DISCUSSION PAPER SERIES

DP15503

FIRMS' EXPOSURES TO GEOGRAPHIC RISKS

Bernard J Dumas, Tymur Gabuniya and Richard C Marston

FINANCIAL ECONOMICS

ISSN 0265-8003

FIRMS' EXPOSURES TO GEOGRAPHIC RISKS

Bernard J Dumas, Tymur Gabuniya and Richard C Marston

Discussion Paper DP15503
Published 29 November 2020
Submitted 23 November 2020

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

FIRMS' EXPOSURES TO GEOGRAPHIC RISKS

Abstract

The distinction between domicile and place of business is becoming more and more relevant as a growing number of firms have activities abroad. In most statistical studies of international stock returns, a firm is included in a country's index if its headquarters are located in that country. This classification scheme ignores the operations of the firm. We propose, instead, to measure the firm's exposures to "geographic zones" according to the place where they conduct business. As a representation of "geographic risks", we synthesize zone factors from all firms in the dataset, be they domestic firms or multinationals. And we show the properties of the exposures to the zone factors.

JEL Classification: C4, F3, F6, G1

Keywords: stock return indexes, stock return exposures, geographic investing, factor models, country factors, expectations-maximization algorithm

Bernard J Dumas - bernard.dumas@insead.edu
INSEAD and CEPR

Tymur Gabuniya - gabuniya.tymur@gmail.com
Birkbeck College, University of London

Richard C Marston - marstonr@wharton.upenn.edu
University of Pennsylvania, Wharton School

Acknowledgements
Dumas's work received the support of a grant from the INSEAD Research Fund. 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 are also grateful for the strong collaboration of Tymur Gabuniya. We thank Philippe Piette of WVB, who generously provided the data, and Winston Dou who suggested subsampling to us.

Firms' Exposures to Geographic Risks*

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

November 22, 2020

Abstract

The distinction between domicile and place of business is becoming more and more relevant as a growing number of firms have activities abroad. In most statistical studies of international stock returns, a firm is included in a country's index if its headquarters are located in that country. This classification scheme ignores the operations of the firm. We propose, instead, to measure the firms's exposures to "geographic zones" according to the place where they conduct business. As a representation of "geographic risks", we synthesize zone factors from all firms in the dataset, be they domestic firms or multinationals. And we show the properties of the exposures to the zone factors.

*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. Marco Del Negro provided us with an Expectation-Maximization computer code; we are grateful to him. We thank for comments participants of an INSEAD Brownbag workshop and of a seminar at Katholieke Universiteit Leuven.

> "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)

How can one estimate the business risks of operating in one country? If a country were completely closed to the outside world, the stock returns of firms in that country would provide a good measure of the risks of operating there. But in an interconnected world, it is not clear which stock returns one should turn to. The traditional way to measure the business risks in a country is to use the stock returns of firms with headquarters in that country. But some firms domiciled in a country may have the bulk of their activities outside of that country. 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 taking place in the Netherlands. These national stock indexes, therefore, reflect the business risks of operating in many countries, not the risks of that single country. National stock indexes also leave out firms based elsewhere that sell and produce goods in that country, even though their stock prices also reflect the business risks of operating in that country. Furthermore, as time goes by, the composition of commonly available country indexes evolves not just as to the list of firms included in the index but also as to the locus of the business conducted by the firms that are included. We aim to use the information on the stock returns of all firms, not just domestic ones, to construct a new type of country stock market index. We develop a way to put together stock market data and data on the firms' operations to create such indexes. We synthesize purely domestic firms from all firms in the dataset. So our geographic indexes are quite different from national stock indexes that include firms in one country on the basis not of their operations, but of their domicile. We will later compare our indexes with indexes of "national firms" based on a given ISIN country label.[1] We adopt our new indexes as definitions of "geographic risks" and estimate individual firms' exposures to (or loadings on) them.

To measure the business risks of operating in a given geographic location, we use the information each firm provides on the share of revenues earned in that location. Admittedly, it should not be the share of revenues alone that serves to measure business risks stemming from that location. We should use the share of free cash flows generated there. But that data, unfortunately, is not available to anyone. Revenues are at best a proxy. For that proxy to be valid at least approximately when capturing the risks of growth rates, we have to be assuming that most operating costs are occurring in the same countries and regions where there are revenues. In that case, the distribution of revenues across countries and regions will be a proxy for the distribution of free cash flows. In short, it is simply assumed, for lack of a better way, that the free cash

[1] The International Securities Identification Number (ISIN) system assigns a two-letter country label to each firm's stock based on their headquarters.

flows of a firm reflect only the risks of the places to which it sells its products, irrespective of its domicile.

In the geographical segment tables that are required in annual reports, firms choose the way to segment their sales internationally. They might choose to identify their main trading partners, but they often group other revenues by region rather than country and sometimes refer to a residual category like "rest of Europe" or "rest of the world". After standardization, we refer to destinations of sales, which are countries or regions, by the generic term "zones." As discussed below, we use whatever information is available to place equality and inequality constraints on some of the loadings of the firms on these zones while we estimate the other loadings. The econometric method is discussed in detail in later sections. Its role is to fill in whichever information on revenues is not available directly.

Our reconstruction should be relevant for at least two purposes. First and principally, corporations contemplating a capital investment in a production or distribution facility in a particular country need to have a proper description of the risks inherent in operating facilities in that country, and not of the risks inherent to being domiciled in the corresponding country.[2] Second, a portfolio manager making an investment decision often does so because he wants to take a view concerning the economic prospects of a country. He may, for instance, see higher growth prospects in that country than other investors do and may accordingly want to pursue a strategy of investing in companies that do business in that country. Investing in the corresponding country's stock market index is not a clean way to implement that strategy if the companies traded on the country's exchange conduct a good deal of business outside the country. In the reverse, an investor may want to invest in a country but fear the form of trading taking place in that 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 choose companies in another country that do a lot of business in the country that is targeted, taking care to hedge away the business that these companies conduct at home and in other countries.[3]

We stress that in the present paper, *except for one generic asset-pricing assumption, which we spell out in Section 3*, we make no specific assumption about asset pricing and/or about the degree of integration of financial markets. And we do not investigate average returns**,** as would be needed to test a particular asset-pricing model. We only estimate the zone risk factors – making some statistical assumptions to be stated below – and their correlations, as well as the firms' exposures to them. Asset pricing theories that have been developed in other papers may come into play when interpreting these correlations. Ours is purely a descriptive statistical model. We are not implying that the risk factors

[2] We are not denying that there may also exist risk premia for being listed in one country. See Froot and Dabora (1999) and Chaieb et al. (2020). At present, we do not have a model to explain such "local-pricing," if it exists. But see the role of "foreign sentiment" in Dumas et al. (2017).

[3] Subsidiarily, our undertaking may enhance the meaningfulness of cross-country correlation studies. See Section 7.2.

we identify are priced factors under any particular asset-pricing theory.[4]

The word "descriptive" is apt. Although the econometric technique we use is powerful enough to allow us to estimate the factor loadings for several thousands of firms each year, statistical tests to accept or reject the model as a whole are out of reach because of sheer size, and would not reject any particular hypothesis about the loadings given the large number of parameters being estimated. Instead, we show that the zone factors themselves are economically reasonable and meaningful and we run several statistical tests to show that the indexes we construct do reflect underlying economic variables significantly.

Three antecedent articles are close to ours in their method and their goals. First in chronological order, Heston and Rouwenhorst (1994) estimated a factor model that decomposed each firm's return into the sum of a country and an industry factor.[5] The loadings only take values of 0 or 1, depending on which country and which industry it belonged to. We use additional, more continuous and quantitative data on revenues, thereby allowing loadings to take values between 0 and 1.

A second reference is Diermeier and Solnik (2001) which examined 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. They have available 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.[6] The statistical analysis can be described as comprising three stages. In a first stage, they construct a domestic 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 the value-weighted average of country stock returns and currency returns for countries that belong to a region. In a second stage, 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 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 more efficient to use all firms to do everything in a single estimation. That is, knowing the geographic distribution of activities of each firm, one should use all the information on all the firms to identify the pure country factors. A Japanese firm, to the degree that it conducts operations

[4] Thus, what we do is not incompatible with the existence of extra risk premia not related to the risks we identify, as in Barrot et al. (2019) or Hoberg and Moon (2019).

[5] See also Griffin and Karolyi (1998).

[6] There are three regions: Asia, Europe and the United States.

in Switzerland, should also help in identifying the Swiss zone factor.

A third reference that is germane to our paper is Brooks and Del Negro (2006, alternatively, 2004). These authors re-estimated the Heston and Rouwenhorst factor model constraining again the loadings of a firm that belonged to a country and an industry to be equal to 1 on those two factors but leaving the other loadings unconstrained and, therefore, not necessarily equal to 0. The resulting number of loadings to be estimated was large. Crucially for us, this feature led the authors to propose a method that could be calculated firm by firm, as opposed to globally for all firms, namely the Expectation-Maximization way of maximizing likelihood (referred to as EM below).

We expand their work by introducing a more flexible factor model that we estimate by means of an innovative econometric approach: using revenue data provided by each firm, we add constraints to the EM algorithm in order to restrict the estimation of factor returns. In that way, our loadings are neither arbitrarily set equal to 0 or 1, nor for all of them mere statistical estimates; many of our loadings reflect actual economic data coming from annual reports. Because of the constraints, our factors are not orthogonal: geographic risks are not independent of each other, so that we can meaningfully discuss their correlations.

Further afield, several authors have, like we do and like the three previous references did, utilized data beyond stock returns to explain their correlation structure. A few studies such as Ammer and Wei (1996), Baele and Soriano (2010), Viceira and Wang (2018), and Akbari et al. (2019) have sought to distinguish between a common cash-flow dynamic (interpreted as economic integration) and a common risk-pricing dynamic (interpreted as financial integration). Viceira and Wang (2018), for example, find that the increase in correlations between stocks cannot be attributed to increased correlation between cash flow shocks and would, therefore, plausibly come from discount rate shocks. The geographic origin of the cash flows is not investigated whereas the origin of revenues is our main focus. 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.[7] A pair of studies by Bekaert et al. (2011, 2013) took a different approach by examining differences in earnings yields rather than their correlations or correlations of stock returns.[8] Finally, Bae et al. (2019) use bilateral export date from developed countries to form emerging market country indexes based on the share prices of developed country firms that sell to emerging markets. The revenue data that we use are broader than export data (because a multinational's foreign sales to an emerging market need not involve any exports at all or at least any exports from its country

[7] Heston and Rouwenhorst (1994) indicate that, even between European financial markets, which were presumably fairly well integrated, country factors as compared to industry factors offer the higher diversification potential. Goetzman et al. (2005) shows that the correlation has been present mostly between "core" (basically developed) countries as opposed to other countries.

[8] They find evidence that markets are becoming more correlated, but like Goetzman et al., they find that is truer for developed than for emerging markets.

of domicile). As Bae et al. explain, however, revenue data are not generally available on a bilateral basis for many countries, so they could not use those to form emerging market country indexes using their method.[9]

The balance of the paper is organized as follows: Sections 1 and 2 describe the dataset of the firms' geographic segments 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 and extended in Section 4. Section 5 addresses the problem of data imbalance. Section 6 analyzes the composition of the geographic indexes, while Section 7.1 compares the behavior of the geographic and traditional indexes across countries and Section 7.2 compares the geographic and traditional indexes across years. Section 8 highlights two key features of the exposures revealed by the geographic indexes. Section 9 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' revenues across geographic segments. We elected to study all the firms in the database provided that the Datastream database contained stock return data for them. Our study covers the years 1999-2014 inclusive. Unfortunately, annual reports do not contain information on the distribution of the firms' purchases across geographic segments. Only revenues are reported, with few exceptions.

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. The most important filter applied to the original dataset deleted all microcaps. That filter alone reduced the number of securities in 2014 from 17,678 to 6,690. The filtering process gave us a set of firms that grows from 1,797 in 1999 to 6,335 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* themselves in the geographical segment tables of 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 revenues database. Firms often list specific countries responsible for major portions of their revenues (e.g., a French firm listing revenues earned in Germany and the U.S.). But firms also list whole regions rather than countries (e.g., Asia). And firms often list "Rest of" segments including "Rest of Europe" in the case of a French firm or "Rest of World", containing countries that *the firm* has not referred to explicitly among its segments. Typically, the annual report also refers to a country called "Home" or to "Domestic revenues".

[9] See also Bodnar and Marston (2002) and Bodnar et al. (2002).

By way of contrast, we call "*zones*" standard geographic entities defined *by us* that are uniform across all the firms of our sample. Twelve of these zones are countries in our sample that are commonly cited by firms in their geographical segment tables. A thirteenth zone is "Rest of World" that contains countries for which *we* have not defined a specific zone. In 2014, the stock markets of the twelve countries specifically identified as zones comprised 74.8% of the total capitalization of the world's stock markets.[10] As an example of the way these zones are referred to in annual reports, consider a firm that has an ISIN starting with the prefix "US" and indicates that its geographic segments are Canada, Mexico, China, and Rest of World. Canada and China are among our twelve zones, and the Rest of World is our residual zone. Since Mexico is not included among our twelve zones, we will classify its revenues as being generated in the "Rest of World" zone.

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 revenues sharpen (or restrict) the estimation of factor returns. The firms from the developed countries have revenues 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 revenue information. Table 1 provides a description of the database. It indicates, for each zone, the number of companies year after year that derive more than 10% of their revenue from 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 derive 10% or more of their revenues from Singapore. So those multinational firms help to determine the Singapore zone factor. In the same year, 105 firms in the database are domiciled in Germany, but there are another 88 firms with a fraction 10% or higher of their revenues from 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 and exposures 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

[10] The source of the world's stock market capitalizations is the World Federation of Exchanges database.

to that zone in all or most years. 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.[11] We return to that problem in Section 5 below.

2 National indexes vs. indexes based on revenues

A stock security's exposure to risks requires obviously a definition of the risk factors, here "geographic risks." For that we need stock market indexes that capture those risks meaningfully.

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 company's headquarters: a firm is included in the index of a country when it is domiciled in that country. These indexes are explicit (as opposed to the one we develop here, which are latent or implicit) 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.[12] We call these "$ISIN$" or "national" indexes and we refer to all firms domiciled in one country as "national firms." As noted, they present the drawback of not accurately reflecting the risk of operating in the country.

Other explicit indexes can be constructed from individual stock returns on the basis of revenue data. We construct a second index, which we call the "70%" index consisting of firms (which we call "domestic firms") with 70% or more of their revenues derived explicitly from one country (which is almost always the country where the firm is domiciled).

Both the national and domestic indexes have the drawback of using fragmentary information. The national indexes are based on the firm's domicile only, while the domestic indexes ignore information about the firms' foreign revenues. We aim to show now that these indexes are not good enough for gauging exposure to geographic risks. 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, e.g., Brazil

[11] The database is compiled in Malaysia.

[12] 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 listings 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 derive more than 10 percent of their revenues from that zone.

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 indexes from sales levels below 70%. Even when some indexes are highly correlated with each other they can lead to quite different exposure coefficients, as we shall see in Section 8. Roll (1977) first pointed out that point of algebra apropos the estimation of the beta of a security against a market index proxy.

Second, as a way to show the potential benefits of using revenue information to inform a risk-factor model, – which the national indexes do not do –, 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: over time the proportion of national firms is falling. 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 revenues 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.

Third, the 70% indexes do capture revenue information, but only the information coming from domestic firms. Below we aim to utilize as well the information from more diversified firms (multinationals) both domestic and foreign, at all levels and not just the 70% level.

The statistical index to be developed in the next sections will allow us to put together all geographic information coming from both stock-returns and revenues, whatever be the form in which revenues are reported: specific destination countries or, in less detail, destination regions such as "Europe", "Middle East" etc..

3 A model of risk exposure

If information were available giving for each firm its share of revenues from 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' revenue fractions. That matrix inversion would directly construct the zone returns from the company returns. In practice, however, the information about geographic segments is not exhaustive. Because of the missing information, the zone 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 factor model is a significant elaboration over Brooks and Del Negro (2004). They considered a model with country factors (and a separate

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
RoW	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: **Correlations between $ISIN$ and 70% indexes.**

	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, which we do not need here since our zone factors are not assumed to be independent), 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 world factor.[13]

Our model specifies the structure from which geographic-zone index returns C and loadings B 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 rates of 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 or *latent* 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. We list below the constraints on B that will enable us to calculate C as a latent factor.

In the model, we identify the exposures of the growth rate of free cash flows to shareholders with the exposures of stock returns. Since stock returns and, therefore, stock prices are involved, we need a generic asset-pricing context that will relate the second moments of returns to the second moments of cash-flow growth. For that we refer to the approximate identity of Campbell and Shiller (1988), which decomposes stock returns into news about cash flows

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

("dividends") and news about future returns, the latter being a factor common to all securities.[14] Our model (1) being a factor model, we endeavor to explain, subject to the constraints, as much as possible of the variance of R_t by means of the second moments of the factors C_t. When we do that, we shall make the assumption that the variance-covariance D of e_t is diagonal, so that, technically, an additional common factor is ruled out. However, the estimated e_t will still contain a very large amount of residual commonality.[15]

Our measurement is not otherwise tied to a particular asset-pricing setting. In particular, we leave aside the implications of the framework for the first moments of returns and we make no assumption about the integration of financial markets, as the framework is equally valid under integration and under segmentation.[16]

We use whatever information we have on firms' activities to set some of the elements of B equal to the corresponding shares of activities. We assume in the model that the (deleveraged) stock returns of a firm reflect 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 third-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 the shares of revenues that SmithKline Beecham receives from the various geographic segments. The firm receives 8% of its revenues from the UK, 12% from Asia, 23.5% from Europe-ex-UK and 46.1% from 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 on the statistical estimation are as follows:

Assumption 1 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 revenues from that zone or region

$$\sum_{k \epsilon j} b_{i,k} = A_{i,j} \tag{2}$$

[14] With recursive utilities, Restoy and Weil (2011) in the domestic context and Dumas et al. (1993) in the international context relate news about future returns to news about future consumption. See Equations (10) and (17) in Dumas et al. (1993).

[15] This remark is relevant because the extant literature shows that the changes in correlations between stock market indices are mainly driven by unobservable factors rather than observable macroeconomic variables (see, e.g., King et al. (1994), Karolyi and Stulz (1996), Ammer and Mei (1996), and Bekaert et al. (2011)). Here we use individual-firm stock returns and revenues, which are microeconomic variables. But we can also expect commonality in the estimated residuals.

[16] See Dumas et al. (1993) for the appplication of the framework separately to the case of integration and to the case of segmentation, the only difference between the two being the consumption terms of the households of each country.

The assumption captures whatever information about country and regional revenues is given in annual reports. For example, $b_{i,k=i's\ country} = A_{i,k=i's\ country}$: the firm's loading on one particular zone factor called "home" is assumed equal to the share of revenues generated at home. But the home country does not play a unique role, as the general form (2) of the restriction shows.[17] Assumption (2) is the statistical rendition of the economic "proxy" assumption that we stated in the introduction: the free cash flows of a firm reflect only the risks of the zones to which it sells its products, irrespective of its domicile.

Assumption 2 *The loadings of a firm i on all the zone factors are non negative*

$$b_{i,k} \geq 0$$

In other words, if the revenues of a firm coming from a particular source grow more than expected, everything else equal, its stock return is higher than expected.

Assumption 3 *The sum of the loadings of a firm i equal* 1

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

Admittedly, firms could be leveraged operationally so that their total risk would be more than what is captured simply by revenues.[18] The assumption holds if costs are almost entirely variable. Since the shares of revenues sum to 1, Assumption 1 forces us to assume that the loadings must sum to 1 as well.[19]

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.

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

The statistical implementation of the model (1) is close in spirit to that of Brooks and Del Negro (2004). The technical aspects are, however, quite innovative since we implement the EM technique in the presence of constraints, including inequality constraints. Appendix B develops the econometric theory under constraints in general terms.[20]

We have described the constraints in Section 3. Because of their presence, zone factors are not independent of each other (which is the reason for which we

[17] 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.

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

[19] By comparing in Section 8 the exposures in the form of the constrained loadings to less constrained exposures, we shall have an opportunity to gauge the reasonableness of these restrictions.

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

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.[21] 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.[22]

Since EM is a form of likelihood maximizations, the following is needed:

Assumption 4 *All random variables are multivariate IID normal within a year.*

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):

$$\begin{aligned} \mathcal{L}\left(B, D, \Omega\right) &= \frac{-NT}{2}\ln\left(2\pi\right) - \frac{T}{2}\ln\left|\Sigma\right| - \frac{T}{2}\text{trace}\left(S\Sigma^{-1}\right) \\ &= \frac{-NT}{2}\ln\left(2\pi\right) - \frac{T}{2}\ln\left|\Sigma\right| - \frac{1}{2}\sum_{t=1}^{T} R_t^{\intercal}\Sigma^{-1}R_t \end{aligned} \tag{3}$$

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).[23] As was pointed out by Brooks and Del 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 be infeasible. Our

[21] Although the estimation will attempt to make the model fit these assumptions, they will not hold for the model estimated from the data. The number of zone factors C is much smaller than the number of 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.

[22] 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.

[23] $\mathbb{E}$ 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 missing (i.e., zero-return) data, as explained in Appendix C.

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 (3) of R, $\ln p\left(R; B, D, \Omega\right)$, to the *joint log-likelihood* of R and C, $\ln p\left(R, C; B, D, \Omega\right)$, 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.[24]

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)):

$$\begin{aligned} LL\left(B, D, \Omega\right) &= -\frac{T}{2}\sum_{j=1}^{N}\ln D_j - \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{N}\frac{\left(R_{j,t} - B_{\text{row }j}C_t\right)^2}{D_j} \\ &\quad -\frac{T}{2}\ln\left|\Omega\right| - \frac{T}{2}\sum_{t}C_t^{\intercal}\Omega^{-1}C_t \\ &= -\frac{T}{2}\ln\left|D\right| - \frac{1}{2}\text{trace}\left\{D^{-1}\left[\sum_{t=1}^{T}R_tR_t^{\intercal}\right.\right. \\ &\quad \left.\left. -2\sum_{t=1}^{T}R_tC_t^{\intercal}B^{\intercal} + B\sum_{t=1}^{T}C_tC_t^{\intercal}B^{\intercal}\right)\right\} \\ &\quad -\frac{T}{2}\ln\left|\Omega\right| - \frac{1}{2}\text{trace}\left(\sum_{t=1}^{T}C_tC_t^{\intercal}\Omega^{-1}\right) \end{aligned} \tag{4}$$

We calculate the expected value of LL given the observations R, at the currently estimated values of the parameters B, D, Ω. That is, we integrate (4) over C. This gives:

$$\begin{aligned} \mathbb{E}\left[LL|R\right] &= -\frac{T}{2}\ln\left|D\right| - \frac{T}{2}\text{trace}\left\{D^{-1}\left(S - 2XB^{\intercal} + BYB^{\intercal}\right)\right\} \\ &\quad -\frac{T}{2}\ln\left|\Omega\right| - \frac{T}{2}\text{trace}\left(Y\Omega^{-1}\right) \end{aligned} \tag{5}$$

where:

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

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

$$Y_{K\times K} \triangleq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[C_tC_t^{\intercal}|R_t\right] \tag{7}$$

[24] See Appendix B.

The sufficient statistics that are contained in (6) and (7) 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 (5) 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 (6) and (7), 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 constraints 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.[25] 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:[26]

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

During the execution of the $\mathbb{EM}$ steps, the full likelihood (3) is calculated periodically to verify that it keeps increasing. That calculation is computationally intensive. $\mathbb{EM}$ theory guarantees that a local maximum is reached. No likelihood maximization algorithm – except in very simple, quadratic cases such as OLS – ever guarantees that a global one is reached.

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

$$\mathbb{E}\left[C_t | R_t\right] = \left(\Omega^{-1} + B^\intercal D^{-1} B\right)^{-1} B^\intercal D^{-1} R_t \tag{8}$$

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.

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).

The convergence of the $\mathbb{EM}$ algorithm is considered to be achieved whenever the largest absolute value of the elasticity of the likelihood (or, equivalently, the relative gradient of the log-likelihood) with respect to any parameter is lower

[25] These optimizations, except for the constraints, are analogous to a time-series regression run for each firm.

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

than 10^{-2}.[27] The maximum number of iterations allowed is 500. For the year 2013, we experimented with a larger number of iterations equal to 1000. The gain in the stability of the estimates was close to null.[28]

Because of time differences around the globe, one might object to the analysis of covariations being conducted on daily stock- return indexes. For that reason, we have entirely redone the estimation for the year 2013, assembling daily returns into non overlapping three-day returns. We could not detect any difference in the results that are reported below.

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, especially since there are so many listed firms from the U.S. and Japan in the database, some of them selling to their own country.

As in any factor analysis, the likelihood maximization itself aims to explain as much of the variance of individual stock returns as possible. That is true also under constraints. 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 revenues from Europe, the volatility of returns of European zones would end up being abnormally large. When we initially ran the algorithm, 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, 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 contains an approximately equal number of firms selling to each and every zone, at each $x\%$ level. We have reproduced in Appendix E the actual set of MatLab instructions that was used.

After that is done, the estimates of the zone indexes C, as per Equation (8),

[27] In that gradient we include the Lagrange multiplier terms, which are due to the constraints; see Appendix B. The log-likelihood is of the order of magnitude of 10^6.

[28] We keep available upon request descriptive statistics on the variability of the estimates of the loadings, compared between the first 500 iterations and the subsequent ones.

are averaged across the subsamples and a single additional iteration of the $\mathbb{M}$ step run on the 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 revenues, 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 indexes of different types. In this section, we comment on 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 refers to the weight given to each firm in the calculation of the estimated zone indexes. It is dictated by formula (8), which we reproduce here:

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

and which can be compared with cross-sectional Generalized Least Squares: $\left(B^\intercal D^{-1} B\right)^{-1} B^\intercal D^{-1} R_t$. 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.[29]

Thus, the weights of the various securities in the various $\mathbb{EM}$ indexes are $\left(\Omega^{-1} + B^\intercal D^{-1} B\right)^{-1} B^\intercal D^{-1}$.[30] 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.

[29] 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.

[30] The weights of the firms in the indexes (a matrix with as many rows as there are zones and as many columns as there are firms) should not be confused with the loadings B of the firms on the indexes (a matrix with as many rows as there are firms and as many columns as there are zones). The loadings satisfy a number of constraints, which we have indicated, including non negativity constraints. The weights can be of any sign, some firms receiving a negative weight in some zone indexes, for reasons to be explained below.

	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: **Correlations between $ISIN$ and $\mathbb{EM}$ indexes.**

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 firms' loadings (Canada, 2014)** 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.

The formula indicates that the relation between the composition of the index 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 revenue database while others are estimated, which complicates the relation between the weights and explicit revenues.

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 revenues from the zone. But, 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 but serve to offset, by means of a negative weight, the influence of the unrelated returns of the second tier.

To illustrate the role of the third tier and firms receiving a negative weight, consider the example of the Canadian zone in 2014. 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. By comparing the two rows, one can remark how the algorithm has been able to choose weights that cancel the unwanted loadings and to focus on Canada alone.

7 The behavior of $\mathbb{EM}$ vs. $ISIN$ indexes

In the coming sections, we produce a series of arguments and statistical tests that show that the zone factors themselves are economically reasonable and meaningful. This is in lieu of formal tests – such as likelihood-ratio tests – of the model as a whole or of specific hypotheses, which are out of reach for lack of knowledge of the sampling distributions and, in any case, would not reject any hypothesis about the loadings because of the sheer number of parameters being estimated.

The purpose of this first section is to show that $\mathbb{EM}$ indexes do capture "geographic risks". To do that, we compare them across zones and we examine their behavior over time.

7.1 Comparisons across zones

The $\mathbb{EM}$ indexes differ from traditional indexes primarily because they incorporate information from the geographic revenue 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 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 not the same.

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 revenues from outside Canada. In other countries like India and Malaysia, more firms receive domestic revenues primarily. In India, for example, only 30.1% of the firms receive substantial revenues from 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 revenue data reported by national firms influence the $\mathbb{EM}$ indexes. To show how revenue 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 (foreign presence). In the case of Canada in 2014, for example, 43.9% of the firms selling in Canada were Canadian. Call this ratio "Inward".

- The ratio of the number of national firms with foreign revenues equal to 30% or more of total revenues to the total number of national firms (foreign activity). In the case of Canada in 2014, 47.7% of its firms are "multinational," defined as firms with 30% or more of their revenues from abroad. Call this ratio "Outward".

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 revenues influence the $\mathbb{EM}$ indexes, consider the link between revenue patterns and the $\mathbb{EM}$ vs. $ISIN$ correlations. The higher the ratio of national firms to all firms selling to that zone the higher should be the correlation between the $\mathbb{EM}$ and $ISIN$ indexes. So the revenue pattern should be positively correlated with the correction between the $\mathbb{EM}$ and $ISIN$ indexes. The regression line for the pooled sixteen years across the twelve zones (not including RoW) is

$$\text{corr}_{\mathbb{EM},ISIN,j,t} = 0.8575 + \underset{(3.507)}{0.1062} \times Inward_{j,t} + \varepsilon_{j,t}$$

where j is a subscript for zones and t a subscript for years. The t statistic in parentheses indicates that the slope is significantly different from zero and positive.

To make better sense of this data, let us focus on two zones with markedly different behavior, Canada and India. Figure 1 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. The higher the ratio of multinationals – the Outward ratio –, the lower should be the correlation between that zone's $\mathbb{EM}$ index and the corresponding $ISIN$ index.

The regression line is

$$\text{corr}_{\mathbb{EM},ISIN,j,t} = 0.9471 - \underset{(-3.0810)}{0.0844} \times Outward_{j,t} + \varepsilon_{j,t}$$

with a significant slope again.

Let us again just compare Canada and India, without a figure. In Canada over the sample period, 43.6% of firms can be described as multinational with 30% or more of their revenues from abroad. In India, only 17.9% of 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).

For the twelve zones, we also run a multiple regression on both ratios:

$$\text{corr}_{\mathbb{EM},ISIN,j,t} = 0.8909 + \underset{(2.7183)}{0.0856} \times Inward_{j,t} - \underset{(-2.1589)}{0.0610} \times Outward_{j,t} + \varepsilon_{j,t}$$

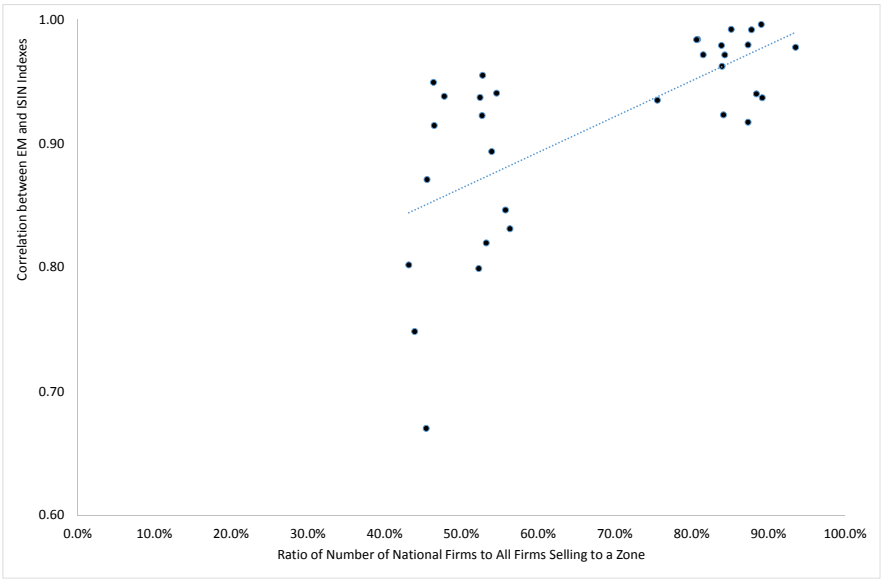

Figure 1: **Relative Importance of a Zone's Firms in Each Zone Index – the case of Canada and India only**: on the x axis is the "Inward" ratio of the number of national firms to all firms selling to a zone. On the y axis is correlation between $\mathbb{EM}$ and $ISIN$ indexes. Each point is a year and a zone (two zones only).

The slopes are both significant.

As the section on the yearly variations approaches, it becomes clear that it is unwarranted to keep the intercept of these regressions the same from one year to the next. Doing so might bias the slopes. When we add control dummies for years, the result is

$$\text{corr}_{\mathbb{EM},ISIN,j,t} = \text{intercept}_t + \underset{(4.8777)}{0.1307} \times Inward_{j,t} - \underset{(-2.1816)}{0.0508} \times Outward_{j,t} + \varepsilon_{j,t}$$

Under all specifications, it is clear that the $\mathbb{EM}$ index differs from the $ISIN$ index in the direction that gives it a geographic meaning that $ISIN$ lacks.

7.2 Comparisons across years

The graphs in Figure 2 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 2 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 focused on one zone each. 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, the bottom graphs of Figure 2 display mean and median statistics of the pairwise correlations between zones for $\mathbb{EM}$ zone indexes on the one hand and for the two explicit country indexes on the other. The difference tends to be negative: $\mathbb{EM}$ indexes are less correlated across zones than are the explicit indexes. The $\mathbb{EM}$ algorithm has been able to remove an undue amount of correlation caused by revenues from 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 goal of our exercise, which was to deconstruct that interpenetration.

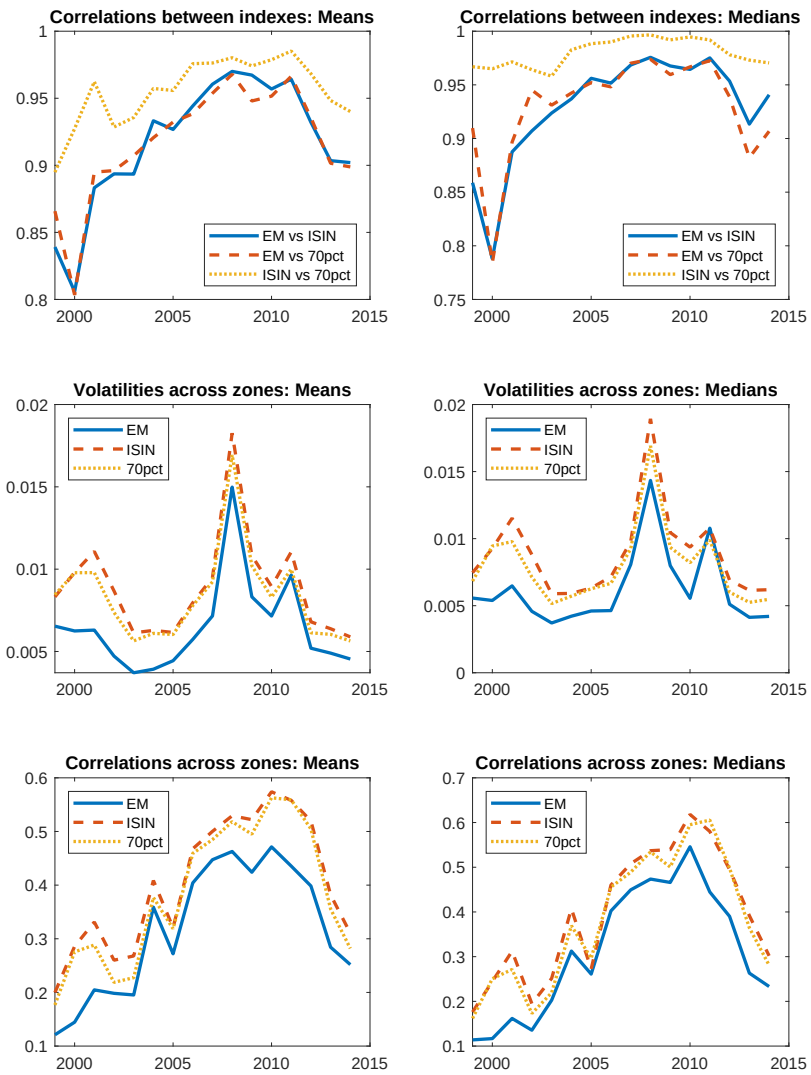

Figure 2: **Comparison and evolution** of second-moment properties of $\mathbb{EM}$, $ISIN$ and 70% indexes. For each year, we obtain the daily standard deviations and pairwise correlations (across indexes in the top pair of graphs, across zones in the bottom pair), of the thirteen zone indexes, of which we take the mean and the median.

We now show evidence that this reduction in zone correlations for the $\mathbb{EM}$ indexes is related to the Inward foreign-presence and Outward foreign-activity ratios introduced earlier. Both ratios capture ways in which firms domiciled in one country generate revenues outside their country thereby causing $ISIN$ indexes to differ from indexes reflecting activities in one country alone. We will show that both of these ratios have less influence on the cross-zone correlations in the $\mathbb{EM}$ regressions than the correlations in the $ISIN$ regressions because the $\mathbb{EM}$ indexes are designed to be more distinctly zone-specific.

In Table 6, we report regressions of the $ISIN$ and $\mathbb{EM}$ cross-zone correlations on the Inward and Outward measures pooling across all years. Consider first the $ISIN$ regressions. Suppose that we compare two country pairs: India and Malaysia, on the one hand, the Inward ratios of which are high and the Outward ratios of which are low; Canada and Singapore, on the other hand, the Inward ratios of which are lower and the Outward ratios are higher. As we move from the second country pair to the first one, the correlation across zones should drop since the economies of the first type are less correlated. We expect the slope coefficients of the Inward ratios to be negative and the slope coefficients of the Outward ratios to be positive. The regressions for $ISIN$ in Table 6 exhibit these very signs.

For the $\mathbb{EM}$ indexes, these effects should be less pronounced since we have constructed the $\mathbb{EM}$ indexes so that they are more distinctly zone-specific than the $ISIN$ ones, $\mathbb{EM}$ having been purged to some extent of the influence of foreign presence and foreign activity. In other words, the $\mathbb{EM}$ cross-zone correlations should be less dependent on the ratios than the $ISIN$ correlations. We do find that the R^2 for $\mathbb{EM}$ is lower than the one for $ISIN$. We also find that the absolute values of three of the four Inward and Outward coefficients are lower in the $\mathbb{EM}$ regressions.[31]

Figure 2 shows that traditional $ISIN$ indexes exhibit an upward trend and then a downward trend in their correlations. After we have controlled for changes in cash flows – revenues as a proxy –, the rise and fall of correlations is still present.

Asset pricing would come into play when interpreting these correlations as being indications of the degree of integration of financial markets.[32] In this

[31] And two of these three show a significant difference between the two index types (third column). These are the slopes on the Outward ratios. The Outward ratios measuring foreign presence in a zone seem to matter the most in distinguishing the $\mathbb{EM}$ from the $ISIN$ indexes. That is consistent with the significance of the slopes in the regressions of Section 7.1.

[32] One must beware that different assets could be driven by different factors, and thus possibly be little correlated, without implying segmentation. It is true, based on dynamic models of international financial-market equilibrium, that, cash flows being kept the same, cross-country correlations are higher if the financial markets are integrated (i.e., if movements of capital take place between countries the same way they do within countries) than if they are segmented. See Dumas et al. (2003). Bekaert et al. (2011) present a similar dynamic model but do not examine correlations of stock returns. 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 (2019)). The catch is the proviso "cash flows being kept the same" As noted in the introduction, as time goes by, the composition of commonly available country

	$ISIN$	$\mathbb{EM}$	$\mathbb{EM}-ISIN$
Intercepts with demeaned regressors	0.37273 (77.195)	0.28674 (48.58)	−0.085988 (−25.653)
Slope coefficients			
Inward ratio one zone	−0.24273 (−10.168)	−0.33572 (−11.504)	−0.089547 (−5.4832)
Outward ratio one zone	0.56521 (17.447)	0.391 (9.8733)	−0.17739 (−7.9502)
Inward ratio other zone	−0.16411 (−4.5455)	−0.16327 (−3.6994)	0.0060439 (0.2672)
Outward ratio other zone	0.17611 (8.1988)	0.052175 (1.987)	−0.12519 (−8.4982)
Year dummies	Yes	Yes	No
Number of observations	1056	1056	1056
R^2	0.55	0.192	0.133

Table 6: **Comparison of pairwise zone correlations of $\mathbb{EM}$ and $ISIN$ indexes**. For each year, we obtain the pairwise correlations across all zone pairs, pool them over the years and regress them against the Inward and Outward zone ratios described in Section 7.1: t-statistics are in parentheses.

paper, as noted in the introduction, we have made no assumption about asset pricing (other than a generic one that relates stock returns to cash flow growth) and/or about the degree of integration of financial markets. We have only calculated the zone factors and their correlations – making some other assumptions, stated above.

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

In this section, we examine firms' exposures to geographic risks. One would justifiably consider to be exposures the loadings, which are contained in the matrix B and which are constrained by Assumptions 1 to 3. However, in the tradition of factor models, it is customary to regress freely the return of a security on a number of factors, thus obtaining its "exposures", in a meaning of the word that is purely based on returns and not on revenues or income.[33] To compare and contrast these two definitons of exposure, we now take the zone indexes – including the $\mathbb{EM}$ indexes – as being given and regress firms' stock returns on them. To maintain the differentiation with the loadings, we call "statistical exposures" the slope coefficients so obtained.

indexes evolves not just as to the list of firms included in the index but also as to the locus of the business conducted by the firms that are included.

[33] See Adler and Dumas (1984).

We regress separately each security's daily return on the daily returns of $\mathbb{EM}$, $ISIN$ and 70% indexes. Even though we know that firms could very well be negatively exposed to a risk factor, we keep the non negativity constraint in place for all indexes, 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' revenues and not about profits or free cash flows. Without the non negativity constraints, the exposures would not be comparable. But, non negativity is now the only set of constraints. Other constraints stated in Assumption 1, relating to sales, and Assumption 3 (summation to 1) are now relaxed. The only information on sales that still plays a role in statistical exposures is encapsulated in the $\mathbb{EM}$ indexes, the composition of which has been designed to reflect sales information (see Section 6), but which contain stock returns only. The calculation of statistical exposures is done yearly by constrained maximum likelihood, the likelihood function being similar to (3) above (but with regressors explicitly given, as opposed to factors to be estimated), where the coefficients of the regression are constrained to be non negative.[34]

We first want to determine whether $\mathbb{EM}$ zone indexes are successful in capturing sales information. To what degree is it true that the sensitivities of stock returns of a firm reflect the risks of the zones to which it sells its products? To answer that question, we compare to actual sales the statistical exposures to twelve $\mathbb{EM}$ zone indexes (ignoring the "rest-of-the-world" zone). This is a generalization of the third-stage comparison performed by Diermeier and Solnik (2001). We do that by regressing the statistical exposures on the composition of sales, using all data points of all firms and all years (75264 observations).[35] The resulting slope coefficient is striking:[36]

$$Exposure_{\mathbb{EM},i,j,t} = Intercept_t + \underset{(145.78)}{0.93774} \times Sales_{i,j,t} + \varepsilon_{i,j,t}$$

where i is a subscript for firms, j is one for zones and t is one for years. The t-statistic of the slope in a test against the value zero is indicated in parentheses and the R^2 is equal to 0.238. We did not expect exposures to reflect sales so starkly as to be equal to them. Indeed, fractions of sales have to be less than 1 while exposures may be above, since constraints have been removed from their estimation, – many definitely are. Furthermore, on the left-hand side of the equation are the exposures to twelve factors only, which are the only messengers of sales information, whereas the right-hand side draws on thousands upon thousands of firms, each with its own sales information. Also, as was explained in Section 3, the construction of the $\mathbb{EM}$ index has used regional information, which could not be included on the right-hand side of the present

[34] We caution that the constraints cause the estimated residuals not to be orthogonal to the regressors. But we do not decompose variance.

[35] Here we use the observation points for which we have an explicit observation of sales to one of the twelve zones.

[36] A regression of the exposures $Exposure_{70\%,i,j,t}$ to the domestic indexes would yield a slope equal to 0.8082, which means that the 70% exposures are lower than what is indicated by sales. But the R^2 would be slightly higher than that of the exposures to the $\mathbb{EM}$ index: 0.29.

	Sum of Foreign Exposures			**Sum of Foreign Exposures /Sum of all Exposures**		
AVERAGE	$\mathbb{EM}$	$ISIN$	70%	$\mathbb{EM}$	$ISIN$	70%
France	1.621	0.459	0.846	0.670	0.385	0.620
Germany	1.774	0.587	1.010	0.697	0.465	0.684
Great B	1.163	0.469	0.651	0.554	0.380	0.460
France*	1.015	0.309	0.489	0.420	0.259	0.358
Germany*	1.167	0.390	0.591	0.459	0.309	0.401
Great B*	0.943	0.362	0.500	0.449	0.293	0.354
Brazil	0.485	0.255	0.289	0.192	0.220	0.244
U.S.	0.869	0.334	0.356	0.441	0.269	0.258
Canada	1.099	0.460	0.571	0.430	0.374	0.421
Australia	0.784	0.390	0.459	0.405	0.310	0.347
Malaysia	0.700	0.303	0.332	0.356	0.252	0.273
Singapore	0.966	0.395	0.570	0.466	0.330	0.435
China	0.657	0.476	0.291	0.298	0.381	0.252
Japan	0.684	0.262	0.319	0.317	0.217	0.244
India	0.652	0.321	0.364	0.314	0.254	0.273
*Sales to countries outside Western Europe						

Table 7: **Exposures to foreign zones in regressions on $\mathbb{EM}$, $ISIN$ and 70% indexes**: Regressions of stock returns of national firms of one country on all $\mathbb{EM}$, $ISIN$ or 70% indexes. In the first three columns, the table shows the sum of the foreign slope coefficients of a firm, averaged across firms and years. The last three columns show the sum of the foreign slope coefficients, averaged across firms and years, divided by the sum of all slope coefficients, averaged across firms and years. The second set of regressions for the European countries (marked with an asterisk) shows coefficients for foreign countries outside Europe.

regression. All this notwithstanding, the slope in that regression is found to be strikingly close to 1.[37]

As our second investigation in this section, we resolve an empirical anomaly of international stock returns. In most empirical estimations, 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. That strikes one as odd in a globalized world. 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 $ISIN$ indexes or domestic 70% indexes.

The question we focus on is whether, in the $\mathbb{EM}$ regressions, the exposures to the foreign factors play a greater role than they do in the $ISIN$ and 70%

[37] The t-statistic in a test of equality of the slope coefficient to the number 1 is equal to -9.73, thus formally rejecting equality. However, one should keep in mind the very large number of observations.

regressions. Table 7 compares the sum of the foreign coefficients for the $\mathbb{EM}$, $ISIN$ and 70% regressions, averaging over all national firms of a country. The table shows the coefficients averaged over the sixteen years of our sample, 1999-2014.[38] The results are clear-cut. The $\mathbb{EM}$ exposures give a much larger role to foreign indexes than do those using the traditional $ISIN$ index or the 70% index. For example, in the U.S. regressions, the average foreign coefficients increase from 0.334 in the $ISIN$ regressions to 0.356 in the 70% regressions to 0.869 in the $\mathbb{EM}$ regressions. Table 7 also shows how much larger a role the foreign factors play relative to the domestic factors in the $\mathbb{EM}$ regressions compared with the two more conventional regressions. The last three columns of the table report the ratios of the average foreign coefficients to the average foreign and domestic coefficients combined. The ratios are larger for the $\mathbb{EM}$ regressions except in the cases of Brazil and China. So, the $\mathbb{EM}$ indexes succeed in highlighting the importance of foreign influences on stock returns. That is to be expected since the $\mathbb{EM}$ indexes reflect the importance of foreign revenues of many of these firms.

9 Conclusion

We have developed, and demonstrated the relevance of, a new technique to identify some of the operational risks that a firm faces. The technique can be useful to corporate managers and asset managers. It is based on the assumption that the free cash flows of a firm and its stock returns reflect the risks of the places to which it sells its products, irrespective of its domicile.

Using the expectations-maximization ($\mathbb{EM}$) method of likelihood maximization, we have generated implicit or latent stock index factor returns based on the geographic zones from which firms receive their free cash flows (proxied by revenues), as opposed to traditional indexes based on firms' headquarters. In an increasingly integrated world economy, commercially speaking, it makes sense to take into account where firms do business rather than just where firms locate their headquarters. As a way better to isolate the risks attached to one zone, the portfolio composition of the $\mathbb{EM}$ index of a zone utilizes information from all firms, be they domestic or multinational, and goes as far as to place some moderately negative weights on some firms that do not sell to that zone.

The resulting indexes differ in their behavior from traditional indexes in accordance with their design and purpose. First, the correlations between the $\mathbb{EM}$ indexes and traditional indexes are lower for a country with many foreign firms selling to it (foreign presence). That is because the revenues of the foreign firms influence that zone's $\mathbb{EM}$ index. Secondly, these correlations are also lower for countries where there are many national firms selling multinationally to foreign countries (foreign activity). That is because the $\mathbb{EM}$ index is able to

[38] The three $\mathbb{EM}$ 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. To adjust for this phenomenon, Table 7 also reports for the European regressions the exposures to zones outside Western Europe.

separate the role of revenues coming from one zone. Thirdly, we show that the $\mathbb{EM}$ indexes are less correlated across zones than traditional indexes. The $\mathbb{EM}$ algorithm has been able to remove an undue amount of correlation caused by revenues from common foreign countries.

Fourthly, we examine the statistical exposures of firms by regressing freely their stock returns on the $\mathbb{EM}$ zone indexes, and verify that they faithfully reflect the firms' foreign revenues. Finally, the firms' statistical exposure coefficients falling on foreign zones are higher in regressions of stock returns on the $\mathbb{EM}$ indexes than in regressions on traditional indexes. That is because the foreign revenues of domestic firms help to determine the foreign $\mathbb{EM}$ indexes.

The $\mathbb{EM}$ factors, therefore, represent a new type of stock index that better reflects the business risks of one geographic zone vs. another.

This work opens the way to more complete factor models that should be investigated. The first priority would be to add to geographic factors local-pricing factors that would capture the eventuality that securities listed on the same stock market correlate excessively.[39] 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 commercially.[40]

[39] See Chaieb et al. (2020). When attempting to exhibit local pricing, it is important to control for the geographic distribution of sources of revenues.

[40] One more topic 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. 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

Adler, M. and B. Dumas, 1984, "Exposure to Currency Risk: Definition and Measurement," *Financial Management*, 13, 41-50.

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: Accessing 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.

Barrot, J.-N., Loualiche, E., Sauvagnat, J., 2019, "The globalization risk premium," *The Journal of Finance*, 74, 2391-2439.

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, 2019. "On the global financial market integration 'swoosh' and the trilemma," *Journal of International Money and Finance*, 94, 227-245.

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.

Bodnar, G. M., B. Dumas and Marston, R. C., 2002, "Pass-through and Exposure," *The Journal of Finance*, 57, 199-232.

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.

Chaieb, I., H. Langlois, and O. Scaillet, 2020, "Factors and Risk Premia in Individual International Stock Returns," working paper n°18-04, Swiss Finance Institute.

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. and A. Karolyi, 1998, "Another Look at the Role of Industrial Structure of Markets for International Diversification Strategies," *Journal of Financial Economics*, 50, 351-373.

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.

Hoberg, G., Moon, S. K., 2019, "The offshoring return premium," Management Science, 65, 2445-2945.

Karolyi, G. A., Stulz, R., 1996, "Why do Markets Move Together? An Investigation of U.S.-Japan Stock Return Comovements," The Journal of Finance, 51, 951-986.

King, M., Sentana, E., Wadhwani, S., 1994, "Volatility and Links between National Stock Markets," Econometrica, 62, 901-933.

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.

Restoy, F. and P. Weil, 2011, "Approximate Equilibrium Asset Prices," *Review of Finance*, 15, 1–28.

Roll, R., 1977, "A critique of the asset pricing theory's tests Part I: On past and potential testability of the theory," *Journal of Financial Economics*, 4, 129–176.

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.

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 revenues 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 are numerous 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 basis of permanent (static) properties obtained from Datastream.
   For ISINs in the WVB database, 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;[41] 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;[42] that leaves 25,441 securities. Then, based on the Worldscope type, Chinese A shares are deleted;[43] that leaves 24,105 securities. Fourth the download contains the name of the company.

[41] ','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'

[42] 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'

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

A first general 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.[44] For the filtered WVB 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 8. 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. For each stock we get the total return index in both the US 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;[45] 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 string of returns for this day if the number of zero returns exceeded a third

[44] 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 10.

[45] 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.

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, we merge into one dataset per year the data on revenues and the data on rates of return. Selection is then done on the basis of revenue data.
   In the process, some securities are lost for some years because WVB, 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 9. Some more filtering is performed. based on revenue 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 a 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 10. 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 choose the one that is preferred according to the rule set up in Table 10 (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 9. 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 listings. 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 8: **Firm count after each stage of the filtering based on properties that vary year by year (Step 2).**

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

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 10: **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).

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". 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, $\int p\left(C|R;\tilde{\psi}\right) dC = 1$):

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

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

$$\begin{aligned} \frac{\partial}{\partial\psi} L\left(R;\tilde{\psi},\phi\right) &= \int \left(\frac{\partial}{\partial\psi} \ln p\left(C,R;\tilde{\psi}\right)\right) p\left(C|R;\tilde{\psi}\right) dC \\ &\quad - \frac{\partial}{\partial\psi} \int p\left(C|R;\tilde{\psi}\right) dC - \phi \cdot \frac{\partial}{\partial\psi} g\left(\tilde{\psi}\right) \\ \frac{\partial}{\partial\psi} L\left(R;\tilde{\psi},\phi\right) &= \int \left(\frac{\partial}{\partial\psi} \ln p\left(C,R;\tilde{\psi}\right)\right) p\left(C|R;\tilde{\psi}\right) dC - 0 \\ &\quad - \phi \cdot \frac{\partial}{\partial\psi} g\left(\tilde{\psi}\right) \end{aligned} \tag{9}$$

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

$$\int \left(\frac{\partial}{\partial\psi} \ln p\left(C,R;\hat{\psi}\right)\right) p\left(C|R;\tilde{\psi}\right) dC - \hat{\phi} \cdot \frac{\partial}{\partial\psi} g\left(\hat{\psi}\right) = 0 \tag{10}$$

$$\hat{\phi} \cdot g\left(\hat{\psi}\right) = 0; \hat{\phi} \geq 0; g\left(\hat{\psi}\right) \geq 0$$

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

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

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 (9)

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

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 (10):

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

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:

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

The differential of that term of the log likelihood are:[46]

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

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

Therefore (transposing, because of row vs. column notation),[47]

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

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 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.

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

D Sufficient statistics of the geographic analysis

Based on the model:

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

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

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

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 $\sum_{t=1}^{T} \mathbb{E}\left[C_t C_t^\intercal | R_t\right]$:

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

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

so that:

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

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*

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

which is a time-series Least Squares (except that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[C_t C_t^{\intercal}|R_t\right]$ 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

```
% second, 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
```