimators $\hat{A}_m$ for each group. The correlations of Pearson's r, Kendall's τ, and Spearman's ρ between each estimated solution $\hat{S}$ and the respective hypothetical solution are shown in boxplots in figure 2

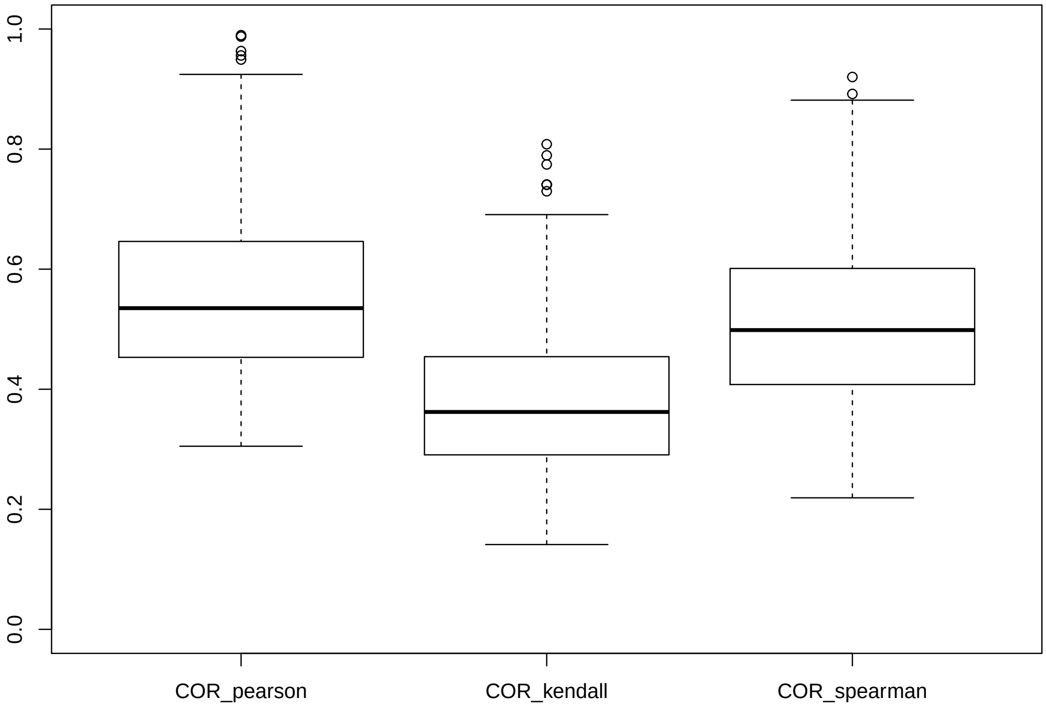

Figure 2: Boxplots of 100 correlations between estimated solution $\hat{S}$ with $m = 500$ and its respective hypothetical solution

Table 4 shows the correlation analysis in more detail.

	Min	25%Q	Median	Mean	75%Q	Max
Pearson	0.3050	0.4534	0.5351	0.5682	0.6461	0.9897
Kendall	0.1413	0.2908	0.3621	0.3953	0.4530	0.8081
Spearman	0.2191	0.4082	0.4985	0.5205	0.6009	0.9201

Table 4: Descriptive statistics of the correlation analysis with $m = 500$

For Pearson's r, we now find correlations with at least a medium effect size with only one correlation being statistically insignificant with $p > 0.1$. Regarding Kendall's τ and Spearman's ρ, we find a higher minimum compared to the estimation with $m = 100$ but five statistical insignicant correlations with $p > 0.1$.

We see that the amount of m is an important setting for the accuracy of our technique. Since the estimated solution $\hat{S}$ is the best fitting estimator $\hat{A}_m$ for each group, the larger selection of possibilities increases or at least does not harm the correlation.

6.5. Testing a Panel Data Analysis

We now extent the single-year hypothetical example as shown in chapter 6.4 to a three-year hypothetical example. We simulate 100 groups as already described and use the same simulated group for three years, pretending that there is no change in the intra-group financing over the three-year period.

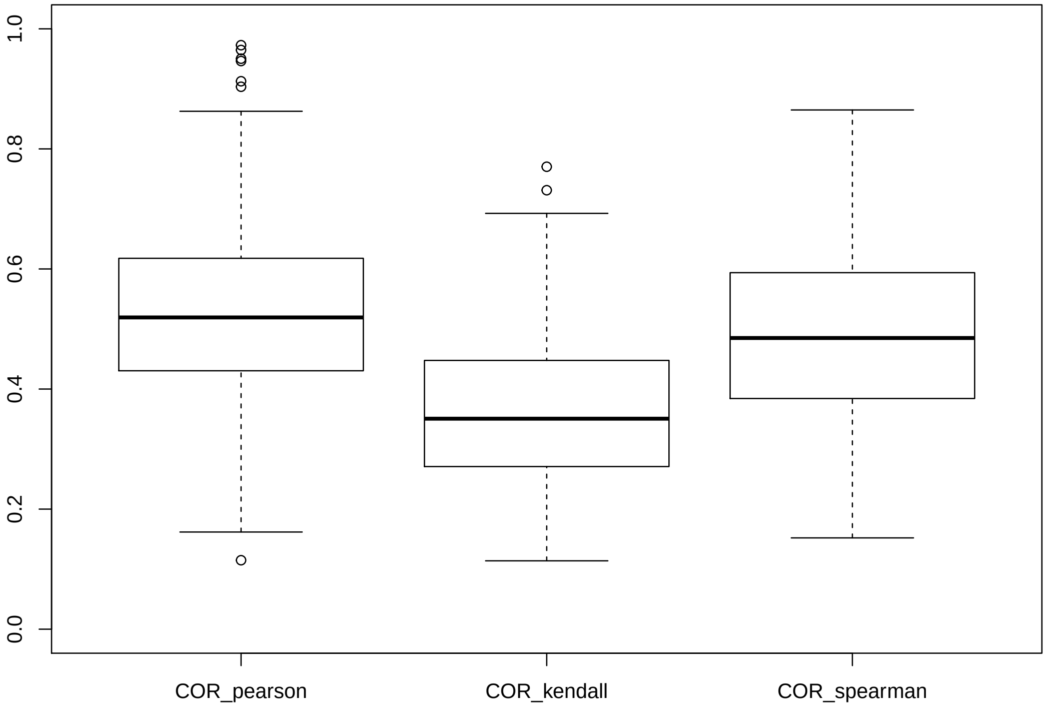

Figure 3: Boxplots of 100 correlations between estimated panel solution $\hat{P}$ with $m = 100$ and its respective hypothetical solution

Table 4 shows the correlation analysis in more detail.

	Min	25%Q	Median	Mean	75%Q	Max
Pearson	0.1149	0.4309	0.5193	0.5376	0.6170	0.9728
Kendall	0.1139	0.2723	0.3506	0.3678	0.4469	0.7703
Spearman	0.1521	0.3843	0.4850	0.4912	0.5934	0.8649

Table 5: Descriptive statistics of the correlation analysis with $m = 100$

In this setting, we find statistical insignificant correlation in only one case with $p > 0.1$ for each Person's r, Kendall's τ, and Spearman's ρ.

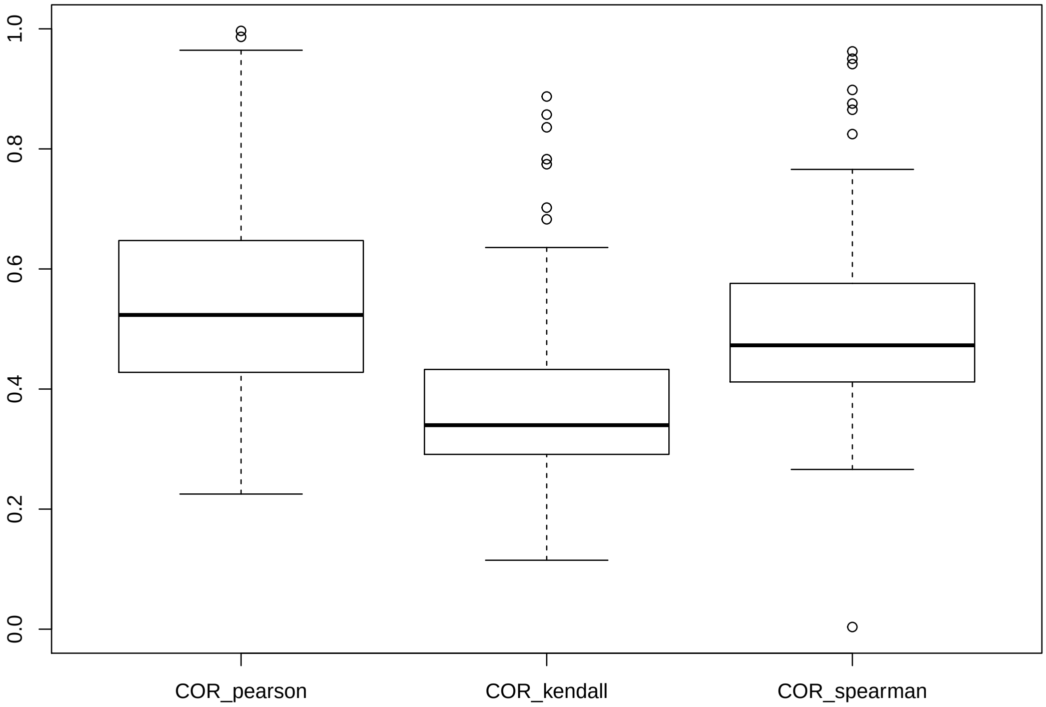

Figure 4: Boxplots of 100 correlations between estimated panel solution $\hat{P}$ with $m = 500$ and its respective hypothetical solution

Table 4 shows the correlation analysis in more detail.

	Min	25%Q	Median	Mean	75%Q	Max
Pearson	0.2251	0.4291	0.5234	0.5518	0.6460	0.9966
Kendall	0.1149	0.2912	0.3398	0.3795	0.4317	0.8873
Spearman	0.0036	0.4125	0.4728	0.5041	0.5751	0.9624

Table 6: Descriptive statistics of the correlation analysis with $m = 500$

Also in this setting, we find statistical insignificant correlation in only one case with $p > 0.1$ for each Person's r, Kendall's τ, and Spearman's ρ. Furthermore, we observe a doubling of the minimum correlation in Pearson's r while correlations of Kendall's τ and Spearman's ρ remain relatively stable (ignoring the outlier for Spearman's ρ).

In conclusion we see our results for the baseline modification and the panel data alteration as first indices that our technology is in the majority of cases able to find estimated (panel) solutions $\hat{S}$ ($\hat{P}$) with statistically significant correlation to the respective hypothetical solution.

6.6. Testing a Correction Data Analysis

As mentioned before, the correction data alteration is still work in progress. To show that a well fitting correction distribution is able to precise the estimated solution $\hat{S}$ we use only one group for a short demonstration. The group is illustrated in Appendix A. As correction data we use the hypothetical solution itself as shown in table A.7.

While only using the baseline model of our technique, we estimate $m = 100$ estimators $\hat{A}_m$. The correlations of Pearson's r, Kendall's τ, and Spearman's ρ of the 100 estimators $\hat{A}_m$ with the hypothetical solution are shown in figure 5.[6]

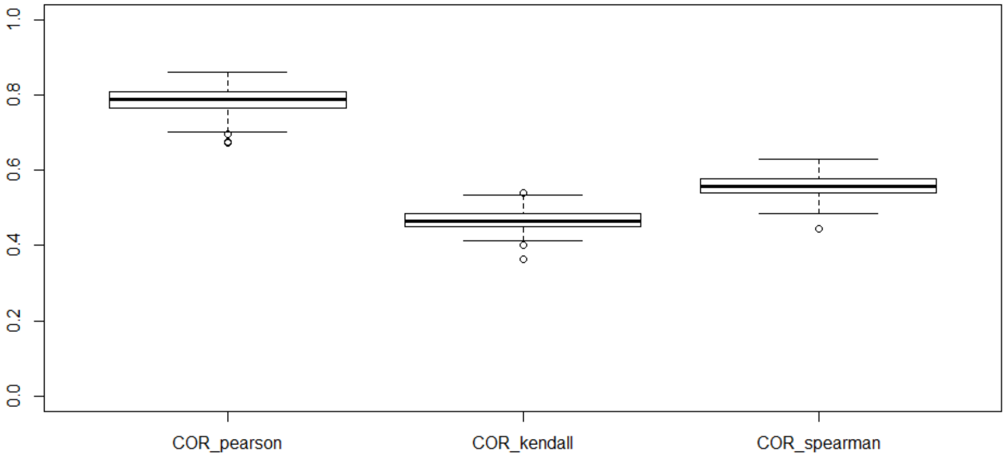

Figure 5: Boxplots of 100 Simulations' Correlations with a correction distribution cf_i included with $k = 10$

[6] Be aware: we now correlate the individual estimators $\hat{A}_m$ with the hypothetical solution and not the estimated solution $\hat{S}$ as done above.

We now include the correction distribution cf_i and use a (very high) weighting factor of $k = 10$ to draw $m = 100$ estimators $\hat{A}_,$. We report corresponding boxplots of Pearson's r, Kendall's τ, and Spearman's ρ in figure 6 of the $m = 100$ estimatiors $\hat{A}_m$ and the hypothetical solution of the hypothetical group as shown in table A.7.

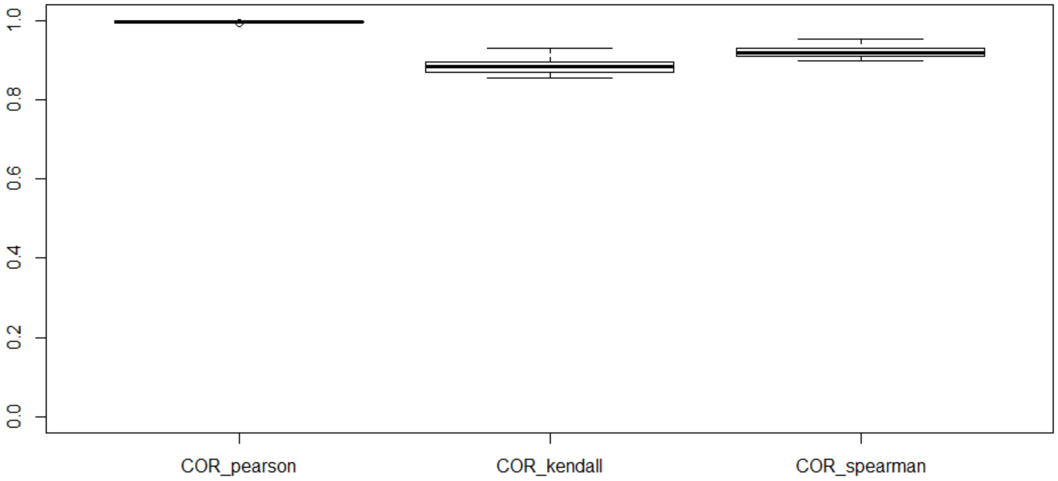

Figure 6: Boxplots of 100 Simulations' Correlations with a correction distribution cf_i included with $k = 10$

This result give a first clue that a well fitting correction distribution cf_i leads to a more accurate estimation results. Regarding Pearson's r, the correlation of all 100 estimated solutions $\hat{S}$ and the hypothetical solution is almost equal to one.

Nevertheless, it is also shown that an overvaluation of the correction distribution cf_i leads to estimators $\hat{A}_m$ and therefor an estimated solution $\hat{S}$ with only limited validity. Therefore, further work has to be done to evaluate the reasonable limitation of the correction data and the effect of a misleading correction distribution cf_i.

7. Summary

We develop a technique to estimate joint frequencies from marginal frequencies in a contingency table of observations between two parties. We demonstrate our technique specifically by estimating data of simulated group firms' unconsolidated financial statements.

We argue that the outcome of our work will result in a valuable empirical construct for future tax and accounting research, which ultimately enables much more specific assessment of country pair tax characteristics (per year) than what researchers are able to do with empirical constructs available today.

In addition to our specific demonstration relying on the intra-group loan volumes, we also argue that the technique developed in this paper can be used for similar settings in other areas of research as well. This would include all potential systems with bilateral relations between observations where only the total of the respective observations are known.

Apart from still necessary further development of the technique itself, we need to examine what happens if errors are included in the data under observance. Furthermore, it is also necessary to theoretically and empirically investigate the array of theoretical constructs that can potentially be captured by the empirical construct of joint frequencies in pairwise settings.

Appendix A. Structure of the Group Simulation Algorithm

This appendix illustrated the operation method of the used simulation algorithm which simulates the hypothetical groups.

First, the simulation algorithm creats a quadratic matrix. This matrix is filled with values by the probability as mentioned above. After this step the diagonal items are set to zero to meet the restriction (7).

	A	B	C	D	E	F	G	Sum
A	0	100	150	75	50	0	0	375
B	15	0	0	0	0	25	0	40
C	0	50	0	50	0	10	0	110
D	10	0	0	0	0	0	0	10
E	0	0	0	0	0	0	0	0
F	0	5	30	0	0	0	0	35
G	0	0	0	100	50	25	0	175
Sum	25	155	180	225	100	60	0	745

Table A.7: Structured illustration of one simulated group

The already shown input vectors $\vec{x}$ and $\vec{z}$ are then computed and saved separately.

$$\vec{x} = \begin{pmatrix} 375 \\ 40 \\ 110 \\ 10 \\ 0 \\ 35 \\ 175 \end{pmatrix} ; \vec{z} = \begin{pmatrix} 25 \\ 155 \\ 180 \\ 225 \\ 100 \\ 60 \\ 0 \end{pmatrix} \tag{A.1}$$

The following figure A.7 shows how the group and its loan volumes would look like:

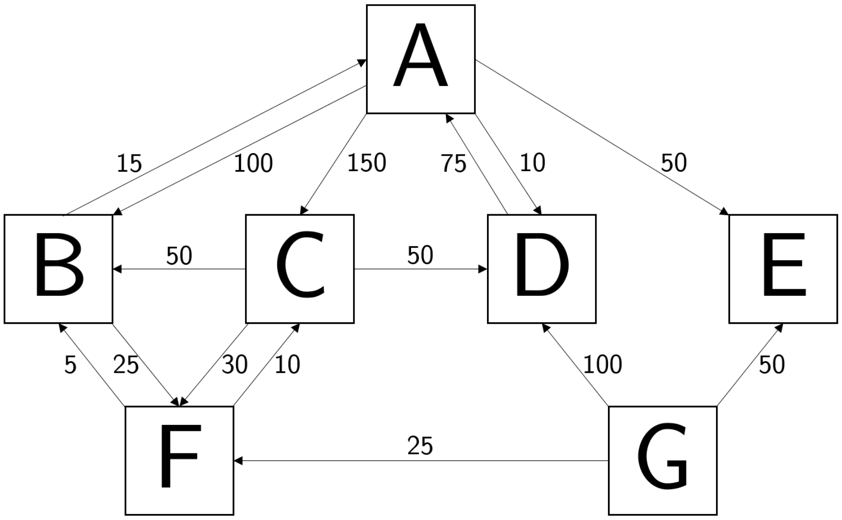

Figure A.7: Illustration of the hypothetical group's loan volumes

References

Amberger, H., Markle, K., and Samuel, D. M. P. (2019). Repatriation Taxes, Internal Agency Conflicts, and Subsidiary-Level Investment Efficiency. *Working Paper*.

Bayes, T. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418.

Buchuk, D., Larrain, B., Muñoz, F., and Urzúa I., F. (2014). The Internal Capital Markets of Business Groups: Evidence from Intra-Group Loans. *Journal of Financial Economics*, 112(2):190–212.

Buettner, T. and Wamser, G. (2013). Internal Debt and Multinational Profit Shifting: Empirical Evidence from Firm-Level Panel Data. *National Tax Journal*, 66(1):63–96.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Erlbaum, Hillsdale, 2. edition.

Datta, B. N. (2010). *Numerical Linear Algebra and Applications*. SIAM, Philadelphia, 2. edition.

De Simone, L., Klassen, K., and Seidman, J. K. (2018). The Effect of Income-Shifting Aggressiveness on Corporate Investment. *Working Paper*.

Deming, W. E. and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, 11(4):427–444.

Hanlon, M. and Heitzman, S. (2010). A Review of Tax Research. *Journal of Accounting and Economics*, 50(2-3):127–178.

Hopland, A. O., Lisowsky, P., Mardan, M., and Schindler, D. (2018). Flexibility in Income Shifting under Losses. *The Accounting Review*, 93(3):163–183.

Huizinga, H., Laeven, L., and Nicodeme, G. (2008). Capital Structure and International Debt Shifting. *Journal of Financial Economics*, 88(1):80–118.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsberechnung*. Springer, Berlin.

Modigliani, F. and Miller, M. H. (1958). The Cost of Capital, Corporation Finance and the Theory of Investment. *The American Economic Review*, 48(3):261–297.

OECD (2013). *Addressing Base Erosion and Profit Shifting.* OECD Publishing, Paris.

Putler, D. S., Kalyanam, K., and Hodges, J. S. (1996). A Bayesian Approach for Estimating Target Market Potential with Limited Geodemographic Information. *Journal of Marketing Research*, 33(2):134–149.

Romeo, C. J. (2005). Estimating Discrete Joint Probability Distributions for Demographic Characteristics at the Store Level Given Store Level Marginal Distributions and a City-Wide Joint Distribution. *Quantitative Marketing and Economics*, 3(1):71–93.

Stein, J. C. (1997). Internal Capital Markets and the Competition for Corporate Resources. *The Journal of Finance*, 52(1):111–133.

Verschueren, I. and Deloof, M. (2006). How Does Intragroup Financing Affect Leverage? Belgian Evidence. *Journal of Accounting, Auditing & Finance*, 21(1):83–108.