Zero-sum correlations – Why they arise and some ways to handle them

Rojan Karakaya, Institute for Futures Studies

About me

- Graduated from the Master's program in Machine Learning and Statistics in 2022
- Worked as a research assistant for Moa Bursell and Magnus Bygren at the Stockholm University department of Sociology in parallel with my studies
- Started working full-time at the Institute for Futures Studies in July 2022

Advice for new students re. studies

- Make sure you have a firm grasp of the basics:
  - Maths, mathematical statistics
  - Philosophical underpinnings

- Without the maths, a lot of machine learning is going to be inscrutable
- Without the underpinnings you'll be in constant confusion

Examples of problems that might arise:

1. Trying to understand VAEs without knowledge of KL-divergence
2. Not knowing what a p-value means in frequentist hypothesis testing

Advice for new students re. <u>early career</u>

- Apply **<u>a lot</u>**
  - It is not uncommon to have to apply ~500 times for first job
  - Apply to ”ancillary” roles as well, e.g. Data Engineering, Data Analyst etc.
- The best time to look for a new job is when you already have one
- Don’t burn yourself out
  - Many early career data scientists feel pressure to work unpaid overtime – leave such roles as soon as you find other work!

Advice for foreign students re. early career

- If you wish to stay in Sweden post-studies: Get in line in the various housing queues for the three metropolitan areas in Sweden!
- Focus on your schoolwork over Swedish classes
  - Most employers in our field use English as their working language
- If you do wish to learn Swedish, SFI is not going to be very helpful
  - Self-study instead

My research at IFFS

- Theme: Discrimination in the labour market and ethnic bigotry
- Projects include
  - Sending fictitious applications to job listings and noting the callback rate for names of different ethnicities and gender
  - Tracing changes to ethnic stereotypes in historical text corpora
  - Investigating real job application data to disaggregate the effects of discrimination from self-sorting

Fixed-sum conditions on outcomes

- Imagine a clustered dataset of e.g. job listings, with outcome data about which applicant was hired within its listing (cluster) along with some covariates for each applicant
- How would one deal with the intra-cluster covariance?

Fixed-sum conditions on outcomes

A priori we have

$$E[y_i] = 1/n$$

$$V[y_i] = E[{y_i}^2] - E[y_i]^2 = (n-1)/n^2$$

Clearly in each cluster (for *n* applicants),

$$\sum_{i=1}^{n} y_i = 1$$

Which means that the covariance within any pair of applicants is

$$Cov(y_i, y_j) = E[y_i y_j] - E[y_i]E[y_j] = 0 - (1/n)^2$$

Fixed-sum conditions on outcomes

…and therefore the covariance matrix for the outcome variable must be

$$\begin{bmatrix} (n-1)/n^2 & -1/n^2 & \cdots & -1/n^2 \\ -1/n^2 & (n-1)/n^2 & \cdots & -1/n^2 \\ \vdots & \vdots & \ddots & \vdots \\ -1/n^2 & -1/n^2 & \cdots & (n-1)/n^2 \end{bmatrix}$$

- This matrix is a special case of a circulant matrix
- For a circulant matrix, one of its eigenvalues must be its row/column sum, i.e. 0!
  - Ergo non-invertible

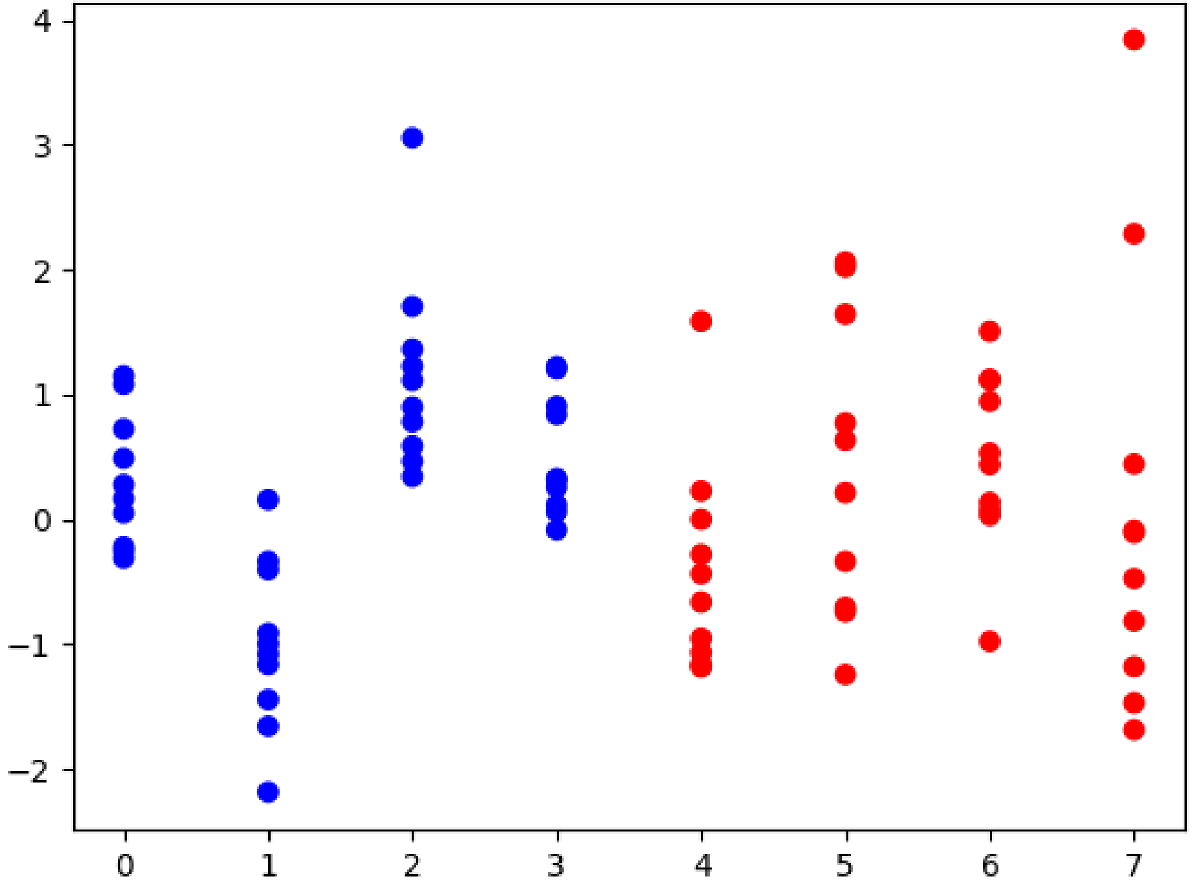
Analogy to multicollinearity

Given the fixed-sum constraint, this non-invertibility should not be too surprising – one row must be a linear combination of the others

- As a consequence, each cluster of *n* observations can contribute a maximum of *n* - 1 degrees of freedom

This is very similar to multicollinearity, except the problem is in the rows rather than columns and in the outcome rather than the covariates

A general point on negative intra-cluster covariances



Solution 1: the Zero-Sum Regression

Clearly some form of transformation that reduces the number of rows in each cluster from n to n – 1 is needed, but which one?

Necessary conditions: transformed observations should be linearly independent

An easy one to remember is

$$R = \begin{bmatrix} 1 & -1/(n-1) & \cdots & & -1/(n-1) \\ 0 & 1 & \cdots & & -1/(n-2) \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -1/2 & -1/2 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}$$

Solution 1: the Zero-Sum Regression

We then do a regular OLS on the transformed data, i.e.

$$\hat{\beta} = (X^T R^T R\, X)^{-1} X^T R^T R \vec{y}$$

Drawbacks of the zero-sum regression

For dichotomous outcomes it shares the same problems as the Linear Probability Model

Estimating actual probabilities from a pool of competitors is very difficult

Good for inferring the effects of the covariates however!

Example 1: ZSR on the Blue Tit Broods dataset

Dataset is from Morales, Achevedo and Machordom (2019)

The aim of the authors was to study how parental care of chicks in a species of bird varies with brood and chick traits

Example 1: ZSR on the Blue Tit Broods dataset

	General Mixed Model (original paper)		Zero-Sum Regression	
Variable	Coefficient	p-value	Coefficient	p-value
Heterozygosity	0.46	**0.048**	0.37	0.067
Relatedness	0.11	0.52	0.26	0.061
Sex (female)	-0.22	**<0.001**	0.26	**<0.001**
Tarsus length	0.55	**<0.001**	0.29	**<0.001**

Solution 2: the repeated Multinomial Logit model without replacement

The basic multinomial logit model is a discrete choice model which gives the probability for choice *i* as

$$P_i = \frac{\exp(\vec{\beta} \cdot \overrightarrow{x_i})}{\sum_{j=1}^{n} \exp(\vec{\beta} \cdot \overrightarrow{x_j})}$$

The Zero-Sum constraint is ”baked in” in this model

Solution 2: the repeated Multinomial Logit model without replacement

For multiple winners and known order, parameter inference and likelihood is straightforward. If options are ordered such that winner 1 has index 1, winner 2 has index 2 etc. then:

$$P_{1,2,3} = \frac{\exp(\vec{\beta} \cdot \overrightarrow{x_1})}{\sum_{j=1}^{n} \exp\left(\vec{\beta} \cdot \overrightarrow{x_j}\right)} \frac{\exp(\vec{\beta} \cdot \overrightarrow{x_2})}{\sum_{j=2}^{n} \exp\left(\vec{\beta} \cdot \overrightarrow{x_j}\right)} \frac{\exp(\vec{\beta} \cdot \overrightarrow{x_3})}{\sum_{j=3}^{n} \exp\left(\vec{\beta} \cdot \overrightarrow{x_j}\right)}$$

Solution 2: the repeated Multinomial Logit model without replacement

For multiple winners and **un**known order, parameter inference and likelihood is *conceptually* straightforward, i.e. the sum of all possible valid combinations

$$P_{\{1,2,3\}} = P_{1,2,3} + P_{2,1,3} + P_{2,3,1} + P_{3,2,1} + P_{1,3,2} + P_{3,1,2}$$

…but note that if the number of winners is *m*, then the number of terms must be *m*!

This can easily become untractable for not-very-large values of *m*

Solution 2: the repeated Multinomial Logit model without replacement

By Bayes's theorem we have:

$$P(\hat{\beta}|unordered\ set)\ P(unordered\ set) = P(unordered\ set\ |\ \hat{\beta})P(\hat{\beta})$$

which in turn can be rewritten as

$$P(unordered\ set\ |\ \hat{\beta})P\left(\hat{\beta}\right) = \sum_{\substack{all\ valid \\ orderings}} P(ordered\ set\ |\ \hat{\beta})P\left(\hat{\beta}\right)$$

Solution 2: the repeated Multinomial Logit model without replacement

The right-hand side sum

$$\sum_{\substack{all\ valid\\ orderings}} P(ordered\ set \mid \hat{\beta})P(\hat{\beta})$$

can be approximated by taking a sample of orders, where sample probability is given by $\hat{\beta}$, and then iteratively updating $\hat{\beta}$

Example 2: Labour market discrimination with Multinomial Logit

A company has given my research team access to almost all of their internal hiring data:

1. Job listings
2. Applicants to each job listing (with covariates)
3. Outcome data (shortlisting, interview and hiring)

Example 2: Labour market discrimination with Multinomial Logit

Our aim is to disentangle the effects of self-sorting (or applicant allocation among listings) and discrimination on unequal labour market outcomes

Example 2: Labour market discrimination with Multinomial Logit

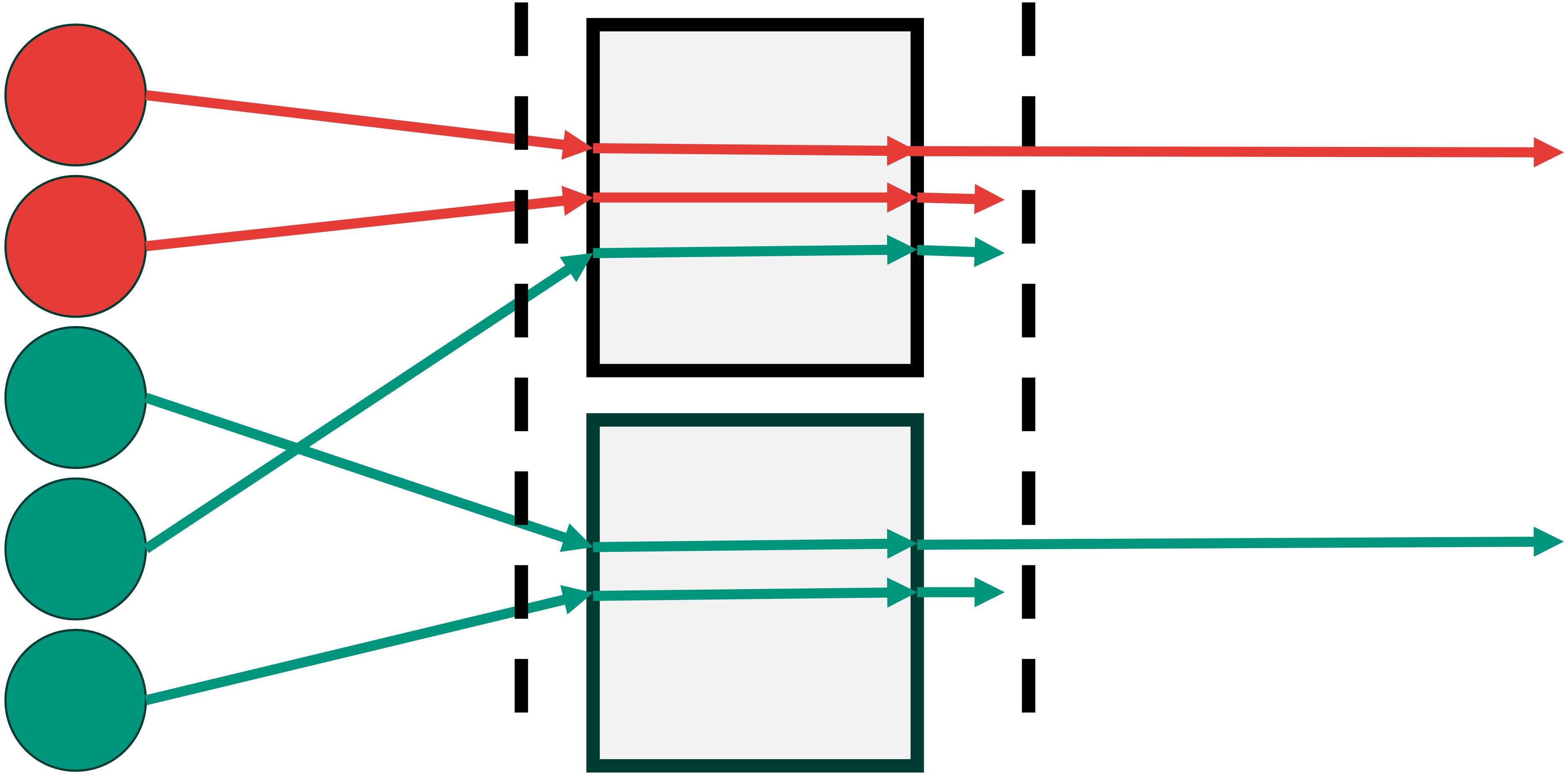
Each individual's probability weight is given by the log-linear employer preference model:

$$W_i = \exp(0.09 * woman_i + 0.25 * european_i + 0.03 * test\ score_i)$$

Individual probabilities are then calculated based on the individual's weight in relation to the probability weights of all the applicants in the pool:

$$P_i = \frac{W_i}{\sum_j W_j}$$

Example 2: Labour market discrimination with Multinomial Logit

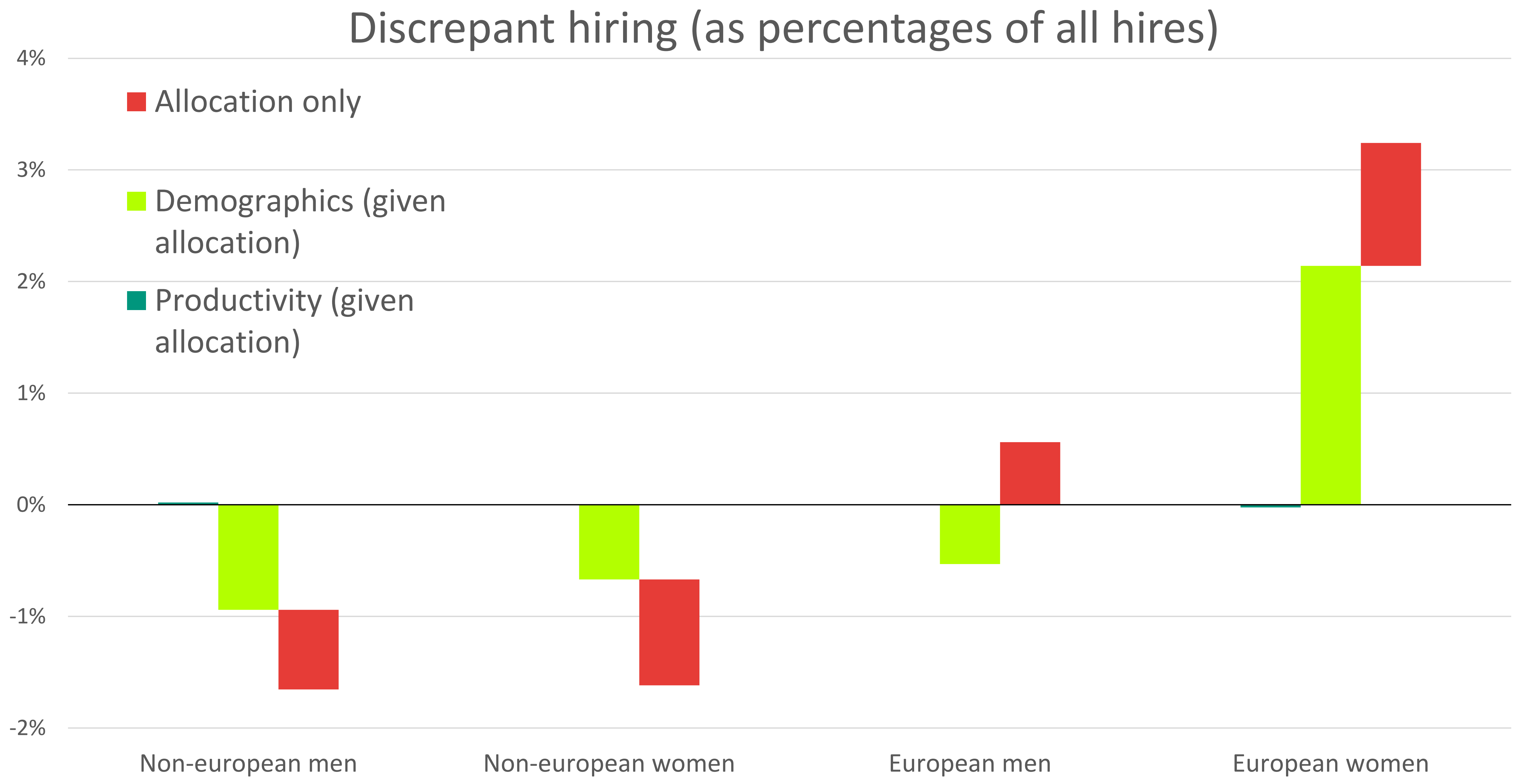


Applicants choose **which listings** to apply to…

… and employers choose among applicants in **their** pool

Example 2: Labour market discrimination with Multinomial Logit

The effect of applicant self-sorting is analysed by way of simulation under as-is allocation, and fully random allocation



Discrepant hiring (as percentages of all hires)
4%
3%
2%
1%
0%
-1%
-2%
Allocation only
Demographics (given allocation)
Productivity (given allocation)
Non-european men
Non-european women
European men
European women

Discrepant hiring (as percentages of own group size)