Geographic Stock Returns Indexes*

Bernard Dumas, INSEAD, University of Torino, NBER and CEPR
Tymur Gabuniya
Richard C. Marston, Wharton School of the University of Pennsylvania and NBER

February 3, 2020

Abstract

In most statistical studies of international stock returns, a firm is included in a country's index if it is listed in the stock market of that country or if its headquarters are located in that country. This classification scheme ignores the operations of the firm. The distinction between domicile and place of business is becoming more and more relevant as a growing number of firms have activities abroad. We propose, instead, to allocate firms to "geographic zones" according to the place where they conduct business and we re-measure zone returns in keeping with that allocation scheme. We synthesize purely domestic firms from all firms in the dataset, where a domestic firm is one that sales to one zone mostly.

*Dumas's work received the support of a grant from the INSEAD Research Fund. Both authors received funding on this project from the INSEAD-Wharton Alliance. We are grateful to Humberto Gomez, a Master of Science student at the University of Lausanne, who set up the MatLab programs for an earlier draft of this paper. Further research assistance was provided by Fiodor Gorokhovik and Pierre Poulain. We thank Olivier Piette of WVB, who generously provided the data, and Winston Dou who suggested subsampling to us.

> "What is changing is that corporations are becoming more and more global in their business activities through increased exports and cross-border M&A." Diermeier and Solnik (2001)

What does it mean for a firm to belong to a country? In most statistical studies of international stock returns, a firm is classified as belonging to a country (i.e., included in the country index) if it is listed in the stock market of that country or if its headquarters are located in that country. This classification scheme ignores the operations of the firm. The distinction between place of stock trading and place of business is becoming more and more relevant as a growing number of firms have activities abroad. For instance, the index of the Amsterdam stock exchange, where many "Dutch" multinationals are traded, is not representative of the risks and returns attached to investing in operations physically taking place in The Netherlands.

We propose, instead, to allocate firms' stock returns to "geographic zones" *according to the place where they conduct business* and we re-measure zone returns in keeping with that allocation scheme. In other words, we synthesize purely domestic firms from all firms in the dataset, where a domestic firm is one that sales to one zone mostly. Such a re-construction should be relevant for at least four purposes. First, financial analysts making an investment decision often do so because they want to take a view concerning the economy and growth prospects of a zone. An analyst, or an investor he or she advises, may see higher growth prospects in that zone than other investors do and may accordingly want to pursue a strategy of investing in companies that do business in the zone. Investing in the corresponding country's stock market index is not a clean way to implement that strategy if and when the companies traded on the country's exchange conduct a good deal of business abroad. Second, an investor may want to invest into a zone but fear the form of trading taking place in the corresponding country's stock market (insider trading, preferential trades and other corrupt practices). In that case, "investing by proxy" may be a good alternative. The investor can invest in companies of another country that do a lot of business in the country one targets, hedging away the business that these companies conduct at home. Foreign companies as opposed to the country's companies can serve to invest in a target zone. Third, corporations contemplating a capital investment in a production or distribution facility in a zone not their own need to have a proper description of the risks inherent to operating facilities in that zone, and not of the risks inherent to being domiciled in the corresponding country.[1]

Fourthly and most importantly for research purposes, our undertaking should enhance the meaningfulness of cross-country correlation studies. It is true in many *dynamic* models of international financial market equilibrium that, everything else equal, the cross-country correlation is higher if the financial markets

[1] We are not denying that there may also exist risk premia for being listed in one country. See Froot and Dabora (1999). At present, we do not have a model to explain such a phenomenon, if it exists.

are integrated (i.e., if movements of capital take place between countries the same way they do within countries) than if they are segmented.[2][3] For that reason, correlations have been used not only to measure diversification potential (Solnik (1974), Heston and Rouwenhorst (1994), Cavaglia et al. (2004)) but also to measure the degree of integration between markets (e.g., Bekaert and Mehl (2017)). The catch, however, is the phrase "everything else equal". As time goes by, the composition of country indexes evolves not just as to the list of firms included in the index but also as to the locus of business conducted by the firms that are included. For that reason, we may be deluding ourselves when we observe that cross-country correlations increase and conclude that the international financial market is growing more integrated. Any rise or fall in correlations can be variously interpreted either as evidence concerning market integration or as the result of the gradual redefinition of stock market indexes due to changing firms' activities.

Diermeier and Solnik (2001) put it plainly in the quote we used as epigraph. The phenomenon they point at may be a form of integration of the market for goods and services but it should not be confused with integration of financial markets. What we take to be a growing common movement between two given country assets may in fact be simply a change in the nature of these assets. In terms of financial theory, the distinction we are making is a key one: it is the distinction between cash flows to be discounted and the state prices used to discount them. Integration of financial markets is a phenomenon by which the state prices become more similar to each other when compared across investors of different countries. If cash flows become more and more similar, this does not qualify as integration. In fact, cash flow similarities make it harder to determine whether or not state prices are becoming more similar.

Most studies have attributed increased market integration to a rise in risk sharing among national economies rather than shifts in economic activity. But a few studies such as Ammer and Wei (1996), Baele and Soriano (2010), Viceira and Wang (2018), and Akbari, Ng, and Solnik (2019) have sought to distinguish between economic integration (a common cash flow dynamic) and financial integration (a common risk pricing dynamic). Viceira and Wang (2018), for example, find that the increase in correlations between stocks can be attributed to increased correlations between discount rate shocks but not increased correlations between cash flow shocks. Cavaglia et al. (2004) shows that the increased correlation between countries is due to country factors becoming more correlated as opposed to industry factors becoming more so. Heston and Rouwenhorst (1994) indicate that, even between European financial markets, which are presumably by now fairly well integrated, countries factors as opposed to industry factors

[2] See Dumas, Harvey and Ruiz (2003). Bekaert et al. (2011) present a similar dynamic model but do not examine correlations of stock returns.

[3] It is important to note the following. In a dynamic, general-equilibrium asset-pricing model with non IID cash flow growth, correlations of stock returns under integration are likelyhigher than they would be under segmentation, because the fluctuations of stochastic discount factors under integration are common to all securities. But no theory has been adduced concerning the correlations that should prevail in intermediate regimes of partial integration. Moreover, in this paper we make the assumption that stock returns are IID within each year.

offer the higher diversification potential. Goetzman et al. (2005) shows that the correlation has been mostly between "core" (basically developed) countries as opposed to other countries. Both observations militate in favor of an interpretation in terms of increased integration of financial markets. A pair of studies by Bekaert et al. (2011, 2013) took a different approach to studying the degree of integration by examining differences in earnings yields rather than correlations. They find evidence that markets are becoming more integrated, but like Goetzman et al., they find integration greater for developed than for emerging markets.

It is the goal of this paper to control for the changes in firms' activities. To reach that goal, we make the assumption that a company bears risks based only on the geographic zones to which it sells its products, irrespective of its domicile. We proxy shares of activities by shares of sales.[4] Admittedly, it should not be the share of sales that would serve to measure the geographic distribution. That is at best a proxy. One should use the share of value that arises from the cash flows generated in the geographic segments. That data, unfortunately, is not available. But for many multinationals, revenue data may not be a bad proxy for overall activity in that zone since many firms both source from and sell to those zones. Multinationals like Volkswagen or Proctor and Gamble have plants in many countries to support sales in those countries.[5]

We stress that, in this paper, we make no assumption about asset pricing and/or the degree of integration of financial markets. We only calculate the zone risk factors – making some other assumptions, to be stated below – and their correlations. Asset pricing would come into play when interpreting these correlations. Ours is only a statistical model.

The literature review on our subject contains primarily one item: the article by Jeff Diermeier and Bruno Solnik, published in the *Financial Analysts Journal*. In it, the authors study monthly stock returns of 1,213 individual companies listed in eight large country stock markets (France, Germany, Italy, Japan, Netherlands, Switzerland, United Kingdom and United States) from July 1989 to January 1999. Diermeier and Solnik have available to them data on stock returns, of course, but also data on the shares of activity (i.e., revenues) of firms in the domestic country and in the three regions of the globe that they have chosen to consider. The statistical analysis can be described as being in three stages.

[4] Bae et al (2019) use bilateral export data from developed countries to form emerging market country indexes based on the share prices of developed country firms that sell to emerging markets. The sales data that we use are broader than export data. A multinational's foreign sales to an emerging market need not involve any exports at all or at least any exports from its country of domicile. As Bae et al explain, however, sales data are not generally available on a bilateral basis for many countries, so they could not use sales data to form emerging market country indexes using their methodology.

[5] A study by Bodnar and Marston (2002) used survey data from 103 U.S. firms to examine the foreign exchange exposure of these firms. Of 103 firms in the sample, 83 firms had between 10% and 50% of their revenue from sales in foreign currency. But 73 of these same firms also had between 10% and 50% of their operating expenses in foreign currency. The majority of firms in the sample thus had substantial expenses as well as sales abroad, with relatively few firms in the sample being pure exporters or importers.

First, they construct a "pure" market index for each country as a value-weighted average return of firms with mostly domestic activities. From these they also calculate regional returns as value-weighted average of country stock returns and currency returns for countries that belong to a "region". There are three regions: Asia, Europe and the United States. Second, they run exposure regressions on all three types of indexes. In the third stage of their study, they ask the key question that motivates the whole undertaking: do these statistical exposures resemble the shares of revenues?

The study by Diermeier and Solnik hits the nail on the head very well but it has two drawbacks. First, it uses a limited number of firms in a limited number of countries. Developing countries in particular are not covered. Second and more importantly, it is implemented in stages. The stage that serves to define pure domestic indexes only uses the firms that have a large share of their activity at home, and a later stage relates the stock-market statistical exposures (mostly of the other firms) to their share of foreign activity. From the point of view of statistical theory, it would be vastly more efficient to use all firms to do everything in a single stage. That is, knowing the geographic distribution of activities of each firm, one should use all the information on all firms to identify the pure zone factors. A Malaysian firm, to the degree that it conducts operations in Switzerland, should also help in identifying the pure Swiss country factor.

The balance of the paper is organized as follows: Sections 1 and 2 describe the dataset of the firms' geographical segments for sales and the differences between traditional indexes and indexes based on sales. Section 3 outlines the statistical model to identify pure country indexes. The statistical technique that will serve to estimate the parameters of the model, namely the EM algorithm, is reviewed in Section 4. Section 5 addresses the problem of data imbalance. Section 6 analyzes the composition of the geographic indexes while Section 7 compares the behavior of the geographic and traditional indexes across countries. Section 8 analyzes the exposures revealed by the geographic indexes while Section 9 compares the geographic and traditional indexes across years. Section 10 states the conclusions.

1 The dataset on firms' geographic segments

The World Vest Base (WVB) database transcribes annual report information for a very large number of firms worldwide. The owner of the database provided us with data on the distribution of the firms' sales across geographic segments. We elected to study all the firms in the database provided the Datastream database contained stock return data for them. Our study covers the years 1999-2014 included. Unfortunately, annual reports do not contain information on the distribution of the firms' purchases across geographic segments.

The selection and filtering of firms based on geographic-segment data and stock-return data is explained in detail in Appendix A. The merging of the two datasets is explained in the same appendix. It left a number of stocks that

grows from 1797 in 1999 to 6335 in 2014.

A note on vocabulary is needed before we go on. We call "*segments*" or destinations, such as countries or regions, the geographic entities that are variously referred to *by firms* in their annual reports, as transcribed in the database. The segment information was the hardest to interpret in the empirical application, due to the non standardized description of regions in the sales database. There are *"Rest of"* segments, which are regions in which firms sell without explicitly giving the name of a country: "Rest of Europe", "Rest of the world" are typical example. These are defined as containing countries that *the firm* has not referred to explicitly among its segments. Typically, the annual report also refers to a country variously called "Home" or "Domestic sales".

By way of contrast we call "*zones*" standard geographic entities defined *by us* that are uniform across all the firms of our sample to reflect the operating risks they take, as reflected in a factor model described below . We consider twelve major zones plus one *"Rest of" zone* that contain the countries for which we have not defined a specific zone. For instance, a firm may have an ISIN starting with US and indicate that its only geographic segment is Zimbabwe. We classify its sales as being made in the "Rest of the World" zone. As another example, the definition of "Home" is simply the zone that the firm's annual report refers to when it refers to "Domestic sales." For each firm, the fraction of home sales is calculated as the ratio of home sales to total sales of the same year.

We compile a large *"dictionary"* that serves to map the very large number of segments or destinations of sales posted in the various annual reports into our standard zones.

The choice of zones to be considered as operating-risk factors is a delicate matter. The purpose of our study is to let data on sales sharpen (or restrict) the estimation of factor returns. The firms from the developed countries have sales that are more diversified than those from developing countries. When a country has few firms, that does not imply that the corresponding zone's index is computed on the basis of these firms only. The algorithm takes into consideration the returns of the firms from other countries that sell in the designated zone. For instance, there are few firms domiciled in Australia in our filtered sample (164 in 2006), but quite a few other firms cite Australia as a sales destination. All the firms that sell in that zone contribute somewhat to the calculation of the zone index. Such is the virtue of the algorithm.

For that reason, we want to choose zones for which we have sufficient sales information. Table 1 provides a description of the database. It indicates, for each zone, the number of companies year after year that sell more than 10% of their sales to that zone, explicitly so. For many countries this number is quite different from the number of firms domiciled in that country. For instance, in 2006, Singapore has 86 firms in our filtered database, but there are another 35 firms that have 10% or more of their sales to Singapore. So those multinational firms help to determine the Singapore zone factor. Germany has 105 domiciled firms in the database, but there are another 88 firms with 10% or higher sales to Germany. So for Germany, there is quite a bit of information coming from non-German multinational firms.

For purposes of estimation reliability, it would be best to let the set of zones vary from year to year. However, since we want to examine the factor correlations over the years, it is imperative that we maintain the same set of zones throughout, which implies a compromise. We choose a set of zones such that for each zone, we have in the database at least 50 firms selling to that zone in every year (or almost every year). For the entire paper, we adopt a *permanent list of thirteen zones*: France, Germany, Great Britain, Brazil, United States, Canada, Australia, Malaysia, Singapore, China, Japan, India and the Rest of the World.

Looking at Table 1 row-wise, we also notice that, unfortunately, our dataset is very unbalanced: many more firms sell to the United States and to Japan than to European countries. The filtered database favors the US and Asia and is less dense on Europe.[6] We return to that problem in Section 5 below.

2 National indexes vs. indexes based on sales

As mentioned in the introduction, classical stock market indexes such as the world and regional indexes published by Datastream and MSCI are based on a stock's place of listing: it is included in the index of a country when it is listed on an exchange of that country. These indexes are explicit and come in two forms: equally weighted and value weighted. In this paper, we focus exclusively on equally weighted indexes. We use the stocks in our database to mimic equally weighted listing-based indexes, using the two-letter prefix of a stock's ISIN number as proxy for place of domicile.[7] We call these "$ISIN$" or "national" indexes and we refer to all firms domiciled in one country as "national firms". They present the drawback of not clearly reflecting the risk of operating in the country.

Other explicit indexes can be constructed from individual stock returns on the basis of sales data: a stock is included in the index of a country if most of its sales (say, 70%) are made to that country. The index is calculated mechanically, also with equal weighting. We call these "70%" or "domestic" indexes as containing firms that sell to their own country almost entirely (which we call "domestic firms"). Both of them have the drawback of using fragmentary information and ignoring the sales information that can come from firms that are diversified geographically.

In this section, we aim to show that national indices are missing something important, quantitatively speaking. First, we show in Table 2 that domestic and national indexes are, indeed, different by displaying the correlation between them year after year and country by country. The reporting habits that explain these correlations are easy to imagine. The correlations are extremely high for,

[6] The database is compiled in Malaysia.

[7] See https://www.isin.org/isin/: "A two-letter country code, drawn from a list (ISO 6166) prepared by the International Organization for Standardization (ISO). This code is assigned according to the location of a company's head office." In Appendix A we describe how multiple domiciles of a firm's securities are eliminated from the database.

ZONES	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
RoW	623	644	725	772	637	555	468	657	619	575	576	569	590	674	635	554
AUSTRALIA	82	93	125	145	145	160	205	218	238	245	180	228	232	213	202	178
BRAZIL						51	59	67	70	71	76	82	86	78	77	83
CANADA	71	88	107	141	161	171	231	262	268	246	204	221	230	205	190	185
CHILE	69	74		71	68	61	78	87	82	83	77	72	71	68	68	63
CHINA	56	80	97	114	158	205	229	242	252	283	280	343	537	564	552	619
FRANCE	70	76	89	83	103	108	118	129	132	125	114	121	132	115	120	105
GERMANY	102	162	177	145	150	149	175	201	205	183	168	166	192	181	182	169
GREAT BRITAIN	171	198	206	190	201	199	249	277	281	256	220	215	217	206	228	231
GREECE						112	101	100	111	98	92	81	70	70	62	
INDIA		66	67	82	129	132	161	253	327	428	338	370	395	361	344	271
INDONESIA				97	71	79	95	84	90	87	102	87	96	98	97	101
ISRAEL							57	63	78	81	60	71	76	73	68	71
ITALY				59	70	66	77	87	90	91	81	84	83	65	72	65
JAPAN	114	144	133	153	197	420	618	662	563	528	854	1034	1070	1110	1288	1304
MALAYSIA	177	186	123	113	133	108	113	397	417	356	369	353	314	286	285	291
MEXICO							53	54	54	53	53	57	62	55	53	58
NEW ZEALAND							69	68	60	62	58	53	54	57	53	52
PAKISTAN										59	66	61	52			54
PHILIPPINES											59	67	65		58	58
POLAND										60	69	78	84	82	60	66
SINGAPORE		82	80	84	96	88	101	121	131	121	132	131	130	119	112	118
SOUTH AFRICA					59	59	62		78	98	80	78	78	82	80	77
SOUTH KOREA						75	89	212	221	225	232	253	406	390	470	478
SPAIN									55	55						64
SWEDEN							70	75	76	74	68	73	73	66		
TAIWAN					92	209	495	507	525	506	473	534	416	438	440	446
THAILAND					59	61	71	80	72	77	70	76	105	86	85	85
TURKEY					70	73	79	94	103	109	105	110	109	109	95	95
UNITED STATES	436	747	751	790	1069	1146	1464	1546	1595	1452	1334	1437	1519	1389	1352	1351

Table 1: **Database description:** Number of companies in the filtered database, for each zone indicated, that sell more than 10 percent of their sales to that zone.

e.g., Brazil and Malaysia because the firms domiciled in Brazil (Malaysia, respectively) are the only ones reporting Brazil (Malaysia) as a 70% sales destination and vice versa. They are a bit lower for France and Germany, for instance, because there are firms not domiciled in the respective country that report sales to it (such as French firms reporting sales to Germany and vice versa). Furthermore, these correlations would have been lower if we had constructed the domestic from sales levels below 70%.

Second, as a way to show the potential benefits of using sales information to inform a factor model, we present in Table 3 the fraction of national firms relative to all firms selling to a zone. The trend is clear in almost all zones: year after year the proportion of non-national firms is rising. For example, in the case of France the percentage of French firms selling to the French zone declines from 52.0% to 36.0% between 2000 and 2014. This trend makes it increasingly imperative to control for sales in setting up zone indexes. The table also shows that some zones remain more national than others. In Malaysia, Japan, and India in 2014, over 70% of the firms selling in those zones are national. In India, for example, there is a trend towards more foreign firms selling to India, but the country's sales remain dominated by Indian firms.

The 70% index captures only the sales information coming from domestic firms. Below we aim to utilize the information from more diversified firms (multinationals) as well. The statistical index to be developed in the next sections will allow us to capture all stock-return and sales information (at all levels and not just the 70% level), whether it is reported for destination countries specifically or, in less detail, for destination regions such as "Europe", "Middle East" etc..

3 The statistical model to be used

If information were available giving for each firm its share of sales to each zone, our goal could simply be reached by computing the (generalized) inverse of a huge matrix, or, equivalently by running a cross-sectional regression of firms' returns on firms' sales.[8] That matrix inversion would directly construct the zone returns from the company returns. In practice, however, the information about geographic segments is not as rich as that. We have at our disposal only partial information about the activities of each firm. Because the country returns cannot be measured directly, they have to be considered as latent (or unobserved, or implicit) factors and the estimation has to be viewed as an exercise in factor analysis.

The following model is adapted from Brooks and Del Negro (2004). They considered a similar factor model with country factors (and a world factor, which we do not have and do not need here) but with restrictions on the loadings (contained in the matrix B below) that differ from ours. They fix the loadings on foreign factors at 0 and they force the country factors to be independent of each other so that all common movements in countries take place through the

[8] $C = (B^\intercal B)^{-1} B^\intercal R$ in the notation of Equation (1) below.

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Rest of the world	0.967	0.967	0.961	0.964	0.958	0.990	0.992	0.994	0.997	0.997	0.997	0.998	0.992	0.985	0.973	0.970
France	0.581	0.760	0.893	0.738	0.835	0.926	0.927	0.957	0.939	0.968	0.941	0.969	0.973	0.936	0.880	0.864
Germany	0.785	0.947	0.944	0.793	0.736	0.814	0.909	0.923	0.898	0.896	0.863	0.889	0.963	0.932	0.849	0.780
Great Britain	0.898	0.928	0.933	0.908	0.942	0.950	0.975	0.977	0.982	0.971	0.982	0.976	0.982	0.978	0.959	0.964
Brazil	0.996	0.990	0.998	1.000	0.986	0.987	0.992	0.998	0.999	1.000	0.999	0.998	0.995	0.994	0.992	0.996
United States	0.980	0.984	0.974	0.985	0.991	0.983	0.993	0.995	0.996	0.997	0.997	0.997	0.997	0.990	0.992	0.994
Canada	0.887	0.910	0.928	0.905	0.945	0.970	0.966	0.985	0.986	0.990	0.983	0.986	0.987	0.962	0.940	0.961
Australia	0.977	0.982	0.985	0.973	0.977	0.991	0.988	0.990	0.997	0.998	0.995	0.997	0.998	0.991	0.990	0.991
Malaysia	0.999	1.000	0.999	0.997	0.995	0.992	0.989	0.996	0.999	0.998	0.996	0.997	0.998	0.995	0.995	0.995
Singapore	0.977	0.964	0.957	0.884	0.909	0.897	0.835	0.944	0.959	0.962	0.960	0.961	0.975	0.952	0.855	0.839
China	0.639	0.668	0.971	0.955	0.926	0.953	0.869	0.930	0.946	0.971	0.963	0.963	0.954	0.906	0.916	0.881
Japan	0.953	0.965	0.975	0.987	0.987	0.999	0.999	0.999	0.998	0.998	0.992	0.995	0.996	0.971	0.990	0.996
India	0.995	0.991	0.993	0.984	0.977	0.993	0.992	0.998	0.997	0.999	0.997	0.998	0.998	0.997	0.997	0.991
Mean	0.895	0.927	0.962	0.929	0.936	0.957	0.956	0.976	0.976	0.980	0.974	0.979	0.985	0.968	0.948	0.940
Median	0.967	0.965	0.971	0.964	0.958	0.983	0.988	0.990	0.996	0.997	0.992	0.995	0.992	0.978	0.973	0.970

Table 2: *ISIN* **vs.** 70% **index correlation.**

	2000	2002	2004	2006	2008	2010	2012	2014
France	52.00%	45.50%	48.50%	45.20%	48.90%	37.70%	37.80%	36.00%
Germany	60.50%	49.20%	42.80%	38.60%	34.80%	31.10%	27.40%	24.60%
Great Britain	51.60%	43.50%	38.60%	41.60%	37.40%	39.00%	36.20%	39.50%
Brazil	67.90%	52.00%	44.80%	45.20%	42.40%	35.30%	29.80%	26.90%
U.S.	73.30%	65.10%	72.50%	73.00%	70.70%	70.00%	63.00%	61.10%
Canada	45.40%	56.30%	52.40%	53.90%	52.80%	47.80%	45.50%	43.90%
Australia	46.50%	49.80%	45.90%	48.90%	48.90%	42.10%	31.60%	26.20%
Malaysia	89.70%	77.30%	70.90%	89.50%	88.70%	86.90%	79.30%	78.40%
Singapore	74.20%	76.70%	75.00%	57.10%	61.30%	51.50%	46.20%	45.40%
China	46.30%	50.00%	53.80%	16.60%	41.30%	30.00%	14.10%	15.10%
Japan	57.80%	52.50%	75.40%	77.40%	75.60%	86.70%	80.80%	84.30%
India	89.20%	88.40%	87.30%	84.10%	89.00%	83.90%	93.50%	75.50%
Mean	62.90%	58.90%	59.00%	55.90%	57.60%	53.50%	48.80%	46.40%

Table 3: **Growing influence of multinationals in a zone's sales:** Ratio of the number of national firms to all firms selling to a zone.

world factor.[9] Our model specifies the structure from which pure geographic-zone index returns C will be calculated:

$$R_t = B \times C_t + e_t \tag{1}$$

where R_t is the realization at time t of the N-vector of time-series demeaned returns (all measured in a common currency) for N stock securities, B is the $N \times K$ matrix that contains the loadings of all firms on all K geographic-zone factors, C_t is the realization at time t of the $K \times 1$ vector of zero-mean returns of *unobserved* zone factors, K being the number of zones and e_t is the realization at time t of the vector of unsystematic residuals of the stock returns.

Without further specification, the model would not be viable because C is not observed so that B is not identified. We impose enough constraints on B to be able to calculate C as a latent factor. We use the fragmentary information we have on firms' activities to set some of the elements of the B equal to the corresponding shares of activities. As mentioned in the introduction, we assume that the (delevered) stock returns of a firm reflect only the risks of the zones to which it sells its products, irrespective of its domicile. That assumption is motivated by the previous study of Diermeier and Solnik (2001). In a specific example of their second-stage result, Diermeier and Solnik cite the example of SmithKline Beecham, a "British" multinational, which has stock-market statistical exposures equal to .17, .08, .31 and .55, respectively, to the UK pure factor, to the Asia factor, to the Europe ex UK factor and to the North American factor. They ask whether statistical exposures to country factors resemble

[9] Similarly, Heston and Rouwenhorst (1994) fix the loading of a firm on its country to be equal to 1.

the shares of revenues that SmithKline Beecham receives from the various geographic segments. The firm makes 8% of its sales to the UK, 12% to Asia, 23.5% to Europe ex UK and 46.1% to North America. It seems that the stock market is broadly aware of the geographic distribution of the activities of the firm.

Specifically, the constraints we impose are as follows:[10]

Assumption 1 *We impose that the sum of the loadings of a firm equal 1*

$$\sum_{k=1}^{K} b_{i,k} = 1$$

That assumption is questionable as firms could be leveraged operationally so that their total risk would be more than what is captured simply by sales.[11] Summation to 1 is an auxiliary assumption we make. Next, we use the data from annual reports in order to write these restrictions:

Assumption 2 The loading of a single zone j or the sum of the loadings of a multiple-zone region j of a specific firm i is equal to the percentage $A_{i,j}$ of activities in that zone or region

$$\sum_{k \epsilon j} b_{i,k} = A_{i,j}$$

Here again, we exploit whatever information about country and regional sales is given in annual reports. In particular, $b_{i,k=i's\ country} = A_{i,k=i's\ country}$: the firm's loading on one particular zone factor called "home" is assumed equal to the share of sales made at home. The home country does not play a unique role, as the more general form of the restriction shows.[12]
It is debatable whether loadings on zone stock markets and share of activities should be set strictly equal to each other or, more broadly, related (perhaps linearly related) to each other. Equality – as opposed to just an increasing relationship – is an auxiliary assumption we make.

Assumption 3 *Non negativity for the loadings of each firm on the zone factors*

$$b_{i,k} > 0$$

As a result of these assumptions and restrictions, some of the elements of the matrix B are observed or "explicit". The others have to be estimated. At this point, we further assume that

Assumption 4 *The loadings $b_{i,k}$, whether observed or estimated are constant over time within a year.*

[10] There are very few antecedents of estimation of factor models with constraints; see Lawley and Maxwell (1971).

[11] Financial leverage is taken care of, albeit imperfectly, by deleveraging the stock returns. See Appendix A.

[12] For "Rest of" zones, the equality constraint is replaced by an inequality because the information in the annual reports often refers to one of the several countries of the zone.

4 Implementation: the $\mathbb{EM}$ indexes

The statistical implementation of the model (1) is very close in spirit to that of Brooks and Del Negro (2004). We have described in Section 3 our own set of restrictions. These do not allow us to assume that zones are independent of each other (which is the reason for which we do not need a world factor). We call Ω the covariance matrix of the unobserved zone factors C and we assume that C and e are independent of each other. The variance-covariance matrix of e, assumed to be diagonal, is denoted D.[13] Were it not for the constraints on B, zone factors could be redefined to be orthogonal to each other. The separate identifications of B and Ω, therefore, is based entirely on the constraints on B, while Ω is chosen freely.[14]

Assumption 5 *All random variables are multivariate IID normal.*

The full log-likelihood $\mathcal{L} \triangleq \ln p\left(R; B, D, \Omega\right)$ of observing R according to the model follows from the multivariate normal distribution, as in Lehmann and Modest (2005):

$$\begin{aligned} \mathcal{L}\left(B, D, \Omega\right) &= \frac{-NT}{2}\ln\left(2\pi\right) - \frac{T}{2}\ln\left|\Sigma\right| - \frac{T}{2}\text{trace}\left(S\Sigma^{-1}\right) \\ &= \frac{-NT}{2}\ln\left(2\pi\right) - \frac{T}{2}\ln\left|\Sigma\right| - \frac{1}{2}\sum_{t=1}^{T} R_t^{\intercal}\Sigma^{-1}R_t \end{aligned} \tag{2}$$

where:

$$\begin{aligned} R &= \left[R_t; t = 1, ..T\right] \\ \Sigma &\triangleq B\Omega B^{\intercal} + D \\ S &\triangleq \frac{1}{T}RR^{\intercal} \\ \left|\Sigma\right| &= \left|\Omega^{-1} + B^{\intercal}D^{-1}B\right|\left|D\right|\left|\Omega\right| \end{aligned}$$

Direct maximization, by equating to zero the gradient of $\mathcal{L}$ with respect to the parameters, yields a huge system of $N \times (K-1) + K \times (K-1)/2 + N$ equations that is nonlinear and hard to solve. Instead, we use an iterative method called the $\mathbb{EM}$ algorithm, which was first proposed to solve missing-data problems by Dempster et al. (1977) and then applied to latent-factor models by Rubin and Thayer (1982).[15] As was pointed out by Brooks and Del

[13] Although the estimation will attempt to make the model fit these assumptions, they will not hold for the estimated data. The number of zone factors C is much smaller than the number firms, so that the estimated variance-covariance matrix of e will not be diagonal. This is standard in factor analysis. Furthermore, the constraints imposed on B by the estimation algorithm will make it impossible to achieve orthogonality between e and C, which is less standard.

[14] Unfortunately, we are not able to provide sufficient conditions for identification. But we verify that the matrix $B\Omega B^{\intercal}$ is of rank K.

[15] $\mathbb{E}$ stands for "expectation" and $\mathbb{M}$ for "maximization". The meaning of that riddle becomes clear below. We apply the $\mathbb{EM}$ algorithm to estimate the latent factors but also to handle mssing (i.e., zero-return) data, as explained in Appendix C.

Negro (2004), the technique brings one very big algorithmic benefit: it allows likelihood optimization, at each stage of the iteration, to be applied to each firm one after the other, as opposed to all of them globally, which would infeasible. Our challenge is to extend the technique to the case in which the factors are not independent of each other and in which, instead, there are restrictions on the loadings, giving rise to a covariance matrix of zone factors.

The $\mathbb{EM}$ method consists in comparing the log-likelihood of R, $\ln p\left(R;B,D,\Omega\right)$, (2) of R to the *joint log-likelihood* of R and C, $\ln pC\left(R,C;B,D,\Omega\right)$, and in showing that, at any given value of the parameters, the gradient of the log-likelihood $\ln p\left(R\right)$ with respect to parameters is equal to the *expected value* of the gradient of the log-likelihood $\ln p\left(R,C\right)$ under the probability distribution of C given R.[16]

Imagining that the latent factors C were observed, the joint log likelihood $LL \triangleq \ln p\left(R,C\right)$ is based on the assumption that both e and C are multivariate normal (see Rubin and Thayer (1982)):

$$\begin{aligned} LL\left(B,D,\Omega\right) &= -\frac{T}{2}\sum_{j=1}^{N}\ln D_j - \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{N}\frac{\left(R_{j,t}-B_{\text{row }j}C_t\right)^2}{D_j} \\ &\quad -\frac{T}{2}\ln\left|\Omega\right| - \frac{T}{2}\sum_{t}C_t^{\intercal}\Omega^{-1}C_t \\ &= -\frac{T}{2}\ln\left|D\right| - \frac{1}{2}\text{trace}\left\{D^{-1}\left[\sum_{t=1}^{T}R_tR_t^{\intercal}\right.\right. \\ &\quad \left.\left. -2\sum_{t=1}^{T}R_tC_t^{\intercal}B^{\intercal} + B\sum_{t=1}^{T}C_tC_t^{\intercal}B^{\intercal}\right)\right\} \\ &\quad -\frac{T}{2}\ln\left|\Omega\right| - \frac{1}{2}\text{trace}\left(\sum_{t=1}^{T}C_tC_t^{\intercal}\Omega^{-1}\right) \end{aligned} \tag{3}$$

We calculate the expected value of LL given the observations R, at the currently estimated values of the parameters B,D,Ω. I.e., we integrate (3) over C. This gives:

$$\begin{aligned} \mathbb{E}\left[LL|R\right] &= -\frac{T}{2}\ln\left|D\right| - \frac{T}{2}\text{trace}\left\{D^{-1}\left(S-2XB^{\intercal}+BYB^{\intercal}\right)\right\} \\ &\quad -\frac{T}{2}\ln\left|\Omega\right| - \frac{T}{2}\text{trace}\left(Y\Omega^{-1}\right) \end{aligned} \tag{4}$$

[16] Appendix B develops the econometric theory under constraints in general terms.

where:

$$S_{N\times N} \triangleq \frac{1}{T} R R^{\intercal}$$

$$X_{N\times K} \triangleq \frac{1}{T}\sum_{t=1}^{T} R_t \mathbb{E}\left[C_t^{\intercal}|R_t\right] \tag{5}$$

$$Y_{K\times K} \triangleq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[C_t C_t^{\intercal}|R_t\right] \tag{6}$$

The sufficient statistics that are contained in (5) and (6) are: $\sum_{t=1}^{T}\mathbb{E}\left[C_t^{\intercal}|R_t\right]$ and $\sum_{t=1}^{T}\mathbb{E}\left[C_t C_t^{\intercal}|R_t\right]$. We compute them in Appendix D, on the basis of the model.

In the $\mathbb{M}$ step of the algorithm, the function (4) is maximized with respect to D, B and Ω, keeping X and Y fixed as computed from the values of the parameters of the previous iteration. At the next iteration, X and Y are recomputed (this is the $\mathbb{E}$ step) from (5) and (6), and the $\mathbb{M}$ part is run anew.

The first-order condition with respect to B need not be written down as, in any case, that optimization is to be done under the restrictions of Section 3, by means of a numerical quadratic-optimization algorithm. The first-order condition with respect to D is simply:

$$D = \text{diagonal}\left[S - 2XB^{\intercal} + BYB^{\intercal}\right]$$

As noted, the optimization with respect to the elements of B and D can fortunately be performed individually firm by firm.[17] That is the major benefit of the $\mathbb{EM}$ algorithm.

With these definitions of X and Y, the estimate of Ω is simply:

$$\Omega = Y$$

Indeed, the first-order condition is:[18]

$$\Omega^{-1} - \Omega^{-1} Y \Omega^{-1} = 0$$

During the execution of the $\mathbb{EM}$ steps, the full likelihood (2) is calculated periodically to verify that it keeps increasing. That calculation is computationally intensive.

Following Appendix D, we define the "$\mathbb{EM}$" zone indexes to be

$$\mathbb{E}\left[C_t|R_t\right] = \left(\Omega^{-1} + B^{\intercal} D^{-1} B\right)^{-1} B^{\intercal} D^{-1} R_t \tag{7}$$

except for the fact that R_t, in this definition, contains the individual firm returns, *not demeaned.* This definition allows us to obtain a measure of the zone indexes at daily frequency.

[17] These optimization, except for the restrictions, are analogous to a time-series regression run for each firm.

[18] Petersen and Pedersen (2007), Page 9, Equation (57) and Page 10, Equation (63).

Using daily stock returns, we perform the estimation year by year (from 1999 to 2014) assuming that all loadings, which are the elements of the matrix B, are constant within a year (Assumption 4 above). Some elements of matrix B are constrained. They can be subject to equality constraints: when each year's geographic shares of sales are explicitly observed, we set them equal to them. They can be subject to inequality constraints for regions, and for non negativity.

The convergence of the $\mathbb{EM}$ algorithm is considered to be achieved whenever the largest absolute value of any element of the relative gradient is lower than 10^{-2}.[19] The maximum number of iterations is 500.

5 The problem of data imbalance

As noted above, our dataset is not balanced: many more firms sell to the United States and to Japan than to European countries. The filtered database favors the US and Asia and is less dense on sales to Europe.

As in any factor analysis, the likelihood maximization itself aims to explain as much of the variance of individual stock returns as possible, the information on the sales of each firm being coded into the constraints of the maximization program. If we implemented the procedure just described without change, the number of individual firm constraints pertaining to each zone would play a critical role in pushing the total return variance into one zone or the other. For instance, since the database contains too little firm information about sales to Europe, the volatility of returns of European zones would end up being abnormally large. When we tried, the volatility of the German zone turned out to be much larger than that of other developed countries and larger than that of some developing countries.

In order to remedy that problem, we resort to subsampling. Subsampling is a technique that has gained a lot of ground in the area of machine learning. See the very lucid survey article by He and Garcia (2009) and also the article by Chen et al. (2013).

Instead of running it once on the whole dataset of each year, we run the algorithm on one hundred subsamples. Each subsample is chosen randomly in a stratified manner. First, for each zone, we count the number of firms for which we have explicit data with a fraction of sales to that destination greater than $x\%$ (with x being successively 70%, 50%, 30% and 10%). For each $x\%$ level, we compute across all zones the minimum number of firms with explicit data, and then randomly select from the firms of the more populated zones a number of firms equal to the minimum number. In this way, each subsample contain an approximately equal number of firms selling to each and every zone, at each $x\%$ level. Because it is not easy to verbalize the procedure exactly, we have reproduced in Appendix E the actual set of MatLab instructions that was used.

Finally, the estimates of the zone indexes, as per Equation (7), are averaged across the subsamples and a single additional iteration of the $\mathbb{M}$ step run on the

[19] In that gradient we include the Lagrange multiplier term, which is due to the constraints. See Appendix B.

entire sample produces our estimates of the loadings B.

6 The composition of $\mathbb{EM}$ indexes

The $\mathbb{EM}$ indexes that we obtain on the basis of sales, differ markedly from the national $ISIN$ indexes based on domicile. Table 4 displays, for each zone and each year, the correlations between the two types of indexes. Some of the correlations are far from being equal to 1 and they are not uniform across countries. Below we devote several sections to a comparison of the two types of indexes. In this section, we comment the composition of the $\mathbb{EM}$ indexes. That will help in understanding differences in their behavior.

Consider the model (1). If C_t were already calculated, one could obtain B by time-series regressions of R_t on C_t. If B were already calculated, on could obtain C_t by cross-sectional regressions of R_t on B. Brooks and Del Negro (2004, 2006) pointed out that the $\mathbb{M}$ step of the algorithm is essentially the within-year (constrained) time-series regression run for each firm, which alternates with the $\mathbb{E}$ step, which is essentially the cross-sectional regression. This section focuses on the latter aspect, while section 8 below focuses on the former.

The composition of the indexes C_t is dictated by formula (7), which we reproduce here:[20]

$$\mathbb{E}\left[C_t|R_t\right] = \left(\Omega^{-1} + B^\intercal D^{-1} B\right)^{-1} B^\intercal D^{-1} R_t$$

and which can be compared with cross-sectional GLS: $\left(B^\intercal D^{-1} B\right)^{-1} B^\intercal D^{-1} R_t$. The term $B^\intercal D^{-1} R_t$ is the weighted covariance at time t of the B loadings of the firms on their stock return, and the term $B^\intercal D^{-1} B + \Omega^{-1}$ can be interpreted as the variance-covariance across firms of the B coefficients.[21]

Thus the weights of the various securities in the various $\mathbb{EM}$ indexes are $\left(\Omega^{-1} + B^\intercal D^{-1} B\right)^{-1} B^\intercal D^{-1}$. These weights do not sum to one over securities. If the Ω^{-1} term were absent (i.e., in the GLS case), a *weighted* sum of them – where the weights in the sum are incorporated by postmultiplication by own B – would sum to 1: $\left(B^\intercal D^{-1} B\right)^{-1} B^\intercal D^{-1} B = I$. The deviation from 1 of the sum allows the algorithm to adjust the variance of the indexes C.

The formula indicates that the relation between the index composition and the loadings B is non linear. Imagining we fixed $\left(\Omega^{-1} + B^\intercal D^{-1} B\right)^{-1}$, the weights would be proportional to the loadings. But some of the loadings are given explicitly by the sales database while others are estimated, which complicates the relation between the weights and explicit sales.

[20] The formula produces an index composition for each subsample (see section 5), each one of them pertaining to a generally different set of firms. One more iteration run on the whole sample constructs the weights we discuss now.

[21] As Brooks and Del Negro (2004) put it, the difference "arises because the $\mathbb{E}$ step estimator treats C_t as a random variable with prior variance Ω while the GLS estimator treats the factor(s) as unknown but fixed coefficients," where "fixed" means "non random". Symbols adapted by us.

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Average
ROW	0.851	0.847	0.919	0.880	0.861	0.907	0.864	0.932	0.958	0.976	0.968	0.957	0.990	0.984	0.901	0.950	0.922
France	0.694	0.674	0.761	0.747	0.830	0.889	0.874	0.935	0.884	0.983	0.966	0.987	0.992	0.977	0.964	0.941	0.881
Germany	0.859	0.729	0.783	0.792	0.598	0.886	0.956	0.960	0.942	0.933	0.970	0.870	0.971	0.937	0.934	0.941	0.879
Great B.	0.684	0.787	0.860	0.879	0.945	0.959	0.968	0.977	0.983	0.977	0.984	0.969	0.983	0.962	0.913	0.939	0.923
Brazil	0.929	0.788	0.875	0.951	0.924	0.916	0.962	0.953	0.968	0.975	0.958	0.947	0.964	0.896	0.867	0.971	0.928
U.S.	0.812	0.706	0.842	0.950	0.971	0.963	0.983	0.981	0.989	0.990	0.995	0.993	0.994	0.986	0.991	0.984	0.946
Canada	0.846	0.670	0.799	0.831	0.820	0.937	0.922	0.893	0.940	0.955	0.949	0.938	0.914	0.871	0.802	0.748	0.865
Australia	0.948	0.903	0.913	0.924	0.935	0.973	0.958	0.952	0.986	0.978	0.944	0.964	0.975	0.915	0.848	0.911	0.939
Malaysia	0.990	0.982	0.988	0.955	0.969	0.939	0.871	0.958	0.993	0.956	0.968	0.978	0.954	0.959	0.976	0.971	0.963
Singapore	0.934	0.926	0.888	0.901	0.901	0.886	0.836	0.951	0.969	0.964	0.971	0.964	0.962	0.951	0.857	0.788	0.915
China	0.524	0.831	0.969	0.907	0.981	0.911	0.901	0.883	0.918	0.939	0.957	0.919	0.865	0.748	0.741	0.664	0.854
Japan	0.866	0.698	0.925	0.960	0.965	0.985	0.980	0.978	0.964	0.988	0.953	0.974	0.987	0.954	0.967	0.985	0.946
India	0.971	0.937	0.962	0.940	0.917	0.979	0.971	0.923	0.992	0.996	0.992	0.979	0.984	0.977	0.984	0.935	0.965
Mean	0.839	0.806	0.883	0.894	0.893	0.933	0.927	0.944	0.960	0.970	0.967	0.957	0.964	0.932	0.903	0.902	
Median	0.859	0.788	0.888	0.907	0.924	0.937	0.956	0.952	0.968	0.976	0.968	0.964	0.975	0.954	0.913	0.941	

Table 4: *ISIN* **vs.** $\mathbb{EM}$ **index correlation.**

Intuitively, we expect the $\mathbb{EM}$ index of a zone to be composed of a combination of firms' returns in three tiers:

- The stock returns of "domestic" firms that sell almost entirely to the zone.
- The returns of other (i.e., "multinational") firms that sell to the zone to varying degrees. These firms bring to the make up of the index their information about sales to the zone. Unfortunately, because they sell to other zones as well, they also introduce into the index the influence of stock returns that are not related to that zone.
- The returns of yet other multinationals that may not sell anything to the zone will serve to offset, by means of a negative weight, the influence of returns of the second tier.

By way of example, we show in Figure 1 for the year 2014 the scatter plot of the weights of the firms in the Canada $\mathbb{EM}$ index against the explicit sales of those firms to Canada. We keep in mind, in this figure as in previous ones, that many firms reporting "North America", for example, and not "Canada" explicitly play a role in the determination of the $\mathbb{EM}$ index but, regrettably, must be shown in the figure as having zero sales to Canada. Zero sales to Canada, therefore, are points in the figure that do not reflect some information that is nonetheless available. The sum of firms' weights in the index is equal to 0.978. Of these Canadian national firms carry a total weight equal to 0.973. Hence non Canadian firms contribute to the index with positive (totaling 0.521) and negative (totaling -0.516) weights that almost sum to 0. Nonetheless, the correlation between the Canadian $ISIN$ (or national) index and the Canadian $\mathbb{EM}$ index is only 0.7482, which means that the influence of the non Canadian national firms is, indeed, felt. The scatter of points shows, as we expect, that firms selling a lot to Canada tend to receive a higher weight and that this is true both for firms with a CA ISIN and for the other firms, so that CA firms selling a lot abroad tend to receive a lower weight in the index. But let us go into some more detail and illustrate the way sales information is combined with stock-returns information.

Of the firms selling 100% to Canada, some receive a much higher weight than others. Among them, Firm CA51925D1069 (The Laurentian Bank) receives the highest weight (0.05861) in the index both because, and as a result of the fact that its correlation with the Canada $\mathbb{EM}$ index is as high as 0.6711, while Firm CA81234D1096 (Sears Canada Inc.) receives the lowest weight (0.0003326) in the index because, despite its being a Canadian firm, its correlation with the Canada $\mathbb{EM}$ index is as low as 0.1687.

We next turn to non Canadian firms. Among them, Firm US8123501061 (Sears Holdings) features the highest fraction of sales to Canada (0.46949). Yet it receives a negligible weight (0.000209) in the index because its correlation with the Canada $\mathbb{EM}$ index is only 0.0386. The non Canadian firm with the highest weight in the index is IL0001260111 (Gazit Globe Ltd, an Israeli firm) because it sells as much as 26.5% of its sales to Canada.

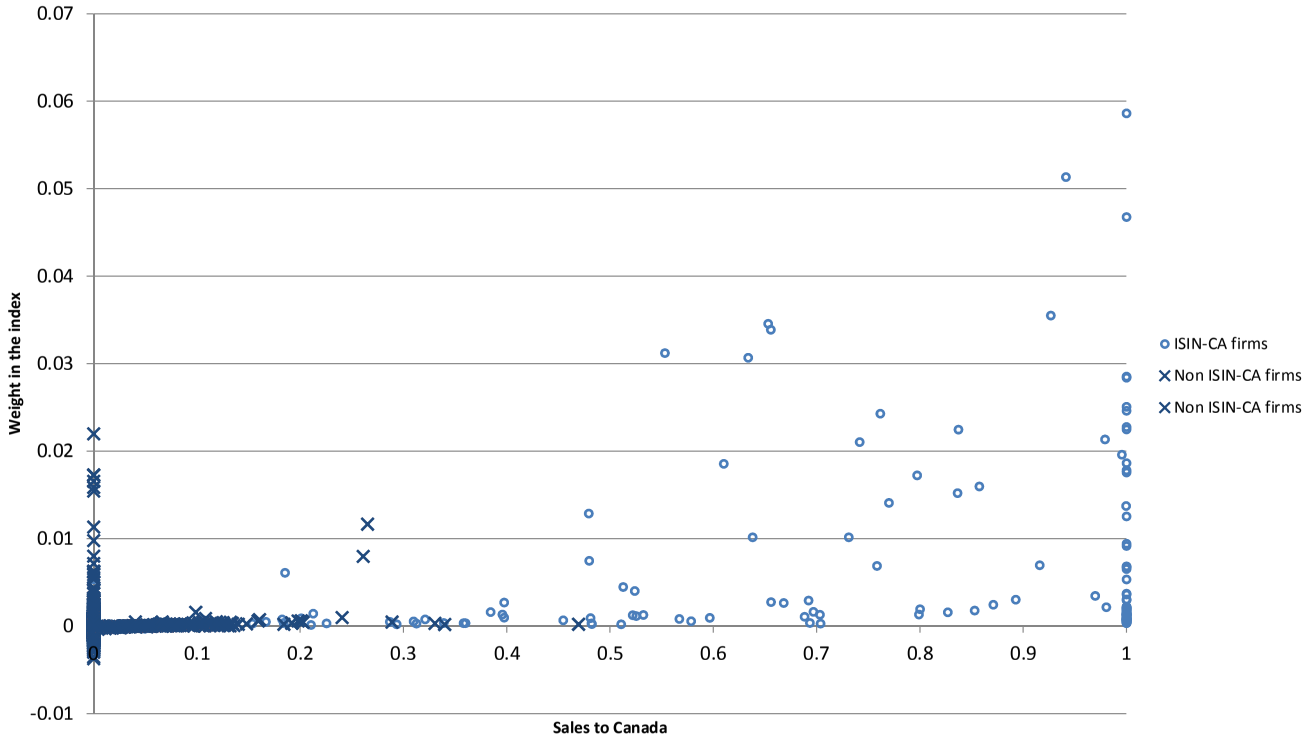

Figure 1: **Composition of the Canada index in 2014**: on the x axis are plotted the shares of sales to Canada of all the firms, and on the y axis are the weights in the EM index. Each point of the cloud of dots represents one security. The $ISIN$-CA firms are shown with a small-circle marker, other firms with a cross.

RoW	FR	DE	GB	BR	US	CA
0.09013	0.050041	0.02128	0.03495	0.03397	0.14146	0.95208
-0.07205	-0.0478	-0.02606	-0.03264	-0.02538	-0.13891	-0.00098
	AU	MY	SG	CN	JP	IN
	0.02975	0.01117	0.0436	0.03121	0.04442	0.01108
	-0.02554	-0.01207	-0.0381	-0.04105	-0.04418	-0.01227

Table 5: **Sums of firm loadings** weighted by the weight in the Canadian $\mathbb{EM}$ index: firms with positive weights in the first row; firms with negative weights in the second row.

Canadian and non Canadian firms that receive a positive weight in the index load (in the sense of the B loadings) on zones other than Canada. The first row of Table 5 gives the sum of these loadings weighted by the firms' weight in the index. Without some offset, the Canada $\mathbb{EM}$ index would be unduly influenced by the non-Canadian sales of multinational firms, as it would contain, for instance, a large total loading of 14% on the U.S. Such is the purpose of the (non Canadian) firms that receive a negative weight (a total of 5220 firms), some of them very small. Their weighted loadings are shown in the second row of the table. It is remarkable how the algorithm has been able to cancel the unwanted loadings and focus on Canada alone.

The above remarks have illustrated the three-tier way in which the $\mathbb{EM}$ index is constructed.

7 The behavior of $\mathbb{EM}$ vs. $ISIN$ indexes – comparisons across countries

The $\mathbb{EM}$ indexes differ from traditional indexes primarily because they incorporate information from the geographical sales data of each firm. To the extent that firms sell abroad, their stock returns should be sensitive to those zone indexes where the sales occur rather than to the zone index where the firms' stocks are domiciled. $\mathbb{EM}$ and $ISIN$ indexes of a zone differ from each other to the degree that national firms and firms selling there are different firms.

The $\mathbb{EM}$ indexes should differ the most for countries where international sales are most important. Countries like France and Canada have many firms with substantial sales abroad. In 2014, for example, 47.7% of Canadian firms had more than 30% of their sales outside Canada. In other countries like India and Malaysia, more firms focus primarily on domestic sales. In India, for example, only 30.1% of the firms had substantial sales abroad in 2014. In Japan in that same year, over 84% of firms selling in Japan were Japanese. For this reason, we should find that the $\mathbb{EM}$ method makes the most difference for countries with many internationally oriented firms and with many foreign firms selling in that country.

This section will explore how international sales data reported by national

firms influence the $\mathbb{EM}$ indexes. To show how sales data affect the $\mathbb{EM}$ indexes, consider two measures of the openness of a country to international sales:

- The ratio of the number of national firms selling domestically to the total number of firms selling to that zone. In the case of Canada in 2014, for example, 43.9% of the firms selling in Canada were Canadian.
- The ratio of the number of national firms with foreign sales equal to 30% or more of total sales to the total number of national firms. In the case of Canada in 2014, 47.7% of its firms are "multinational," defined as firms with 30% or more of their sales abroad.

If a country has many foreign firms selling to it, the $\mathbb{EM}$ index should reflect the influence of these sales on that country's $\mathbb{EM}$ index. If a country has many national firms selling abroad, these firms should exert influence on the $\mathbb{EM}$ indexes for other zones. In both cases, the correlation between the $\mathbb{EM}$ and $ISIN$ indexes should be relatively low.

To illustrate how international sales influence the $\mathbb{EM}$ indexes, consider two sets of charts showing a link between sales patterns and the $\mathbb{EM}$ vs. $ISIN$ correlations. Figure 2 shows for all firms$\times$years the relative importance of a zone's national firms in each zone index, on the x axis, and the correlation between the $\mathbb{EM}$ and $ISIN$ indexes, on the y axis. The lower the ratio of national firms to all firms selling to that zone the lower the correlation between the $\mathbb{EM}$ and $ISIN$ indexes. So the sales pattern should be positively correlated with the correction between the $\mathbb{EM}$ and $ISIN$ indexes. The chart shows results for the sixteen years across the twelve zones (not including RoW).

To make better sense of this data, let us focus on two countries with markedly different behavior, Canada and India. Figure 3 shows the data for these two countries only. In Canada, the percentage of national firms selling in Canada is only 50.1% on average over the sample period, 1999 to 2014. In India, that percentage is 85.1%. As a result, the correlation between the $\mathbb{EM}$ and $ISIN$ indexes should be lower for Canada than for India. Indeed, that is the case since Canada has an average correlation of 0.865 in contrast to a 0.965 correlation for India. The sixteen observations for India are clustered in the northeast quadrant of the diagram while the Canada observations are clustered in the center of the chart. So the $\mathbb{EM}$ method makes more difference for a country like Canada where sales by foreign firms are relatively important.

Sales by national firms to foreign markets are also important in making the $\mathbb{EM}$ indexes different from the $ISIN$ indexes. Figure 4 focuses on the relative importance of a zone's multinational firms defined as the ratio of national firms in that zone with at least 30% of their sales outside that zone. The higher the ratio of multinationals, the lower are the correlation between that zone's $\mathbb{EM}$ index and the corresponding $ISIN$ index.

Without the support of a figure this time, let us again compare Canada and India. In Canada over the sample period, 43.6% of firms can be described as multinational with 30% or more of their sales abroad. In India, only 17.9% of

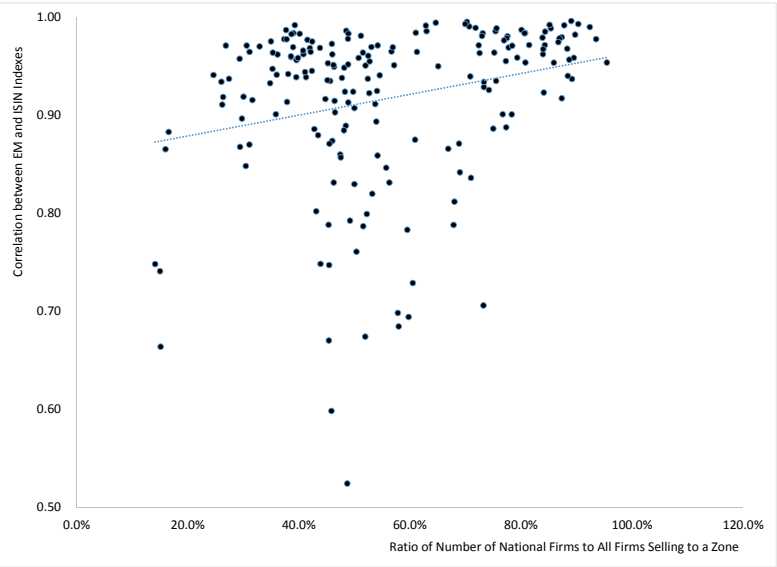

Figure 2: **Relative Importance of a Zone's Firms in Each Zone Index**: on the x axis is the ratio of the number of national firms to all firms selling to a zone. On the y axis is correlation between $\mathbb{EM}$ and $ISIN$ indexes. Each point is a year and a zone. The regression line is: $y = 0.1085 \times x + 0.858$.

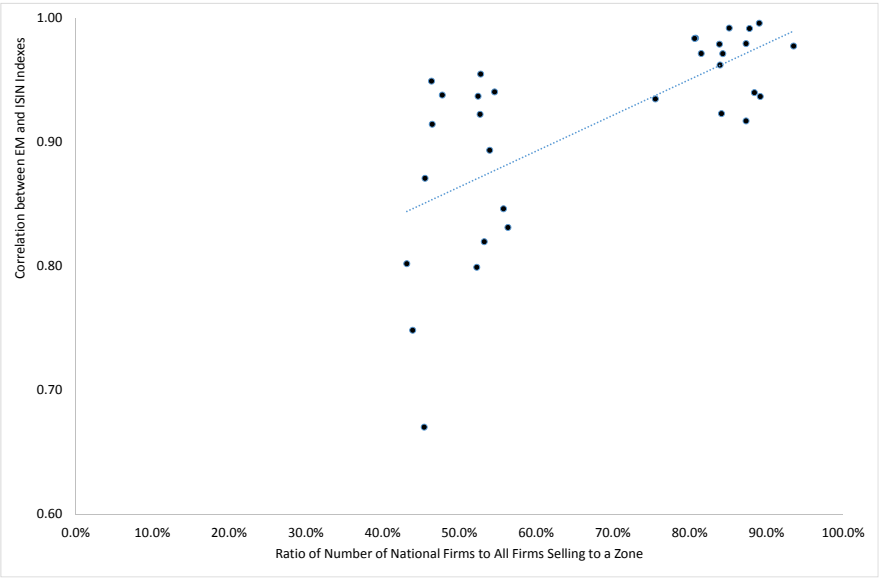

Figure 3: **Relative Importance of a Zone's Firms in Each Zone Index – the case of Canada and India only**: on the x axis is the ratio of the number of national firms to all firms selling to a zone. On the y axis is correlation between $\mathbb{EM}$ and $ISIN$ indexes. Each point is a year and a zone. The regression line is: $y = 0.2885 \times x + 0.7249$.

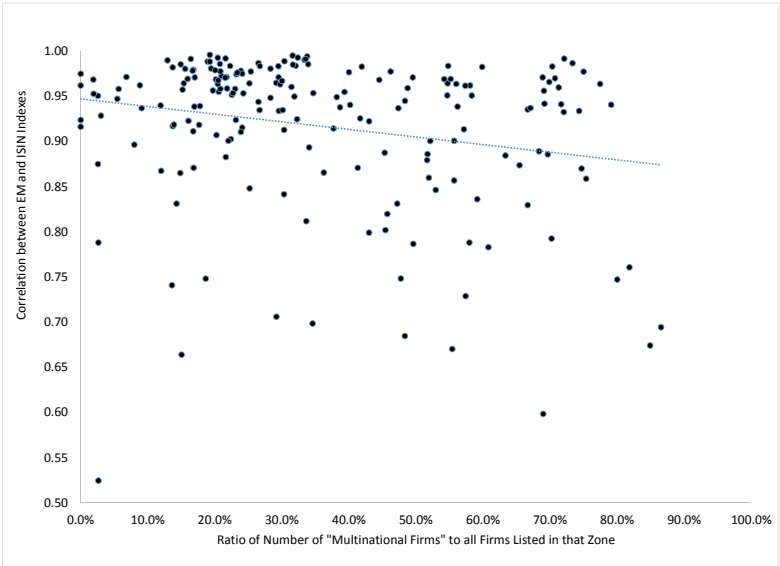

Figure 4: **Relative Importance of a Zone's "Multinationals"**: on the x axis is the ratio of the number of "multinational firms" to all firms domiciled in that zone. On the y axis is correlation between $\mathbb{EM}$ and $ISIN$ indexes. Each point is a year and a zone. The regression line is: $y = -0.1126 \times x + 0.9765$.

firms are multinational. The correlation between the $\mathbb{EM}$ and $ISIN$ indexes for Canada is accordingly lower for Canada (0.865) than it is for India (0.965).

8 The exposures revealed by $\mathbb{EM}$ indexes

In the tradition of factor models, it is customary to regress the return of a security on a number of factors, thus obtaining their "exposures", in a meaning of the word that is purely based on returns and not on sales. To maintain the differentiation, we call them "statistical exposures". In most empirical estimations of that type, it is found that the returns of firm domiciled in a country generate a statistical exposure to their national stock market index that is much larger than the one to the world stock market, which strikes one as odd in an integrated world financial market. One reason could be that the country's index is improperly defined. We now examine the statistical exposures that are generated from $\mathbb{EM}$ indexes, instead of national or $ISIN$ indexes.

For that purpose, we regress separately each security's daily return on the daily returns of each of the two explicit sets of regressors. This is done by constrained maximum likelihood, the likelihood function being similar to (2) above (but with regressors explicitly given, as opposed to factors to be estimated), where the coefficients of the regression are here again constrained to be non negative.[22] Even though we know that firm could very well be negatively exposed to a risk factor, we retain the non negativity constraint, in order to take into account the fact that, in the construction of $\mathbb{EM}$ indexes, we have only been able to use information about firms' sales and not about profits. Without the constraints, the exposures would not be comparable.

The question we will focus on is whether in the $\mathbb{EM}$ regressions the exposures to the foreign factors play a greater role than they do in the $ISIN$ regressions. Table 6 compares the sum of the foreign coefficients for the $\mathbb{EM}$ and $ISIN$ regressions, averaging over all national firms of a country, for each year as well as the average coefficients over the sixteen years from 1999 to 2014. One drawback of the $\mathbb{EM}$ method is that the three European zone indices for France, Germany, and Great Britain are highly correlated. So the returns of firms in those zones tend to load on all three zone indexes rather than primarily on the own index. French firms should load on the German index to the extent that they sell to that zone, but these firms seem to load more than would be justified based on sales data. To adjust for this phenomena, Table 6 also reports for the European regressions the exposures to zones outside Western Europe.

The results in Table 6 are clear-cut. The $\mathbb{EM}$ regressions give a much larger role to foreign indexes than do regressions using traditional indexes like the $ISIN$ indexes. For example, in the U.S. regressions the average foreign coefficients increase from 0.334 to 0.869. In the case of Canada, foreign coefficients increase from 0.460 to 1.099. The only consistent exception is India where the

[22] We caution that the constraints cause the estimated residuals not to be orthogonal to the regressors, so that the mean squared residuals do not truly capture the unexplained variance. A covariance term would introduce a correction.

$\mathbb{EM}$	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Average
France	1.363	3.208	3.619	2.837	2.045	1.517	1.200	1.115	1.183	1.220	0.950	1.199	0.778	1.220	1.323	1.158	1.621
Germany	1.732	3.587	4.305	2.191	2.132	1.592	1.382	1.578	1.138	0.828	1.815	1.261	0.808	1.444	1.508	1.080	1.774
Great B.	1.430	2.795	2.709	1.838	1.096	1.074	0.800	0.893	1.054	0.676	0.664	0.491	0.577	0.852	0.817	0.841	1.163
France*	0.934	2.670	2.988	1.739	1.355	0.820	0.770	0.647	0.496	0.395	0.651	0.505	0.427	0.544	0.581	0.725	1.015
Germany*	0.957	3.288	3.269	1.534	1.481	1.004	0.845	0.776	0.512	0.524	0.687	0.919	0.548	0.785	0.788	0.759	1.167
Great B*	1.097	2.569	2.426	1.398	0.940	0.672	0.679	0.644	0.426	0.558	0.571	0.425	0.545	0.685	0.666	0.781	0.943
Brazil	0.728	0.491	0.596	0.425	0.531	0.610	0.553	0.505	0.440	0.238	0.370	0.382	0.268	0.332	0.665	0.634	0.485
U.S.	1.071	4.113	1.508	0.769	0.769	0.669	0.606	0.566	0.520	0.468	0.510	0.448	0.340	0.541	0.478	0.529	0.869
Canada	1.078	2.571	1.498	1.206	1.273	1.055	1.008	0.859	0.854	0.600	0.894	0.926	0.794	1.237	0.961	0.776	1.099
Australia	0.616	0.994	1.166	0.801	1.015	0.774	0.767	0.703	0.746	0.676	0.703	0.485	0.539	1.046	0.788	0.729	0.784
Malaysia	0.850	0.895	0.946	0.682	0.899	0.684	1.095	0.540	0.816	0.360	0.563	0.546	0.416	0.625	0.667	0.622	0.700
Singapore	1.052	1.823	1.500	1.018	1.042	1.233	1.132	0.871	0.807	0.611	0.805	0.615	0.479	0.834	0.779	0.854	0.966
China	0.327	1.678	0.941	0.137	0.862	0.841	0.585	0.507	0.392	0.383	0.452	0.219	0.506	0.869	0.645	0.401	0.609
Japan	0.885	1.494	0.923	0.924	0.751	0.719	0.551	0.439	0.515	0.326	0.657	0.445	0.348	0.635	0.689	0.645	0.684
India	0.956	0.902	1.422	0.741	1.001	0.692	0.638	0.435	0.519	0.315	0.465	0.254	0.365	0.569	0.539	0.616	0.652
$ISIN$	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Average
France	0.462	0.489	0.823	0.672	0.603	0.394	0.474	0.395	0.352	0.360	0.365	0.356	0.339	0.385	0.390	0.483	0.459
Germany	0.539	0.707	1.078	0.764	0.725	0.669	0.560	0.547	0.442	0.423	0.551	0.441	0.441	0.533	0.495	0.470	0.587
Great B	0.498	0.651	0.853	0.683	0.566	0.417	0.426	0.419	0.375	0.397	0.332	0.301	0.312	0.407	0.382	0.486	0.469
France*	0.327	0.364	0.714	0.505	0.452	0.267	0.286	0.236	0.220	0.204	0.227	0.179	0.196	0.226	0.228	0.316	0.309
Germany*	0.391	0.498	0.877	0.571	0.556	0.355	0.376	0.313	0.260	0.246	0.342	0.258	0.253	0.325	0.279	0.333	0.390
Great B*	0.391	0.498	0.877	0.569	0.556	0.355	0.376	0.313	0.260	0.246	0.342	0.221	0.253	0.325	0.279	0.364	0.389
Brazil	0.367	0.293	0.191	0.209	0.230	0.201	0.309	0.319	0.374	0.215	0.230	0.214	0.179	0.206	0.272	0.273	0.255
U.S.	0.546	0.649	0.613	0.484	0.395	0.261	0.311	0.242	0.206	0.225	0.257	0.178	0.183	0.227	0.226	0.337	0.334
Canada	0.467	0.490	0.563	0.550	0.470	0.355	0.525	0.357	0.369	0.408	0.449	0.420	0.446	0.537	0.496	0.449	0.460
Australia	0.328	0.352	0.338	0.336	0.415	0.328	0.436	0.304	0.351	0.429	0.408	0.396	0.381	0.530	0.463	0.438	0.390
Malaysia	0.373	0.306	0.355	0.335	0.422	0.298	0.393	0.238	0.361	0.145	0.244	0.255	0.172	0.265	0.262	0.418	0.303
Singapore	0.482	0.473	0.465	0.379	0.479	0.369	0.430	0.369	0.378	0.375	0.380	0.347	0.289	0.353	0.400	0.360	0.395
China	1.161	0.825	0.414	0.134	0.411	0.597	0.501	0.375	0.473	0.466	0.317	0.224	0.296	0.350	0.332	0.217	0.443
Japan	0.322	0.334	0.379	0.330	0.372	0.299	0.361	0.321	0.174	0.170	0.162	0.122	0.136	0.179	0.252	0.279	0.262
India	1.052	1.823	1.500	1.018	1.042	1.233	1.132	0.871	0.807	0.611	0.805	0.615	0.479	0.834	0.779	0.854	0.966

* Sales to countries outside Western Europe

Table 6: **Sum of exposures to foreign zones in regressions on $\mathbb{EM}$ and $ISIN$ indexes**

$\mathbb{EM}$ regressions actually have lower foreign coefficients. So the $\mathbb{EM}$ indexes succeed in enhancing the importance of foreign influences on stock returns. That is to be expected since the $\mathbb{EM}$ indexes reflect the importance of foreign sales to many of these firms.

9 The behavior of $\mathbb{EM}$ vs. $ISIN$ indexes - comparisons across years

The graphs in Figure 5 compare over the years and across countries the second moments of the $\mathbb{EM}$ indexes with those of the $ISIN$ and 70% indexes. For each year, we obtain a cross-section of standard deviations and pairwise correlations of individual zone returns and display the mean and the median. The second moments of all three indexes are obtained from the daily returns.

The pair of graphs at the top of the figure displays the three pairwise correlations between the daily stock returns of the three indexes, the mean and the median being calculated over the thirteen zones. The three indexes are strongly correlated with each other. They become more so around 2008, as can be expected on the occasion of a market crash (although the opposite seems to occur in 2000 for two of the three pairwise correlations). But they drop again.

The middle pair of graphs in Figure 5 is based on the thirteen standard deviations of daily stock returns of which, each year, we take the mean and the median. Not surprisingly all the volatilities rise with the two stock market crashes of 2001 and 2008. The two graphs reveal that, for most countries, it is the case that the $\mathbb{EM}$ indexes are less volatile than the two explicit indexes. This comes as a surprise since the observed indexes are diversified across zones while the $\mathbb{EM}$ indexes, by construction, are not. One interpretation is that some of the volatility of developed-country stock exchange indexes (which are more numerous in our sample) arises from their firms' involvement in developing country zones, which are more volatile. Indeed, most of the increase in internationalization reflects the penetration of developing markets by developed-country firms.

Finally, with the bottom graphs of Figure 5, we reach the destination of our work. They display mean and median statistics of the pairwise correlations between $\mathbb{EM}$ zone indexes on the one hand and the two explicit country indexes on the other. The difference tends to be negative: $\mathbb{EM}$ indexes are less correlated across countries than are the explicit indexes. The $\mathbb{EM}$ algorithm has been able to remove an undue amount of correlation caused by sales to common foreign countries. This is consistent with the hypothesis that some of the correlation between traditional stock market indexes arises from the interpenetration of corporate activity across countries. We have achieved the principal goal of our exercise, which was to deconstruct that interpenetration.

Traditional $ISIN$ indexes exhibit an upward trend and then a downward trend in their correlations. That, however, was not convincing evidence, if at

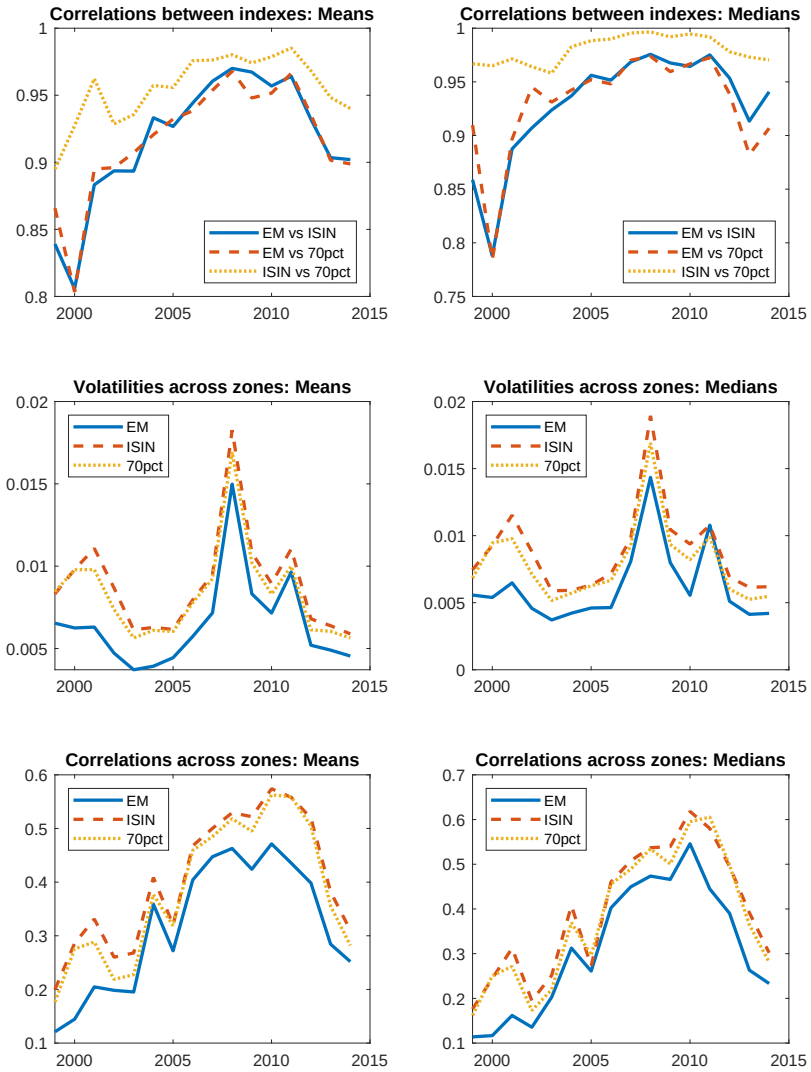

Figure 5: **Comparison and evolution** of second-moment properties of $\mathbb{EM}$, $ISIN$ and 70% indexes. For each year, we obtain the daily standard deviations and pairwise correlations (across indexes in the top pair of graphs, across zones in the bottom pair), of the thirteen zone indexes, of which we take the mean and the median.

all,[23] of the rising and falling integration of financial markets because the firms domiciled in each separate country were becoming more and more similar in their cash flows. Once we have controlled for changes in cash flows – sales as a proxy –, the rise and fall of correlations is still present. In this paper, as noted in the introduction, we have made no assumption about asset pricing and/or the degree of integration of financial markets. We have only calculated the zone factors – making some other assumptions, as explained above – and their correlations. Asset pricing comes into play when interpreting these correlations.

Overall the two explicit indexes $ISIN$ and 70% behave more similarly with each than they do with the implicit $\mathbb{EM}$ index.

10 Conclusion

We have identified implicit or latent stock index factor returns based on the geographic zones in which firms have their activity, as opposed to explicit, traditional indexes constructed according to the country stockmarket in which they are domiciled.

We have provided more convincing evidence, than was thusfar available, that financial markets tended to become integrated until 2008 but became more fragmented following the crisis.

This work opens the way to more complete factor models, which should be investigated. The first priority would be to add to geographic zone indexes local-pricing factors reflecting the hypothesis that securities listed on the same stock market correlate excessively. If that factor played a significant role over and beyond the factors that reflect actual sales activity, it would be evidence of some form of irrationality.

The second priority would be to add industry factors and to test the hypothesis that zones are no more than portfolios of industries, as they should be in a world that is integrated both financially and by way of worldwide trading of goods and services.

One more topic of research of research should be contemplated. We would need a technique to introduce a weighting of the firms so that one could compare, for instance, value-weighted indexes of the explicit and implicit kinds.[24]

[23] See Footnote 3.

[24] It is straightforward to introduce a weighting in the likelihood function. But the way to weigh the constraints pertaining to the various firms has escaped us.

Bibliography

Akbari, A., Ng, L. and B. Solnik, 2019, "Emerging Markets are Catching Up: Economic or Financial Integration?," Journal of Financial & Quantitative Analysis, forthcoming.

Ammer, J. and Mei, J., 1996, "Measuring international economic linkages with stock market data," *The Journal of Finance*, 51, 1743–1763.

Bae, J. W., R. Elkamhi, and M. Simutin, 2019, "The Best of Both Worlds: Assessing Emerging Economies by Investing in Developed Markets," *The Journal of Finance*, 74, 2579-2617.

Baele, L., and P. Soriano, 2010, "The Determinants of Increasing Equity Market Comovement: Economic or Financial Integration?," *Review of World Economics*, 146, 573-589.

Bekaert, G., C. R. Harvey, C. T. Lundblad, and S. Siegel, 2011, "What Segments Equity Markets?," *Review of Financial Studies*, 24, 3841-3890.

Bekaert, G., C. R. Harvey, C. T. Lundblad, and S. Siegel, 2013, "The European Union, the Euro, and Equity Market Integration," *Journal of Financial Economics*, 109, 583-603

Bekaert, G. and A. Mehl, 2017, "On the Global Financial Market Integration 'Swoosh' and the Trilemma," working paper, Columbia Business School.

Bodnar, G. M. and Marston, R. C., 2002, "A Simple Model of Foreign Exchange Exposure," in T. Negishi, R. Ramachandran and K. Mino (ed), *Economic Theory, Dynamics and Markets: Essays in Honor of Ryuzo Sato*, Kluwer Academic Press.

Brooks, R. and M. Del Negro, 2004, "A Latent Factor Model with Global, Country and Industry Shocks for International Stock Returns," working paper, Federal Reserve Bank of Atlanta, previously circulated as: "International Diversification Strategies," 2002-23b.

Brooks, R. and M. Del Negro, 2006, "Firm-Level Evidence on International Stock Market Comovement," *Review of Finance*, 10, 69-98.

Cappiello, L., Engle, R.F. and Sheppard, K., 2006,."Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns," *Journal of Financial Econometrics*, 4, 537–572.

Cavaglia, S., J. Diermeier, V. Moroz, and S. De Zordo 2004, "Investing in Global Equities," *The Journal of Portfolio Management,* 30, 3, 88-94.

Chen, L., W. W. Dou and Z. Qiao, 2013, "Ensemble Subsampling for Imbalanced Multivariate Two-Sample Tests," *Journal of the American Statistical Association*, 1308-1323.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, B39, 1-38.

Diermeier, J. and B. Solnik, 2001, "Global Pricing of Equity," *Financial Analysts Journal*, 57, 3, July/August.

Dumas, B., C. R. Harvey and P. Ruiz, 2003, "Are Correlations in International Stock Returns Justified by Subsequent Changes in National Outputs?" *The Journal of International Money and Finance*, 22, 777-811.

Froot, K. A and Dabora, E. M., 1999, "How are stock prices affected by the location of trade?," *Journal of Financial Economics*, Elsevier, 53, 189-2016,

Goetzmann, W. N., L. Li and K. G. Rouwenhorst, 2005, "Long-Term Global Market Correlations," *Journal of Business*,.78, 1-38.

Griffin, J. M., P. J. Kelly and F. Nardari, 2010, "Do Market Efficiency Measures Yield Correct Inferences? A Comparison of Developed and Emerging Markets" *Review of Financial Studies*, 23, 3225–3277.

He, H. and E. A. Garcia, 2009, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, 21, 1263-1284.

Heston, S. L. and K. G. Rouwenhorst, 1994, "Does Industrial Structure Explain the Benefits of International Diversification," *Journal of Financial Economics*, 36, 3-27.

Lawley, D. N and A. E. Maxwell, 1971, *Factor analysis as a statistical method,* Elsevier.

Lehmann, B. N. and D. M. Modest,.2005, "Diversification and the Optimal Construction of Basis Portfolios," *Management Science,* 51, 581-598.

Petersen, K. B. and M. S. Pedersen, 2007, *The Matrix Cookbook*, on the web.

Rubin, D. B. and D. T. Thayer, 1982, "EM Algorithms for ML Factor Analysis," *Psychometrika*, 59, 69-76.

Solnik, B., 1974, "Why Not Diversify Internationally Rather than Domestically?" *Financial Analysts Journal*, 30, 48-54.

Viceira, L. M. and Z. Wang, 2018, "Global Portfolio Diversification for Long-Horizon Investors," working paper, Harvard Business School.

Zakoian, J.-M., 1994, "Threshold heteroskedastic models," *Journal of Economic Dynamics and Control*, 18, 931–955.

Appendixes

A Description of data filtering

The proxy for foreign sales activity of firms is the geographical breakdown of revenues. *World'Vest Base* has extensive information about sales activities of a large number of companies (98% of global capitalization in 2003) and is our source of the revenue breakdown information. The data on geographical distribution of sales was given to us in 2016 for the years 1999-2014. A total of 77,184 security ISINs is available. In some cases, it is impossible to interpret a sales destination reported in WVB. Additionally there is a number of negative, zero and missing revenue values. To circumvent the first problem, we restrict the analysis to a list of 937 sales destinations that unambiguously refer to countries or geographical regions. We address the second problem by eliminating the records which correspond to negative, zero or missing revenue values. The filtering is done in the following steps:

1. First, firms are selected from the entire multi-year sample on the basic of permanent (static) properties obtained from Datastream.
   For WVB ISINs, a "static" download from *Datastream* is performed. It contains first the stock type ("ordinary" vs. others) according to Worldscope. When that piece of data is "NA", the security is deleted; that leaves 47,370 ISINs. Second, it contains the average market capitalization over all the years. When that piece of data is "NA", the security is deleted; that leaves 34,460 ISINs. Third, it contains the "TRCS code," which is one indication on security type. When that piece of data is "NA", the security is deleted; that leaves 27,391 ISINs. Based on the TRCS code, a number of security categories (other than ordinary shares) are deleted;[25] that leaves 26,738 securities. Then we restrict securities by their country of origin as indicated in the first two letters of their ISIN; the restriction is to 56 countries;[26] that leaves 25,441 securities. Then, based on the Worldscope type, Chinese A shares are deleted;[27] that leaves 24,105 securities. Fourth the download contains the name of the company. A first general

[25] ','ABS','ADR','BD','BDIND','BWT','CF','CMD','CON','CPRF','CV','EC','EQIND','ES',
'ET','EWT','EX','FT','FUN','GDR','GSH','INT','INVT','JDC','KDC','LIST','OP','OWT',
'PREF','PREFI','PRFI','SWAPS','UC','UCIND','UT'
and 'ADR','BDR','SWEDDR','TRAD','CICNPPRF','NONCUM','PART','SUBSRTS',
'ENHTRUST','INDEXLN','CEF','CHESS','COWNT','CPR','CUM','DEBENT',
'DRC','EDR','ETF','ETN','INVESTSHAR','OPF','PREFERRED','PRF','GDR',
'GENUS','INTERDR','NVDR','SWEDR','REDEEM','REI','RTS','SAVE',
'STAPLED','STKDIV','UNT','OPT','PARTPAID','DVR'

[26] Developed countries: 'AU'; 'AT'; 'BE'; 'CA';'CY'; 'DK'; 'FI'; 'FR'; 'DE'; 'GR'; 'HK'; 'IE'; 'IL'; 'IT'; 'JP'; 'LU'; 'NL'; 'NZ'; 'NO'; 'PT'; 'SG'; 'KR'; 'ES'; 'SE'; 'CH'; 'TW'; 'GB'; 'US';
Developing countries: 'AR'; 'BD'; 'BR'; 'BG'; 'CL'; 'CN'; 'CO'; 'CZ'; 'EG'; 'HU'; 'IN'; 'ID'; 'KE'; 'LT'; 'MY'; 'MX'; 'MA'; 'PK'; 'PE'; 'PH'; 'PL'; 'RO'; 'ZA'; 'LK'; 'TH'; 'TR'; 'VE'; 'ZW'

[27] In fact, the following types are kept: 'B Gu'; 'H Gu'; 'N Gu'; 'S Gu'; 'L Gu'

cleaning recommended by Griffin et al. (2010) based on the name eliminates 5 securities. But a more comprehensive cleaning also recommended by Griffin et al. (2010) eliminates country-specific types, which are too many to list; that leaves 23,588 securities.

2. Further selection is based on properties that vary year by year but are assumed to stay the same within a year
   For each year, data are extracted from WVB: the list of ISIN numbers for the companies that are present during that year, the names of these companies, the year end of their annual report, the sales destinations, the revenues from each of the sales destinations, the currency unit in which these revenues are expressed and the report type.[28] For the filtered BWV sample of company ISINs of each year, data is downloaded from Datastream regarding market capitalization and leverage at the beginning of the year. The number of securities available in each year is indicated in column (2) of Table 7. When "NA" appears for the market capitalization or leverage entry of security, that security is dropped; that leaves each year the number of securities indicated in column (3). Securities that in each country have a market capitalization below the 97th percentile of the country's capitalizations are eliminated as "microcaps;" that leaves each year the number of securities indicated in column (4).

3. A selection that requires stock returns is performed.
   Securities' daily return indexes during the year were downloaded from Datastream, both in a common currency, the US dollar, and in local currency. For each stock we get the total return index in both, the USA dollar and the home currency. While the estimation is done using the dollar returns, the home currency returns are used for filtering. The reason behind this choice is that our filtering requires calculation of the number of non-zero returns and the dollar non-zero returns may be a result of the exchange rate variation rather than the price variation. A few securities suffer from an entry of "NA" or "#ERROR" in this download; the number of remaining securities is indicated in column (5). Daily rates of return are calculated in US dollar and deleveraged, so that rates of return from now on are rates of returns on companies' assets. If the rate of return in original-currency unit is equal to zero, this is an indication of thin trading; within each year, we delete securities for which there are abnormally high returns and for which there are days of thin trading;[29] the number of remaining securities is indicated in column (6). We remove holidays common to most of the countries: based on the dollar returns, we count the number of zero returns for each day and remove the whole

[28] According to the WVB Data Manual, multiple records with different report types may exist due to a change in the accounting standards, in the income statement format, due to reclassification of items or to changes in the fiscal year end. See Table 9.

[29] We divide a year into sub-periods of 20 trading days (with the last two sub-periods overlapping) and require all stocks in the sample to have at least one non-zero return within each of the sub-periods with zero-returns calculated on the basis of home currency prices.

string of returns for this day if the number of zero returns exceeded a third of the total number of stocks. While the number of days in a year with an observation is reduced, the number of securities is not affected.

4. Finally, the selected Datastream data are merged with the BWV data on sales. Selection is then done on the basis of sales data.
   We merge into one dataset per year the data on sales and the data on rates of return. In the process, some securities are lost because BWV, while showing the revenues of a particular company for some years, may not show them for a specific year; the number of securities remaining is shown in column (2) of Table 8. Some more filtering is performed. based on sales data. First we choose "report type." A firm may restate or reclassify its revenues within a year. Once the information is available, WVB updates the records on the revenue breakdown and, at the same time, retains the old records. As the result, a given firm may have multiple records. Therefore, using the data without additional filtering may bias the geographical distribution of the sales activities. We try to solve this problem by a filtering procedure elaborated below. According to the WVB Data Manual, multiple records may exist due to a change in the accounting standards, in the income statement format, due to reclassification of items or to changes in the fiscal year end. Consequently, for filtering we use information on the report type, currency and on the fiscal year end. Using the report type code, we select the report that belongs to I, II priority groups in Table 9: consolidated and covering a period of 12 months and consolidated and covering a period of less than 12 months. In order to keep the maximum number of records, for a given firm we choose the most detailed financial record type. Moreover, if the number of records grouped by the financial data header is the same, we chose the one which is preferred according to the rule set up in Table 9 (the most preferred report type being at the top of the list and the least preferred at the bottom). We then look at the firms which have multiple records corresponding to the same geographical regions and choose the records with the latest fiscal year end date. Some firms have multiple records from the same report type and fiscal year end date but stated in different currencies. For these firms we keep revenues stated in one arbitrarily chosen currency. Finally, to control for a possibility that a firm without multiple records referring to the same geographical segment, reports revenues in different currencies, we convert revenue data into US dollar. That process leaves the number of firms indicated in column (3) of Table 8. Then we eliminate any repeated sales destinations, which does not affect the number of securities. That process may leave some companies with no explicit sales exposure to any of the sales destinations; these are eliminated; column (4) indicates how many remain. Finally, several securities with different ISINs have the same name. These are multiple domiciles. We choose the ISIN that has the largest market cap. That leaves the final number of securities in each indicated in column (5).

Year	**# securities** (2)	**NA in mcap or debt col. del'd** (3)	**microcaps deleted** (4)	**NA and #ERROR del'd** (5)	**Thin trading del'd** (6)
1999	6852	5244	2233	2232	2124
2000	10242	8244	3106	3104	2898
2001	11619	10020	3642	3593	3013
2002	12980	11048	3985	3983	3572
2003	14640	12611	4721	4713	4287
2004	17344	14190	5365	5357	4896
2005	18612	16530	6814	6808	6269
2006	20042	19516	7803	7790	7627
2007	20580	20117	7940	7937	7786
2008	20321	19915	7635	7632	7515
2009	19895	19459	7103	7096	6976
2010	19520	19147	7292	7289	7159
2011	19178	18809	7251	7243	7143
2012	18546	18233	6906	6899	6764
2013	18338	18005	6851	6839	6680
2014	17678	17264	6690	6681	6498

Table 7: **Firm count after each stage of the filtering based on properties that vary year by year (Step 2).**

B $\mathbb{EM}$ theory with (possibly inequality) constraints

Call ψ the collection of parameters to be estimated. The constraints we are considering are constraints on the values of some parameters. Lagrange multipliers are special "parameters". But the Lagrangian is not a log-likelihood. Let the constraints be $g(\psi) \geq 0$. Our goal is to maximize

$$L(R;\psi,\phi) \triangleq \ln p(R;\psi) - \phi \cdot g(\psi)$$

But:

$$\begin{aligned} L(R;\psi,\phi) &= \ln p(C,R;\psi) - \ln p(C|R;\psi) - \phi \cdot g(\psi) \\ \frac{\partial}{\partial \psi} L(R;\psi,\phi) &= \frac{\partial}{\partial \psi} \ln p(C,R;\psi) - \frac{\frac{\partial}{\partial \psi} p(C|R;\psi)}{p(C|R;\psi)} - \phi \cdot \frac{\partial}{\partial \psi} g(\psi) \end{aligned}$$

Now, take the expected value under the conditional probability distribution with any given parameter value $\tilde{\psi}$ (naturally, $\int p\left(C|R;\tilde{\psi}\right) dC = 1$):

$$\begin{aligned} \frac{\partial}{\partial \psi} L(R;\psi,\phi) &= \int \left(\frac{\partial}{\partial \psi} \ln p(C,R;\psi) \right) p\left(C|R;\tilde{\psi}\right) dC \\ &\quad - \int \frac{\frac{\partial}{\partial \psi} p(C|R;\psi)}{p(C|R;\psi)} p\left(C|R;\tilde{\psi}\right) dC - \phi \cdot \frac{\partial}{\partial \psi} g(\psi) \end{aligned}$$

Year # securities	**After splicing** (2)	**After choosing report type** (3)	**Firms with** B (4)	**After multiple domic. del'd** (5)
1999	1987	1893	1880	1797
2000	2721	2500	2486	2397
2001	2790	2480	2462	2378
2002	3265	2819	2809	2725
2003	3868	3401	3384	3297
2004	4501	4020	4001	3908
2005	5701	5197	5187	5088
2006	6889	6123	6104	6000
2007	7130	6344	6306	6200
2008	6873	6241	6221	6117
2009	6482	6136	6127	6025
2010	6779	6531	6521	6424
2011	6806	6645	6635	6520
2012	6476	6317	6305	6210
2013	6590	6444	6427	6335
2014	6480	6342	6328	6241

Table 8: **Firm count after each stage of the filtering done on the basis of sales data (Step 4).**

At the point $\psi = \tilde{\psi}$ at which we took the expected value,

$$\begin{aligned}
\frac{\partial}{\partial \psi} L\left(R; \tilde{\psi}, \phi\right) &= \int \left(\frac{\partial}{\partial \psi} \ln p\left(C, R; \tilde{\psi}\right)\right) p\left(C|R; \tilde{\psi}\right) dC \\
&\quad - \frac{\partial}{\partial \psi} \int p\left(C|R; \tilde{\psi}\right) dC - \phi \cdot \frac{\partial}{\partial \psi} g\left(\tilde{\psi}\right) \\
\frac{\partial}{\partial \psi} L\left(R; \tilde{\psi}, \phi\right) &= \int \left(\frac{\partial}{\partial \psi} \ln p\left(C, R; \tilde{\psi}\right)\right) p\left(C|R; \tilde{\psi}\right) dC - 0 \qquad (8) \\
&\quad - \phi \cdot \frac{\partial}{\partial \psi} g\left(\tilde{\psi}\right)
\end{aligned}$$

At each stage, the maximization in the $\mathbb{EM}$ algorithm picks the value $\hat{\psi}$ and $\hat{\phi}$ such that:

$$\int \left(\frac{\partial}{\partial \psi} \ln p\left(C, R; \hat{\psi}\right)\right) p\left(C|R; \tilde{\psi}\right) dC - \hat{\phi} \cdot \frac{\partial}{\partial \psi} g\left(\hat{\psi}\right) = 0 \qquad (9)$$
$$\hat{\phi} \cdot g\left(\hat{\psi}\right) = 0; \hat{\phi} \geq 0; g\left(\hat{\psi}\right) \geq 0$$

Priority	WVB header	Description
I	C	Consolidated report covering 12-months period
	CR	Consolidated report containing restated data
	CS	Consolidado legislacao secretaria (Brazil)
	CC	Consolidado em moeda de podre aquisitivo constante (Brazil)
	I	Consolidated report meeting international standards covering a 12-months period
	IR	Consolidated report meeting international standards and containing restated data
II	IP	Consolidated report meeting international standards covering a period less than 12 months
	CP	Consolidated report covering a period less than 12 months
	CU	Consolidated report with preliminary/summary data
	CQ	Quarter/Interim report

Table 9: **Record type priority:** Multiple records are eliminated according to the priority rule described in this table (from the most preferred in the first row to the least preferred in the last row).

If $\hat{\psi} = \tilde{\psi} \triangleq \psi^*$ and $\hat{\phi} = \phi \triangleq \phi^*$ are reached,

$$\begin{aligned} \frac{\partial}{\partial \psi} L\left(R; \psi^*, \phi^*\right) &= \int \left(\frac{\partial}{\partial \psi} \ln p\left(C, R; \psi^*\right) \right) p\left(C|R; \psi^*\right) dC \\ -\phi^* \cdot \frac{\partial}{\partial \psi} g\left(\psi^*\right) &= 0 \end{aligned}$$

which is an optimum for $L\left(R; \psi, \phi\right)$.

Gradients with respect to ψ with (possibly inequality) constraints The following equality, which is a special case of equation (8)

$$\frac{\partial}{\partial \psi} L\left(R; \hat{\psi}, \hat{\phi}\right) = \int \left(\frac{\partial}{\partial \psi} \ln p\left(C, R; \hat{\psi}\right) \right) p\left(C|R; \hat{\psi}\right) dC - \hat{\phi} \cdot \frac{\partial}{\partial \psi} g\left(\hat{\psi}\right)$$

provides two alternative ways to compute the value of the gradients that is reached at the end of each iteration. The expression on the right-hand side is very convenient because the Lagrange multipliers can be easily substituted out of it. Indeed, based on (9):

$$\begin{aligned} \frac{\partial}{\partial \psi} L\left(R; \hat{\psi}, \hat{\phi}\right) &= \int \left(\frac{\partial}{\partial \psi} \ln p\left(C, R; \hat{\psi}\right) \right) p\left(C|R; \hat{\psi}\right) dC \\ &\quad - \int \left(\frac{\partial}{\partial \psi} \ln p\left(C, R; \hat{\psi}\right) \right) p\left(C|R; \tilde{\psi}\right) dC \end{aligned}$$

To perform that calculation, the only step needed is the updating of the expected value that is the first term on the right-hand side.

C Gradients for *missing data* using the expected value of the joint log-likelihood

Let R be now a matrix of returns that are actually observed, with zeros where the missing values are located, and let $\mathcal{R}$ be a matrix containing zeros everywhere except where there are missing values, which are entered as unknowns. Then the matrix of return is $R + \mathcal{R}$ so that:

$$\begin{aligned} S &= \frac{1}{T}\left(\mathcal{R}R^\intercal + R\mathcal{R}^\intercal + RR^\intercal + \mathcal{R}\mathcal{R}^\intercal\right) \\ \operatorname{trace}\left(S\Sigma^{-1}\right) &= \frac{1}{T}\operatorname{trace}\left(\left(\mathcal{R}R^\intercal + R\mathcal{R}^\intercal + RR^\intercal + \mathcal{R}\mathcal{R}^\intercal\right)\Sigma^{-1}\right) \end{aligned}$$

The differential of that term of the log likelihood are:[30]

$$\begin{aligned}
\partial\mathcal{R} \quad : \quad & d\text{trace}\left(S\Sigma^{-1}\right) = \frac{1}{T}\text{trace}\left(\left(d\mathcal{R}R^{\intercal} + Rd\mathcal{R}^{\intercal} + d\mathcal{R}\mathcal{R}^{\intercal} + \mathcal{R}d\mathcal{R}^{\intercal}\right)\Sigma^{-1}\right) \\
= \quad & \frac{1}{T}\text{trace}\left(d\mathcal{R}R^{\intercal}\Sigma^{-1} + Rd\mathcal{R}^{\intercal}\Sigma^{-1} + d\mathcal{R}\mathcal{R}^{\intercal}\Sigma^{-1} + \mathcal{R}d\mathcal{R}^{\intercal}\Sigma^{-1}\right) \\
= \quad & \frac{1}{T}\text{trace}\left(R^{\intercal}\Sigma^{-1}d\mathcal{R} + d\mathcal{R}^{\intercal}\Sigma^{-1}R + \mathcal{R}^{\intercal}\Sigma^{-1}d\mathcal{R} + d\mathcal{R}^{\intercal}\Sigma^{-1}\mathcal{R}\right) \\
= \quad & \frac{1}{T}\text{trace}\left(2R^{\intercal}\Sigma^{-1}d\mathcal{R} + 2\mathcal{R}^{\intercal}\Sigma^{-1}d\mathcal{R}\right) \\
= \quad & \frac{2}{T}\text{trace}\left(\left(R^{\intercal}\Sigma^{-1} + \mathcal{R}^{\intercal}\Sigma^{-1}\right)d\mathcal{R}\right)
\end{aligned}$$

Therefore (transposing, because of row vs. column notation),[31]

$$\frac{\partial\text{trace}\left(S\Sigma^{-1}\right)}{\partial\mathcal{R}} = \frac{2}{T}\left(\Sigma^{-1}R + \Sigma^{-1}\mathcal{R}\right)$$

Besides, the partial derivatives of the $\ln|\Sigma|$ term are equal to zero.

However, most of the elements of $\mathcal{R}$ are constants. For a single element (i,t),

$$\begin{aligned}
\partial\mathcal{R}_{i,t} \quad : \quad & \frac{2}{T}\text{trace}\left(\left(R^{\intercal}\Sigma^{-1} + \mathcal{R}^{\intercal}\Sigma^{-1}\right)J^{i,t}\right) \\
= \quad & \frac{2}{T}\left(\Sigma^{-1}R + \Sigma^{-1}\mathcal{R}\right)_{i,t} \\
= \quad & \frac{2}{T}\left[\left(\Sigma^{-1}R\right)_{i,t} + \left(\Sigma^{-1}\mathcal{R}\right)_{i,t}\right] \\
= \quad & \frac{2}{T}\left[{}_{i}\left(\Sigma^{-1}\right)R_t + {}_{i}\left(\Sigma^{-1}\right)\mathcal{R}_t\right]
\end{aligned}$$

Imagine only one missing (i) at time t:

$$\begin{aligned}
{}_{i}\left(\Sigma^{-1}\right)R_t + {}_{i}\left(\Sigma^{-1}\right)^{i}\mathcal{R}_{i,t} &= 0 \\
\mathcal{R}_{i,t} &= -\frac{{}_{i}\left(\Sigma^{-1}\right)R_t}{{}_{i}\left(\Sigma^{-1}\right)^{i}}
\end{aligned}$$

If there are several missing $(\{i\})$ at time t:

$$\begin{aligned}
{}_{\{i\}}\Sigma^{-1}R_t + {}_{\{i\}}\left(\Sigma^{-1}\right)^{\{i\}}\mathcal{R}_{\{i\},t} &= 0 \\
\mathcal{R}_{\{i\},t} &= -\left({}_{\{i\}}\left(\Sigma^{-1}\right)^{\{i\}}\right)^{-1}\cdot{}_{\{i\}}\Sigma^{-1}R_t
\end{aligned}$$

What is the purpose of the minus sign? It is to offset the deviations from Σ of

[30] https://en.wikipedia.org/wiki/Matrix_calculus, section "Scalar-by-matrix identities"

[31] Verified on: http://www.matrixcalculus.org/matrixCalculus

the other elements of S. However, changing one element will change the mean:

$$\begin{aligned}
\text{trace}\left(S\Sigma^{-1}\right) &= \frac{1}{T}\sum_t\sum_k\sum_j\left(R_{k,t}-\frac{1}{T}\sum_\tau R_{k,\tau}\right) \\
&\quad {}_k\left(\Sigma^{-1}\right)_j\left(R_{j,t}-\frac{1}{T}\sum_\tau R_{j,\tau}\right) \\
\frac{\partial\text{trace}\left(S\Sigma^{-1}\right)}{R_{i,t}} &= \frac{2}{T}\left(1-\frac{1}{T}\right)\sum_j {}_i\left(\Sigma^{-1}\right)_j\left(R_{j,t}-\frac{1}{T}\sum_\tau R_{j,\tau}\right)
\end{aligned}$$

That makes no difference to the replacement rule. But we make sure that R contains zero at the missing values.

D Sufficient statistics of the geographic analysis

Based on the model:

$$\text{cov}\left[C_t, R_t^\intercal\right] = \text{cov}\left[C_t, C_t^\intercal B^\intercal + e_t^\intercal\right] = \Omega B^\intercal$$

we compute $\mathbb{E}\left[C_t^\intercal|R_t\right]$ and $\mathbb{E}\left[C_t C_t^\intercal|R_t\right]$:

$$\begin{aligned}
\mathbb{E}\left[C_t|R_t\right] &= \text{cov}\left[C_t, R_t^\intercal\right]\left[\text{var}\left[R_t\right]\right]^{-1}R_t \\
&= \Omega B^\intercal\left[B\Omega B^\intercal + D\right]^{-1}R_t
\end{aligned}$$

But

$$\left[B\Omega B^\intercal + D\right]^{-1} = D^{-1} - D^{-1}B\left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1}B^\intercal D^{-1}$$

the great benefit of this transformation being that the matrix to be inverted is only as large as the number of zones, as opposed to being as large as the number of firms.

$$\begin{aligned}
\mathbb{E}\left[C_t|R_t\right] &= \Omega B^\intercal\left(D^{-1} - D^{-1}B\left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1}B^\intercal D^{-1}\right)R_t \\
&= \left(\Omega - \Omega B^\intercal D^{-1}B\left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1}\right)B^\intercal D^{-1}R_t \\
&= \left(\Omega\left(\Omega^{-1} + B^\intercal D^{-1}B\right) - \Omega B^\intercal D^{-1}B\right)\left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1} \\
&\quad \cdot B^\intercal D^{-1}R_t \\
&= \left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1}B^\intercal D^{-1}R_t \\
&\triangleq \delta R_t
\end{aligned}$$

Remark 1 *Compare with cross-sectional GLS:* $\left(B^\intercal D^{-1}B\right)^{-1}B^\intercal D^{-1}R_t$.

Next, we handle the second sufficient statistic $\sum_{t=1}^T \mathbb{E}\left[C_t C_t^\intercal|R_t\right]$:

$$\mathbb{E}\left[C_t C_t^\intercal|R_t\right] = \text{var}\left[C_t|R_t\right] + \mathbb{E}\left[C_t|R_t\right]\mathbb{E}\left[C_t^\intercal|R_t\right]$$

$$
\begin{aligned}
\operatorname{var}[C_t|R_t] &= \Omega - \operatorname{cov}[C_t, R_t^\intercal]\left[\operatorname{var}[R_t]\right]^{-1}\operatorname{cov}[R_t, C_t^\intercal] \\
&= \Omega - \Omega B^\intercal\left[B\Omega B^\intercal + D\right]^{-1} B\Omega \\
&= \Omega - \Omega B^\intercal\left(D^{-1} - D^{-1}B\left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1}B^\intercal D^{-1}\right)B\Omega \\
&= \Omega - \left(\Omega - \Omega B^\intercal D^{-1}B\left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1}\right)B^\intercal D^{-1}B\Omega \\
&= \Omega - \left(\Omega\left(\Omega^{-1} + B^\intercal D^{-1}B\right) - \Omega B^\intercal D^{-1}B\right) \\
&\quad \cdot \left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1}B^\intercal D^{-1}B\Omega \\
&= \Omega - \underbrace{\left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1}B^\intercal D^{-1}}_{\delta^\triangleq}B\Omega \\
&= \left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1}\left(\left(\Omega^{-1} + B^\intercal D^{-1}B\right)\Omega - B^\intercal D^{-1}B\Omega\right) \\
&= \left(\Omega^{-1} + B^\intercal D^{-1}B\right)^{-1} \\
&\triangleq \Delta
\end{aligned}
$$

so that:

$$\mathbb{E}\left[C_tC_t^\intercal|R_t\right] = \Delta + \delta R_t R_t^\intercal \delta^\intercal \triangleq Y_t$$

Hence:

$$
\begin{aligned}
X_{N\times K} &= \frac{1}{T}\sum_{t=1}^{T} R_t R_t^\intercal \delta^\intercal = S\delta^\intercal \\
Y_{K\times K} &= \frac{1}{T}\sum_{t=1}^{T}\left(\Delta + \delta R_t R_t^\intercal \delta^\intercal\right) = \Delta + \delta S\delta^\intercal
\end{aligned}
$$

Remark 2 *If there were no constraints, the estimate for B would be*

$$B = XY^{-1} = \frac{1}{T}\sum_{t=1}^{T} R_t\mathbb{E}\left[C_t^\intercal|R\right]\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[C_tC_t^\intercal|R_t\right]\right]^{-1}$$

which is a time-series LS (except that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[C_tC_t^\intercal|R_t\right]$ contains but is not equal to $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[C_t|R_t\right]\mathbb{E}\left[C_t^\intercal|R_t\right]$).

E MatLab code for stratifying the dataset and for subsampling

E.1 Stratifying

```
%% Stratifying the sample
   % first, find out how many firms in each zone at several sales levels
   zones_d07=[];
   for i=1:number_of_zones
```

```
d07=find(firm_individual_exposure(i,:)>=0.7);
zones_d07=[zones_d07 length(d07)];
end
zones_d05=[];
for i=1:number_of_zones
d05=find(firm_individual_exposure(i,:)>=0.5 & firm_individual_exposure(i,:)<0.7);
zones_d05=[zones_d05 length(d05)];
end
zones_d03=[];
for i=1:number_of_zones
d03=find(firm_individual_exposure(i,:)>=0.3 & firm_individual_exposure(i,:)<0.5);
zones_d03=[zones_d03 length(d03)];
end
zones_d01=[];
for i=1:number_of_zones
d01=find(firm_individual_exposure(i,:)>=0.1 & firm_individual_exposure(i,:)<0.3);
zones_d01=[zones_d01 length(d01)];
end
zones_d00=[];
for i=1:number_of_zones
d00=find(firm_individual_exposure(i,:)>0 & firm_individual_exposure(i,:)<0.1);
zones_d00=[zones_d00 length(d00)];
end
min_zones_d07=min(zones_d07);
min_zones_d05=min(zones_d05);min_zones_d03=min(zones_d03);
min_zones_d01=min(zones_d01);
min_zones_d00=min(zones_d00);
```

E.2 Subsampling

```
% Select firms randomly to equate the number of firms in each zone
at several sales levels
   rnd07=[];
   for i=1:number_of_zones
   d07=find(firm_individual_exposure(i,:)>=0.7);
   rnd07=[rnd07 randsample(d07,min_zones_d07)];
   end
   rnd05=[];
   for i=1:number_of_zones
   d05=find(firm_individual_exposure(i,:)>=0.5 & firm_individual_exposure(i,:)<0.7);
   rnd05=[rnd05 randsample(d05,min_zones_d05)];
   end
   rnd03=[];
   for i=1:number_of_zones
   d03=find(firm_individual_exposure(i,:)>=0.3 & firm_individual_exposure(i,:)<0.5);
   rnd03=[rnd03 randsample(d03,min_zones_d03)];
```

```
end
rnd01=[];
for i=1:number_of_zones
d01=find(firm_individual_exposure(i,:)>=0.1 & firm_individual_exposure(i,:)<0.3);
rnd01=[rnd01 randsample(d01,min_zones_d01)];
end
rnd00=[];
for i=1:number_of_zones
d00=find(firm_individual_exposure(i,:)>0 & firm_individual_exposure(i,:)<0.1);
rnd00=[rnd00 randsample(d00,min_zones_d00)];
end
% third, remove all other firms
```