Partitioning specialisation in multilevel textual corpora

A hierarchical model of pairwise similarities

Carlo Debernardi

CS2Italy2025

16/01/2025

Background

Recent advances in text analysis
(Mikolov et al. 2013, Vaswani et al. 2017, Reimers & Gurevych 2019)

	Text Embedding	Generative LLMs
Architecture	Encoder	Decoder
Output	Vectors	Text
Example	BERT	GPT

Close relatives, but one is vastly more popular then the other.
Not all potential applications have been explored yet.

Background

Typical application of text-embedding:
dimensionality reduction and clustering
(e.g: BERTopic - Grootendorst 2022)

Great for exploration but:

- Approximate lower dimensionality
- Dichotomization (technical and theoretical issue)
- Difficult to integrate into inferential analysis

Background

How to use the output of text embedding algorithms in inferential analyses?

Output: high dimensional vectors, usually optimized to encode relations as cosine similarities

- As a covariate: to be explored (Gaussian processes?)
- As the outcome: focus of this work

I.e. are texts, on average, more (or less) similar as a function of x?

A case study

Objective

Study the specialisation in the scientific literature, disentangling field and journal-level effects.

Specialisation operationalised as the average similarity among pairs of abstracts (i.e. higher average similarity = higher specialisation).

Ideally we want to be able to introduce covariates (e.g. time trend).

Note: the modeling exercise forces us to take care of details that turn out to be interesting in and of themselves.

The data

Data collection:

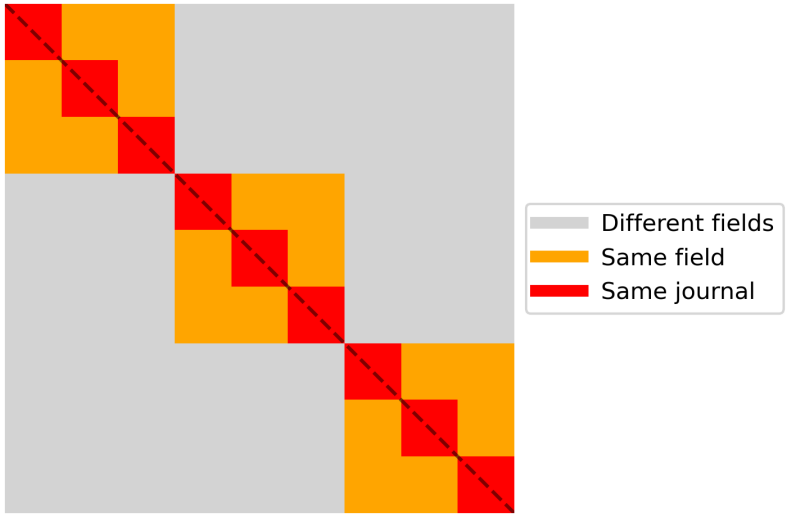
- Data collected from OpenAlex (CC0 license!)
- 5 fields (ASJC classification, Scopus)
- 20 journals per field (high rank, at least 100 pubs)
- 100 random papers per journal (timeframe 2010-2020)
- Results in 10k abstracts

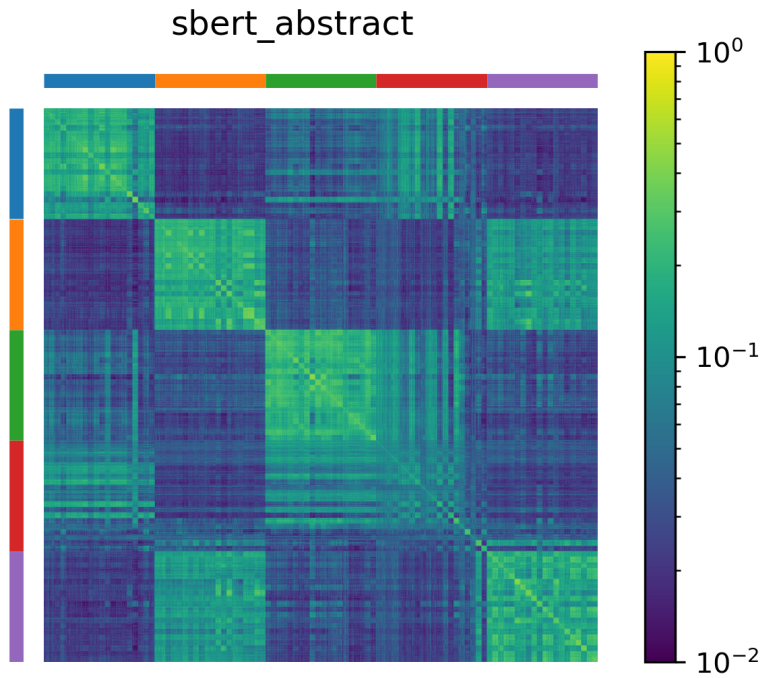
Note: random subsampling in testing phase, 1.9M obs per field!

Fields:

- Condensed Matter Physics
- Economics and Econometrics
- Genetics
- Multidisciplinary
- Sociology and Political Science

Structure in the data





Highly structured data! $\rightarrow$ Hierarchical model
Non-independence & in/out journal

The model

$$Y_{ij} \sim Beta(\mu_{ij}, \phi)$$

Beta model to respect rescaled outcome domain [0-1]
Varying intercept as a baseline version

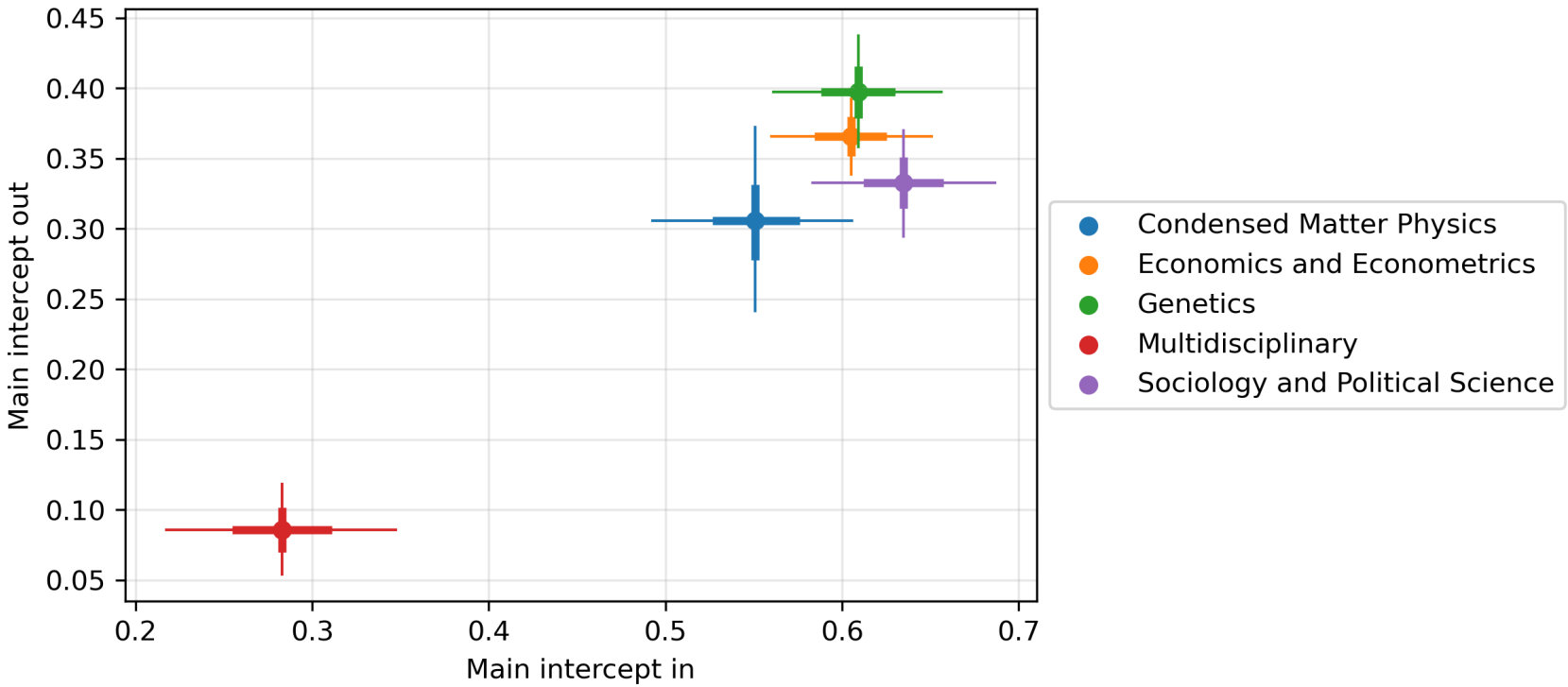
$$\mu_{ij} = logit^{-1}(F + J + P_i + P_j),\ if\ J_i = J_j$$

Avg journal spec | Journal spec | Paper centrality (in journal)

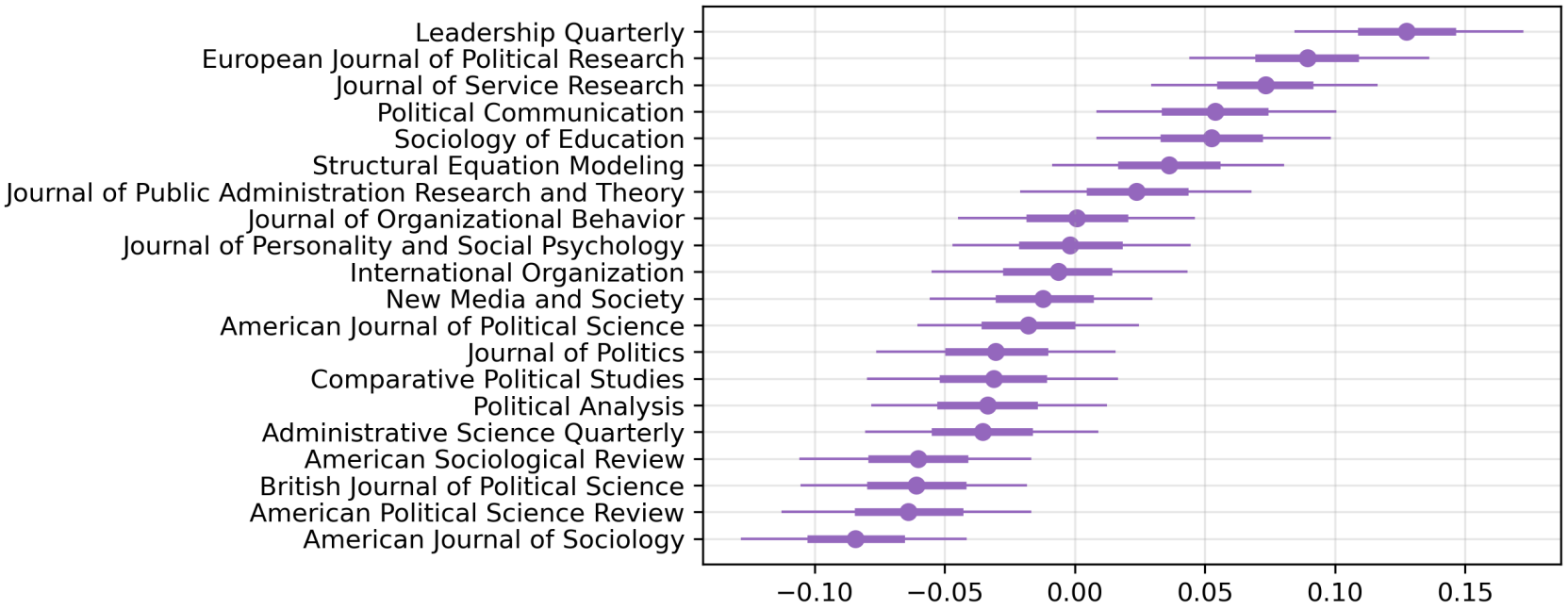
$$\mu_{ij} = logit^{-1}(F + J_i + J_j + P_i + P_j),\ if\ J_i \neq J_j$$

Field spec | Journal centrality (in field) | Paper centrality (in field)

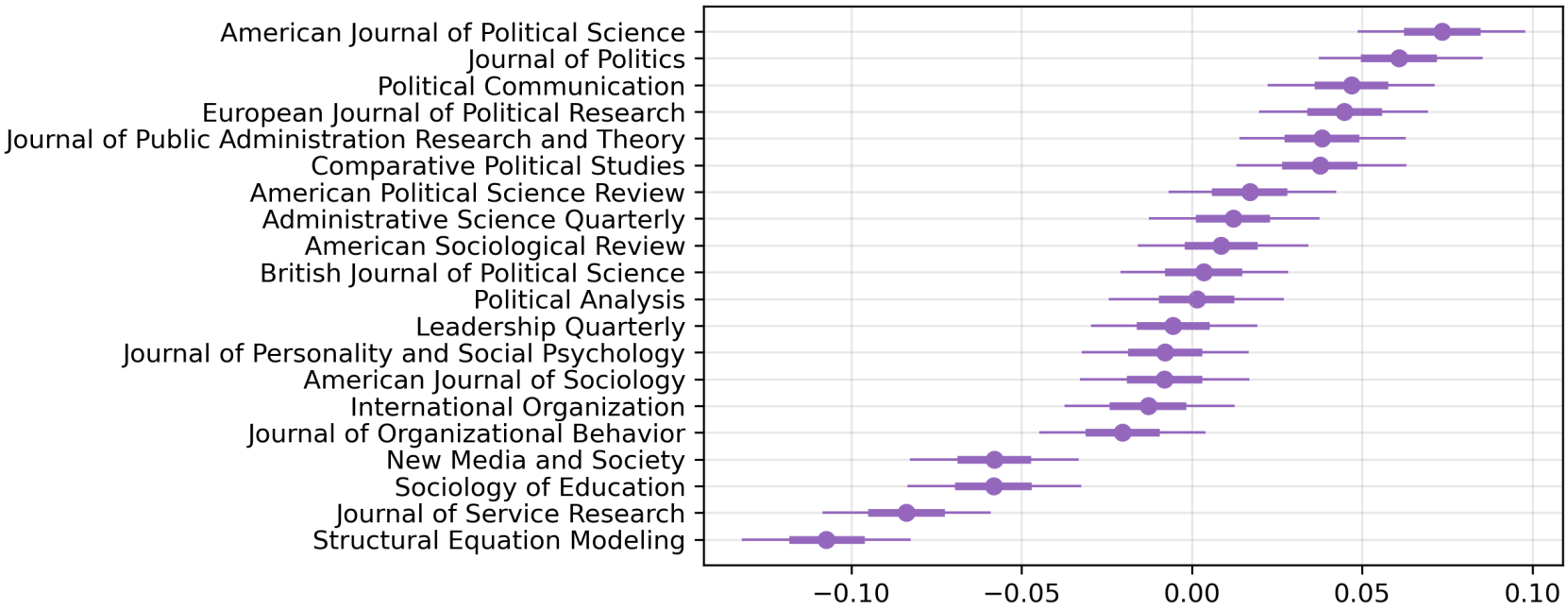
Field intercepts



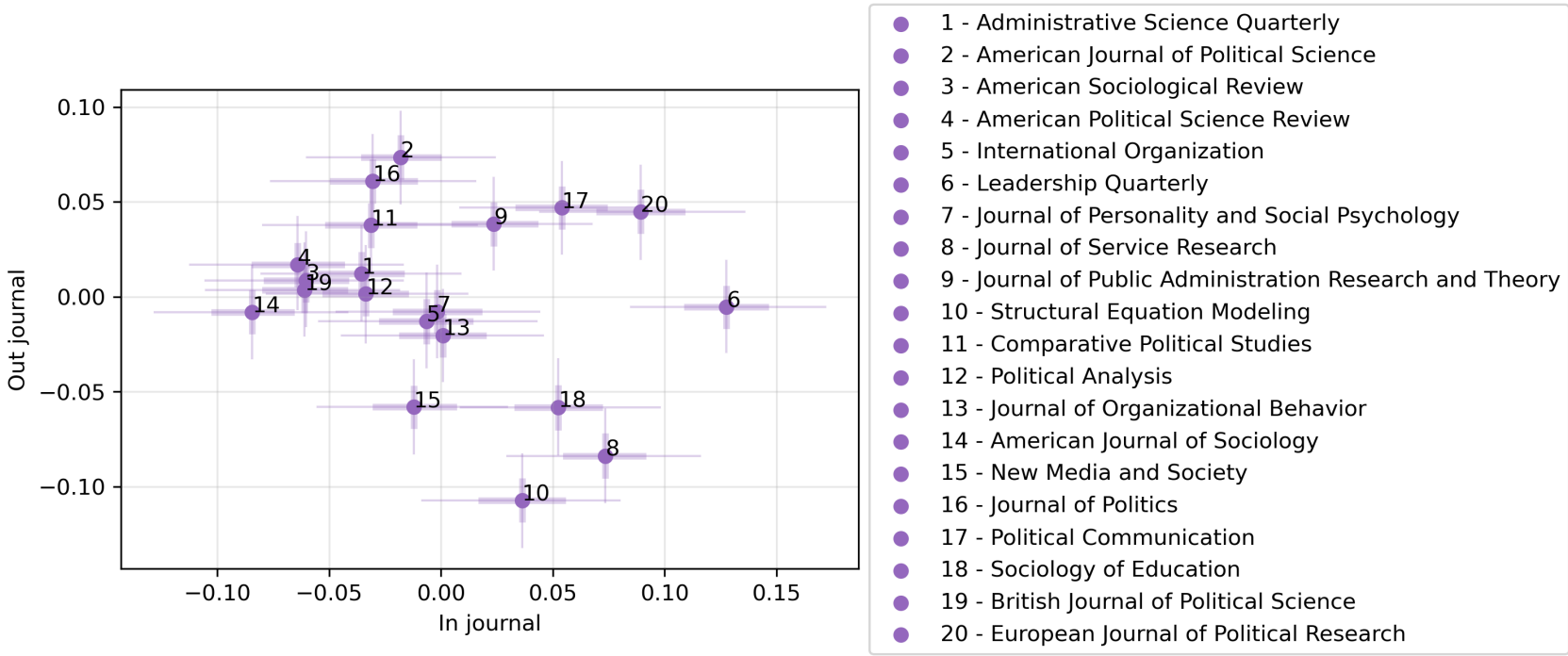
Journal intercepts - specialisation



Journal intercepts - centrality



Journal intercepts



Summary

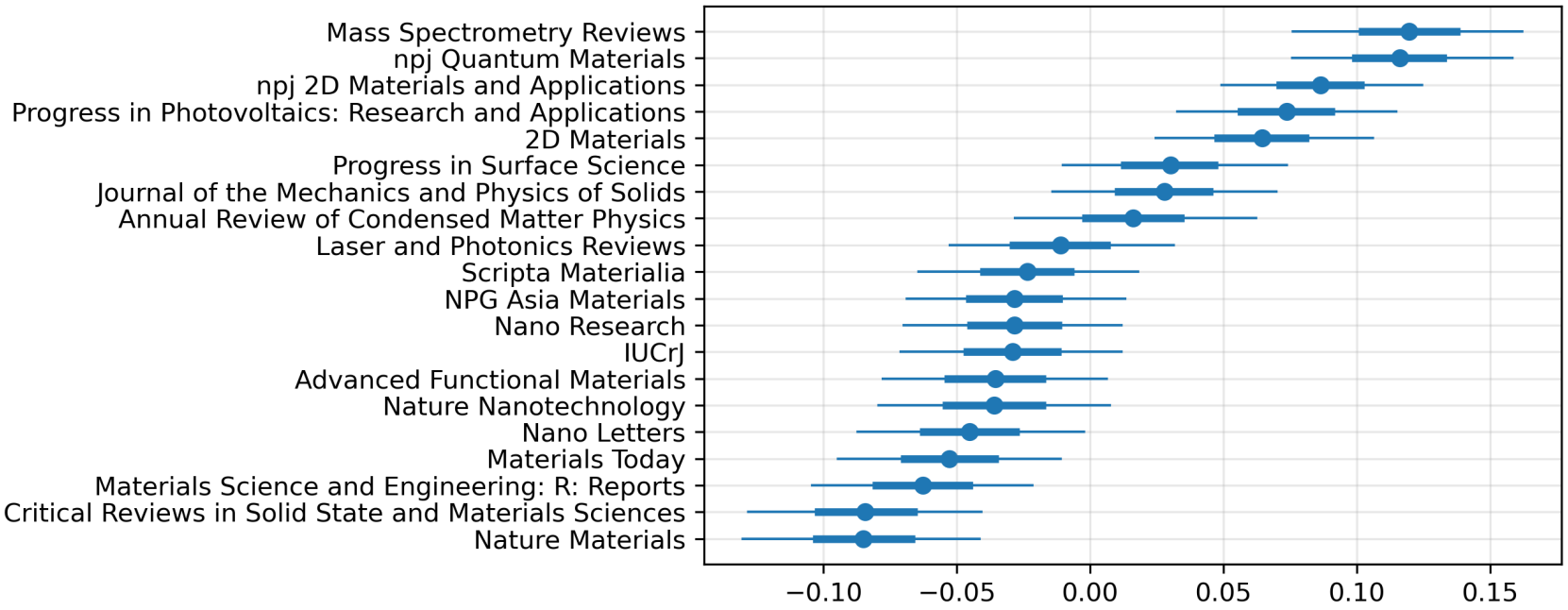
To sum up:

- There are still unexplored applications of text embedding
- It can be employed together with inferential analysis
- We gain the ability to estimate marginal quantities
- We are forced to deal with the structure of the data
- It is possible to disentangle structural levels that contribute to specialisation at varying degree

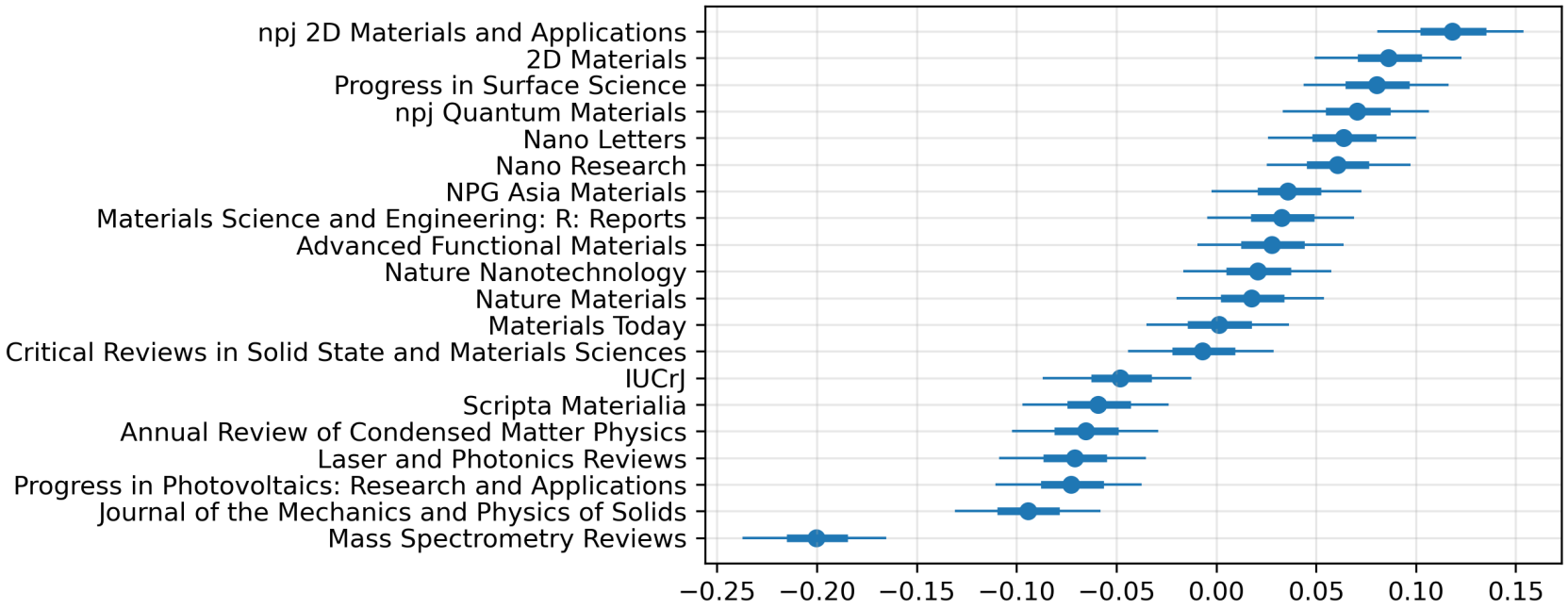
Thanks for the attention

carlo.debernardi@unisi.it

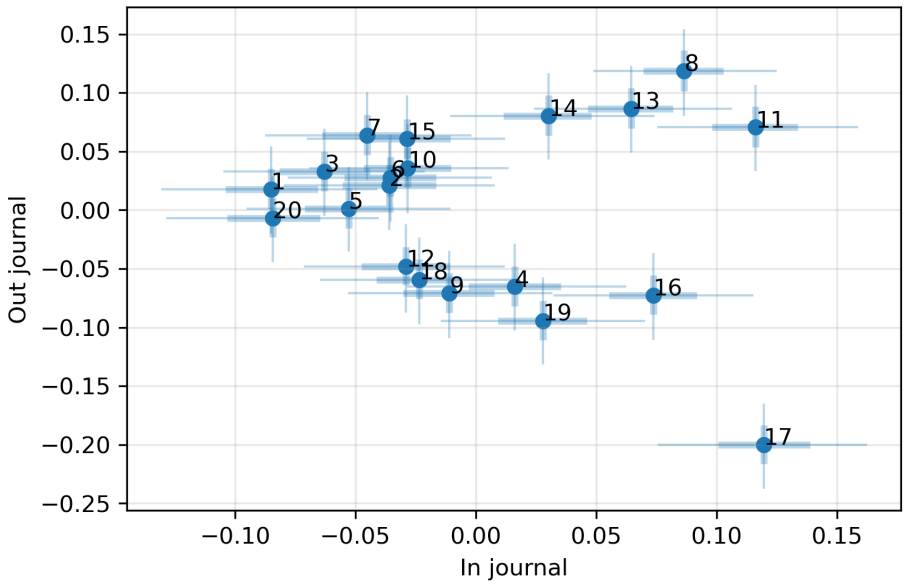
Journal intercepts - specialisation



Journal intercepts - centrality

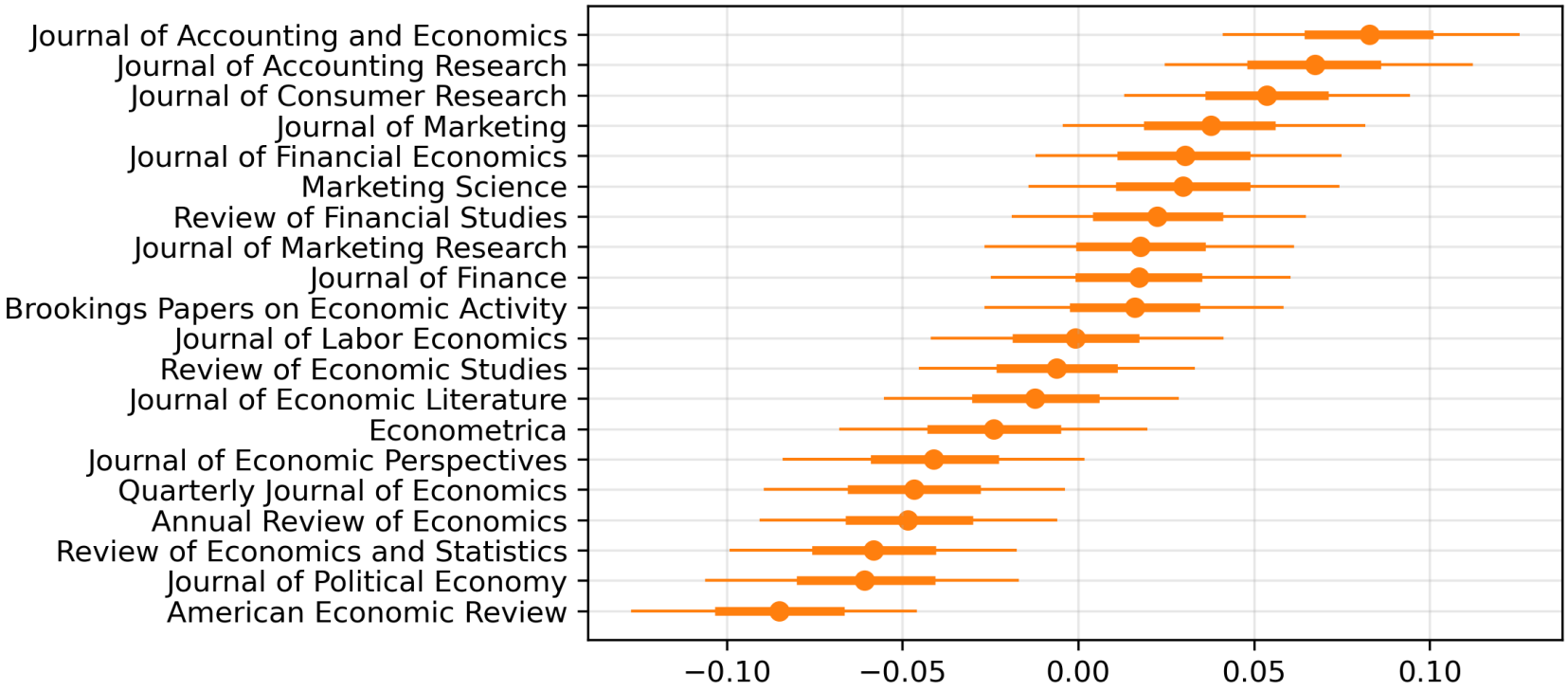


Journal intercepts

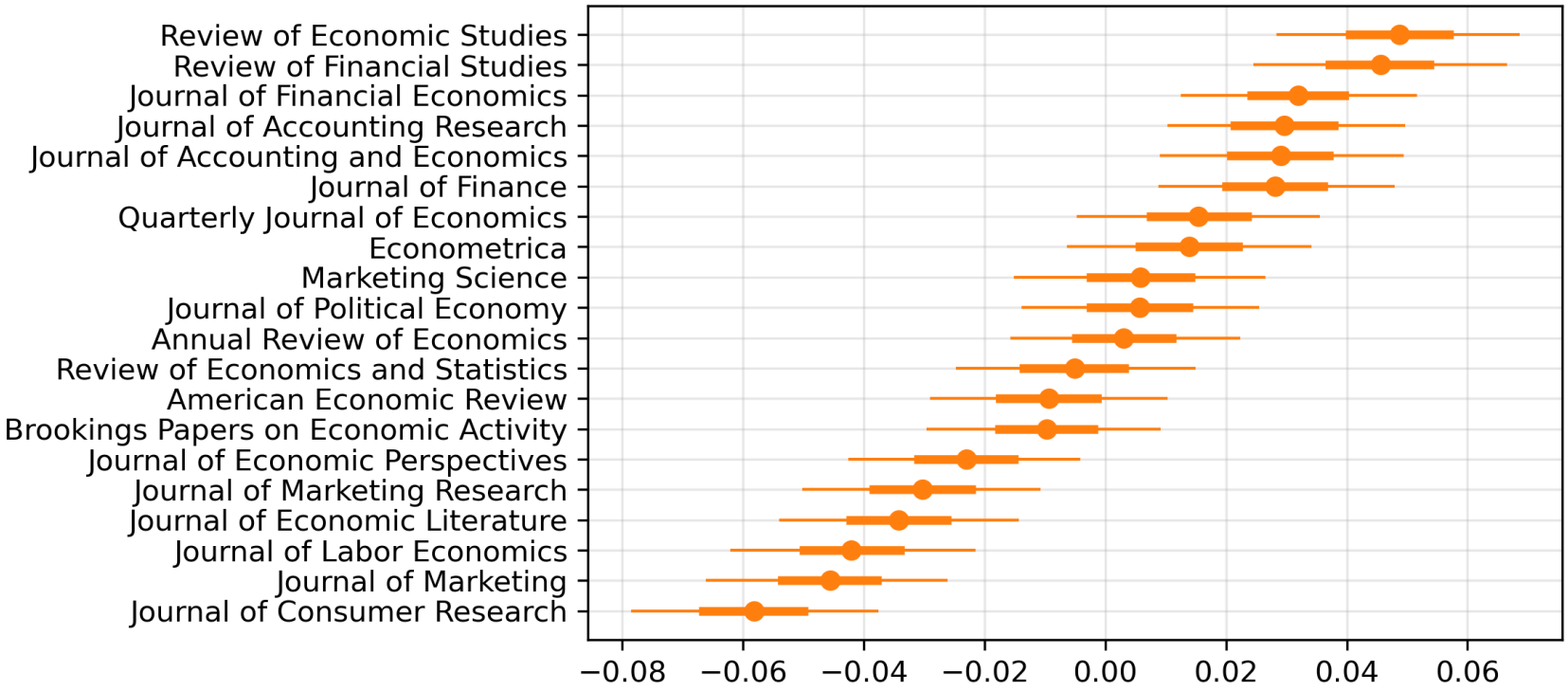


1 - Nature Materials
2 - Nature Nanotechnology
3 - Materials Science and Engineering: R: Reports
4 - Annual Review of Condensed Matter Physics
5 - Materials Today
6 - Advanced Functional Materials
7 - Nano Letters
8 - npj 2D Materials and Applications
9 - Laser and Photonics Reviews
10 - NPG Asia Materials
11 - npj Quantum Materials
12 - IUCrJ
13 - 2D Materials
14 - Progress in Surface Science
15 - Nano Research
16 - Progress in Photovoltaics: Research and Applications
17 - Mass Spectrometry Reviews
18 - Scripta Materialia
19 - Journal of the Mechanics and Physics of Solids
20 - Critical Reviews in Solid State and Materials Sciences

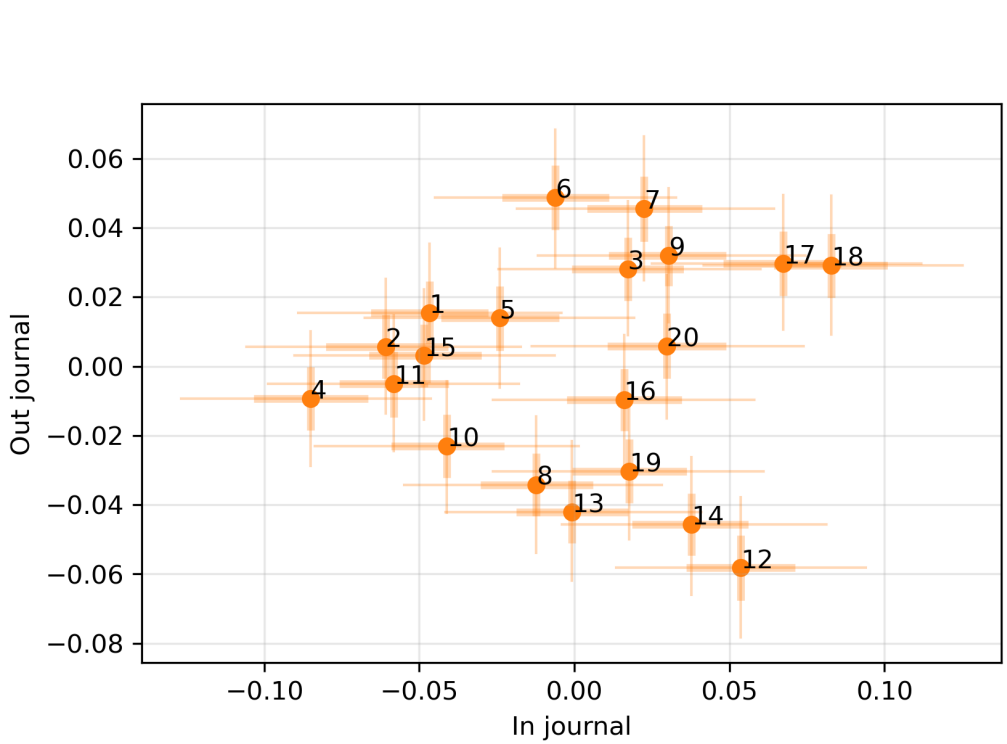
Journal intercepts - specialisation



Journal intercepts - centrality

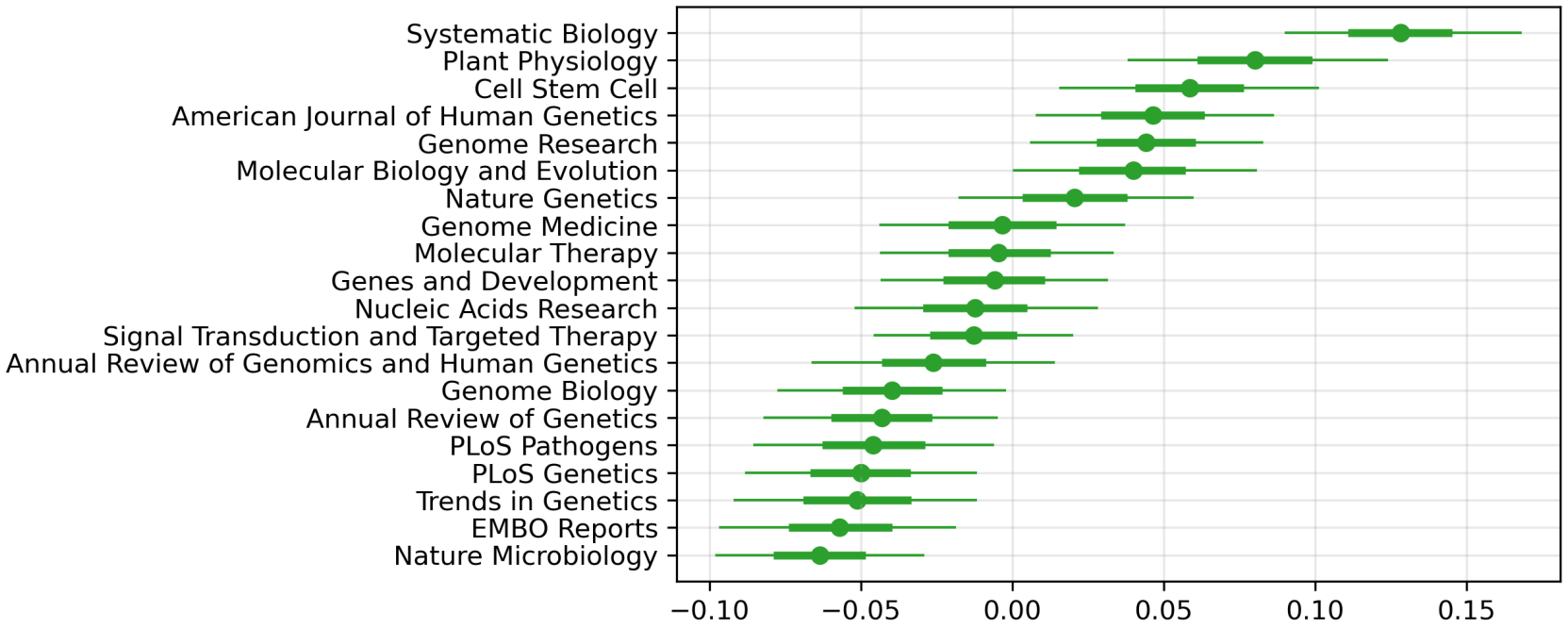


Journal intercepts

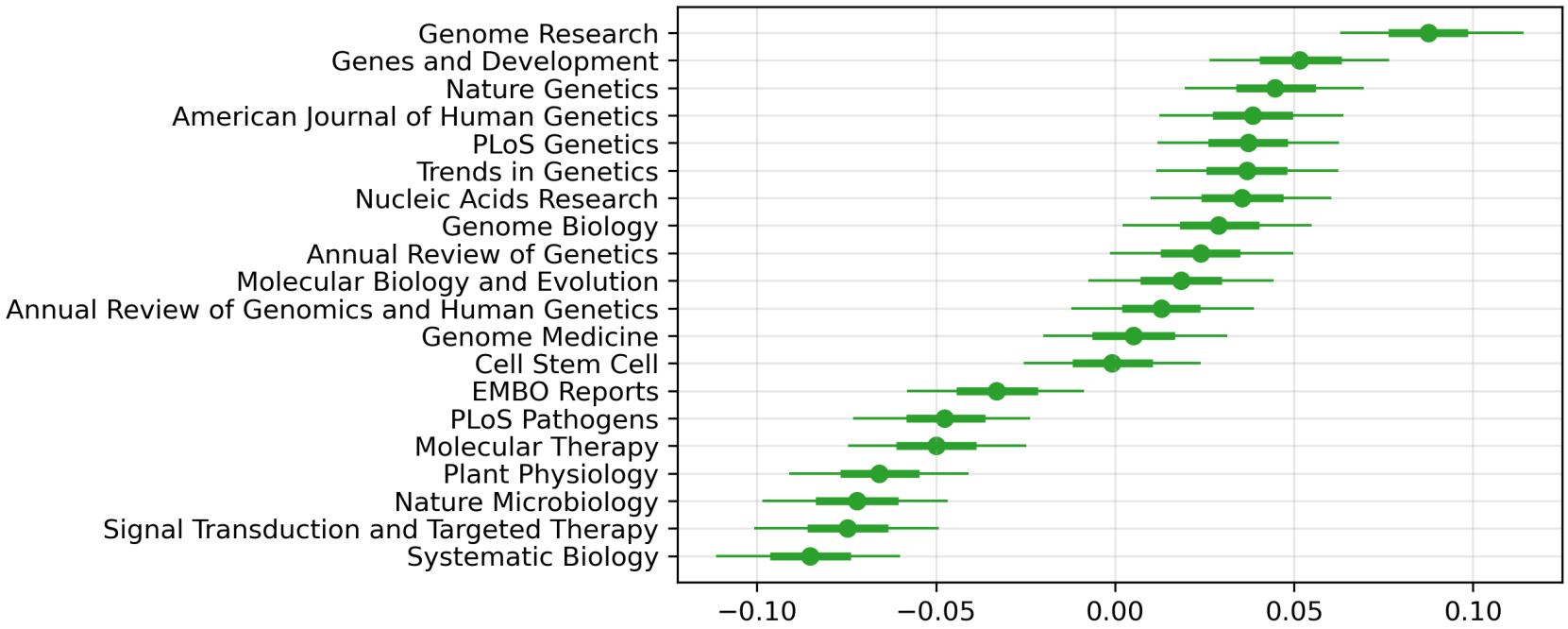


- 1 - Quarterly Journal of Economics
- 2 - Journal of Political Economy
- 3 - Journal of Finance
- 4 - American Economic Review
- 5 - Econometrica
- 6 - Review of Economic Studies
- 7 - Review of Financial Studies
- 8 - Journal of Economic Literature
- 9 - Journal of Financial Economics
- 10 - Journal of Economic Perspectives
- 11 - Review of Economics and Statistics
- 12 - Journal of Consumer Research
- 13 - Journal of Labor Economics
- 14 - Journal of Marketing
- 15 - Annual Review of Economics
- 16 - Brookings Papers on Economic Activity
- 17 - Journal of Accounting Research
- 18 - Journal of Accounting and Economics
- 19 - Journal of Marketing Research
- 20 - Marketing Science

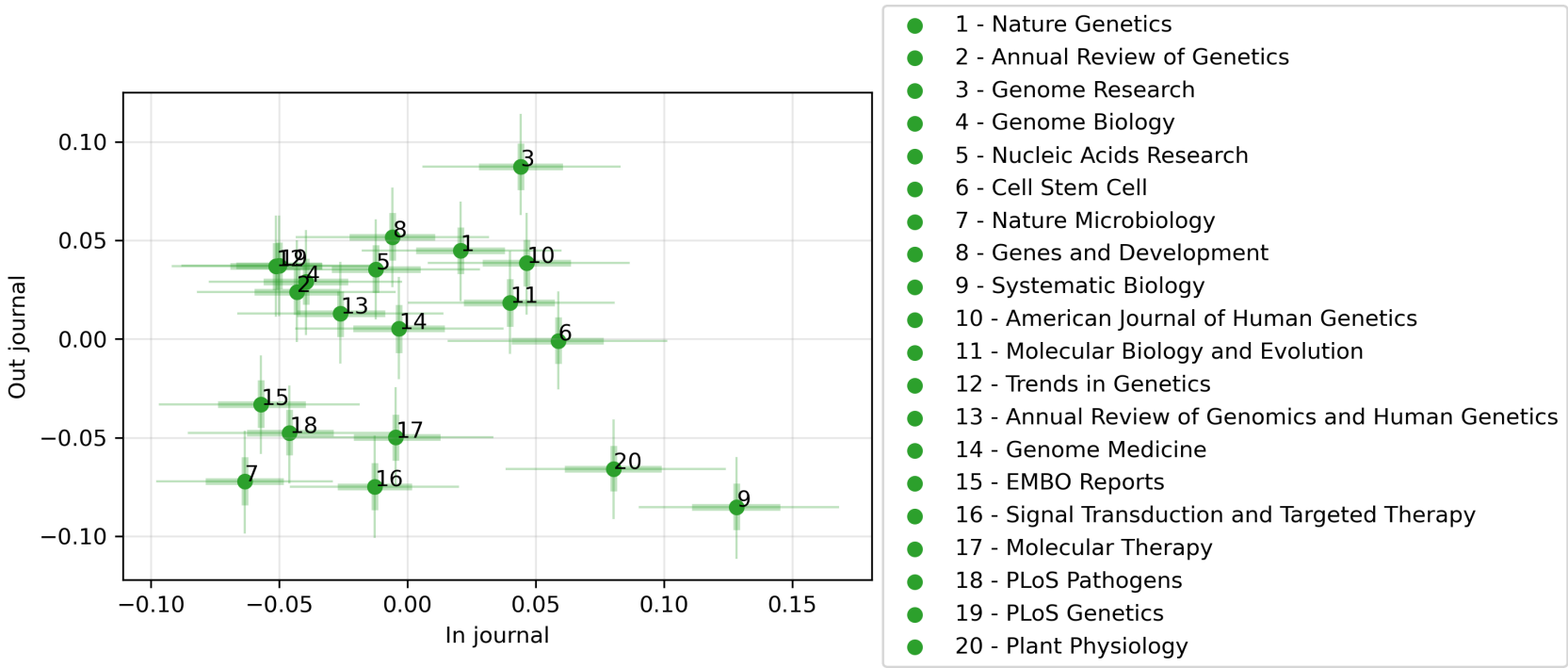
Journal intercepts - specialisation



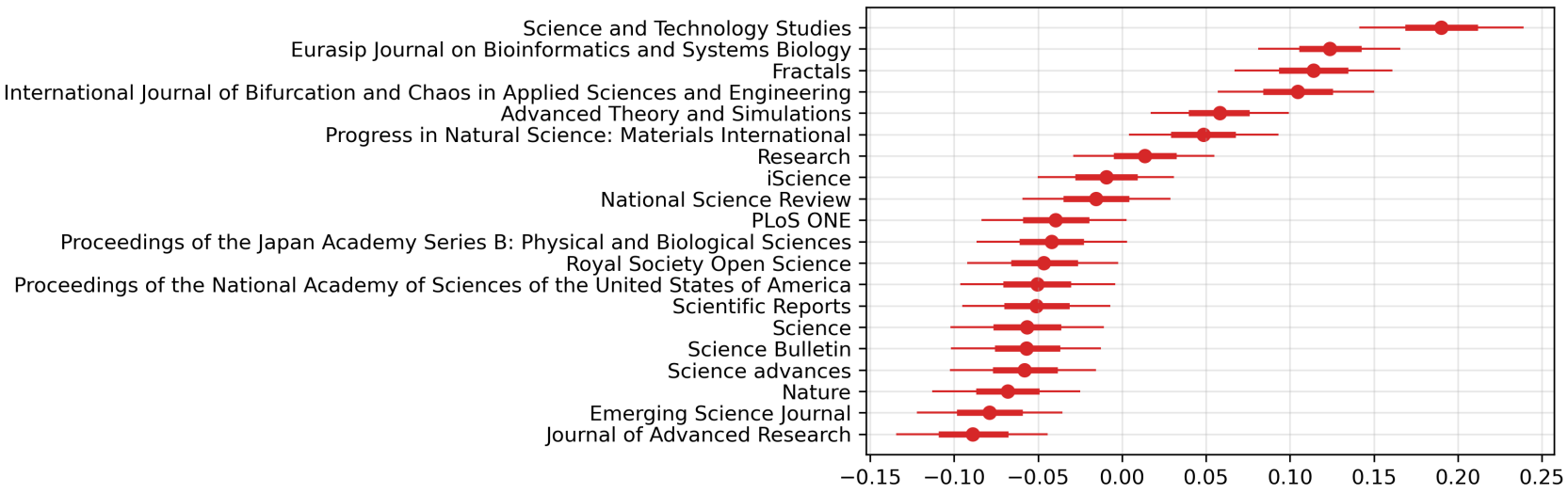
Journal intercepts - centrality



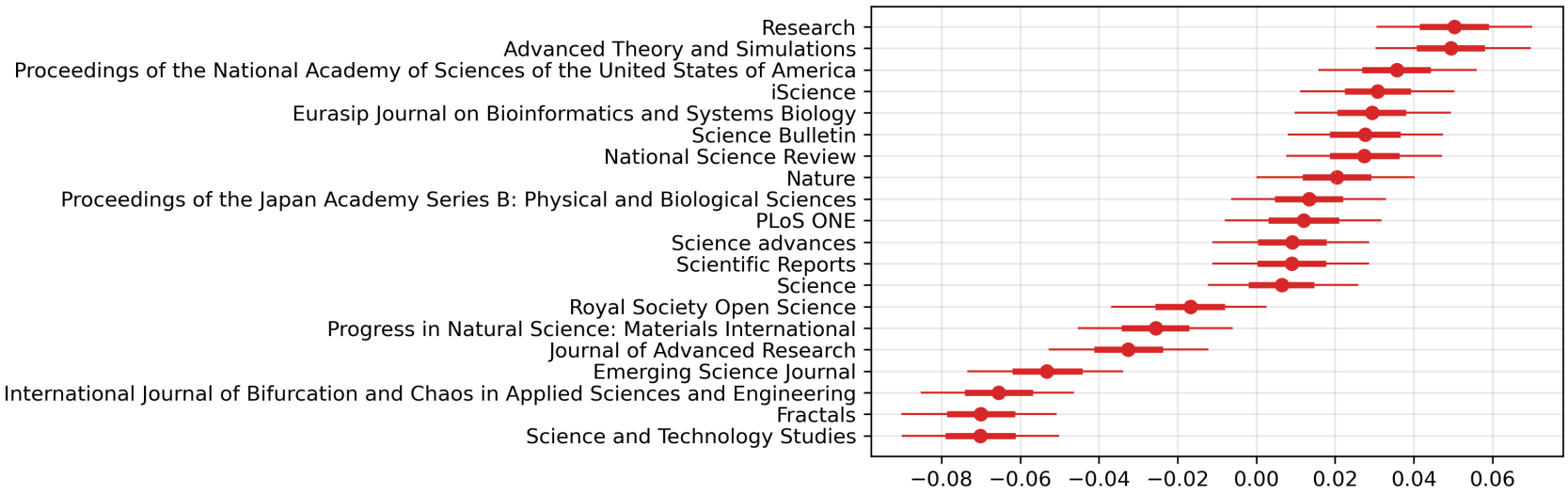
Journal intercepts



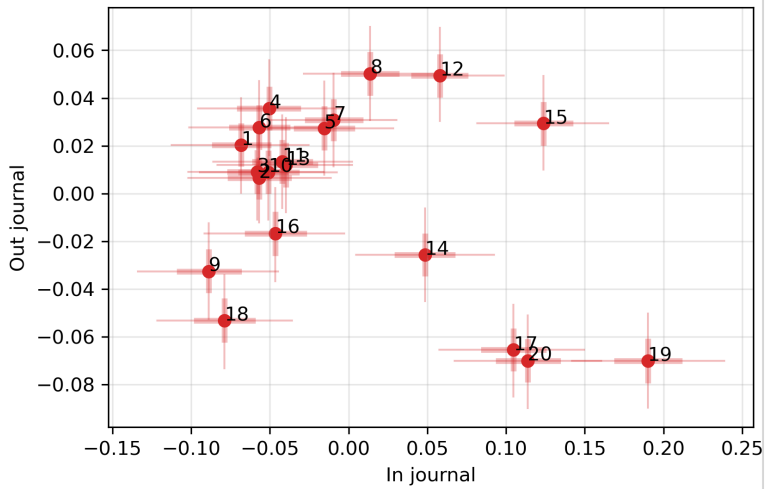
Journal intercepts - specialisation



Journal intercepts - centrality



Journal intercepts



1 - Nature
2 - Science
3 - Science advances
4 - Proceedings of the National Academy of Sciences of the United States of America
5 - National Science Review
6 - Science Bulletin
7 - iScience
8 - Research
9 - Journal of Advanced Research
10 - Scientific Reports
11 - Proceedings of the Japan Academy Series B: Physical and Biological Sciences
12 - Advanced Theory and Simulations
13 - PLoS ONE
14 - Progress in Natural Science: Materials International
15 - Eurasip Journal on Bioinformatics and Systems Biology
16 - Royal Society Open Science
17 - International Journal of Bifurcation and Chaos in Applied Sciences and Engineering
18 - Emerging Science Journal
19 - Science and Technology Studies
20 - Fractals