Received: 19 July 2022 | Revised: 29 August 2022 | Accepted: 1 September 2022

DOI: 10.1111/jors.12621

RESEARCH ARTICLE

JOURNAL OF REGIONAL SCIENCE WILEY

An information-theoretic approach to the analysis of location and colocation patterns

Alje van Dam[1,2] | Andres Gomez-Lievano[3] | Frank Neffke[3,4] | Koen Frenken[1]

[1]Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands

[2]Centre for Complex Systems Studies, Utrecht University, Utrecht, The Netherlands

[3]Growth Lab, Center for International Development, Harvard University, Cambridge, Massachusetts, USA

[4]Complexity Science Hub Vienna, Vienna, Austria

Correspondence

Koen Frenken, Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands.
Email: k.frenken@uu.nl

Funding information

A.D and K.F are funded by the Netherlands Organisation for Scientific Research (NWO) under the Vici scheme, number 453-14-014. F.N. received financial support from the Austrian Research Agency (FFG), project #873927 (ESSENCSE)

Abstract

The study of location and colocation of economic activities lies at the heart of economic geography and related disciplines, but the indices used to quantify these patterns are often defined ad hoc and lack a clear statistical foundation. We propose a statistical framework to quantify location and colocation associations of economic activities using information-theoretic measures. We relate the resulting measures to existing measures of revealed comparative advantage, localization, specialization, and coagglomeration and show how different measures derive from the same general framework. To support the use of these measures in hypothesis testing and statistical inference, we develop a Bayesian estimation approach to provide measures of uncertainty and statistical significance of the estimated quantities. We illustrate this framework in an application to an analysis of location and colocation patterns of occupations in US cities.

KEYWORDS

coagglomeration, location quotient, pointwise mutual information, relatedness, revealed comparative advantage

1 | INTRODUCTION

The recognition of differential specialization patterns lies at the heart of economics since the works of Adam Smith and David Ricardo. Economists studying task assignments (Roy, 1951; Sattinger, 1993), urban economies (Ellison & Glaeser, 1997; Ellison et al., 2010), or international trade (Balassa, 1965; Krugman, 1991), all stress the fact that different economic entities specialize in different activities. Scholars in each of these fields have relied on indices that quantify, for example, the revealed comparative advantage of exports, the specialization of regions, and the extent of localization and (co-)agglomeration of industries. However, currently a plethora of such measures exist, without a principled way of deriving them nor of determining the amount of uncertainty in their measurement, and they are often validated only by their predictive or explanatory power. Here we take a different approach, motivated by the idea that the "best" measure is the one that reflects most closely what it intends to measure, given some underlying model of the data. In this paper, we propose a statistical framework from which measures of location and colocation can be derived. Although the methodology generalizes immediately to other contexts, to fix ideas, we focus on economic geography and derive measures of location (the prevalence of an activity in a location), specialization (the concentration of a location on few activities), localization (the concentration of an activity in few locations) and colocation (the degree to which different activities are found in the same locations) from a single statistical framework, revealing the internal connections between these concepts.

To do so, we treat location as the realization of an event that can be described by random variables like a location and an economic activity. For example, we can consider the city and occupation of a randomly sampled worker. We use the Pointwise Mutual Information (PMI) to express the association between a location and the economic activity of a sample in terms of the *information* that the economic activity of a sample (e.g., a worker's occupation) gives about its location (e.g., the city where that person works). Next, we show how the PMI can be used, in turn, to quantify the association between *pairs* of economic activity, in terms of how much information observing a particular activity gives about observing another activity in the same location. That is, if we randomly sample a pair of workers from the same city, how much information does the occupation of one of them provide about the likely occupation of the other?

Metrics based on Information Theory such as the PMI have found various applications in economics (Theil, 1967), and are uniquely derived from axioms about how information can be gained from probability distributions (Cover & Thomas, 2005; Shannon, 1948). One of their key properties is that, by taking expectations, they can be aggregated and decomposed to form well-defined measures that have an interpretation in terms of information. This allows the use of the PMI as a building block of information-theoretic measures that describe properties at the location, activity, or even economy level.

The resulting measures can be related to well-known existing indices of localization and specialization. In particular, at the level of location-activity pairs—as exemplified in country-product or city-industry data—our metric of association, the PMI, is a transformation of the widely used index of revealed comparative advantage (RCA) (Balassa, 1965), also known as the Location Quotient in regional science (Isard, 1960).

Despite its widespread application, there has been ongoing debate about the theoretical foundations and empirical properties of the RCA index, some of which complicate empirical analysis based on these indices (Ballance et al., 1987; Kunimoto, 1977; Vollrath, 1991; Yeats, 1985). One of the main issues is the distributional skew of the index, which has led to the proposal of several alternative indices that aim to overcome this issue (Hoen & Oosterhaven, 2006; Laursen, 2015; Yu et al., 2009). We provide an information-theoretic motivation for considering the logarithm of the RCA index and a Bayesian approach to estimating this quantity. This not only resolves the issue of distributional skew, but also ensures that the measure always attains finite values, and suggests a natural measure of uncertainty for the estimates.

Building on the PMI, we derive measures on more aggregate levels of analysis, leading to measures of the localization of economic activities that capture the degree to which economic activities are geographically constrained. We do so by calculating an activity's expected PMI (i.e., the expected association of the activity with a given location) over all locations. This quantity is called Kullback–Leibler divergence, and has been proposed as a measure of localization by Mori et al. (2005). Likewise, we can calculate the expected location-activity PMI of a particular location across all economic activities. This average association of a location with given activities provides a measure of specialization that is conceptually similar to Krugman's specialization index (Krugman, 1991).

The PMI can be applied to any probability distribution representing joint occurrences. Thus, we show how it can be applied to the distribution of colocated pairs of economic activity, providing a measure of spatial association between economic activities. Colocation patterns of economic activities have received increasing attention in urban economics (Ellison et al., 2010) and studies on the diversification patterns of national economies (Hidalgo et al., 2007). In the former field, authors have used colocation patterns as a dependent variable to test theories related to Marshallian externalities (Marshall, 1920). In this literature, the coagglomeration index of Ellison et al. (2010) has become a de facto standard (Diodato et al., 2018; Faggio et al., 2017). The latter field consists of a growing literature that uses colocation (or, more generally, co-occurrence) patterns to construct measures of technological relatedness or similarity, which are used as independent variables to study the diversification patterns of economies (Hidalgo et al., 2018, 2007; Neffke et al., 2011). The “proximity” measure introduced in the seminal paper by Hidalgo et al. (2007) has been applied to study a wide variety of topics, including regional economic development (Boschma et al., 2013), technological innovation (Boschma et al., 2015), the labor market (Alabdulkareem et al., 2018), the green economy (Mealy & Teytelboym, 2020) and global science (Miao et al., 2022). Here, we show how information theory can be used to derive an alternative measure for colocation from first principles, clarifying its underlying assumptions and statistical properties. We will discuss relations between the PMI, the coagglomeration index, and the proximity measure, and compare these measures empirically to one another.

As in the case of location–activity pairs, marginalizing the PMIs of colocated activity–activity pairs yields meaningful measures on the level of activities. Accordingly, the expected spatial association of an activity with all other activities gives a measure of the spatial “codependence” of an activity. This measure reveals how “picky” activities are in their tendencies to colocate with other activities. This spatial codependence is low for activities that locate independently of other activities, whereas codependence is high for activities that are preferentially found in the presence of certain other activities.

As an empirical illustration, we apply the proposed measures to US city-occupation employment data, showing the associations between cities and occupations, and between pairs of occupations. The data reveals clear (co-) location patterns. For instance, occupations related to manufacturing strongly concentrate in smaller cities and are negatively associated to most other occupations, whereas knowledge-intensive services mostly occur in the largest cities and are associated with each other. We furthermore compare our colocation association with the coagglomeration index of Ellison et al. (2010) and the proximity measure of Hidalgo et al. (2007), which are both widely used in their respective fields. We find substantial differences between each of these measures. Although the coagglomeration index gives a similar pattern to the PMI for occupation data on an aggregate level, the measures differ when considering more fine-grained data, leading to a different ranking for pairs occupations with the highest colocation. Our measure also significantly differs from the proximity measure. For instance, the latter tends to assign high proximities to pairs of occupations that have neutral associations according to the PMI. We show where these differences originate and discuss their implications.

The probabilistic basis that underlies the PMI ensures that the framework is explicit about the null models, priors, and data-generating processes we assume. This puts the measurement of location and colocation on a rigorous statistical footing. Furthermore, we show how the PMI can be estimated in practice. To do so, we use a Bayesian framework that assumes that the data on the presence of units of economic activities across locations are drawn from a multinomial distribution. This Bayesian estimation framework resolves some well-known measurement issues and provides a measure of uncertainty for the estimated quantities, allowing a test of significance for the resulting associations. Building on Wolpert and Wolf (1995) and Hutter and Zaffalon (2005), we provide analytical approximations of the posterior mean and variance of all the measures we propose. We also make available a Python class that enable users to easily compute all proposed measures (https://github.com/aljevandam/Colocation).

2 | INFORMATION-THEORETIC MEASURES OF LOCATION

2.1 | Notation

Consider data on the location of economic activities in the form of an $N_c \times N_i$ dimensional matrix or contingency table **Q**, where N_c and N_i are the number of locations and economic activities in the classifications of the data, respectively. The matrix **Q** contains the counts of each cell in the matrix, its entries q_{ci} denoting the number of occurrences of activity type i in location c. For instance, q_{ci} can be the number of workers employed in a particular occupation i in a city c, the number of establishments of industry i in region c, or the number of dollars of product i exported by country c. The total amount of activity of type i and the total activity in location c are given by the column sums $q_i = \sum_c q_{ci}$ and row sums $q_c = \sum_i q_{ci}$, respectively. Total economic activity is given by $q = \sum_{c,i} q_{ci}$.

We will consider the matrix **Q** to be the outcome of an independent sampling process from the underlying distribution **p** with probabilities

$$p_{ci} = P(C(w) = c, A(w) = i) \tag{1}$$

that a sample w (i.e., a worker, an establishment, a dollar) is part of activity i in location c. The matrix **Q** is considered to be the outcome of sampling the categorical random variable wq times, where each sample has location $C(w)$ and activity type $A(w)$. The marginal probabilities are given by $p_i = \sum_c p_{ci} = P(A(w) = i)$ and $p_c = \sum_i p_{ci} = P(C(w) = c)$. The location-activity probabilities p_{ci} will be the main object of interest as they hold information on the associations between locations and activities (Section 2.2).

A similar approach can be taken to study the colocation of economic activities. To this end, we examine the probabilities p_{ij} that a randomly sampled *pair* of economic activities from the same location has types i and j (Section 3.1).

2.2 | Location association

The dependencies hidden in the joint probabilities p_{ci} can now be used to measure the association between an activity and a location. Information theory provides a framework to quantify these associations explicitly in units of information (e.g., bits). The association between $C(w) = c$ and $A(w) = i$ is given by their pointwise mutual information $PMI(p_{ci})$ (Fano, 1961). The PMI quantifies the association between two outcomes by assessing the information content of the realization $(C(w) = c, A(w) = i)$ relative to the information content when the realization comes from a null model in which $C(w)$ and $A(w)$ are independent random variables, that is, $p_{ci} = p_c\, p_i$. Intuitively, PMI provides an answer to the question *how much information does observing c provide about the presence of i?* PMI has been used in several fields, including economics (Theil, 1967), administrative sciences (Theil, 1972), and linguistics (Church & Hanks, 1989).

In information theory, the information content or "surprise" of an outcome i is defined as $\log\left(\frac{1}{p_i}\right)$. Observing an event that occurs with small probability leads to a high information content or surprise, whereas highly likely events contain little information. The difference between the information contents of p_{ci} and $p_c\, p_i$ gives a measure of the surprise of observing p_{ci} while expecting $p_c\, p_i$. Depending on the base of the logarithm, PMI measures association in units of *bits* (base 2) or *nats* (natural logarithm). This *association* between outcomes c and i is given by the logarithm of ratio of the joint probabilities and the null model:

$$PMI(p_{ci}) = \log\left(\frac{p_{ci}}{p_c\, p_i}\right). \tag{2}$$

$PMI(p_{ci})$ will be positive when it is more likely to observe c and i together than expected under independence, that is, $p_{ci} > p_c\, p_i$, whereas $PMI(p_{ci})$ takes negative values when c and i are less likely to occur together than expected under the null model of independence, that is, $p_{ci} < p_c\, p_i$. $PMI(p_{ci}) = 0$ if and only if $p_{ci} = p_c\, p_i$, indicating that c and i

are independent (i.e., the incidence of an activity is independent of the place). Note that the ratio $\frac{p_{ci}}{p_c p_i}$ in the PMI can be interpreted in two alternative ways using conditional probabilities: as $\frac{p_{i|c}}{p_i}$ or $\frac{p_{c|i}}{p_c}$. That is, the probability of observing i conditional on knowing c relative to the same probability without knowing c, or vice versa, the probability of observing c conditional on knowing i relative to the same probability not knowing i. The largest values these ratios can achieve is when $p_{i|c} = 1$ (in which case $p_{ci} = p_c$) or $p_{c|i} = 1$ (in which case $p_{ci} = p_i$). Thus, the maximum value of $PMI(p_{ci})$ is given by $\max\{\log\left(\frac{1}{p_i}\right), \log\left(\frac{1}{p_c}\right)\} = \log\left(\frac{1}{p_{ci}}\right)$, which is attained either when activity i always occurs in location c, or when activity i is the only activity in location c. $PMI(p_{ci})$ is not bounded from below, as it tends to $-\infty$ as the joint probability p_{ci} tends to 0.

2.3 | Localization and specialization

Many questions are better answered at more aggregate levels of analysis than the level of location–activity pairs. For instance, one may want to know which activities are most localized in space, or which locations are most specialized in terms of their economic activities.

Measuring location associations using PMI naturally leads to measures that describe such associations at higher levels of aggregation, which can be interpreted as measures of *localization* (one quantity for each activity) and *specialization* (for each location). In the following, we show how such measures follow from the information-theoretic framework. Table 1 summarizes each of the measures based on the location probabilities p_{ci} and the relations among them.

The *localization* of an activity can be defined as the degree of dissimilarity between the activity's own geographical distribution and the distribution of the population or of total economic activity across all locations (Hoover, 1936; Mori et al., 2005). Highly localized activities will be distributed across locations in a very different way than what one would expect from locations' sizes. Activities with a low degree of localization will be distributed in proportion to the population of locations.

This can be quantified by comparing how much, on average, the probability that a unit of activity of type i is located in a location differs from the probability that *any* unit of activity is located there. Consider, for example, all the associations of a particular activity with every location. Let $p_{c|i} = p_{ci}/p_i$ be the probability that a sample of activity i is located in location c, and recall that the probability that a random sample (i.e., regardless of its economic activity) is located in c is given by p_c. The expected location association of a sample with activity i is given by

$$\begin{aligned} KL(p_{c|i}|p_c) &= \sum_c p_{c|i} PMI(p_{ci}) \\ &= \sum_c p_{c|i} \log(p_{c|i}/p_c). \end{aligned}$$

Here, KL denotes the Kullback–Leibler divergence (Kullback & Leibler, 1951), which measures the deviation between the distribution across all locations of a specific activity, given by probabilities $p_{c|i}$, and the overall distribution of locations, given by the probabilities p_c. The resulting metric gives the expected association of a

TABLE 1 Overview of measures following from location probabilities p_{ci}

Unit of analysis	Measure	Formula
Location–activity	Association	$PMI(p_{ci})$
Activity	Localization	$KL(p_{c\|i}\|p_c) = \mathbb{E}_{p_{c\|i}}[PMI(p_{ci})]$
Location	Specialization	$KL(p_{i\|c}\|p_i) = \mathbb{E}_{p_{i\|c}}[PMI(p_{ci})]$
System	Overall specialization	$MI(C(w), A(w)) = \mathbb{E}_{p_{ci}}[PMI(p_{ci})]$

location with a particular activity type across all locations. This expected association can be interpreted as a measure of localization.

A localization that is close to zero indicates that knowing that sample has activity type i does, on average, not provide much information on where it is located. In other words, the probability distribution for the location of samples with activity i, $p_{c|i}$, is not very different from the overall distribution of locations p_c. In contrast, a high localization for an activity i shows that observing activity type i provides a lot of information on where we are most likely to find it, implying that the distribution of probabilities $p_{c|i}$ is very different from that of the probabilities p_c.

A measure of *specialization* can be obtained in the same way, by considering the expected PMI across all locations. More precisely, the expected association of a sample with a particular location can be quantified as the dissimilarity between the distribution of activities given a location c, $p_{i|c}$, and the overall distribution of activity types p_i. Aggregating the $PMI(p_{ci})$ to the location level thus leads to a measure of specialization given by

$$KL(p_{i|c}|p_i) = \sum_i p_{i|c} PMI(p_{ci}).$$

It is possible to aggregate even further: taking the expectation over both locations and activities yields the *expected* association across location–activity pairs, or equivalently as either the expected localization of activities or the expected specialization of locations. This can be interpreted as a measure for the overall specialization at the system level. The resulting quantity is known as the mutual information (MI) (Cover & Thomas, 2005) and quantifies the dependence between two random variables. In this case, it measures the dependence between the random variables $C(w)$ and $A(w)$, which describe the type and location of a randomly sampled unit of activity. It is given by

$$MI(C(w), A(w)) = \sum_{c,i} p_{ci} PMI(p_{ci}) \tag{3}$$

$$= \sum_i p_i KL(p_{c|i}|p_c) \tag{4}$$

$$= \sum_c p_c KL(p_{i|c}|p_i). \tag{5}$$

When $MI(C(w), A(w)) = 0$, the location of a randomly sampled unit is independent of its activity type, which implies that all economic activity is distributed proportionally to location size, or equivalently that every location has an identical distribution of activities. In this situation, there is no specialization in the system in the sense that all locations have identical activity mixes. The maximum value of $MI(C(w), A(w))$ is reached when each location has its own unique activity, so that each location is maximally specialized and each activity is maximally localized. The mutual information measure may be used to compare specialization across different systems (e.g., comparing the degree of overall specialization across countries), or to track the changes over time (e.g., comparing the degree of overall specialization before and after the establishment of a trade agreement).

2.4 | Relations to existing measures

The information-theoretic approach allows us to understand comparative advantage, localization, and specialization as part of the same framework. Furthermore, it resolves some known methodological problems of existing measures.

For instance, the location association is equivalent to the logarithmic transformation of the widely used RCA index (Balassa, 1965). This is easily seen by considering the definition of the RCA. The *RCA* of a location–activity pair is given by the ratio of the share of activity i within location c to the share of activity i in the overall economy:

$$RCA(c, i) = \frac{q_{ci}}{q_c} \Big/ \frac{q_i}{q}. \tag{6}$$

That is, the RCA compares the observed share of activity i in location c in the numerator to share of i in the economy as a whole in the denominator. Since q_i and q_c are interchangeable in (6), $RCA(c, i)$ can be interpreted in two ways: as a measure of “localization” of activity i in location c, or as a measure of “specialization“ of location c in activity i. The neutral value is given by $RCA(c, i) = 1$, where the share of activity i in location c is equal to the total share of activity i across all locations.

Consider now the maximum-likelihood estimate for the multinomial probabilities $\hat{p}_{ci} = \frac{q_{ci}}{q}$. We can now express PMI as:

$$\begin{aligned} PMI(p_{ci}) &= \log\left(\frac{\hat{p}_{ci}}{\hat{p}_c\,\hat{p}_i}\right) \\ &= \log\left(\frac{q_{ci}}{q_c}\Big/\frac{q_i}{q}\right) \\ &= \log(RCA(c, i)), \end{aligned}$$

showing that, conceptually, the PMI equals the logarithm of the RCA index. Our approach stands therefore as a generalization of the RCA index, showing that there is an information-theoretic notion of association underlying the RCA.

One of the downsides of the RCA index when computed from data is that it is heavily skewed and asymmetric around its neutral value. A logarithmic transformation of the index has been suggested as a possible solution. This makes the metric symmetric around a neutral value of 0 (Vollrath, 1991). However, this transformed index becomes undefined whenever $q_{ci} = 0$.

However, seen through the lens of probabilities, the practical problem of having to take the logarithm of zero when $q_{ci} = 0$ becomes a problem related to miss-estimating p_{ci}. in Section 4, we show how to overcome this problem using a Bayesian approach to estimate probabilities p_{ci} that are always strictly positive.

It is further noteworthy that the localization $KL(p_{c|i}|p_c)$ has the exact same functional form as the measure of industrial localization put forward by Mori et al. (2005), although the null model implicit in their metric is based on a location's area. That is, the authors take the probabilities p_c to be proportional to the area of that location as opposed to its population size (i.e., q_c).

Ignoring differences in how the distributions are estimated, the localization of an economic activity i can be written as $KL(p_{c|i}|p_c) = \mathbb{E}_{p_{c|i}}[\log(RCA(c, i))]$. This shows that localization can be understood as the expected value of the logarithm of the RCA of the activity over all locations in which it occurs (under the probability distribution $p_{c|i}$). This holds regardless of the “null model” considered. Hence, one could follow Mori et al. (2005) and use their area-based null model to define a measure on the location–activity level that is analogous to the RCA index.

In a similar way, specialization of a location c can be seen as the expected value of the logarithm of the RCA, but now over activities within the given location: $KL(p_{i|c}|p_i) = \mathbb{E}_{p_{i|c}}[\log(RCA(c, i))]$. This measure is akin to Krugman's specialization index (Krugman, 1991), which is given by $K(c) = \sum_i |p_{i|c} - p_i|$. Like $KL(p_{i|c}|p_i)$, Krugman's index also considers the “average deviation” of $p_{i|c}$ to p_i, but expresses this deviation in terms of absolute differences.

In our framework, the localization of activities and specialization of locations are essentially the same measures, defined for different units of analysis, and all quantified by aggregating bits of information. The functional forms we use are dictated by information theory and lead to easy-to-interpret measures that are consistently defined across different units of analysis.

3 | INFORMATION-THEORETIC MEASURES OF COLOCATION

3.1 | Colocation association

So far, we have studied the probabilities p_{ci}, which summarize location patterns of economic activity. Our framework can however readily be extended to study more complex patterns. Here, we show how it can be applied to the distribution of colocated pairs of economic activity, providing a measure of spatial association between economic activities.

To study colocations within our framework, we examine the probabilities p_{ij} that two random samples, w_1 and w_2, from the same location have types i and j, respectively. Consider the conditional probability of sampling activity j in the same location as a given sample of activity i, that is,

$$\begin{aligned} p_{j|i} &= P(A(w_2) = j | A(w_1) = i, C(w_2) = C(w_1)) \\ &= \sum_c P(A(w_2) = j, C(w_1) = c | A(w_1) = i, C(w_2) = c) \\ &= \sum_c P(A(w_2) = j | A(w_1) = i, C(w_2) = c) P(C(w_1) = c | A(w_1) = i, C(w_2) = c) \\ &= \sum_c P(A(w_2) = j | C(w_2) = c) P(C(w_1) = c | A(w_1) = i) \\ &= \sum_c p_{j|c} p_{c|i}. \end{aligned}$$

This means the probability of sampling activities i and j with both samples having the same location is given by

$$p_{ij} = p_{j|i} p_i = p_i \sum_c p_{j|c} p_{c|i} = \sum_c p_{j|c} p_{ci}.$$

Note that the marginal probabilities are given by $\sum_j p_{ij} = p_i$ and $\sum_i p_{ij} = p_j$. Furthermore, $p_{ij} \neq p_i p_j$ in general, meaning that the probabilities of finding occupations i and j are not independent, and this dependence arises from the requirement that both samples are drawn from the same location. It is these dependencies that we will quantify in the following to construct a measure of association between pairs of economic activities.

As with the location–activity associations, the association between activity types can be quantified by the PMI. The association between two activities is then defined as

$$PMI(p_{ij}) = \log\left(\frac{p_{ij}}{p_i p_j}\right), \tag{7}$$

where $p_i p_j$ is the null model that describes a situation where i and j are distributed independently of each other. What $PMI(p_{ij})$ captures is that the presence of some activities may increase or decrease the probability that other activities are present in the same location. Hence, observing a particular type of economic activity holds information about the likelihood of observing other types of activities in the same location. Economic activities that are more likely to occur together ("co-occur") than expected under independence will have a positive association, whereas activities that are less likely to co-occur than expected under independence will have a negative association. Another way of seeing this, is by noting that $PMI(p_{ij})$ is positive when $p_{j|i} > p_j$, that is, when we observe that type i increases the probability of observing type j when sampling units of activity from the same location. Likewise, negative associations indicate that conditional on observing i, the probability of sampling a unit of activity j in the same location decreases. The $PMI(p_{ij})$ is symmetric, since $p_{ij} = p_{ji}$. Computing this measure for all pairs of activity types thus leads to a symmetric, square matrix that has as entries the colocation association $PMI(p_{ij})$.

The diagonal entries of this matrix hold "self-associations" $PMI(p_{ii})$. Self-association is high when observing an activity of type i in a particular region increases the likelihood that a second randomly sampled unit in that location is also of type i. This is the case when the probability of observing i is above average in a few locations, and below average in others. The self-association can thus be interpreted as a measure of geographical concentration. Note that the self-association is always positive, that is, $PMI(p_{ii}) \geq 0$, since observing a unit of activity of type i can never lower the probability of finding another unit of activity of type i (we sample with replacement). The matrix of colocation associations thus provides a joint estimate of geographic concentration and colocation.

3.2 | Aggregate measures of colocation

The measures for colocation association are identical to those proposed for location associations, only applied to different probabilities. Hence, aggregate measures can be obtained in a similar way as for the location associations. A measure of the average association of an activity j with any other activity is obtained by taking the expectation over all other activities, leading to

$$KL(p_{j|i}|p_j) = \sum_j p_{j|i} PMI(p_{ij}). \tag{8}$$

We call this measure the codependence of a particular activity. It quantifies the deviation of the distribution of activity types conditional on having observed activity type i, $p_{j|i}$, with respect to the unconditional distribution of probabilities p_j. When activity type i has, on average, strong colocation associations with other activity types, this deviation will be large. In other words, activity i “cares” about the type of activity it colocates with. A low value of $KL(p_{j|i}|p_j)$ on the other hand implies that the distribution of probabilities $p_{j|i}$ does not differ much from the distribution of p_j, meaning that activity i is uninformative for the type of activities it colocates with. This implies that activity i colocates with the “average” distribution of activity types, suggesting it is indifferent to the other activities in a location.

Taking the expectation of the codependence over all activity types, or equivalently taking the expectation of the colocation association over all activity pairs leads to the mutual information

$$\begin{aligned} MI(A(w_1), A(w_2)) &= \sum_i p_i KL(p_{j|i}|p_j) \\ &= \sum_{ij} p_{ij} PMI(p_{ij}). \end{aligned}$$

This is a measure of dependence between $A(w_1)$ and $A(w_2)$, which each describe the activity types of the member of a randomly sampled pair from the same location. The overall codependence is thus a system-level variable that describes how the types of a randomly sampled pair are on average (spatially) associated. Such a measure may, for example, help understand how the overall strength of coagglomeration externalities differs across economies or changes over time. Table 2 gives a summary of the measures that follow from analysis of the colocation distribution p_{ij}.

3.3 | Coagglomeration

Colocation patterns have received increasing attention as a dependent variable to test theories on Marshallian externalities (Diodato et al., 2018; Ellison et al., 2010; Faggio et al., 2017). Here we briefly describe how Ellison et al. (2010) derive their coagglomeration index. These authors present a location choice model for profit-maximizing plants (Ellison & Glaeser, 1997; Ellison et al., 2010) in which the (combined) effects of natural advantage and spillovers between activity types determine coagglomeration patterns. The authors propose the following pairwise coagglomeration index:

TABLE 2 Overview of measures following from colocation probabilities p_{ij}

Unit of analysis	Measure	Formula
Activity–activity	Colocation association	$PMI(p_{ij})$
Activity–activity	Geographic concentration	$PMI(p_{ii})$
Activity	Codependence	$KL(p_{j\|i}\|p_j) = \mathbb{E}_{p_{j\|i}}[PMI(p_{ij})]$
System	Overall codependence	$MI(A(w_1), A(w_2)) = \mathbb{E}_{p_{ij}}[PMI(p_{ij})]$

$$\gamma_{ij} = \frac{\sum_c (p_{c|i} - p_c)(p_{c|j} - p_c)}{1 - \sum_c p_c^2}. \tag{9}$$

Note that in our notation, activity shares $\frac{q_{ci}}{q_i}$ and $\frac{q_c}{q}$ are replaced by probabilities $p_{c|i}$ and p_c. This makes specific that we regard the former shares as maximum likelihood estimates of the latter probabilities.

The coagglomeration of all activity pairs can be collected in a matrix with entries γ_{ij}, completely analogous to the $PMI(p_{ij})$ in Section 3.1. The diagonal entries γ_{ii} contain the agglomeration index of a single activity (Ellison & Glaeser, 1997), when neglecting effects of the plant size distribution. Mori et al. (2005) show that the agglomeration index of Ellison and Glaeser (1997) can be written as $\gamma_i = a_i G_i - b_i \approx \frac{\sum_c (p_{c|i} - p_c)^2}{1 - \sum_c p_c^2}$. This approximation is valid when plants are reasonably uniformly distributed, in which case the plant size effect is negligible. The plant size distribution determines the size of the chunks in which the activity counts are generated by the data generating process. Quantifying the dependencies that arise from such a data-generating process is an interesting direction for future research, but for now, we focus on the simpler case in which information on the chunk sizes (e.g., the plant size distribution) is unavailable.

To facilitate the comparison of the coagglomeration index in Equation (9) to our own colocation association metric, we rewrite the latter as

$$\begin{aligned} PMI(p_{ij}) &= \log\left(\frac{\sum_c p_{i|c}\, p_{j|c}\, p_c}{p_i\, p_j}\right) \\ &= \log\left(\sum_c \left(\frac{p_{c|i}}{p_c}\right)\left(\frac{p_{c|j}}{p_c}\right) p_c\right). \end{aligned}$$

This expression shows that both indices capture how different activities covary in space. In either case, the intensity of spatial colocation may be generated by a location choice model akin to the one by Ellison and Glaeser (1997). The difference lies, however, in the functional form used to measure the deviation from the reference distribution. The colocation association compares probabilities by taking ratios $p_{i|c}/p_c$, whereas the coagglomeration index considers differences $p_{i|c} - p_c$. Furthermore, the colocation association weights each of the differences by p_c.

Although the coagglomeration index is derived from an economic model, the measure of concentration that lies at its heart enters the derivation as an assumption. Our framework provides a principled way to quantify these deviations, by leveraging information theory. The advantage of such an approach is that it gives insight into the underlying assumptions on the data generating process, the used reference distribution, and the estimation procedure with its corresponding uncertainties. For instance, the literature is not entirely consistent in the choice of the reference distribution that is used in (co-)agglomeration indices. In some work, the reference distribution is taken to be the share of total employment in location c, which we denote by p_c (Ellison & Glaeser, 1997, 1999; Faggio et al., 2017). In other work, the reference distribution is given by the average share of employment in industry i in a location, given by $\hat{p}_{c|i} = \frac{1}{N_i}\sum_i p_{c|i}$ (Diodato et al., 2018; Ellison et al., 2010).

3.4 | Proximity

Other studies make use of colocation patterns as an independent variable (often referred to as “proximity” or “relatedness”) to study the diversification patterns of economies in terms of trade, production, technology, and jobs (Boschma et al., 2013, 2015; Hidalgo et al., 2007; Muneepeerakul et al., 2013; Neffke et al., 2011). Relatedness has been shown to be predictive of how economies will develop in the future. That is, economic activities that enter a system are typically related to the ones that are already present (Hidalgo et al., 2018). Relatedness has also been use to construct network representations of these economic systems that reveal these potential diversification patterns (Hidalgo et al., 2007).

One of the measures that has become standard is the proximity measure proposed in Hidalgo et al. (2007). As opposed to considering the relative frequencies of economic activities, this proximity is based on the presence or absence of economic activities. The presence or absence of an economic activity in a location is defined using the RCA index, resulting in the presence–absence matrix **M** that is defined by

$$M_{ci} = \begin{cases} 1 & \text{if } RCA(c, i) > 1 \\ 0 & \text{if } RCA(c, i) \le 1, \end{cases}$$

The entries of this matrix are subsequently modeled as binary random variables X_{ci} which denote the presence or absence of activity i in location c. The co-occurrence of economic activities is then quantified by the conditional probability that an activity j is present given that i is present, that is

$$P(X_i = 1|X_j = 1) = \frac{\sum_c M_{ci} M_{cj}}{M_j},$$

where $M_j = \sum_c M_{cj}$ denotes the number of locations that specialize in activity j. To obtain a symmetric proximity matrix, the proximity between activities i and j is then defined as the minimum of two conditional probabilities

$$\phi_{ij} = \min\{P(X_i = 1|X_j = 1), P(X_j = 1|X_i = 1)\}.$$

At this point we can distinguish two major differences of this approach to the proposed approach using PMI. First, the underlying model is different, as the proximity measure is based on a presence-absence matrix, implying a binary random variable X_{ci} as opposed to a multinomial variable q_{ci}. The latter takes into account the intensity of economic activities (i.e., their quantity), while the former distinguishes only between presence and absence. The motivations behind thresholding the data using RCA is to appropriately normalize data and to reduce noise (Hidalgo, 2021).

We note that the thresholding procedure can also introduce noise, as activities that are not localized (i.e., those with $RCA(c, i) \approx 1$ for every location), may arbitrarily be set to 1 or 0 as their counts may be just over or under the threshold. As a consequence, nonlocalized activities will have a presence in about half of all locations, which is the maximal number of presences possible under the RCA measure.

A second difference is that the proximity measure is based on probabilities (and thus takes values between 0 and 1), whereas the PMI is based on information (and thus also takes negative values, and zero is a natural point of reference indicating the absence of association). Using conditional probabilities as a measure of colocation leads to the questionable property that the proximity measure is biased toward activities with many presences. This can be seen by considering two activities that are independent, such that $P(X_i|X_j) = P(X_i)$ and $P(X_j|X_i) = P(X_j)$. In that case, the proximity equals $\phi_{ij} = \min\{P(X_i = 1), P(X_j = 1)\} \propto \min\{M_i, M_j\}$ (Muneepeerakul et al., 2013). Hence, even though they are independent, the proximity between activities with many presences can be high. The PMI considers instead not the conditional probability but the difference in the information content of the conditional and the unconditional probability. Applied to the presence–absence matrix, this would lead to the measure

$$PMI(X_i = 1, X_j = 1) = \log(P(X_i = 1|X_j = 1)) - \log(P(X_i = 1)),$$

which is inherently symmetric and does not need to be symmetrized using the minimum.

4 | BAYESIAN ESTIMATION

Computing the information-theoretic measures of colocation requires estimates of the probabilities p_{ci} and p_{ij}. A straightforward way to estimate these probabilities is to consider the share of every location–activity pair, corresponding to the maximum-likelihood estimate $\hat{p}_{ci} = \frac{q_{ci}}{q}$. Here we estimate p_{ci} using a Bayesian framework, which has two major advantages over the maximum-likelihood approach. First, the Bayesian approach always returns nonzero probability estimates. This is crucial for the computation of the PMI, as it ensures all values will be finite. Second, the Bayesian framework yields a full posterior distribution for the estimated probabilities as opposed to a point estimate. The posterior distribution provides a natural description of the uncertainty in the estimated parameter values. These can be used, for example, to assess whether estimated associations are significantly nonzero.

Assuming that $\mathbf{Q}$ is generated by an independent sampling process, the probability of its realization is given by a multinomial distribution

$$P(\mathbf{Q}|\mathbf{p}) = \frac{\Gamma(q+1)}{\prod_{c,i}\Gamma(q_{ci}+1)}\prod_{c,i} p_{ci}^{q_{ci}},$$

where $\mathbf{p}$ is the matrix containing probabilities p_{ci}, with $\sum_{c,i} p_{ci} = 1$.

Applying Bayes' rule, the posterior distribution for the matrix of probabilities $\mathbf{p}$ is then given by

$$P(\mathbf{p}|\mathbf{Q}) \propto P(\mathbf{Q}|\mathbf{p})P(\mathbf{p}),$$

where $P(\mathbf{p})$ represents the prior distribution. A conjugate prior for the multinomial distribution is the Dirichlet distribution

$$P(\mathbf{p}|\boldsymbol{\alpha}) \sim Dir(\boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\prod_{c,i}\Gamma(\alpha_{ci})}\prod_{c,i} p_{ci}^{\alpha_{ci}-1},$$

where $\alpha = \sum_{c,i}\alpha_{ci}$. This gives the distribution of $\mathbf{p}$ given hyperparameter $\boldsymbol{\alpha}$. The posterior distribution for $\mathbf{p}$ given the data $\mathbf{Q}$ and hyperparameter $\boldsymbol{\alpha}$ is then given by

$$P(\mathbf{p}|\mathbf{Q},\boldsymbol{\alpha}) \sim Dir(\mathbf{Q}+\boldsymbol{\alpha}) \propto \prod_{ci} p_{ci}^{q_{ci}+\alpha_{ci}-1}.$$

An estimate for the parameters p_{ci} is then given by the expectation of the marginals of the posterior distribution, so that

$$\hat{p}_{ci} = \mathbb{E}[p_{ci}|\mathbf{Q},\boldsymbol{\alpha}] = \frac{q_{ci}+\alpha_{ci}}{q+\alpha} = \frac{\tilde{q}_{ci}}{\tilde{q}},$$

where we write $\tilde{q}_{ci} = q_{ci} + \alpha_{ci}$ and $\tilde{q} = \alpha + q$. The hyperparameter $\boldsymbol{\alpha}$ can be interpreted as a matrix of "pseudocounts", giving the assumed number of observed units of activity for every c, i pair before seeing the data $\mathbf{Q}$.

4.1 | Estimation of the posterior mean and variance of PMI

For the measurement of (co-)location, one could use the estimate $\hat{p}_{ci} = \mathbb{E}[p_{ci}]$ directly to compute $PMI(\hat{p}_{ci})$ and $PMI(\hat{p}_{ij})$. However, this will induce a systematic bias due to Jensen's inequality, which states that $\mathbb{E}[PMI(p_{ci})] \gtreqless PMI(\mathbb{E}[p_{ci}])$ depending on whether $PMI(p_{ci})$ is concave or convex. One needs instead an estimate of $PMI(p_{ci})$, which itself is a random variable the distribution of which is determined by the posterior distribution of p_{ci}. To this end, we approximate the mean and variance of the posterior distribution of $PMI(p_{ci})$, which will yield estimates of the expected location–activity association and the uncertainty around that estimate, respectively. Our approach is based on Wolpert and Wolf (1995) and Hutter and Zaffalon (2005), who discuss the estimation of information-theoretic quantities using a Bayesian approach in depth.

First, we approximate the posterior distribution of $PMI(p_{ci})$ by using its Taylor expansion around the mean $\hat{p}_{ci}$. Letting $\Delta_{ci} = p_{ci} - \hat{p}_{ci}$, and noting the fact that $|\Delta_{ci}| < 1$, yields

$$\begin{aligned}PMI(p_{ci}) &= PMI(\hat{p}_{ci}) + \sum_{c'j} \Delta_{c'j} \frac{\partial}{\partial p_{c'j}} PMI(p_{ci}) \\ &+ \sum_{c'j} \frac{\Delta_{ci}^2}{2} \frac{\partial^2}{\partial p_{c'j} \partial p_{c''k}} PMI(p_{ci}) + O\left(\Delta_{ci}^3\right)\end{aligned} \tag{10}$$

Note that $\mathbb{E}[\Delta_{ci}] = 0$ and thus $\mathbb{E}[\Delta_{ci}\Delta_{c'j}] = \mathrm{Cov}[p_{ci}\, p_{c'j}]$, where expectations are taken with respect to the posterior distribution of p_{ci}. Furthermore, Hutter and Zaffalon (2005) show that $\mathbb{E}[\Delta_{ci}^3] = O(\tilde{q}^{-2})$. It follows that (see Appendix A.1.1)

$$\begin{aligned}\mathbb{E}[PMI(p_{ci})] &\approx PMI(\hat{p}_{ci}) + \sum_{c'jc''k} \frac{\mathrm{Cov}[p_{c'j}\, p_{c''k}]}{2} \frac{\partial^2}{\partial p_{c'j} \partial p_{c''k}} PMI(p_{ci}) \\ &= PMI(\hat{p}_{ci}) + \frac{1}{2(q+1)} \left(\frac{1}{\hat{p}_c} + \frac{1}{\hat{p}_i} - \frac{1}{\hat{p}_{ci}} - 1 \right).\end{aligned} \tag{11}$$

The second term accounts for systematic bias in the estimate of $PMI(p_{ci})$.

The variance of $PMI(p_{ci})$ can be obtained by subtracting (11) from (10), leading to (see Appendix A.1.1)

$$\begin{aligned}\mathrm{Var}[PMI(p_{ci})] &= \mathbb{E}[(PMI(p_{ci}) - \mathbb{E}[PMI(p_{ci})])^2] \\ &\approx \frac{\mathrm{Var}[p_{ci}]}{\hat{p}_{ci}^2} + \frac{\mathrm{Var}[p_c]}{\hat{p}_c^2} + \frac{\mathrm{Var}[p_i]}{\hat{p}_i^2} \\ &= \frac{1}{\tilde{q}+1} \left(\frac{1}{\hat{p}_{ci}} + \frac{1}{\hat{p}_c} + \frac{1}{\hat{p}_i} - 3 \right).\end{aligned} \tag{12}$$

Equation (12) provides a measure for the uncertainty around the point estimate $\mathbb{E}[PMI(p_{ci})]$. In Section 4.3, we show how we evaluate the significance of an estimate using this information.

Approximations for the posterior expectation and variance of the KL divergence and MI are obtained using a similar approach. The full derivations are provided in Appendix A, and the results for location measures are shown in Table A1.

The estimates for the posterior distribution of $PMI(p_{ij})$ can be obtained by replacing p_{ci} with p_{ij}, although the computation of $\mathrm{Var}[p_{ij}]$ is more involved as p_{ij} is not Dirichlet distributed. The results for colocation measure are shown in Table A2.

We compare the analytical approximations to numerical simulations in Appendix B.2, showing the accuracy of the approximations of the posterior expectation and variance. Python code enabling computation of the posterior expectation and variance of all proposed information-theoretic quantities is available at https://github.com/aljevandam/Colocation.

4.2 | Choice of prior

The Bayesian estimation requires a choice for the prior distribution. Given the Dirichlet prior amounts to choosing a suitable matrix of pseudocounts $\boldsymbol{\alpha}$. Before going into the shape of the prior, it is worth noting that there is a close relation between the Bayesian estimate $\hat{p}_{ci}$ and the maximum likelihood estimate $\left(\text{given by } \frac{q_{ci}}{q}\right)$. This becomes clear by rewriting $\hat{p}_{ci}$ as

$$\hat{p}_{ci} = \frac{q}{q+\alpha}\left(\frac{q_{ci}}{q}\right) + \frac{\alpha}{q+\alpha}\left(\frac{\alpha_{ci}}{\alpha}\right),$$

showing that for large sample sizes relative to the total number of pseudocounts, that is, $\alpha \ll q$, the Bayesian estimate and the maximum-likelihood estimate are nearly identical. The importance of our prior in the final estimate is given by $\frac{\alpha}{\alpha+q}$, a quantity we will refer to as the prior weight.

However, even for low prior weights the effect of the prior for a particular $\hat{p}_{ci}$ may be substantial if $q_{ci} \ll \alpha_{ci}$, and may lead to considerable differences between the Bayesian estimate and the maximum-likelihood estimate for cells with a low number of observations. Note that in many applications, this is not uncommon because typical examples of the matrix **Q** are very sparse. That is, they contain few cells with many counts and many cells with no or very low counts.

What is an appropriate prior for the problem at hand? A popular choice is to set all $\alpha_{ci} = 1$, leading to a uniform prior. The uniform prior can be considered to be an uninformative prior as it gives equal probability to any probability distribution **p**. However, in count-data that cover several orders of magnitude (like the occupational data considered in the current paper), adding a constant pseudocount to cells with the fewest observations can have a large effect on the resulting estimates.

Here, we opt instead for a prior that assumes that the pseudocounts hold no information on *associations* in the data, that is, we choose a prior distribution such that $PMI(\hat{p}_{ci}) = 0$ and $PMI(\hat{p}_{ij}) = 0$ before seeing any data. This is accomplished by setting $\alpha_{ci} = \alpha\frac{q_c q_i}{q^2}$, so that $\mathbb{E}[p_{ci}|\boldsymbol{\alpha}] = \frac{q_c q_i}{q^2}$. We will refer to this as the "proportional prior" as it sets the number of pseudocounts proportional to the product of the marginals q_c and q_i in each cell. The parameter α controls the weight of the prior.

In Appendix C we study the effect of both the uniform prior and the proportional prior empirically, along with the effect of different prior weights. We find that the proportional prior exhibits practical properties: it allows smoothing the data by adding counts to the cells with zero observations, while keeping the estimated associations for other cells mostly in place. This allows to select a prior weight such that the estimated associations for cells with zero observations are of the same order of magnitude as the associations of the other cells.

4.3 | Significance testing

The Bayesian approach provides a measure of uncertainty for each estimate through the variance of the posterior distribution, allowing for statistical inference. For example, we can determine which of the estimated (co-)location associations is significantly nonzero using the "probability of direction", which can be considered as the Bayesian equivalent of the p value (Makowski et al., 2019).

The probability of direction determines the probability that an association is strictly positive or negative, which is given by proportion of the posterior distribution that is of the median's sign (Makowski et al., 2019). We compute this probability by assuming normality of the posterior distribution, and take an association to be significantly nonzero if this probability is less than some threshold ϵ. In Appendix B.3 we explain this in more detail and justify the normal approximation numerically.

The significance of an estimate will thus depend on the variance of the posterior distribution, which in turn depends on the assumptions made regarding the data-generating process. The more fine-grained the counts, the less variance in the estimated quantities (this is shown mathematically for p_{ci} in Appendix B.3). The reason is that the data generating process is assumed to create the data at the level of counts, so that more-fine grained units represent more observations.

Hence, the assumed units in the data-generating process will determine the significance of estimates. In the context of (co-)agglomeration of industries, for example, the relevant unit of analysis ideally matches the one at which location decisions are made. A plausible candidate for this is the plant, suggesting an analysis of data containing plant counts by industry for a given location. However, the *relative* uncertainty of estimates $\hat{p}_{ci}$ is independent of the units of **Q**, as the variance is affected by the granularity of the data in the same way across activities and locations if units have uniform sizes.

5 | EMPIRICAL EXAMPLE

As an example of an application of the proposed (co-)location measures, we apply them to US employment data from 2016, provided by the Bureau of Labor Statistics. These data are available at https://www.bls.gov/oes/special.requests/oesm16ma.zip. Using these data, we construct a matrix **Q** that contains the number of employees q_{ci} in a particular occupation group *i* for every Metropoliton Statistical Area (MSA) *c*, where the occupations groups are defined at two levels of aggregation (major and detailed). After excluding all MSA's located in Puerto Rico, all occupations in the major group “Farming, Fishing and Forestry”, and removing the 10% smallest detailed occupations groups, the data consists of 387 MSA's, 21 major occupations groups, and 720 detailed occupation groups. In the following, we will refer to MSA's and occupations groups as “cities” and “occupations”, respectively, distinguishing between major en detailed occupation groups where necessary.

The minimal number of counts found in the data equals 10. This suggests a data-generating process that assigns occupations per 10 workers. To this end, we divide the counts in the data by 10 to illustrate how the sizes of sampled units matter. The prior is taken to be the “proportional” prior as discussed in Section 4.2, setting the prior weight such that it represents 5% of the total counts. With this prior strength, the location associations are all of the same order of magnitude (see Appendix C).

5.1 | (Co-)location patterns of occupations in the US

Figure 1a shows the location association of each city and major occupation group. The rows and columns are sorted by size of each city and occupation respectively, putting the largest occupations in the bottom rows and the largest cities in the columns on the right-hand side. Recall that apart from differences in the estimation procedure, these associations are equivalent to the log-transformed RCA or Location Quotient. Indeed, the Spearman rank correlation between the two measures is 0.99. Figure 1b shows the localization of each occupation, that is, the average location association of each occupation, given by $KL(p_{c|i})$. The localization shows for each occupation, represented by a row of the association matrix, how much its distribution deviates from the city-size distribution. The matrix of associations shows in which cities these deviations take place.

The largest occupations are generally the least localized and have associations close to zero for all cities. This implies that these occupations are distributed proportional to city size, that is, their relative frequency is equal in each city. These occupations consist mostly of nontraded services, including “Office and Administrative Support”, “Sales and Related” and “Food preparation and serving.” The occupations with high localization show varying patterns of association, showing that the nature of localization can differ across occupations. For example, “Production” has a strong association with small- and middle-sized cities, and has weak or negative associations with the largest cities. “Computer and Mathematical” on the other hand is mainly associated with the largest cities. A similar pattern is found for other localized occupations such as “Arts, Design, Entertainment, Sports and Media” and “Legal.” These occupations represent typical “big city” occupations, which seem to consist mostly of knowledge-intensive services. “Life, Physical and Social Sciences” occupations are highly localized but associated with cities of varying size—a pattern that is possibly driven by university towns of varying sizes.

Figure 1c,d show the colocation associations and codependence of each occupation, revealing a pattern that is consistent with the location patterns. The most codependent occupations are “Production”, “Computer and Mathematical” and “Life, Physical and Social Sciences.” “Production” has positive associations only to “Transportation and Materials Moving and Installation”, “Maintenance and Repair” and “Architecture and Engineering”, and negative associations with most other occupations. “Computer and Mathematical” is positively associated with other knowledge-intensive services such as “Business and Financial Operations”, “Life, Physical and Social Sciences” and “Architecture and Engineering.” “Life, Physical and Social Sciences” has the strongest self-association, although it is also positively associated with most other knowledge-intensive services. The least codependent occupations, that is, those having on average a neutral association with other occupations, consist mostly of nontraded services such as “Protective service”, ‘Food Preparation and Serving’ and ‘Personal Care and Service’.

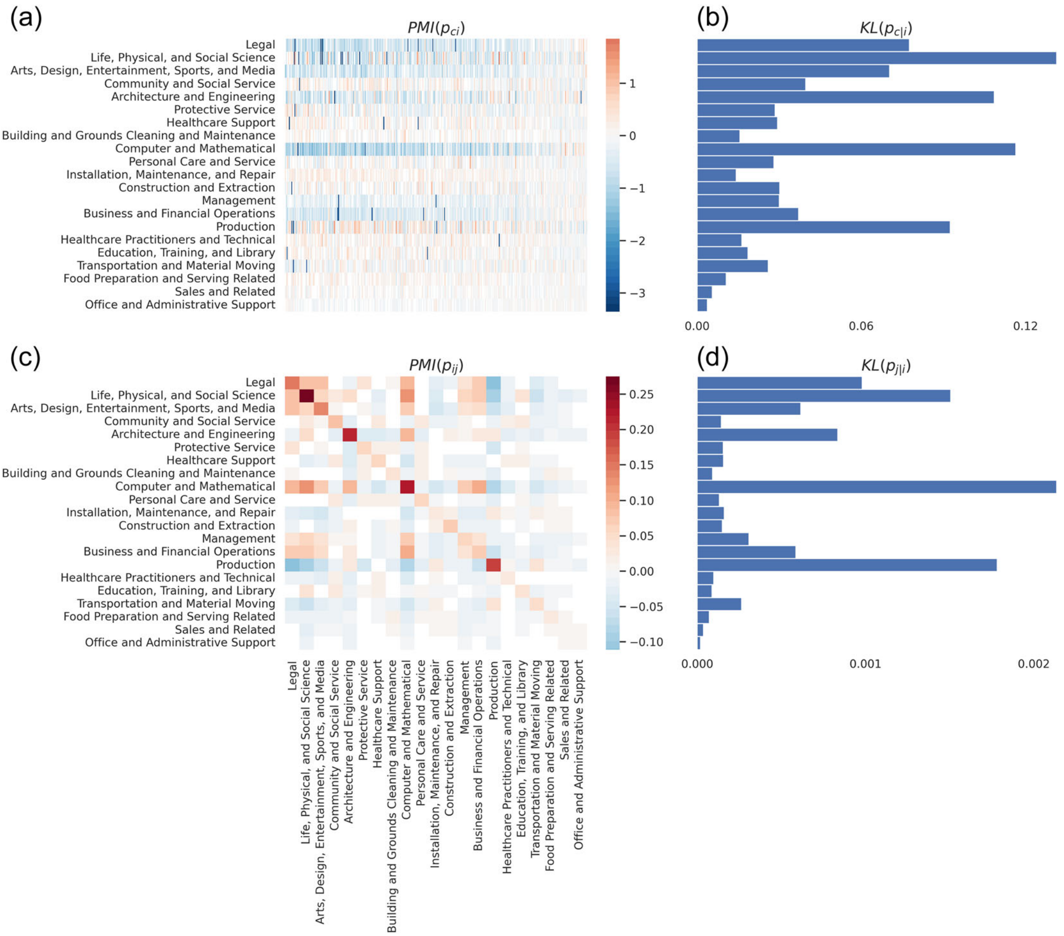

FIGURE 1 (a) Location associations of city–occupation pairs. Columns represent cities, sorted by size with the largest city on the right. Occupations sorted by size with smallest on top. (b) Localization of occupations. (c) Colocation association for occupation pairs. (d) Codependence of occupations.

5.2 | Comparison to existing colocation measures

How do these results compare to other measures of colocation? Figure 2 shows the colocation patterns of major occupation groups using PMI (A), the Ellison–Glaeser (EG) coagglomeration measure (B), and Hidalgo et al.'s proximity measure (C). The colocation association and coagglomeration measures give at first sight a similar pattern, whereas the proximity measure yields rather different results. Proximity assigns high values to occupations that generally have low colocation associations, such as “Community and Social Service”, “Personal Care and Service”, and “Healthcare Practitioners.” Furthermore, the distinct pattern for “Production”, showing a negative association to most other occupations, seems to be washed out by the thresholding procedure and is not visible in the proximity measure, where it shows a pattern that is similar to the nontraded services.

Figure 3 compares the three measures for both major occupations groups (top row) and detailed occupation groups (bottom row). Even though the PMI and EG give a similar pattern overall (their rank correlation is 0.84), results for specific

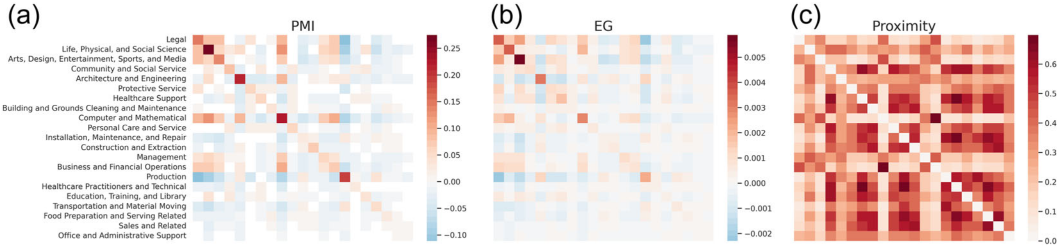

FIGURE 2 Heatmaps of PMI (a), Ellison-Glaeser coagglomeration (b), and proximity (c). Occupations are sorted by size (smallest on top/left).

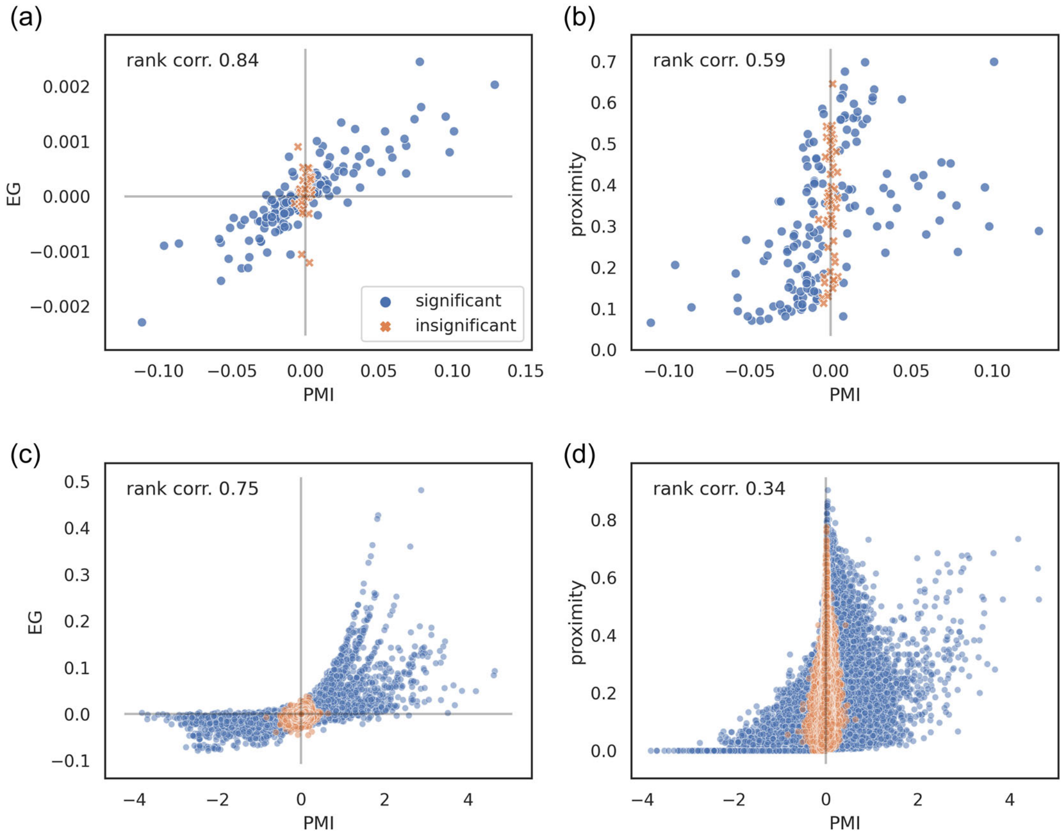

FIGURE 3 Scatter plots of the colocation of occupation pairs, excluding self-associations. (a) PMI against the coagglomeration for major occupation groups. (b) PMI against proximity for major occupation groups. (c) PMI against the coagglomeration for major detailed occupation groups. (d) PMI against proximity for detailed occupation groups. Orange markers indicate significance according to the Bayesian estimation procedure. Insets in top left show the Spearman rank correlation.

occupation pairs can differ substantially between the two measures. Figure 3a shows that there exist occupations pairs with a positive colocation association but a negative coagglomeration and vice versa. Furthermore, Table 3 shows that the top and bottom ranking occupation pairs can differ a lot for both measures. There are even some pairs whose association is insignificant according to the PMI, but rank quite high in the EG measure (see Table 4).

TABLE 3 Top 5 and bottom 5 colocation pairs by PMI, excluding self-associations

Occupation 1	Occupation 2	*PMI* (p_{ij})	EG	Proximity	Rank PMI	Rank EG	Rank proximity	Significant
Computer and mathematical	Life, physical, and social science	0.126	0.002	0.288	1	2	118	Yes
Business and financial operations	Computer and mathematical	0.099	0.001	0.699	2	8	1	Yes
Computer and mathematical	Architecture and engineering	0.096	0.001	0.299	3	18	114	Yes
Computer and mathematical	Legal	0.093	0.001	0.394	4	4	68	Yes
Life, physical, and social science	Legal	0.077	0.002	0.237	5	3	137	Yes
Production	Business and financial operations	−0.056	−0.001	0.126	206	198	183	Yes
Transportation and material Moving	Life, physical, and social science	−0.057	−0.001	0.185	207	194	152	Yes
Production	Computer and mathematical	−0.084	−0.001	0.103	208	199	194	Yes
Production	Life, physical, and social science	−0.093	−0.001	0.206	209	200	147	Yes
Production	Legal	−0.108	−0.002	0.065	210	210	210	Yes

TABLE 4 Top 5 colocation pairs by EG for insignificant associations, excluding self-associations

Occupation 1	Occupation 2	$PMI(p_{ij})$	EG	Proximity	Rank PMI	Rank EG	Rank proximity	Significant
Healthcare support	Legal	−0.005	0.001	0.126	118	14	182	No
Education, training, and library	Legal	−0.002	0.001	0.129	102	29	180	No
Education, training, and library	Arts, design, entertainment, sports, and media	0.002	0.001	0.169	69	30	161	No
Community and social service	Legal	0.003	0.001	0.163	64	31	164	No
Installation, maintenance, and repair	Architecture and engineering	0.000	0.000	0.304	86	37	110	No

TABLE 5 Top 10 colocation pairs by EG for detailed occupations, excluding self-associations

Occupation 1	Occupation 2	$PMI(p_{ij})$	EG	Proximity
Economists	Political scientists	2.873	0.481	0.067
Rail-track laying and maintenance equipment operators	Railroad conductors and yardmasters	1.844	0.427	0.111
Subway and streetcar operators	Railroad conductors and yardmasters	1.821	0.419	0.333
Barbers	Railroad conductors and yardmasters	1.707	0.363	0.250
Political scientists	Artists and related workers, all other	2.613	0.360	0.133
Fashion designers	Railroad conductors and yardmasters	1.668	0.339	0.091
Railroad conductors and yardmasters	Costume attendants	1.611	0.325	0.053
Grounds maintenance workers, all other	Railroad conductors and yardmasters	1.500	0.280	0.043
Railroad conductors and yardmasters	Fabric and apparel patternmakers	1.479	0.263	0.111
Fabric and apparel patternmakers	Makeup artists, theatrical and performance	1.753	0.260	0.222

Figure 3c shows that the discrepancy between the colocation association and coagglomeration becomes even more pronounced for the detailed occupation groups. Both measures produce a completely different top 10 (Tables 5 and 6). The highest colocation associations are assigned to occupations in the textiles industry, while the highest coagglomeration is assigned to pairs involving “Political Scientists” and “Railroad Conductors and Yardmasters.” These differences may have profound consequences for empirical analyses.

A similar analysis is shown for the proximity measure (Figure 3b,d). With rank correlations of 0.58 for major occupations groups and 0.34 for the detailed occupation groups, the PMI and proximity are very different. Major occupation groups with insignificant associations have a wide range of proximities, and among them are some of the occupation pairs with the highest proximity (see Table 7). Also, some of the pairs with the strongest associations have low proximities, as shown in Table 3.

The fact that occupation pairs with small colocation associations get assigned high proximities can be understood through the properties of the proximity measure described in Section 3.4. Occupation pairs consisting of occupations with low localization, that is, $PMI(p_{ci}) \approx 0$ for most locations, will also have low colocation association. That is, they offer little surprise in the information-theoretic sense. Yet when constructing the presence matrix, these occupations may be assigned many presences by the thresholding procedure used (since $RCA(c, i) \approx 1$). Figure 4a shows that this is the case empirically for the detailed occupation groups. These presences can in turn lead to high proximities, since the conditional probability of occurrence for an independent occupation equals the marginal probability of occurrence, which is proportional to the number of presences of an occupation. Indeed, Figure 4b shows that occupation pairs with many presences have high proximities. Hence, the least localized occupations can end up being the most proximate. This effect is clearly visible in Figure 2, in which occupations pairs with neutral colocation association are assigned high proximities.

This effect can be particularly important when considering network representations based on the proximity measure that show only the edges with highest proximity: for data containing many activities with low localization, these networks will emphasize precisely the edges between activities that have a neutral association as measured by PMI.

TABLE 6 Top 10 colocation pairs by PMI for detailed occupations, excluding self-associations

Occupation 1	Occupation 2	$PMI(p_{ij})$	EG	Proximity
Textile winding, twisting, and drawing out machine setters, operators, and tenders	Extruding and forming machine setters, operators, and tenders, synthetic and glass fibers	4.634	0.092	0.524
Textile winding, twisting, and drawing out machine setters, operators, and tenders	Textile knitting and weaving machine setters, operators, and tenders	4.611	0.083	0.632
Textile winding, twisting, and drawing out machine setters, operators, and tenders	Textile bleaching and dyeing machine operators and tenders	4.183	0.049	0.733
Textile knitting and weaving machine setters, operators, and tenders	Extruding and forming machine setters, operators, and tenders, synthetic and glass fibers	3.850	0.050	0.524
Textile winding, twisting, and drawing out machine setters, operators, and tenders	Textile cutting machine setters, operators, and tenders	3.678	0.019	0.281
Textile knitting and weaving machine setters, operators, and tenders	Textile bleaching and dyeing machine operators and tenders	3.646	0.062	0.684
Textile Winding, Twisting, And Drawing Out Machine Setters, Operators, And Tenders	Textile, apparel, and furnishings workers, all other	3.512	0.002	0.300
Helpers–extraction workers	Derrick operators, oil and gas	3.443	0.156	0.476
Extruding and forming machine setters, operators, and tenders, synthetic and glass fibers	Textile bleaching and dyeing machine operators and tenders	3.433	0.056	0.524
Helpers–extraction workers	Wellhead pumpers	3.406	0.123	0.429

TABLE 7 Top 5 colocation pairs by proximity for insignificant associations, excluding self-associations

Occupation 1	Occupation 2	*PMI* (p_{ij})	EG	Proximity	Rank PMI	Rank EG	Rank proximity	Significant
Food preparation and serving related	Healthcare practitioners and technical	0.001	0.000	0.645	75	77	4	No
Sales and related	Healthcare practitioners and technical	0.001	0.000	0.544	76	89	24	No
Production	Healthcare support	−0.002	−0.001	0.542	107	202	25	No
Installation, maintenance, and repair	Healthcare support	−0.000	−0.000	0.537	89	163	26	No
Education, training, and library	Building and grounds cleaning and maintenance	0.001	0.000	0.522	77	67	30	No

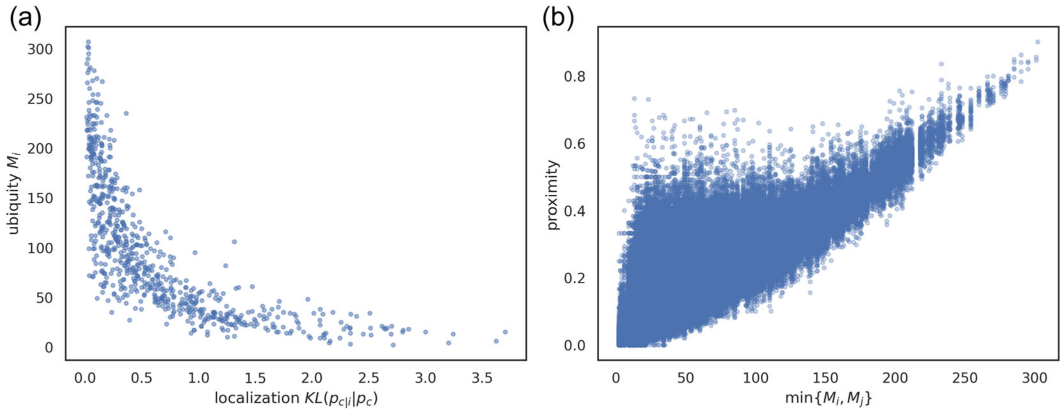

FIGURE 4 Relation between localization and ubiquities (a), and presences and proximities (b) for detailed occupation groups

6 | DISCUSSION

Information theory offers a unified way to estimate location and colocation associations using PMI. This yields measures that are similar to the well-known RCA index Balassa (1965) or Location Quotient (Isard, 1960) and the coagglomeration index (Ellison et al., 2010). However, our measures have important advantages over these existing measures.

First, by deriving these metrics from a unified framework, we were able to show the intrinsic connections between hitherto seemingly unrelated measures. This is not only satisfying from a methodological point of view, but allows exploring the relations between concepts like revealed comparative advantage, specialization, localization, concentration, and colocation.

Second, the proposed measures are derived from a formal framework (information theory) in a way that is explicit in the assumed data generating process, the chosen null models, and the estimation procedures. Different choices for these assumptions lead to different results. However, the afforded transparency allows constructing arguments against and in favor of such alternatives that take into consideration aspects of the specific context at hand. Such a discussion can be framed in terms of an underlying model, rather than ad hoc specificities of a particular index. For instance, we used a null model based on the assumption that neutral associations imply a distribution of location-activity pairs that is proportional to the sizes of locations and activities (Hoover, 1936). The assumption that activities are distributed proportional to the area of a location (Mori et al., 2005) leads to a different null model. Another possibility is to determine the expected number of (co-)occurrences on the basis of external factors that could drive the distribution of activities over locations, using for instance a regression model (Jara-Figueroa et al., 2018; Neffke et al., 2011).

Third, the framework provides uncertainty estimates for all the information-theoretic quantities involved which can be used to make statistical inferences. The significance test presented here is just one example of many possibilities. Most currently used indices are applied without any notion of uncertainty. Using these uncertainties in practice however may present some challenges, as the Bayesian estimation procedure leaves room for the selection of different priors, prior weights, and granularity of the data generating process. Here, for reasons of practicality, we applied a Dirichlet prior with parameters chosen such that the associations before seeing data are zero. However, in some contexts, alternative priors may be natural choices. An example of this is the maximum entropy prior (Wolpert & Wolf, 1995). When choosing non-Dirichlet priors, posterior distributions may be obtained through numerical simulation. Setting the granularity of the data (i.e., the total number of observations) will determine the absolute magnitude of the uncertainty in the estimates. This simply reiterates that inferences should always be made with an underlying data-generating process in mind. In spite of this, we can still make statements about the relative magnitudes of uncertainties, which are independent of the granularity of the data generating process. Future research could expand the underlying probabilistic model to incorporate more specific models of location choice, including heterogeneity in the "chunk size" of observational units in the data generating process, as in the case of the nonuniform plant size distribution of Ellison and Glaeser (1997).

Finally, it is important to note that the information-theoretic approach can be readily extended to move beyond an analysis of pairwise colocations, as it also allows analyzing multivariate associations. For instance, one could analyze associations between multiple variables (e.g., occupations, cities, and industries) or multiway colocations (such as the colocation of triplets instead of pairs of activities). The PMI between three economic activities i, j, k is given by $PMI(p_{ijk}) = \log\left(\frac{p_{ijk}}{p_i p_j p_k}\right)$. Such higher-order associations could be further analyzed using the information-theoretic concepts of redundancy and synergy (Finn & Lizier, 2018). This may help disentangle different types of associations, corresponding to different economic interactions. For instance, the association between a pair of economic activities could be conditional on the presence of (a specific combination of) other activities, or be driven by the mutual dependence on a (combination of) other economic activities or on some external variable such as the presence of a natural resource. Further development of this analytical framework could help reveal such higher-order relations among economic activities.

ACKNOWLEDGMENTS

A.D. and K.F. are funded by the Netherlands Organisation for Scientific Research (NWO) under the Vici scheme, number 453-14-014. F.N. received financial support from the Austrian Research Agency (FFG), project #873927 (ESSENCSE).

ORCID

Frank Neffke http://orcid.org/0000-0002-3924-6636

Koen Frenken http://orcid.org/0000-0003-4731-0201

REFERENCES

Alabdulkareem, A., Frank, M. R., Sun, L., AlShebli, B., Hidalgo, C., & Rahwan, I. (2018). Unpacking the polarization of workplace skills. *Science Advances*, *4*(7), eaao6030. https://doi.org/10.1126/sciadv.aao6030

Balassa, B. (1965). Trade liberalisation and "revealed" comparative advantage. *The Manchester School*, *33*(2), 99–123. https://doi.org/10.1111/j.1467-9957.1965.tb00050.x

Ballance, R., Forstner, H., & Murray, T. (1987). Consistency tests of alternative measures of comparative advantage. *The Review of Economics and Statistics*, *69*(1), 157. https://doi.org/10.2307/1937915

Boschma, R., Minondo, A., & Navarro, M. (2013). The emergence of new industries at the regional level in Spain: A proximity approach based on product relatedness. *Economic Geography*, *89*(1), 29–51. https://doi.org/10.1111/j.1944-8287.2012.01170.x

Boschma, R., Balland, P. A., & Kogler, D. F. (2015). Relatedness and technological change in cities: The rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Industrial and Corporate Change*, *24*(1), 223–250. https://doi.org/10.1093/icc/dtu012

Church, K., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics* (Vol. 16, pp. 76–83). Association for Computational Linguistics. https://doi.org/10.3115/981623.981633

Cover, T., & Thomas, J. (2005). *Elements of information theory*. Wiley. https://doi.org/10.1002/047174882X

Diodato, D., Neffke, F., & O'Clery, N. (2018). Why do industries coagglomerate? How marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics*, *106*, 1–26. https://doi.org/10.1016/j.jue.2018.05.002

Ellison, G., & Glaeser, E. (1997). Geographic concentration in U.S. manufacturing industries: A dartboard approach. *Journal of Political Economy*, *105*(5), 889–927. https://doi.org/10.1086/262098

Ellison, G., & Glaeser, E. (1999). The geographic concentration of industry: Does natural advantage explain agglomeration? *American Economic Review*, *89*(2), 311–316. https://doi.org/10.1257/aer.89.2.311

Ellison, G., Glaeser, E., & Kerr, W. (2010). What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review*, *100*(3), 1195–1213. https://doi.org/10.1257/aer.100.3.1195

Faggio, G., Silva, O., & Strange, W. (2017). Heterogeneous agglomeration. *Review of Economics and Statistics*, *99*(1), 80–94. https://doi.org/10.1162/REST_a_00604

Fano, R. (1961). *Transmission of information: A statistical theory of communications*. Wiley.

Finn, C., & Lizier, J. (2018). Pointwise partial information decomposition using the specificity and ambiguity lattices. *Entropy*, *20*(4), 297. https://doi.org/10.3390/e20040297

Hidalgo, C. A. (2021). Economic complexity theory and applications. *Nature Reviews Physics*, *3*(2), 92–113. https://doi.org/10.1038/s42254-020-00275-1

Hidalgo, C. A., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., Glaeser, E., He, C., Kogler, D. F., Morrison, A., Neffke, F., Rigby, D., Stern, S., Zheng, S., & Zhu, S. (2018). The principle of relatedness. In A. J. Morales, C. Gershenson, D. Braha, A. A. Minai, & Y. Bar-Yam (Eds.), *Unifying themes in complex systems IX* (pp. 451–457). Springer International Publishing. https://doi.org/10.1007/978-3-319-96661-8_46

Hidalgo, C. H., Klinger, B., Barabási, A.-L., & Hausmann, R. (2007). The product space conditions the development of nations. *Science*, *317*(5837), 482–487. https://doi.org/10.1126/science.1144581

Hoen, A. R., & Oosterhaven, J. (2006). On the measurement of comparative advantage. *Annals of Regional Science*, *40*(3), 677–691. https://doi.org/10.1007/s00168-006-0076-4

Hoover, E. (1936). The measurement of industrial localization. *The Review of Economics and Statistics*, *18*(4), 162. https://doi.org/10.2307/1927875

Hutter, M., & Zaffalon, M. (2005). Distribution of mutual information from complete and incomplete data. *Computational Statistics and Data Analysis*, *48*(3), 633–657. https://doi.org/10.1016/j.csda.2004.03.010

Isard, W. (1960). *Methods of regional analysis: An introduction to regional science*. MIT Press.

Jara-Figueroa, C., Jun, B., Glaeser, E. L., & Hidalgo, C. A. (2018). The role of industry-specific, occupation-specific, and location-specific knowledge in the growth and survival of new firms. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(50), 12646–12653. https://doi.org/10.1073/pnas.1800475115

Krugman, P. (1991). *Geography and trade*. MIT Press.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Kunimoto, K. (1977). Typology of trade intensity indices. *Hitotsubashi Journal of Economics*, *17*(2), 15–32. https://doi.org/10.15057/7981

Laursen, K. (2015). Revealed comparative advantage and the alternatives as measures of international specialization. *Eurasian Business Review*, *5*(1), 99–115. https://doi.org/10.1007/s40821-015-0017-1

Makowski, D., Ben-Shachar, M. S., Chen, S. H., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*, 1–14. https://doi.org/10.3389/fpsyg.2019.02767

Marshall, A. (1920). *Principles of economics* (8th ed.). MacMillan.

Mealy, P., & Teytelboym, A. (2020). Economic complexity and the green economy. *Research Policy*. Advance online publication. https://doi.org/10.1016/j.respol.2020.103948

Miao, L., Murray, D., Jung, W.-S., Larivière, V., Sugimoto, C. R., & Ahn, Y.-Y. (2022). The latent structure of global scientific development. *Nature Human Behaviour*. Advance online publication. https://doi.org/10.1038/s41562-022-01367-x

Mori, T., Nishikimi, K., & Smith, T. (2005). A divergence statistic for industrial localization. *Review of Economics and Statistics*, *87*(4), 635–651. https://doi.org/10.1162/003465305775098170

Muneepeerakul, R., Lobo, J., Shutters, S. T., Goméz-Liévano, A., & Qubbaj, M. R. (2013). Urban economies and occupation space: Can they get "there" from "here"? *PLoS ONE*, *8*(9), e73676. https://doi.org/10.1371/journal.pone.0073676

Neffke, F., Henning, M., & Boschma, R. (2011). How do regions diversify over time? industry relatedness and the development of new growth paths in regions. *Economic Geography*, *87*(3), 237–265. https://doi.org/10.1111/j.1944-8287.2011.01121.x

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*(2), 135–146.

Sattinger, M. (1993). Assignment models of the distribution of earnings. *Journal of Economic Literature*, *31*(2), 831–880.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Theil, H. (1967). *Economics and information theory*. North-Holland Publishing Company.

Theil, H. (1972). *Statistical decomposition analysis: With applications in the social and administrative sciences*. North-Holland Publishing Company.

Vollrath, T. (1991). A theoretical evaluation of alternative trade intensity measures of revealed comparative advantage. *Weltwirtschaftliches Archiv*, *127*(2), 265–280. https://doi.org/10.1007/BF02707986

Wolpert, D., & Wolf, D. (1995). Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, *52*(6), 6841–6854. https://doi.org/10.1103/PhysRevE.52.6841

Yeats, A. (1985). On the appropriate interpretation of the revealed comparative advantage index: Implications of a methodology based on industry sector analysis. *Weltwirtschaftliches Archiv*, *121*(1), 61–73. https://doi.org/10.1007/BF02705840

Yu, R., Cai, J., & Leung, P. S. (2009). The normalized revealed comparative advantage index. *Annals of Regional Science*, *43*(1), 267–282. https://doi.org/10.1007/s00168-008-0213-3

How to cite this article: Dam, A., Gomez-Lievano, A., Neffke, F., & Frenken, K. (2023). An information-theoretic approach to the analysis of location and colocation patterns. *Journal of Regional Science*, 63, 173–213. https://doi.org/10.1111/jors.12621

APPENDIX A: DERIVATIONS OF POSTERIOR MEAN AND VARIANCE

A.1. Location estimates

As described in the main text, the Dirichlet prior leads to a Dirichlet posterior for the matrix of probabilities $\hat{p}_{ci}$, for which exact values of the mean and variance of the marginals are easily written down. In the following, we show how we approximate the mean and variance of the posterior distribution of the information-theoretic measures applied to these probabilities, following the approach suggested in Wolpert and Wolf (1995) and Hutter and Zaffalon (2005). An overview of the results that follow is given in Table A1.

It should be noted that although we approximate the moments of the posterior distributions of $PMI(p_{ci})$ here, it is in this specific case possible to obtain exact results for the moments of the posterior distributions, since the moments of logarithmically transformed Dirichlet random variables are known and given by the digamma function (see Wolpert & Wolf, 1995 and Hutter & Zaffalon, 2005). For example, the expectation is given by $\mathbb{E}[PMI(p_{ci})] = \psi(\tilde{q}_{ci}) - \psi(\tilde{q}_c) - \psi(\tilde{q}_i) + \psi(\tilde{q})$. Nevertheless, we use the approximation approach here since it generalizes to the case where the variables are not Dirichlet distributed, such as in case of colocation probabilities p_{ij} or when using non-Dirichlet priors.

A.1.1 The posterior mean and variance of *PMI* ($\mathbf{p}_{ci}$)

To obtain an approximation for the posterior distribution of $PMI(p_{ci})$, we closely follow Hutter and Zaffalon (2005), computing its Taylor expansion around the $\hat{p}_{ci}$. Writing $\Delta_{ci} = p_{ci} - \hat{p}_{ci}$, and noting the fact that $|\Delta_{ci}| < 1$, this gives

$$\begin{aligned}
PMI(p_{ci}) &= PMI(\hat{p}_{ci}) + \sum_{c'j} \Delta_{c'j} \tfrac{\partial}{\partial p_{c'j}} PMI(p_{ci}) + \sum_{c'j} \frac{\Delta_{ci}^2}{2} \frac{\partial^2}{\partial p_{c'j} \partial p_{c''k}} PMI(p_{ci}) + \mathcal{O}\left(\Delta_{ci}^3\right) \\
&= PMI(\hat{p}_{ci}) + \sum_{c'j} \Delta_{c'j} \left(\tfrac{\delta_{cc'}\delta_{ij}}{\hat{p}_{ci}} - \tfrac{\delta_{cc'}}{\hat{p}_c} - \tfrac{\delta_{ij}}{\hat{p}_i}\right) \\
&\quad + \sum_{c'jc''k} \frac{\Delta_{c'j}\Delta_{c''k}}{2} \left(-\frac{\delta_{cc'c''}\delta_{ijk}}{\hat{p}_{ci}^2} + \frac{\delta_{cc'c''}}{\hat{p}_c^2} + \frac{\delta_{ijk}}{\hat{p}_i^2}\right) + \mathcal{O}\left(\Delta_{ci}^3\right).
\end{aligned} \tag{A1}$$

TABLE A1 Overview of approximations of the posterior mean and variance for location measures

Variable	Posterior mean	Posterior variance
p_{ci}	$\hat{p}_{ci} = \frac{\tilde{q}_{ci}}{\tilde{q}}$	$\frac{\hat{p}_{ci}(1 - \hat{p}_{ci})}{(\tilde{q} + 1)}$
$p_{c\mid i}$	$\hat{p}_{c\mid i} = \frac{\tilde{q}_{ci}}{\tilde{q}_i}$	$\frac{\hat{p}_{c\mid i}(1 - \hat{p}_{c\mid i})}{\tilde{q}_i + 1}$
$PMI(p_{ci})$	$PMI(\hat{p}_{ci}) + \frac{1}{2(q + 1)}\left(\frac{1}{\hat{p}_c} + \frac{1}{\hat{p}_i} - \frac{1}{\hat{p}_{ci}} - 1\right)$	$\frac{1}{q + 1}\left(\frac{1}{\hat{p}_{ci}} + \frac{1}{\hat{p}_c} + \frac{1}{\hat{p}_i} - 3\right)$
$KL(p_{c\mid i})$	$KL(\hat{p}_{c\mid i} \mid \hat{p}_c) + \frac{N_c - 1}{2(\tilde{q}_i + 1)} + \frac{1}{2(q + 1)}\left(\sum_c \frac{\hat{p}_{c\mid i}}{\hat{p}_c} - 1\right)$	$\frac{1}{\tilde{q}_i + 1}\left(\sum_c \hat{p}_{c\mid i} PMI(\hat{p}_{ci})^2 - KL(\hat{p}_{c\mid i} \mid \hat{p}_c)^2\right) + \frac{1}{q + 1}\left(\sum_c \frac{\hat{p}_{c\mid i}^2}{\hat{p}_c} - 1\right)$
$MI(p_{ci})$	$MI(\hat{p}_{ci}) + \frac{(N_c - 1)(N_i - 1)}{2(\tilde{q} + 1)}$	$\frac{1}{\tilde{q} + 1}\left(\sum_{ci} \hat{p}_{ci} PMI(\hat{p}_{ci})^2 - MI(\hat{p}_{ci})^2\right)$

Note that $\mathbb{E}[\Delta_{ci}] = 0$ and thus $\mathbb{E}[\Delta_{ci}^2] = \mathrm{Var}[p_{ci}]$ and $\mathbb{E}[\Delta_{ci}\Delta_{c'j}] = \mathrm{Cov}[p_{ci}, p_{c'j}]$, where expectations are taken with respect to the posterior distribution of p_{ci}. Furthermore, (Hutter & Zaffalon, 2005) show that $\mathbb{E}[\Delta_{ci}^3] = \mathcal{O}(\tilde{q}^{-2})$. It follows that

$$\begin{aligned}\mathbb{E}[PMI(p_{ci})] &\approx PMI(\hat{p}_{ci}) + \sum_{c'jc''k} \frac{\mathrm{Cov}[p_{c'j}, p_{c''k}]}{2}\left(-\frac{\delta_{cc'c''}\delta_{ijk}}{\hat{p}_{ci}^2} + \frac{\delta_{cc'c''}}{\hat{p}_c^2} + \frac{\delta_{ijk}}{\hat{p}_i^2}\right)\\ &= PMI(\hat{p}_{ci}) + \sum_{c'jc''k} \frac{\delta_{c'c''}\delta_{jk}\hat{p}_{c'j} - \hat{p}_{c'j}\hat{p}_{c''k}}{2(\tilde{q}+1)}\left(-\frac{\delta_{cc'c''}\delta_{ijk}}{\hat{p}_{ci}^2} + \frac{\delta_{cc'c''}}{\hat{p}_c^2} + \frac{\delta_{ijk}}{\hat{p}_i^2}\right)\\ &= PMI(\hat{p}_{ci}) + \frac{1}{2(\tilde{q}+1)}\left(-\frac{\hat{p}_{ci} - \hat{p}_{ci}^2}{\hat{p}_{ci}^2} + \sum_{jk}\frac{\delta_{jk}\hat{p}_{cj} - \hat{p}_{cj}\hat{p}_{ck}}{\hat{p}_c^2} + \sum_{c'c''}\frac{\delta_{c'c''}\hat{p}_{c'i} - \hat{p}_{c'i}\hat{p}_{c''i}}{\hat{p}_i{}^2}\right)\\ &= PMI(\hat{p}_{ci}) + \frac{1}{2(\tilde{q}+1)}\left(\frac{1}{\hat{p}_c} + \frac{1}{\hat{p}_i} - \frac{1}{\hat{p}_{ci}} - 1\right),\end{aligned} \tag{A2}$$

where we used that $\mathrm{Cov}[p_{c'j}, p_{c''k}] = \frac{\delta_{c'c''}\delta_{jk}\hat{p}_{c'j} - \hat{p}_{c'j}\hat{p}_{c''k}}{\tilde{q}+1}$ since the matrix of p_{ci} Dirichlet distributed.

The variance of $PMI(p_{ci})$ can be obtained by subtracting (A1) from (A1):

$$\begin{aligned}\mathrm{Var}[PMI(p_{ci})] &= \mathbb{E}[(PMI(p_{ci}) - \mathbb{E}[PMI(p_{ci})])^2]\\ &\approx \mathbb{E}\left[\left(\sum_{c'j}\Delta_{c'j}\left(\frac{\delta_{cc'}\delta_{ij}}{\hat{p}_{ci}} - \frac{\delta_{cc'}}{\hat{p}_c} - \frac{\delta_{ij}}{\hat{p}_i}\right)\right)^2\right]\\ &= \mathbb{E}\left[\left(\frac{\Delta_{ci}}{\hat{p}_{ci}} - \frac{\sum_j\Delta_{cj}}{\hat{p}_c} - \frac{\sum_{c'}\Delta_{c'i}}{\hat{p}_i}\right)^2\right]\\ &= \mathbb{E}\left[\left(\frac{\Delta_{ci}}{\hat{p}_{ci}} - \frac{\Delta_c}{\hat{p}_c} - \frac{\Delta_i}{\hat{p}_i}\right)^2\right].\end{aligned} \tag{A3}$$

Writing out all the terms within the above expectation yields product of Δ_{ci}, Δ_i, and Δ_c. Since p_{ci}, p_i, and p_c are mutually independent we have that $\mathbb{E}[\Delta_{ci}\Delta_c] = \mathbb{E}[\Delta_{ci}]\mathbb{E}[\Delta_c] = 0$, and this holds for all cross-terms. For the square terms we have $\mathbb{E}[\Delta_{ci}^2] = \mathrm{Var}[p_{ci}]$ and likewise for Δ_c and Δ_i, so that

$$\begin{aligned}\mathrm{Var}[PMI(p_{ci})] &\approx \frac{\mathrm{Var}[p_{ci}]}{\hat{p}_{ci}^2} + \frac{\mathrm{Var}[p_c]}{\hat{p}_c^2} + \frac{\mathrm{Var}[p_i]}{\hat{p}_i^2}\\ &= \frac{1}{\tilde{q}+1}\left(\frac{\hat{p}_{ci}(1-\hat{p}_{ci})}{\hat{p}_{ci}^2} + \frac{\hat{p}_c(1-\hat{p}_c)}{\hat{p}_c^2} + \frac{\hat{p}_i(1-\hat{p}_i)}{\hat{p}_i^2}\right)\\ &= \frac{1}{\tilde{q}+1}\left(\frac{1}{\hat{p}_{ci}} + \frac{1}{\hat{p}_c} + \frac{1}{\hat{p}_i} - 3\right).\end{aligned}$$

A.1.2 The posterior mean and variance of $KL(p_{c|i}|p_c)$

To examine the localization $KL(p_{c|i}|p_c)$, we compute the Taylor expansion around $\hat{p}_{c|i}$ and p_c. Note that these variables are independent. First consider the posterior of $p_{c|i}$, the conditional probability that a sample has location c given it has activity type i. Following the same Bayesian estimation procedure as for p_{ci}, it is readily seen that the vector of $p_{c|i}$'s for given i follows a Dirichlet distribution with as parameter the vector of $\tilde{q}_{ci}$'s for a given i. We then have

$$\hat{p}_{c|i} = \frac{\tilde{q}_{ci}}{\tilde{q}_i}$$

and the (co-)variance given by

$$\mathrm{Cov}[p_{c|i}, p_{c'|i}] = \frac{\delta_{cc'}\hat{p}_{c|i} - \hat{p}_{c|i}\hat{p}_{c'|i}}{(\tilde{q}_i + 1)}. \tag{A4}$$

Now computing the first and second derivatives of $KL(p_{c|i}|p_c)$ with respect to $p_{c|i}$ gives

$$\begin{aligned}\frac{\partial KL(p_{c|i}|p_c)}{\partial p_{c|i}} &= \frac{\partial}{\partial p_{c|i}}\left(\sum_{c'} p_{c'|i}\log\left(\frac{p_{c'|i}}{p_{c'}}\right)\right) = \log\left(\frac{p_{c|i}}{p_c}\right) + 1\\ &= PMI(p_{c|i}) + 1,\end{aligned}$$

where $PMI(p_{c|i}) = \log(p_{c|i}) - \log(p_i)$. The second derivative is given by

$$\frac{\partial^2 KL(p_{c|i}|p_c)}{\partial p_{c'|i}\, p_{c|i}} = \frac{\partial}{\partial p_{c|i}}\log\left(\frac{p_{c'|i}}{p_{c'}} + 1\right) = \frac{\delta_{cc'}}{p_{c|i}}.$$

The derivatives with respect to p_c are given by

$$\frac{\partial KL(p_{c|i}|p_c)}{\partial p_c} = \frac{\partial}{\partial p_c}\sum_{c'} p_{c'|i}\log\left(\frac{p_{c'|i}}{p_{c'}}\right) = -\frac{p_{c|i}}{p_c}$$

and

$$\frac{\partial^2 KL(p_{c|i}|p_c)}{\partial p_{c'}\, p_c} = -\frac{\partial}{\partial p_c}\frac{p_{c'|i}}{p_{c'}} = \frac{\delta_{cc'}\, p_{c|i}}{p_c^2}.$$

Writing $\Delta_{c|i} = p_{c|i} - \hat{p}_{c|i}$ and $\Delta_c = p_c - \hat{p}_c$, the Taylor expansion is given by

$$\begin{aligned}KL(p_{c|i}|p_c) = {} & KL(\hat{p}_{c|i}|\hat{p}_c) + \sum_c \Delta_{c|i} PMI(\hat{p}_{ci}) - \sum_c \Delta_c \frac{\hat{p}_{c|i}}{\hat{p}_c}\\ & + \frac{1}{2}\sum_{c,c'} \Delta_{c|i}\Delta_{c'|i}\frac{\delta_{cc'}}{\hat{p}_{c|i}} + \frac{1}{2}\sum_{c,c'} \Delta_c\Delta_{c'}\frac{\delta_{cc'}\,\hat{p}_{c|i}}{\hat{p}_c^2} + \mathcal{O}\left(\Delta_{c|i}^3\right)\\ = {} & KL(\hat{p}_{c|i}|\hat{p}_i) + \sum_c \Delta_{c|i} PMI(\hat{p}_{c|i}) - \sum_c \Delta_c \frac{\hat{p}_{c|i}}{\hat{p}_c} + \frac{1}{2}\sum_c \frac{\Delta_{c|i}^2}{\hat{p}_{c|i}}\\ & + \frac{1}{2}\sum_c \frac{\Delta_c^2 \hat{p}_{c|i}}{\hat{p}_c^2} + \mathcal{O}\left(\Delta_{c|i}^3\right).\end{aligned}$$

Taking expectations then leads to

$$\begin{aligned}\mathbb{E}[KL(p_{c|i}|p_c)] &\approx KL(\hat{p}_{c|i}|\hat{p}_c) + \frac{1}{2}\sum_c \frac{\mathrm{Var}[p_{c|i}]}{\hat{p}_{c|i}} + \frac{1}{2}\sum_c \mathrm{Var}[p_c]\frac{\hat{p}_{c|i}}{\hat{p}_c^2}\\ &= KL(\hat{p}_{c|i}|\hat{p}_c) + \frac{1}{2(\tilde{q}_i + 1)}\sum_c (1 - \hat{p}_{c|i}) + \frac{1}{2(\tilde{q} + 1)}\sum_c \left(\frac{\hat{p}_{c|i}}{\hat{p}_c} - \hat{p}_{c|i}\right)\\ &= KL(\hat{p}_{c|i}|\hat{p}_c) + \frac{N_c - 1}{2(\tilde{q}_i + 1)} + \frac{1}{2(\tilde{q} + 1)}\left(\sum_c \frac{\hat{p}_{c|i}}{\hat{p}_c} - 1\right)\end{aligned} \tag{A5}$$

The variance is given by

$$\begin{aligned}\mathrm{Var}[KL(p_{c|i}|p_c)] &= \mathbb{E}\left[\left(KL(p_{c|i}|p_c) - \mathbb{E}[KL(p_{c|i}|p_c)]\right)^2\right]\\ &= \mathbb{E}\left[\left(\sum_c \Delta_{c|i} PMI(\hat{p}_{c|i}) - \sum_c \Delta_c \frac{\hat{p}_{c|i}}{\hat{p}_c}\right)^2\right]\\ &= \mathbb{E}\left[\sum_{c,c'} \Delta_{c|i}\Delta_{c'|i} PMI(\hat{p}_{c|i}) PMI(\hat{p}_{c'|i}) + \sum_{c,c'} \Delta_c\Delta_{c'}\frac{\hat{p}_{c|i}\,\hat{p}_{c'|i}}{\hat{p}_c\,\hat{p}_c^{'}}\right]\\ &= \sum_{c,c'} \mathrm{Cov}[p_{c|i}, p_{c'|i}] PMI(\hat{p}_{c|i}) PMI(\hat{p}_{c'|i}) + \sum_{c,c'} \mathrm{Cov}[p_c, p_{c'}]\frac{\hat{p}_{c|i}\,\hat{p}_{c'|i}}{\hat{p}_c\,\hat{p}_c^{'}}\\ &= \frac{1}{\tilde{q}_i + 1}\left(\sum_c \hat{p}_{c|i} PMI(\hat{p}_{ci})^2 - KL(\hat{p}_{c|i}|\hat{p}_c)^2\right) + \frac{1}{\tilde{q} + 1}\left(\sum_c \frac{\hat{p}_{c|i}^2}{\hat{p}_c} - 1\right),\end{aligned} \tag{A6}$$

where we used that $\mathbb{E}[\Delta_{c|i}\Delta_c] = 0$. The results for $KL(p_{i|c}|p_i)$ are obtained by symmetry in the indices c and i.

A.1.3 The posterior mean and variance of $MI(p_{ci})$

For estimating the mutual information we follow the same strategy as for the pointwise mutual information, following the derivation given in Hutter and Zaffalon (2005). Note that we have

$$\begin{aligned}\frac{\partial MI(p_{ci})}{\partial p_{ci}} &= \frac{\partial}{\partial p_{ci}}\left(\sum_{c'j} p_{c'j}\log(p_{c'j}) - \sum_{c'} p_{c'}\log(p_{c'}) - \sum_{j} p_j \log(p_j)\right)\\ &= \log(p_{ci}) + 1 - \log(p_c) - 1 - \log(p_i) - 1\\ &= \log\left(\frac{p_{ci}}{p_c\, p_i}\right) - 1\end{aligned}$$

and

$$\frac{\partial^2 MI(p_{ci})}{\partial p_{c'j}\partial p_{ci}} = \frac{\delta_{cc'}\delta_{ij}}{p_{ci}} - \frac{\delta_{cc'}}{p_c} - \frac{\delta_{ij}}{p_i},$$

where $\delta_{cc'}$ is the Kronecker delta.

Again writing $\Delta_{ci} = \hat{p}_{ci} - p_{ci}$, the Taylor expansion around $\hat{p}_{ci}$ is given by

$$MI(p_{ci}) = MI(\hat{p}_{ci}) + \sum_{ci}\Delta_{ci}PMI(\hat{p}_{ci}) + \frac{1}{2}\sum_{cic'j}\Delta_{ci}\Delta_{c'j}\left(\frac{\delta_{cc'}\delta_{ij}}{\hat{p}_{ci}} - \frac{\delta_{cc'}}{\hat{p}_c} - \frac{\delta_{ij}}{\hat{p}_i}\right) + \mathcal{O}\left(\Delta_{ci}^3\right).$$

Now noting that $\mathbb{E}[\Delta_{ci}] = 0$ and $\mathbb{E}[\Delta_{ci}\Delta_{c'j}] = \mathrm{Cov}[p_{ci}, p_{c'j}]$, taking expectations gives

$$\begin{aligned}\mathbb{E}[MI(p_{ci})] &\approx MI(\hat{p}_{ci}) + \frac{1}{2}\sum_{cic'j}\mathrm{Cov}[p_{ci}, p_{c'j}]\left(\frac{\delta_{cc'}\delta_{ij}}{\hat{p}_{ci}} - \frac{\delta_{cc'}}{\hat{p}_c} - \frac{\delta_{ij}}{\hat{p}_j}\right)\\ &= MI(\hat{p}_{ci}) + \frac{(N_c - 1)(N_i - 1)}{2(\tilde{q} + 1)}.\end{aligned}$$

The variance can be obtained by computing

$$\begin{aligned}\mathrm{Var}[MI(p_{ci})] &= \mathbb{E}[(MI(p_{ci}) - \mathbb{E}[MI(p_{ci})])^2]\\ &\approx \mathbb{E}\left[\left(\sum_{ci}\Delta_{ci}PMI(\hat{p}_{ci})\right)^2\right]\\ &= \mathbb{E}\left[\sum_{cic'j}\Delta_{ci}\Delta_{c'j}PMI(\hat{p}_{ci})PMI(\hat{p}_{c'i})\right]\\ &= \sum_{cic'j}\mathrm{Cov}[p_{ci}, p_{c'j}]PMI(\hat{p}_{ci})PMI(\hat{p}_{c'i})\\ &= \frac{1}{\tilde{q} + 1}\left(\sum_{ci}\frac{\tilde{q}_{ci}}{\tilde{q}}\log\left(\frac{\tilde{q}\tilde{q}_{ci}}{\tilde{q}_c\tilde{q}_i}\right)^2 - \left(\sum_{ci}\frac{\tilde{q}_{ci}}{\tilde{q}}\log\left(\frac{\tilde{q}\tilde{q}_{ci}}{\tilde{q}_c\tilde{q}_i}\right)\right)^2\right)\\ &= \frac{1}{\tilde{q} + 1}\left(\sum_{ci}\hat{p}_{ci}PMI(\hat{p}_{ci})^2 - MI(\hat{p}_{ci})^2\right).\end{aligned}$$

A.2. Colocation estimates

Estimation of quantities involving the colocation probabilities p_{ij} is more involved than those based on the location probabilities p_{ci}, since the p_{ij} are not Dirichlet distributed, so we do not have expressions for their mean and variance. In the following, we first show how to obtain $\mathbb{E}[p_{ij}]$ and $\mathrm{Var}[p_{ij}]$ by expressing them in terms of Dirichlet distributed variables. We then use these quantities to obtain estimates for information-theoretic measures of colocation. An overview of the results is given in Table A2.

TABLE A2 Overview of approximations of the posterior mean and variance for colocation measures

RV	Posterior mean	Posterior variance
p_{ij}	$\hat{p}_{ij} = \frac{1}{\tilde{q}}\sum_c \frac{\tilde{q}_{ci}(\tilde{q}_{cj} + \delta_{ij})}{\tilde{q}_c + 1}$	$\frac{1}{\tilde{q}(\tilde{q}+1)}\left(\sum_c \frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)(\tilde{q}_{cj}+2\delta_{ij})(\tilde{q}_{cj}+1+2\delta_{ij})}{(\tilde{q}_c+2)(\tilde{q}_c+3)} - \sum_c \frac{\tilde{q}_{ci}^2(\tilde{q}_{cj}+\delta_{ij})^2}{(\tilde{q}_c+1)^2} - \tilde{q}\hat{p}_{ij}^2\right)$
$p_{j\mid i}$	$\hat{p}_{j\mid i} = \frac{1}{q_i}\sum_c \frac{q_{ci}q_{cj}}{q_c}$	$\frac{1}{\tilde{q}_i(\tilde{q}_i+1)}\left(\sum_c \frac{\tilde{q}_{cj}(\tilde{q}_{cj}+1)\tilde{q}_{ci}(\tilde{q}_{ci}+1)}{\tilde{q}_c(\tilde{q}_c+1)} - \sum_c \frac{\tilde{q}_{cj}^2\tilde{q}_{ci}^2}{\tilde{q}_c^2} - \tilde{q}_i\hat{p}_{j\mid i}^2\right)$
$PMI(p_{ij})$	$PMI(\hat{p}_{ij}) + \frac{\mathrm{Var}[p_{ij}]}{2}\left(-\frac{1}{\hat{p}_{ij}^2} + \frac{1}{\hat{p}_i^2} + \frac{1}{\hat{p}_j^2}\right)$	$\mathrm{Var}[p_{ij}]\left(\frac{1}{\hat{p}_{ij}} - \frac{1}{\hat{p}_i} - \frac{1}{\hat{p}_j}\right)^2$
$KL(p_{j\mid i})$	$KL(\hat{p}_{j\mid i}\mid\hat{p}_j) + \frac{1}{2}\sum_j \frac{\mathrm{Var}[p_{j\mid i}]}{\hat{p}_{j\mid i}} + \frac{1}{2(q+1)}\left(\sum_j \frac{\hat{p}_{j\mid i}}{\hat{p}_j} - 1\right)$	$\frac{1}{\tilde{q}_i+1}\left(\sum_c \hat{p}_{c\mid i}PMI(\hat{p}_{ci})^2 - KL(\hat{p}_{c\mid i}\mid\hat{p}_c)^2\right) + \frac{1}{q+1}\left(\sum_c \frac{\hat{p}_{c\mid i}^2}{\hat{p}_c} - 1\right)$
$MI(p_{ij})$	$MI(\hat{p}_{ij}) + \frac{1}{2}\sum_{ij} \frac{\mathrm{Var}[p_{ij}]}{\hat{p}_{ij}} - \frac{N_i - 1}{\tilde{q}+1}$	$\sum_{ijkl}\mathrm{Cov}[p_{ij}, p_{kl}]PMI(\hat{p}_{ij})PMI(\hat{p}_{kl})$

To obtain an explicit expression for $\mathbb{E}[p_{ij}]$, we rewrite it as

$$\begin{aligned}
\hat{p}_{ij} = \mathbb{E}[p_{ij}] &= \mathbb{E}\left[\sum_c p_{j\mid c}\, p_{ci}\right] \\
&= \sum_c \mathbb{E}[p_{j\mid c}\, p_{ci}] \\
&= \sum_c \mathbb{E}[p_{j\mid c}]E[p_{ci}] \\
&= \frac{1}{\tilde{q}}\sum_c \frac{\tilde{q}_{ci}\tilde{q}_{cj}}{\tilde{q}_c}.
\end{aligned}$$

Here we used the fact that $p_{j\mid c}$ and p_{ci} are independent and Dirichlet distributed.
We compute $\mathrm{Var}[p_{ij}]$ as

$$\begin{aligned}
\mathrm{Var}[p_{ij}] &= \mathbb{E}\left[p_{ij}^2\right] - \mathbb{E}[p_{ij}]^2 \\
&= \mathbb{E}\left[\sum_c p_{j\mid c}\, p_{ci} \sum_{c'} p_{j\mid c'}\, p_{c'i}\right] - \hat{p}_{ij}^2 \\
&= \sum_{c,c'} \mathbb{E}[p_{j\mid c}\, p_{j\mid c'}]\mathbb{E}[p_{ci}\, p_{c'i}] - \hat{p}_{ij}^2 \\
&= \frac{1}{\tilde{q}(\tilde{q}+1)}\left(\sum_{c\neq c'} \frac{\tilde{q}_{ci}\tilde{q}_{cj}\tilde{q}_{c'i}\tilde{q}_{c'j}}{\tilde{q}_c^2} + \sum_c \frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)\tilde{q}_{cj}(\tilde{q}_{cj}+1)}{\tilde{q}_c(\tilde{q}_c+1)}\right) - \hat{p}_{ij}^2 \\
&= \frac{1}{\tilde{q}(\tilde{q}+1)}\left(\sum_{c,c'} \frac{\tilde{q}_{ci}\tilde{q}_{cj}\tilde{q}_{c'i}\tilde{q}_{c'j}}{\tilde{q}_c^2} - \sum_c \frac{\tilde{q}_{ci}^2\tilde{q}_{cj}^2}{\tilde{q}_c^2} + \sum_c \frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)\tilde{q}_{cj}(\tilde{q}_{cj}+1)}{\tilde{q}_c(\tilde{q}_c+1)}\right) - \hat{p}_{ij}^2 \\
&= \frac{1}{\tilde{q}(\tilde{q}+1)}\left(\tilde{q}^2\hat{p}_{ij}^2 - \sum_c \frac{\tilde{q}_{ci}^2\tilde{q}_{cj}^2}{\tilde{q}_c^2} + \sum_c \frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)\tilde{q}_{cj}(\tilde{q}_{cj}+1)}{\tilde{q}_c(\tilde{q}_c+1)}\right) - \hat{p}_{ij}^2 \\
&= \frac{1}{\tilde{q}(\tilde{q}+1)}\left(\sum_c \frac{\tilde{q}_{ci}(\tilde{q}_{ci}+1)\tilde{q}_{cj}(\tilde{q}_{cj}+1)}{\tilde{q}_c(\tilde{q}_c+1)} - \sum_c \frac{\tilde{q}_{ci}^2\tilde{q}_{cj}^2}{\tilde{q}_c^2} - \tilde{q}\hat{p}_{ij}^2\right).
\end{aligned} \tag{A7}$$

Likewise, the covariance $\mathrm{Cov}[p_{ij}, p_{kl}]$ can be computed as

$$
\begin{aligned}
\mathrm{Cov}[p_{ij}, p_{kl}] &= \mathbb{E}[p_{ij}\,p_{kl}] - \hat{p}_{ij}\,\hat{p}_{kl} \\
&= \sum_{c,c'} \mathbb{E}[p_{j|c}\,p_{l|c'}]\mathbb{E}[p_{ci}\,p_{c'k}] - \hat{p}_{ij}\,\hat{p}_{kl} \\
&= \frac{1}{\tilde{q}(\tilde{q}+1)}\left(\sum_{c,c'} \frac{\tilde{q}_{ci}\tilde{q}_{cj}\tilde{q}_{c'k}\tilde{q}_{c'l}}{\tilde{q}_c^2} - \sum_c \frac{\tilde{q}_{ci}\tilde{q}_{cj}\tilde{q}_{ck}\tilde{q}_{cl}}{\tilde{q}_c^2}\right. \\
&\quad \left. + \sum_c \frac{\tilde{q}_{ci}\tilde{q}_{cj}(\tilde{q}_{ck}+\delta_{ik})(\tilde{q}_{cl}+\delta_{jl})}{\tilde{q}_c(\tilde{q}_c+1)}\right) - \hat{p}_{ij}\,\hat{p}_{kl} \\
&= \frac{1}{\tilde{q}(\tilde{q}+1)}\left(\sum_c \frac{\tilde{q}_{ci}\tilde{q}_{cj}(\tilde{q}_{ck}+\delta_{ik})(\tilde{q}_{cl}+\delta_{jl})}{\tilde{q}_c(\tilde{q}_c+1)} - \sum_c \frac{\tilde{q}_{ci}\tilde{q}_{cj}\tilde{q}_{ck}\tilde{q}_{cl}}{\tilde{q}_c^2} - \tilde{q}\hat{p}_{ij}\,\hat{p}_{kl}\right).
\end{aligned}
$$

A.2.1 The posterior mean and variance of $p_{j|i}$

For the conditional probabilities, we have

$$
\hat{p}_{j|i} = \mathbb{E}\left[\sum_c p_{j|c}\,p_{c|i}\right] = \sum_c \mathbb{E}[p_{j|c}\,p_{c|i}] = \frac{1}{\tilde{q}_i}\sum_c \frac{\tilde{q}_{ci}\tilde{q}_{cj}}{\tilde{q}_c}
$$

The variance is given by

$$
\begin{aligned}
\mathrm{Var}[p_{j|i}] &= \mathbb{E}\left[p_{j|i}^2\right] - \hat{p}_{j|i}^2 = \mathbb{E}\left[\left(\sum_c p_{j|c}\,p_{c|i}\right)^2\right] - \hat{p}_{j|i}^2 \\
&= \mathbb{E}\left[\sum_{c,c'} p_{j|c}\,p_{j|c'}\,p_{c|i}\,p_{c'|i}\right] - \hat{p}_{j|i}^2 = \sum_{c,c'} \mathbb{E}[p_{j|c}\,p_{j|c'}]\mathbb{E}[p_{c|i}\,p_{c'|i}] - \hat{p}_{j|i}^2 \\
&= \sum_{c,c'} \frac{\tilde{q}_{cj}(\tilde{q}_{c'j}+\delta_{cc'})\tilde{q}_{ci}(\tilde{q}_{c'i}+\delta_{cc'})}{\tilde{q}_c(\tilde{q}_{c'}+\delta_{cc'})\tilde{q}_i(\tilde{q}_i+1)} - \hat{p}_{j|i}^2 \\
&= \frac{1}{\tilde{q}_i(\tilde{q}_i+1)}\left(\sum_{c,c'} \frac{\tilde{q}_{ci}\tilde{q}_{c'i}\tilde{q}_{cj}\tilde{q}_{c'j}}{\tilde{q}_c\tilde{q}_{c'}} - \sum_c \frac{\tilde{q}_{cj}^2\tilde{q}_{ci}^2}{\tilde{q}_c^2}\right. \\
&\quad \left. + \sum_c \frac{\tilde{q}_{cj}(\tilde{q}_{cj}+1)\tilde{q}_{ci}(\tilde{q}_{ci}+1)}{\tilde{q}_c(\tilde{q}_c+1)}\right) - \hat{p}_{j|i}^2 \\
&= \frac{1}{\tilde{q}_i(\tilde{q}_i+1)}\left(\tilde{q}_i^2\hat{p}_{ji}^2 - \sum_c \frac{\tilde{q}_{cj}^2\tilde{q}_{ci}^2}{\tilde{q}_c^2}\right. \\
&\quad \left. + \sum_c \frac{\tilde{q}_{cj}(\tilde{q}_{cj}+1)\tilde{q}_{ci}(\tilde{q}_{ci}+1)}{\tilde{q}_c(\tilde{q}_c+1)}\right) - \hat{p}_{j|i}^2 \\
&= \frac{1}{\tilde{q}_i(\tilde{q}_i+1)}\left(\sum_c \frac{\tilde{q}_{cj}(\tilde{q}_{cj}+1)\tilde{q}_{ci}(\tilde{q}_{ci}+1)}{\tilde{q}_c(\tilde{q}_c+1)} - \sum_c \frac{\tilde{q}_{cj}^2\tilde{q}_{ci}^2}{\tilde{q}_c^2} - \tilde{q}_i\,\hat{p}_{j|i}^2\right),
\end{aligned}
$$

and likewise

$$
\begin{aligned}
\mathrm{Cov}[p_{j|i}\,p_{k|i}] &= \mathbb{E}[p_{j|i}\,p_{k|i}] - \hat{p}_{j|i}\,\hat{p}_{k|i} \\
&= \sum_{c,c'} \mathbb{E}[p_{j|c}\,p_{k|c'}\,p_{c|i}\,p_{c'|i}] - \hat{p}_{j|i}\,\hat{p}_{k|i} \\
&= \sum_{c,c'} \mathbb{E}[p_{j|c}\,p_{k|c'}]\mathbb{E}[p_{c|i}\,p_{c'|i}] - \hat{p}_{j|i}\,\hat{p}_{k|i} \\
&= \sum_{c,c'} \frac{\tilde{q}_{cj}(\tilde{q}_{c'k}+\delta_{jk}\delta_{cc'})\tilde{q}_{ci}(\tilde{q}_{c'i}+\delta_{cc'})}{\tilde{q}_c(\tilde{q}_{c'}+\delta_{cc'})\tilde{q}_i(\tilde{q}_i+1)} - \hat{p}_{j|i}\,\hat{p}_{k|i} \\
&= \frac{1}{\tilde{q}_i(\tilde{q}_i+1)}\left(\sum_{c,c'} \frac{\tilde{q}_{cj}\tilde{q}_{c'k}\tilde{q}_{ci}\tilde{q}_{c'i}}{\tilde{q}_c\tilde{q}_{c'}} - \sum_c \frac{\tilde{q}_{cj}\tilde{q}_{ck}\tilde{q}_{ci}^2}{\tilde{q}_c^2}\right. \\
&\quad \left. + \sum_c \frac{\tilde{q}_{cj}(\tilde{q}_{ck}+\delta_{jk})\tilde{q}_{ci}(\tilde{q}_{ci}+1)}{\tilde{q}_c(\tilde{q}_c+1)}\right) - \hat{p}_{j|i}\,\hat{p}_{k|i} \\
&= \frac{1}{\tilde{q}_i(\tilde{q}_i+1)}\left(\sum_c \frac{\tilde{q}_{cj}(\tilde{q}_{ck}+\delta_{jk})\tilde{q}_{ci}(\tilde{q}_{ci}+1)}{\tilde{q}_c(\tilde{q}_c+1)} - \sum_c \frac{\tilde{q}_{cj}\tilde{q}_{ck}\tilde{q}_{ci}^2}{\tilde{q}_c^2} - \tilde{q}_i\,\hat{p}_{j|i}\,\hat{p}_{k|i}\right).
\end{aligned}
$$

A.2.2 The posterior mean and variance of $PMI(p_{ij})$

The estimation of $PMI(p_{ij})$ is analogous to that of $PMI(p_{ci})$, leading to

$$\begin{aligned}\mathbb{E}[PMI(p_{ij})] &\approx PMI(\hat{p}_{ij}) + \frac{\mathrm{Cov}[p_{kl}, p_{k'l'}]}{2}\left(-\frac{\delta_{ikk'}\delta_{jll'}}{\hat{p}_{ij}^2} + \frac{\delta_{ikk'}}{\hat{p}_i^2} + \frac{\delta_{jll'}}{\hat{p}_j^2}\right)\\ &= PMI(\hat{p}_{ij}) + \frac{1}{2}\left(-\frac{\mathrm{Var}[p_{ij}]}{\hat{p}_{ij}^2} + \sum_{ll'}\frac{\mathrm{Cov}[p_{il}, p_{il'}]}{\hat{p}_i^2} + \sum_{kk'}\frac{\mathrm{Cov}[p_{kj}, p_{k'j}]}{\hat{p}_j^2}\right)\\ &= PMI(\hat{p}_{ij}) - \frac{\mathrm{Var}[p_{ij}]}{2\hat{p}_{ij}^2} + \frac{1}{2(\tilde{q}+1)}\left(\frac{1}{\hat{p}_i} + \frac{1}{\hat{p}_j}\right).\end{aligned} \tag{A8}$$

The variance is derived in a similar manner as in (A3), leading to

$$\begin{aligned}\mathrm{Var}[PMI(p_{ci})] &= \mathbb{E}[(PMI(p_{ij}) - \mathbb{E}[PMI(p_{ij})])^2]\\ &= \mathbb{E}\left[\left(\frac{\Delta_{ij}}{\hat{p}_{ij}} - \frac{\Delta_i}{\hat{p}_i} - \frac{\Delta_j}{\hat{p}_j}\right)^2\right].\end{aligned}$$

Again, the cross terms $\mathbb{E}[\Delta_{ij}\Delta_i]$ and $\mathbb{E}[\Delta_{ij}\Delta_j]$ are equal to zero but we now have that $\mathbb{E}[\Delta_i\Delta_i] = \mathrm{Cov}[p_i, p_j]$. This leads to

$$\begin{aligned}\mathrm{Var}[PMI(p_{ij})] &\approx \frac{\mathrm{Var}[p_{ij}]}{\hat{p}_{ij}^2} + \frac{\mathrm{Var}[p_i]}{\hat{p}_i^2} + \frac{\mathrm{Var}[p_j]}{\hat{p}_j^2} + \frac{2\mathrm{Cov}[p_i p_j]}{\hat{p}_i\hat{p}_j}\\ &= \frac{\mathrm{Var}[p_{ij}]}{\hat{p}_{ij}^2} + \frac{1}{\tilde{q}+1}\left(\frac{\hat{p}_i(1-\hat{p}_i)}{\hat{p}_i^2} + \frac{\hat{p}_j(1-\hat{p}_j)}{\hat{p}_j^2} + \frac{2\delta_{ij}\hat{p}_i - \hat{p}_i\hat{p}_j}{\hat{p}_i\hat{p}_j}\right)\\ &= \frac{\mathrm{Var}[p_{ij}]}{\hat{p}_{ij}^2} + \frac{1}{\tilde{q}+1}\left(\frac{1+2\delta_{ij}}{\hat{p}_i} + \frac{1}{\hat{p}_j} - 4\right).\end{aligned} \tag{A9}$$

A.2.3 The posterior mean and variance of $KL(p_{j|i}|p_j)$

For the colocation case, following (A5) gives

$$\begin{aligned}\mathbb{E}[KL(p_{j|i}|p_j)] &\approx KL(\hat{p}_{j|i}|\hat{p}_j) + \frac{1}{2}\sum_j\frac{\mathrm{Var}[p_{j|i}]}{\hat{p}_{j|i}} + \frac{1}{2}\sum_j\mathrm{Var}[p_j]\frac{\hat{p}_{j|i}}{\hat{p}_j^2}\\ &= KL(\hat{p}_{j|i}|\hat{p}_j) + \frac{1}{2}\sum_j\frac{\mathrm{Var}[p_{j|i}]}{\hat{p}_{j|i}} + \frac{1}{2(\tilde{q}+1)}\left(\sum_j\frac{\hat{p}_{j|i}}{\hat{p}_j} - 1\right).\end{aligned}$$

For the variance, we follow (A6) and obtain

$$\begin{aligned}\mathrm{Var}[KL(p_{j|i}|p_j)] &= \sum_{j,k}\mathrm{Cov}[p_{j|i}, p_{k|i}]PMI(\hat{p}_{j|i})PMI(\hat{p}_{k|i}) + \sum_{j,k}\mathrm{Cov}[p_j, p_k]\frac{\hat{p}_{j|i}\hat{p}_{k|i}}{\hat{p}_j\hat{p}_k}\\ &= \frac{1}{\tilde{q}_i+1}\left(\sum_c\hat{p}_{c|i}PMI(\hat{p}_{ci})^2 - KL(\hat{p}_{c|i}|\hat{p}_c)^2\right) + \frac{1}{\tilde{q}+1}\left(\sum_c\frac{\hat{p}_{c|i}^2}{\hat{p}_c} - 1\right),\end{aligned}$$

where $PMI(\hat{p}_{j|i}) = \log\left(\frac{\hat{p}_{j|i}}{\hat{p}_j}\right)$.

A.2.4 The posterior mean and variance of $MI(A(w_1), A(w_2))$

For the case of colocation, we get

$$\mathbb{E}[MI(p_{ij}] \approx MI(\hat{p}_{ij}) + \frac{1}{2}\sum_{ijkl} \mathrm{Cov}[p_{ij}, p_{kl}]\left(\frac{\delta_{ik}\delta_{jl}}{\hat{p}_{ij}} - \frac{\delta_{ik}}{\hat{p}_i} - \frac{\delta_{jl}}{\hat{p}_j}\right)$$

$$= MI(\hat{p}_{ij}) + \frac{1}{2}\sum_{ij} \frac{\mathrm{Var}[p_{ij}]}{\hat{p}_{ij}} - \frac{N_i - 1}{\tilde{q} + 1}$$

and

$$\mathrm{Var}[MI(p_{ij})] \approx \sum_{ijkl} \mathrm{Cov}[p_{ij}, p_{kl}] PMI(\hat{p}_{ij}) PMI(\hat{p}_{kl}).$$

APPENDIX B: NUMERICAL RESULTS

B.1. Python class

We make available the Python code enabling computation of the proposed (co-)location measures at https://github.com/aljevandam/Colocation. The posterior expectation and variance of all proposed information-theoretic quantities are available, with exception of the standard deviation of $MI(A(w_1), A(w_2))$ as this becomes computationally costly for large data sets. We leave the efficient computation of $MI(A(w_1), A(w_2))$ as a topic for future research.

B.2. Numerical simulations of posterior estimates

Here we compare the analytical approximations to simulated posterior distributions. Figures B1 and B2 compare the analytical approximations for the posterior mean and variance to the mean and variance of a numerically sampled posterior distribution, for a generated data set containing multinomially distributed counts. Figures B3 and B4 show the same comparison for the BLS major occupations groups described in the main text. The code that generated these figures is available at https://github.com/aljevandam/Colocation.

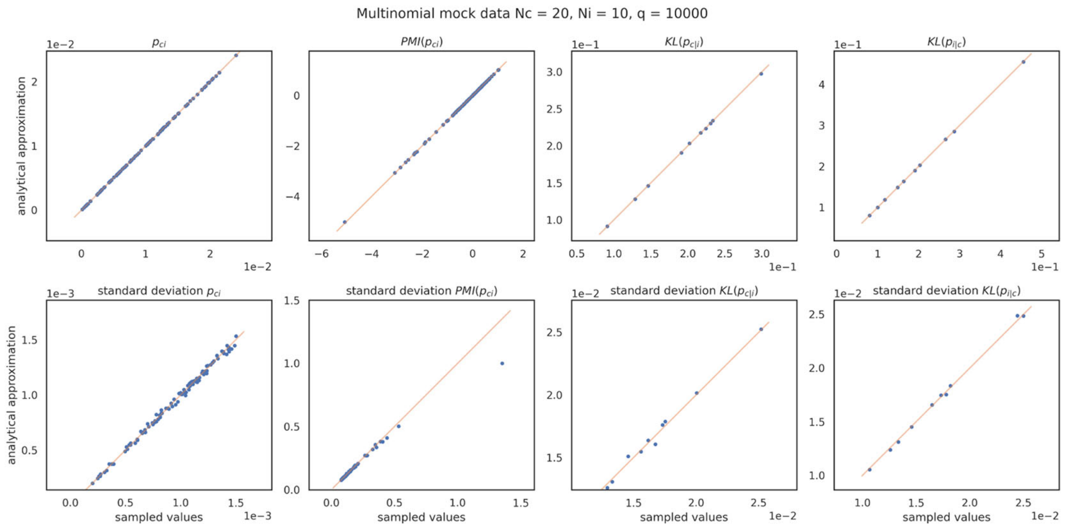

FIGURE B1 Comparison of analytical approximations with numerical simulations for multinomial mock data. Location measures. Zero prior.

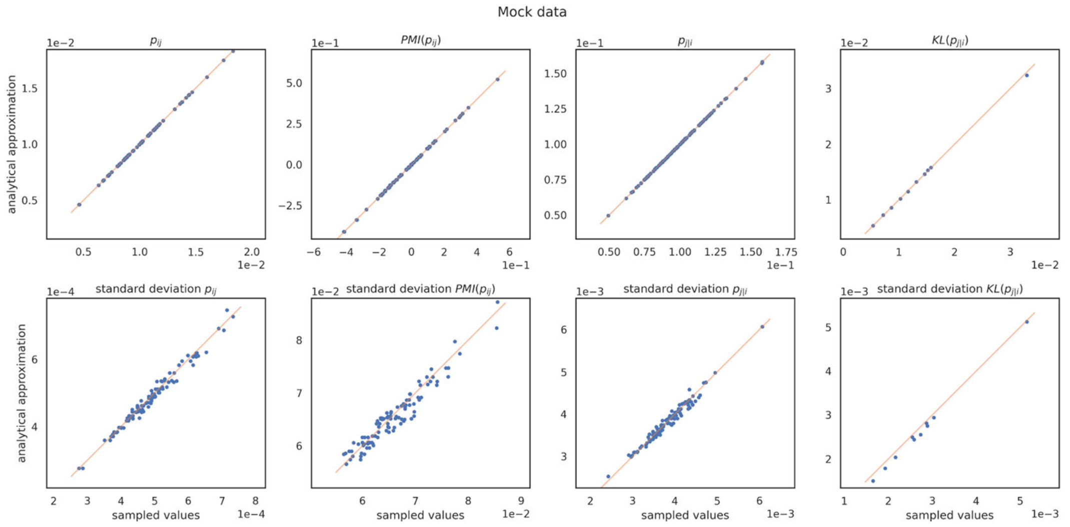

FIGURE B2 Comparison of analytical approximations with numerical simulations for multinomial mock data. Colocation measures. Zero prior.

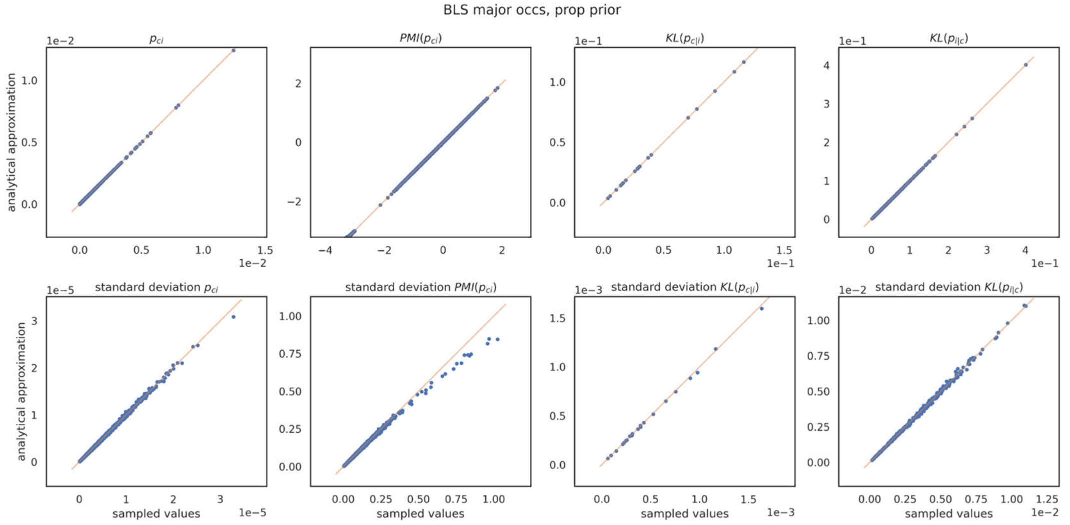

FIGURE B3 Comparison of analytical approximations with numerical simulations for BLS major occupation groups. Location measures. Prior strength 0.05.

B.3. Significance test

As mentioned in the main text, we evaluate the significance of associations using the "probability of direction", which is defined as the proportion of the posterior distribution that is of the median's sign, and can be interpreted as the probability that an estimated parameter is strictly positive or negative (Makowski et al., 2019).

To compute the probability of direction, first consider the shape of the posterior distribution. From our approximations we have the estimates for the posterior $\mathbb{E}[PMI(p_{ij})] = \mu$ and $\mathrm{Var}[PMI(p_{ij})] = \sigma^2$. Under the Bernstein-von Mises theorem we expect the posterior of $PMI(p_{ij})$ to be approximately normally distributed with

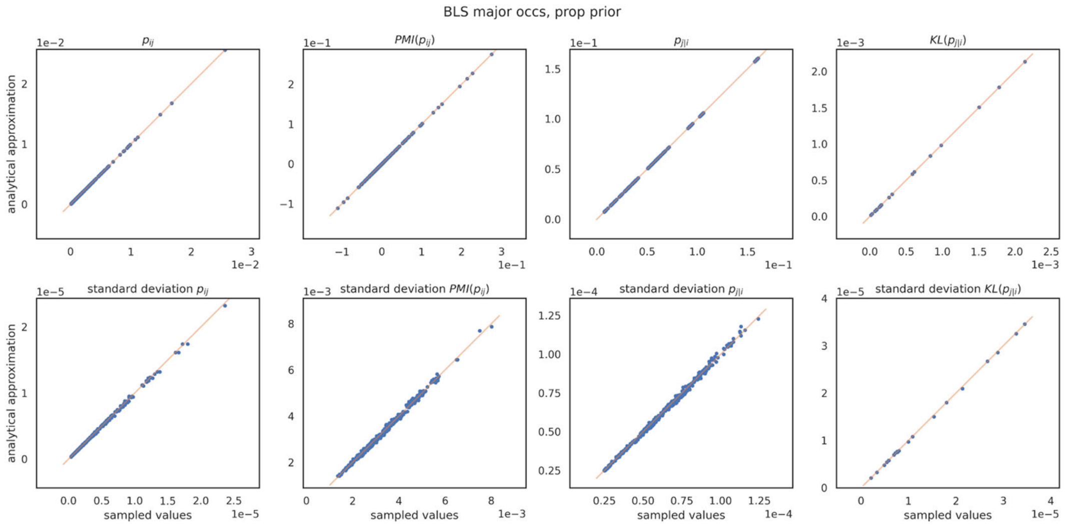

FIGURE B4 Comparison of analytical approximations with numerical simulations for BLS major occupation groups. Colocation measures. Prior strength 0.05

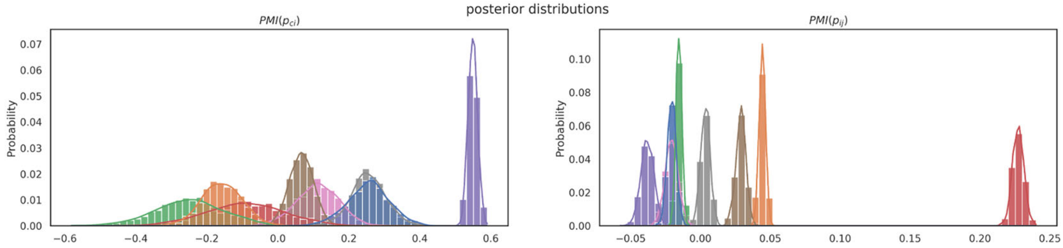

FIGURE B5 Eight randomly selected simulated posterior distribution of location associations (left panel) and colocation association (right panel).

mean μ and variance σ^2. We confirm this normality numerically in Figure B5, which shows a random selection of the sampled posterior distributions of (co-)location associations. All posteriors show a bell-curved shape.

The probability of direction gives the probability that $PMI(p_{ij})$ is of a different sign than μ. We then state that $PMI(p_{ij})$ is significantly nonzero if for a given threshold ϵ we have that $P(PMI(p_{ij}) < 0) < \epsilon$ when $\mu > 0$ and $P(PMI(p_{ij}) > 0) < \epsilon$ when $\mu < 0$. This implies that for a posterior distribution with mean close to zero and a large variance, the estimate μ will turn out to be insignificant, whereas the μ will be significant when the variance is small relative to the absolute value of the mean.

APPENDIX C: PRIOR DISTRIBUTION

We analyze the effect of the prior of the resulting estimates empirically using the data on major occupation groups. We compare three cases:

- using a "zero prior", that is, setting the pesudocounts $\alpha_{ci} = 0$ everywhere. This leads to the maximum-likelihood estimate $\hat{p}_{ci} = \frac{q_{ci}}{q}$.
- using a uniform prior, setting $\alpha_{ci} = \frac{\alpha}{N_c N_i}$.

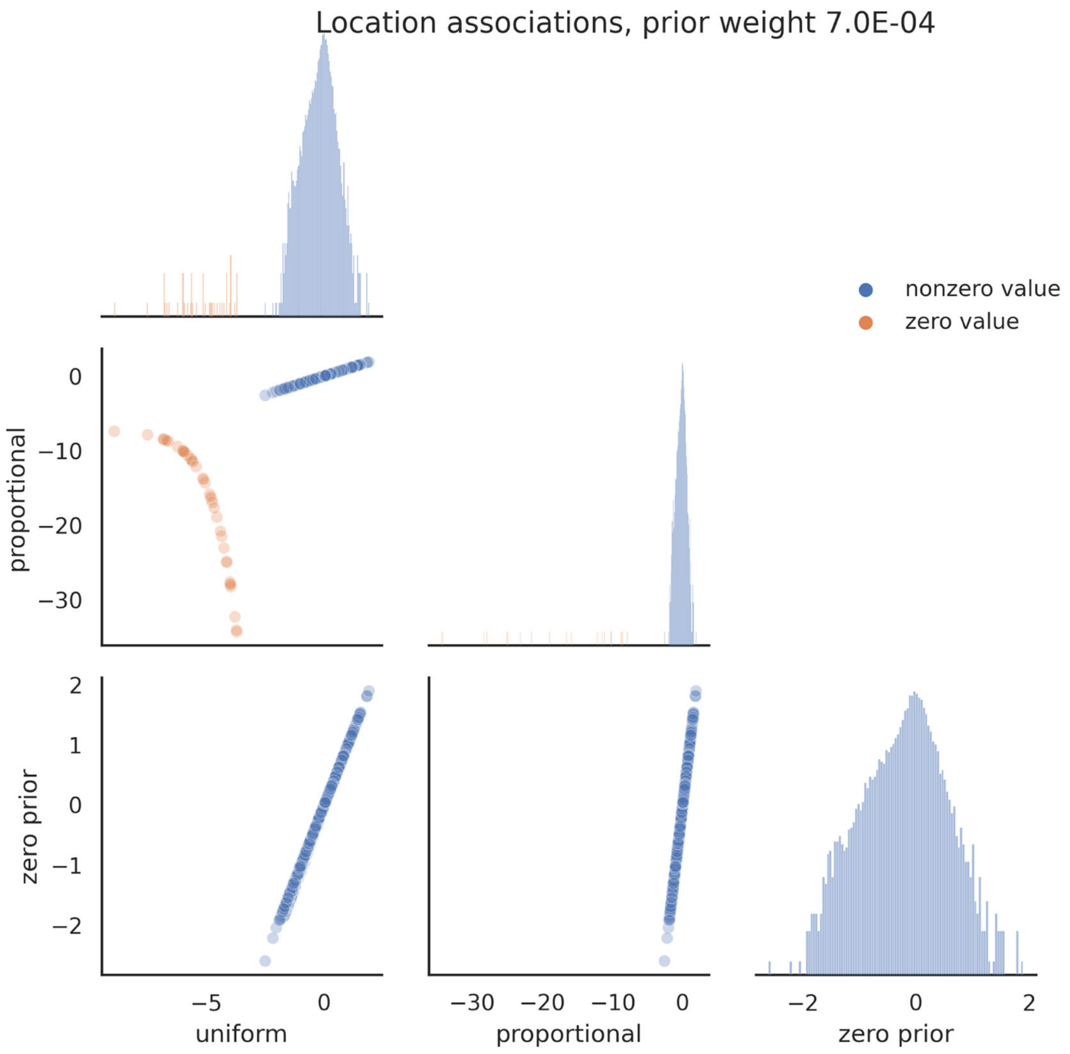

FIGURE C1 Location associations of major occupations groups for different priors. The vertical axes of the histograms have a log scale. Colors indicate which cells have zero observations in the data. The prior weight is 0.0007.

- using the “proportional” prior, setting $\alpha_{ci} = \alpha\frac{q_c q_i}{q^2}$.

Here, α determines the prior strength, and equals the total number of pesudocounts in the prior. Setting $\alpha = N_c N_i$ amounts to a single observation for each cell (which makes the uniform prior uninformative in the sense that is gives equal probability to any probability distribution). For the data set we consider here, $\alpha = N_c N_i$ amounts to a prior strength of 0.0007, meaning it represents 0.07% of the total amount of counts $q + \alpha$ (Figure C1).

To get an idea of how the prior strength affects results, we consider three prior strengths:

- $\alpha \approx N_c N_i \approx 0.0007$, so that the uniform prior is uninformative.
- $\alpha = 0.05$, which is the value we used in the empirical analysis and makes sure that the lowest location associations are of the same order of magnitude as other associations.

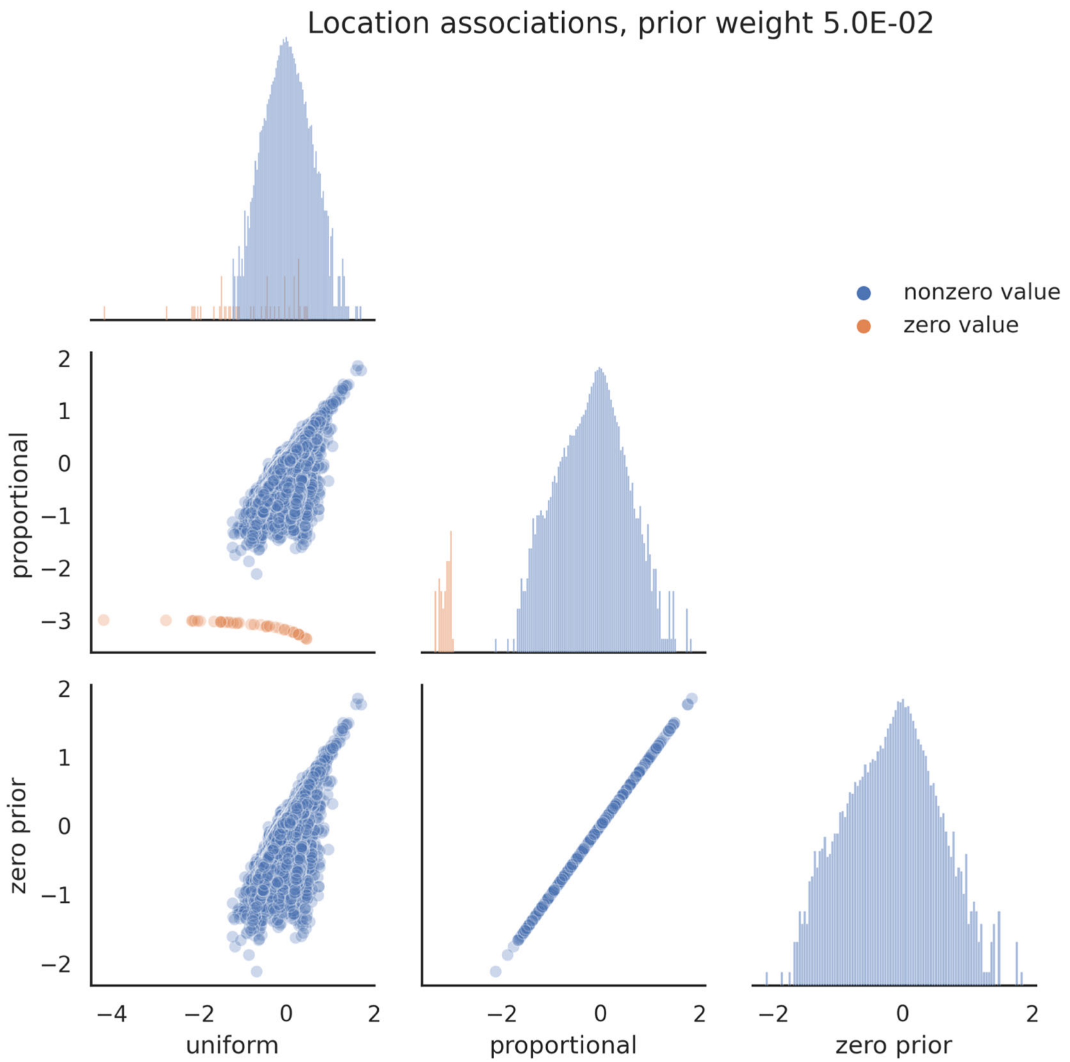

FIGURE C2 Location associations of major occupations groups for different priors. The vertical axes of the histograms have a log scale. Colors indicate which cells have zero observations in the data. The prior weight is 0.05.

- $\alpha = 0.5$, so that the prior information and the data weigh equally.

Figures C2, C3, and C4 show the results for location association for different priors and prior weights. Cells for which $q_{ci} = 0$, in which the counts are entirely determined by the prior, are colored blue in each figure. For the case of the zero prior, values of log(0) are omitted.

The scatter plots show the influence of the prior on the estimated associations. For low prior strengths, the difference with the zero prior is negligible as the prior gets washed out by the data. The cells without observations in the data however can be seen to differ for the uniform and proportional prior. As the prior strength increases, the estimates for the uniform prior becomes increasingly different from the maximum likelihood estimate. The proportional prior can be seen to keep the estimates in cells with observations mostly in place.

The diagonals show histograms of the associations. The vertical axis has a log scale, to better show the bins with few counts. For small prior strengths, the zero-valued cells can be seen to have strongly negative associations.

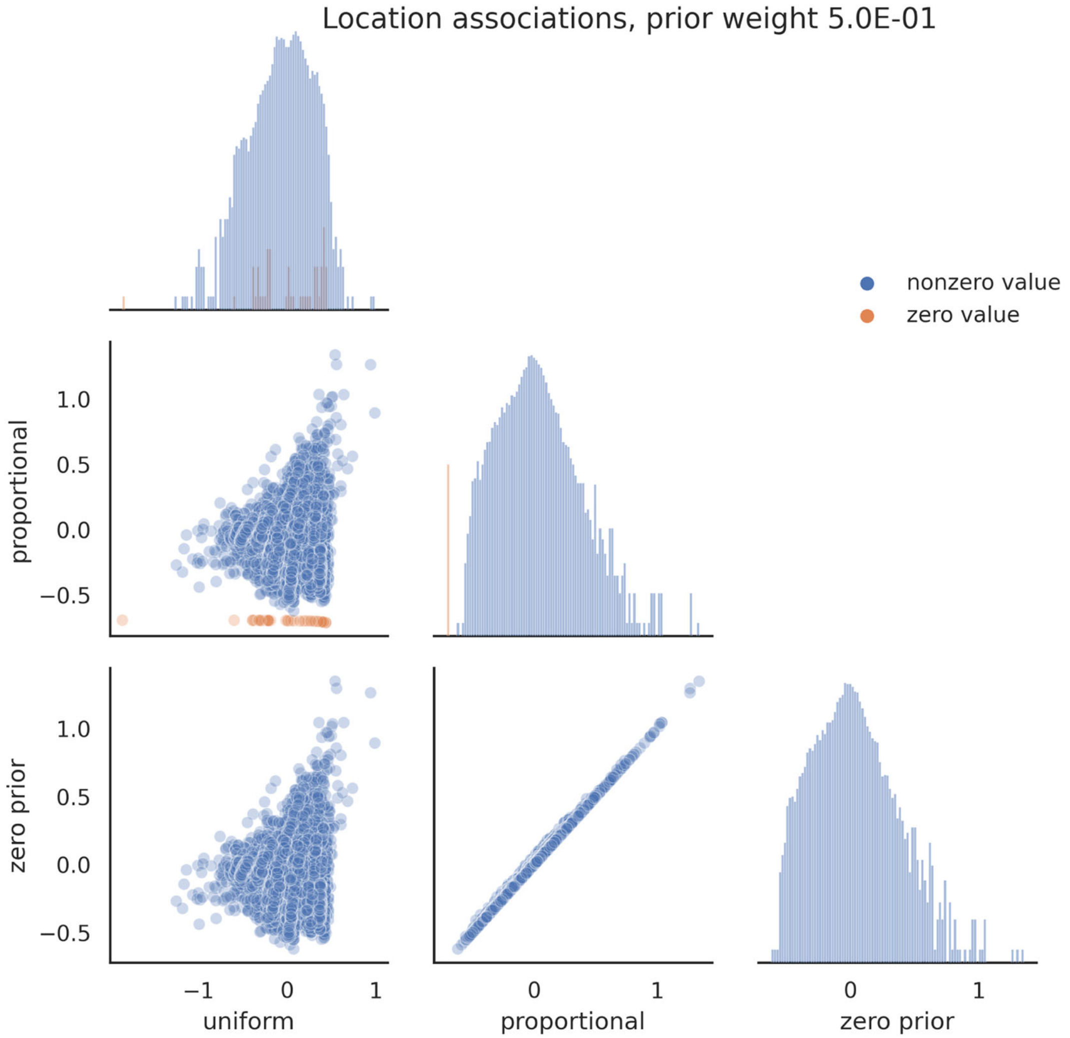

FIGURE C3 Location associations of major occupations groups for different priors. The vertical axes of the histograms have a log scale. Colors indicate which cells have zero observations in the data. The prior weight is 0.5.

For the zero prior, their theoretical value is minus infinity. As the prior strength increases, the associations in the cells get spread get spread out over a wide range of values under the uniform prior, as for small cells the effect of adding a constant pseudocounts may be very large. The proportional prior adds pseudocounts proportional to the marginals in each cell, moving them closer to the other observations but keeping them below the associations for which we have data.

It seems that the proportional prior has practical properties for the estimation of associations: it keeps the associations for which we have plenty observation in place even for large prior strengths, while enabling us to deal with cells in which we have very few or no observations.

For the colocation associations, we show a similar analysis. For the major occupations groups we do not have a case where $\hat{p}_{ij} = 0$ for the zero prior—hence one could use the zero prior without any difficulties (but this is in general not the case). Again, we find that the proportional prior keeps most values in place, even for large prior strengths, while the uniform prior shows increasingly large deviations from the maximum likelihood estimate (Figures C5 and C6).

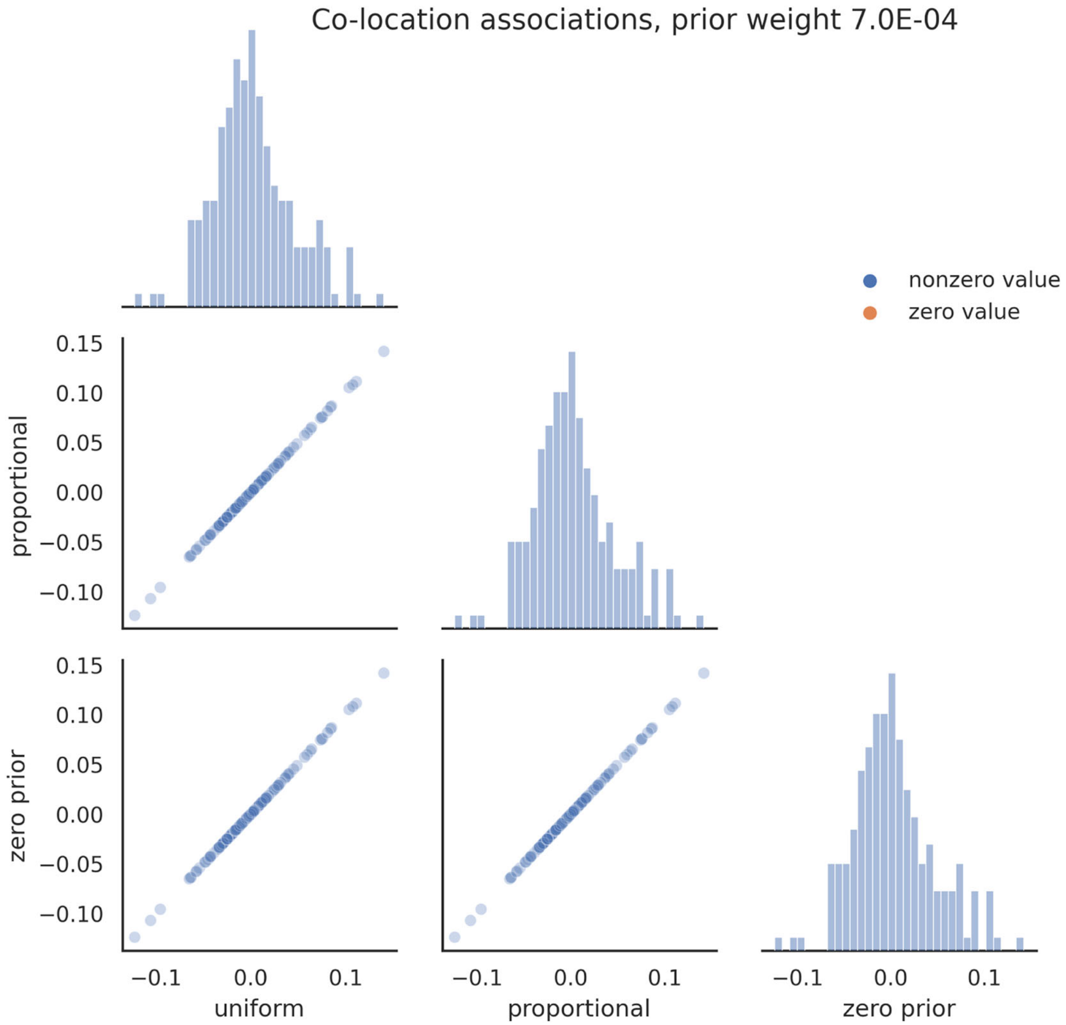

FIGURE C4 Colocation associations of major occupations groups for different priors. The vertical axis of the histograms has a log scale. The prior weight is 0.0007.

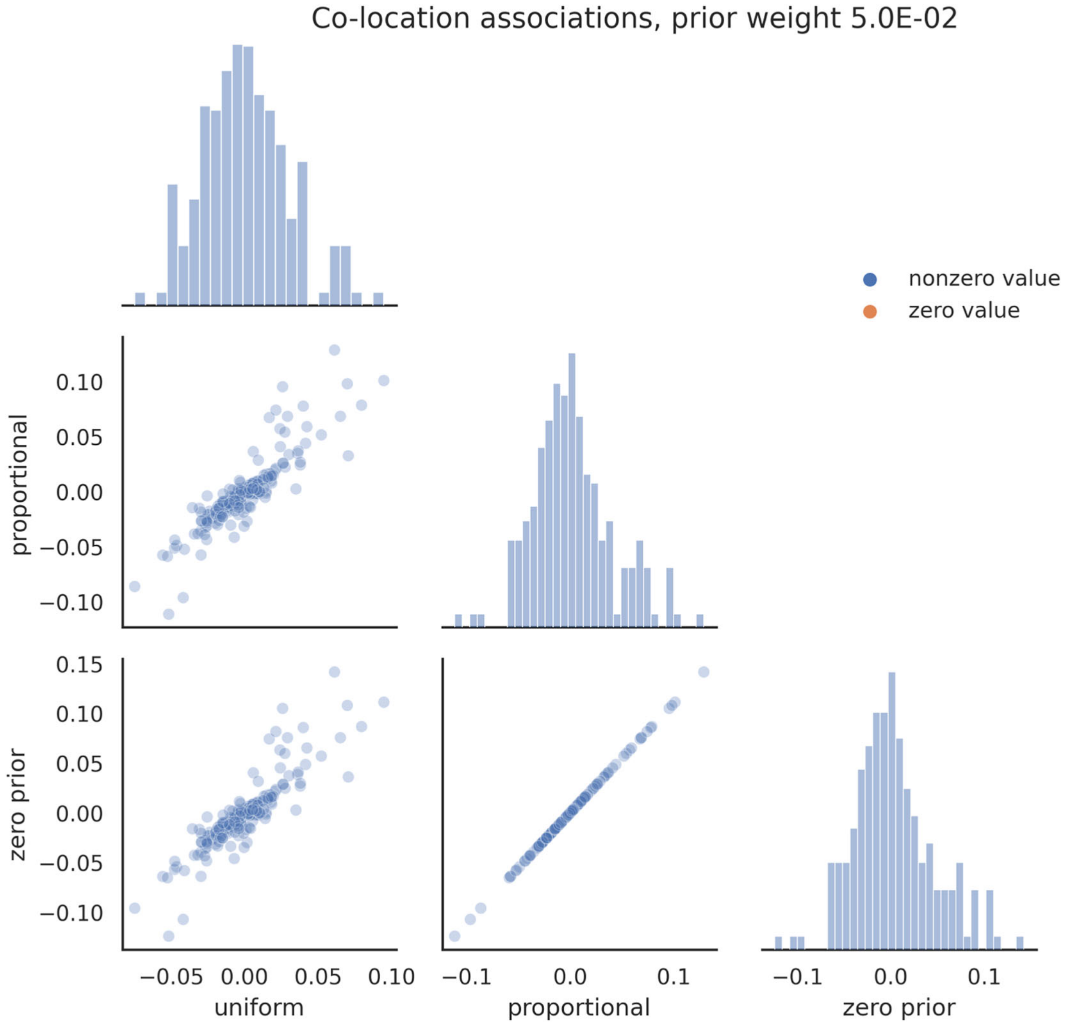

FIGURE C5 Colocation associations of major occupations groups for different priors. The vertical axis of the histograms has a log scale. The prior weight is 0.05.

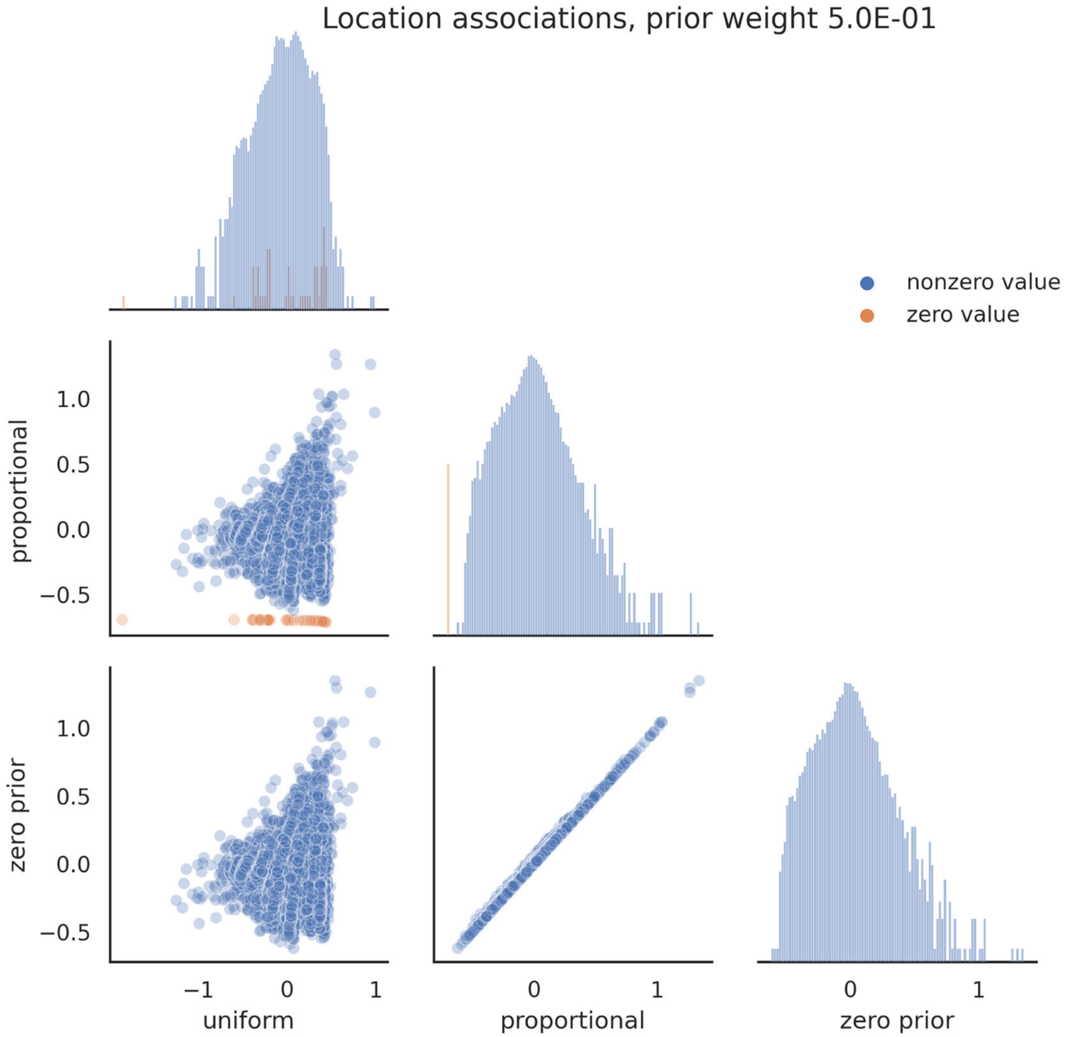

FIGURE C6 Colocation associations of major occupations groups for different priors. The vertical axis of the histograms has a log scale. The prior weight is 0.5.