Statistical Inference for Maximin Effects: Identifying Stable Associations across Multiple Studies

Zijian Guo

Rutgers University

Overview of talk

1. Group Distributionally Robust Models
2. Identification and Inference Challenge
3. Method and Theory
4. Numerical Studies

Multi-source data is everywhere

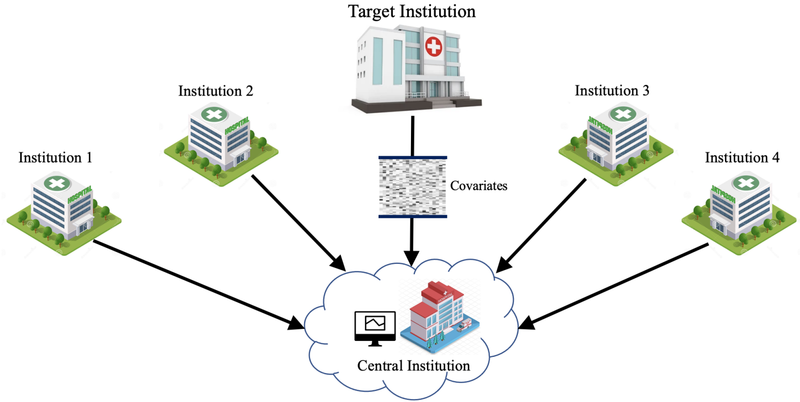


Figure: Figure credited to the Cai Lab at Harvard.

Invariance from Mutli-source Data

- Data heterogeneity: sub-populations, healthcare centers and environments.
- Standard model averaging: data merging, meta-analysis, and federated learning
- Today's focus: group distributionally robust models
  - Heterogeneity: allowing for distributional shifts
  - Generalizability: capturing stable associations shared across different sites

Setups

For $1 \leq l \leq L$, the training data $\{X_i^{(l)}, Y_i^{(l)}\}_{1 \leq i \leq n_l}$ follows,

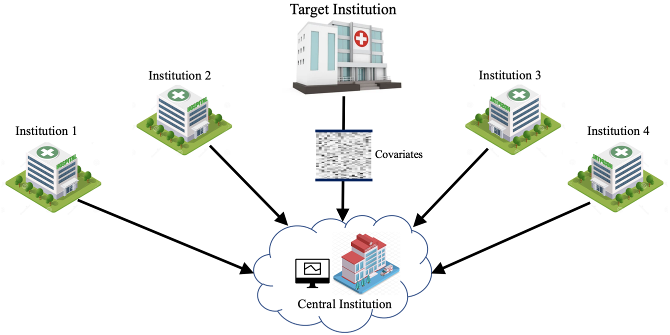
$$X_i^{(l)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X^{(l)}, \quad Y_i^{(l)} \mid X_i^{(l)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{Y|X}^{(l)} \quad \text{for} \quad 1 \leq i \leq n_l.$$

We consider the target population,

$$X_i^{\mathbb{Q}} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_X, \quad Y_i^{\mathbb{Q}} \mid X_i^{\mathbb{Q}} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_{Y|X} \quad \text{for} \quad 1 \leq i \leq N_{\mathbb{Q}},$$

1. No observations of $\{Y_i^{\mathbb{Q}}\}_{1 \leq i \leq N_{\mathbb{Q}}}$.
2. Covariate shift: $\mathbb{Q}_X$ is different from $\{\mathbb{P}_X^{(l)}\}_{1 \leq l \leq L}$.
3. $\mathbb{Q}_{Y|X}$ is different from $\{\mathbb{P}_{Y|X}^{(l)}\}_{1 \leq l \leq L}$.

A class $\mathcal{C}(\mathbb{Q}_X)$ of target populations



$$\mathcal{C}(\mathbb{Q}_X) = \left\{ \mathbb{T} = (\mathbb{Q}_X, \mathbb{T}_{Y|X}) : \mathbb{T}_{Y|X} = \sum_{l=1}^{L} q_l \cdot \mathbb{P}^{(l)}_{Y|X} \quad \text{with} \quad q \in \Delta_L \right\},$$

where $\Delta_L := \left\{ q \succ 0 : \sum_{l=1}^{L} q_l = 1 \right\}$.

Group Distributionally Robust Models

If the test data $\{X_i, Y_i\} \sim \mathbb{T}$, define the reward function for β,

$$\mathbf{E}_{\mathbb{T}} Y_i^2 - \mathbf{E}_{\mathbb{T}} (Y_i - X_i^{\intercal}\beta)^2.$$

The worst-case or adversarial reward,

$$R_{\mathbb{Q}}(\beta) = \min_{\mathbb{T} \in \mathcal{C}(\mathbb{Q}_X)} \left\{ \mathbf{E}_{\mathbb{T}} Y_i^2 - \mathbf{E}_{\mathbb{T}} (Y_i - X_i^{\intercal}\beta)^2 \right\}.$$

The group distributionally robust model

$$\beta^*(\mathbb{Q}) := \arg\max_{\beta \in \mathbb{R}^p} R_{\mathbb{Q}}(\beta).$$

Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

Covariate-shift Maximin Effect

If $\mathbb{P}_X^{(l)} = \mathbb{Q}_X$, then $\beta^*(\mathbb{Q})$ is reduced to the maximin effect

$$\beta^* := \arg\max_{\beta \in \mathbb{R}^p} R(\beta) \tag{1}$$

$$R(\beta) = \min_{1 \le l \le L} \left\{ \mathbf{E}[Y_1^{(l)}]^2 - \mathbf{E}[Y_1^{(l)} - (X_1^{(l)})^\intercal \beta]^2 \right\}.$$

Meinshausen, N., & Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. The Annals of Statistics, 43(4), 1801-1830.

Overview of talk

1. Group Distributionally Robust Models
2. Identification and Inference Challenge
3. Method and Theory
4. Numerical Studies

Identification

$$Y^{(l)}_{n_l\times 1} = X^{(l)}_{n_l\times p} b^{(l)}_{p\times 1} + \epsilon^{(l)}_{n_l\times 1}, \quad \text{for} \quad 1 \leq l \leq L.$$

Theorem (Meinshausen & Bühlmann (2015), G. (2020))

If $\lambda_{\min}(\Sigma^{\mathbb{Q}}) > 0$, *then* $\beta^*(\mathbb{Q})$ *is identified as*

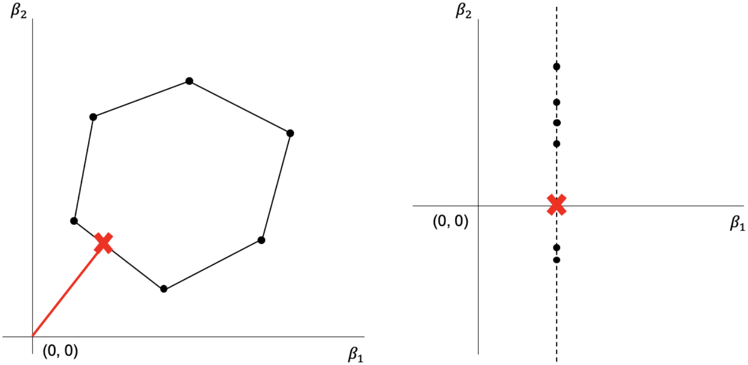
$$\beta^*(\mathbb{Q}) = \sum_{l=1}^{L} [\gamma^*(\mathbb{Q})]_l b^{(l)} \quad \textit{with} \quad \gamma^*(\mathbb{Q}) = \arg\min_{\gamma\in\Delta^L} \gamma^{\intercal}\Gamma^{\mathbb{Q}}\gamma \qquad (2)$$

where $\Delta^L = \{\gamma \in \mathbb{R}^L : \gamma_j \geq 0, \ \sum_{j=1}^{L} \gamma_j = 1\}$ *is the simplex over* $\mathbb{R}^L$ *and*

$$\Gamma^{\mathbb{Q}}_{lk} = (b^{(l)})^{\intercal}\Sigma^{\mathbb{Q}} b^{(k)} \quad \textit{for} \quad 1 \leq l, k \leq L,$$

with $\Sigma^{\mathbb{Q}} = \mathbf{E} X_i^{\mathbb{Q}} [X_i^{\mathbb{Q}}]^{\intercal}$.

Generalizability: stable associations



Meinshausen & Bühlmann (2015)

Asymptotic normality fails

For two groups, the optimal aggregation weight $(\gamma_1^*, 1-\gamma_1^*)^\intercal$ with

$$\gamma_1^* = \min\left\{\max\left\{\frac{\Gamma^{\mathbb{Q}}_{22}-\Gamma^{\mathbb{Q}}_{12}}{\Gamma^{\mathbb{Q}}_{11}+\Gamma^{\mathbb{Q}}_{22}-2\Gamma^{\mathbb{Q}}_{12}}, 0\right\}, 1\right\}.$$

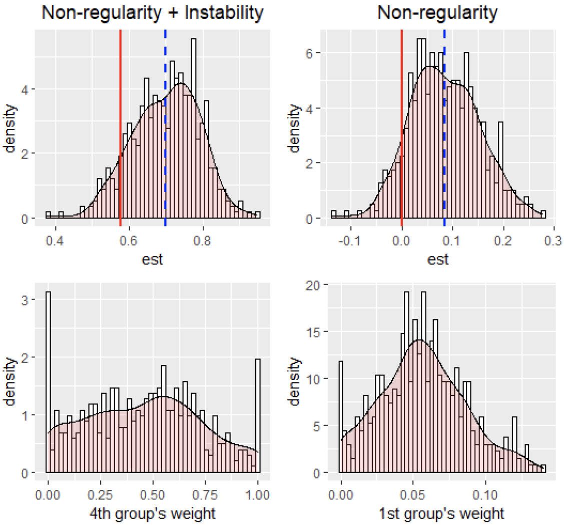
$$\widehat{\gamma}_1 = \min\left\{\max\left\{\frac{\widehat{\Gamma}^{\mathbb{Q}}_{22}-\widehat{\Gamma}^{\mathbb{Q}}_{12}}{\widehat{\Gamma}^{\mathbb{Q}}_{11}+\widehat{\Gamma}^{\mathbb{Q}}_{22}-2\widehat{\Gamma}^{\mathbb{Q}}_{12}}, 0\right\}, 1\right\}.$$

1. Non-regularity:

$$\frac{\Gamma^{\mathbb{Q}}_{22}-\Gamma^{\mathbb{Q}}_{12}}{\Gamma^{\mathbb{Q}}_{11}+\Gamma^{\mathbb{Q}}_{22}-2\Gamma^{\mathbb{Q}}_{12}} \approx 0 \quad \text{or} \quad 1.$$

2. Instability:

$$\Gamma^{\mathbb{Q}}_{11}+\Gamma^{\mathbb{Q}}_{22}-2\Gamma^{\mathbb{Q}}_{12} \approx 0.$$



Non-regularity + Instability

Non-regularity

Overview of talk

1. Group Distributionally Robust Models
2. Identification and Inference Challenge
3. Method and Theory
4. Numerical Studies

Two Sources of Randomness

$$\text{Target parameter:} \quad x_{\text{new}}^{\intercal}\beta^{*} = \sum_{l=1}^{L} \gamma_l^{*} \cdot x_{\text{new}}^{\intercal} b^{(l)}.$$

$$\text{Plugin estimator:} \quad \widehat{x_{\text{new}}^{\intercal}\beta^{*}} = \sum_{l=1}^{L} \widehat{\gamma}_l \cdot \widehat{x_{\text{new}}^{\intercal} b^{(l)}}$$

1. $\widehat{x_{\text{new}}^{\intercal} b^{(l)}} - x_{\text{new}}^{\intercal} b^{(l)}$ is asymptotically normal
2. Even if $\widehat{\Gamma}^{\mathbb{Q}} - \Gamma^{\mathbb{Q}}$ is asymptotically normal, $\widehat{\gamma}_l$ may be non-normal

$$\widehat{\gamma} = \arg\min_{\gamma \in \Delta^{L}} \gamma^{\intercal} \widehat{\Gamma}^{\mathbb{Q}} \gamma$$

ReSampling

Note that

$$\mathrm{vecl}(\widehat{\Gamma}^{\mathbb{Q}}) - \mathrm{vecl}(\Gamma^{\mathbb{Q}}) \overset{d}{\approx} \mathcal{N}(\mathbf{0}, \widehat{\mathbf{V}}).$$

We resample

$$\left\{\mathrm{vecl}(\widehat{\Gamma}^{[m]})\right\}_{1\leq m\leq M} \overset{i.i.d.}{\sim} N\left(\mathrm{vecl}(\widehat{\Gamma}^{\mathbb{Q}}), \widehat{V} + d_0/n \cdot \mathbf{I}\right),$$

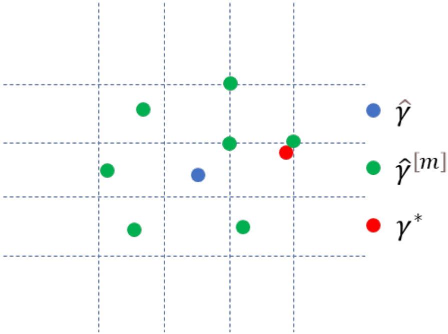
with $d_0 = \max\left\{\tau \cdot n\|\widehat{V}\|_{\infty}, 1\right\}$.

Construct the sampled weight vector as

$$\widehat{\gamma}^{[m]} = \arg\min_{\gamma\in\Delta^L} \gamma^{\intercal}\widehat{\Gamma}^{[m]}_{+}\gamma.$$

Xie, M. G., & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. International Statistical Review, 81(1), 3-39.

Sampling property



There exists $1 \leq m^{*} \leq M$ such that

$$\widehat{\gamma}^{[m^{*}]} \approx \gamma^{*}.$$

Sampling property

$$\widehat{x_{\rm new}^{\intercal}\beta}^{[m]} = \sum_{l=1}^{L} \widehat{\gamma}_l^{[m]} \cdot \widehat{x_{\rm new}^{\intercal} b^{(l)}} \quad \text{for} \quad 1 \leq m \leq M,$$

$$\widehat{x_{\rm new}^{\intercal}\beta}^{[m^*]} - x_{\rm new}^{\intercal}\beta = \sum_{l=1}^{L} (\widehat{\gamma}_l^{[m^*]} - \gamma_l^*) \cdot \widehat{x_{\rm new}^{\intercal} b^{(l)}} + \sum_{l=1}^{L} \gamma_l^* \cdot (\widehat{x_{\rm new}^{\intercal} b^{(l)}} - x_{\rm new}^{\intercal} b^{(l)}).$$

With $m = m^*$,

- The uncertainty of $\sum_{l=1}^{L} (\widehat{\gamma}_l^{[m^*]} - \gamma_l^*) \cdot \widehat{x_{\rm new}^{\intercal} b^{(l)}}$ is negligible
- Quantify the uncertainty of $\sum_{l=1}^{L} \gamma_l^* \cdot (\widehat{x_{\rm new}^{\intercal} b^{(l)}} - x_{\rm new}^{\intercal} b^{(l)})$

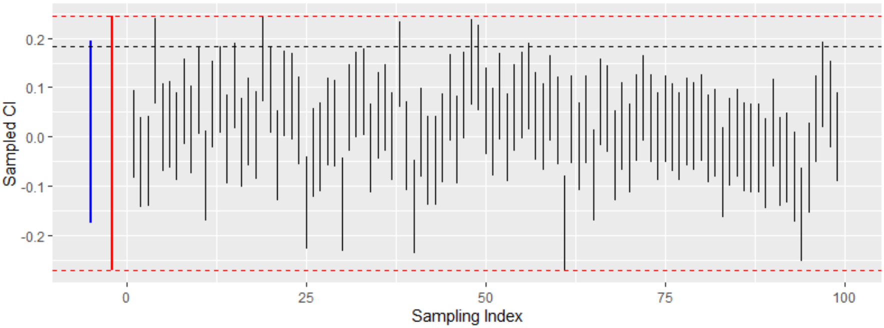
Proposed CI

For $1 \leq m \leq M$,

$$\mathrm{Int}_{\alpha}^{[m]}(x_{\mathrm{new}}) = \left(\widehat{x_{\mathrm{new}}^{\intercal}\beta}^{[m]} - z_{1-\alpha/2}\widehat{\mathrm{se}}^{[m]}(x_{\mathrm{new}}), \widehat{x_{\mathrm{new}}^{\intercal}\beta}^{[m]} + z_{1-\alpha/2}\widehat{\mathrm{se}}^{[m]}(x_{\mathrm{new}})\right)$$

$$\widehat{\mathrm{se}}^{[m]}(x_{\mathrm{new}}) = 1.01\sqrt{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}.$$

$$\mathrm{CI}_{\alpha}\left(x_{\mathrm{new}}^{\intercal}\beta^{*}\right) = \cup_m \mathrm{Int}_{\alpha}^{[m]}(x_{\mathrm{new}}).$$



- red: our proposal.
- blue: asymptotic normal CI with the oracle SE.
- **The confidence interval is of a parametric length.**

Overview of talk

1. Group Distributionally Robust Models
2. Identification and Inference Challenge
3. Method and Theory
4. **Numerical Studies**

Normality CI

We introduce the normality CI as

$$\left(\widehat{x_{\rm new}^{\intercal}\beta^{*}} - 1.96 \cdot \widehat{\rm SE}, \widehat{x_{\rm new}^{\intercal}\beta^{*}} + 1.96 \cdot \widehat{\rm SE}\right), \tag{3}$$

where $\widehat{\rm SE}$ denotes the sample SE calculated based on 500 simulations.

		Coverage		Length		
Setting	$\mathbb{I}(\delta)$	normality	Proposed	normality	Proposed	Length Ratio
(I-1)	3.368	0.700	0.960	0.352	0.597	1.693
(I-2)	3.707	0.818	0.978	0.320	0.543	1.699
(I-3)	3.182	0.748	0.970	0.352	0.588	1.673
(I-4)	1.732	0.770	0.956	0.520	0.796	1.532
(I-5)	1.857	0.796	0.978	0.445	0.710	1.594
(I-6)	1.987	0.710	0.980	0.480	0.832	1.732
(I-7)	0.029	0.848	0.985	0.250	0.507	2.028
(I-8)	0.031	0.758	0.981	0.262	0.530	2.020
(I-9)	0.010	0.830	0.988	0.690	1.264	1.832
(I-10)	0.030	0.940	0.988	0.232	0.315	1.354

Bootstrap or Sub-sampling

Setting		m-out-of-n subsampling				m-out-of-n bootstrap					Proposed	
$p = 30$	normality	$m = 200$	$m = 300$	$m = 400$	$m = 500$	$m = 200$	$m = 300$	$m = 400$	$m = 500$	$m = 1000$	Cov	L-ratio
(I-1)	0.686	0.380	0.390	0.380	0.408	0.432	0.460	0.502	0.490	0.536	0.956	1.678
(I-2)	0.808	0.418	0.456	0.474	0.446	0.456	0.510	0.532	0.546	0.602	0.990	1.723
(I-3)	0.770	0.392	0.494	0.464	0.440	0.482	0.480	0.516	0.544	0.634	0.984	1.766
(I-4)	0.816	0.620	0.668	0.670	0.668	0.672	0.694	0.710	0.700	0.794	0.990	1.686
(I-5)	0.790	0.612	0.626	0.626	0.594	0.628	0.664	0.702	0.732	0.732	1.000	1.843
(I-6)	0.806	0.590	0.636	0.654	0.632	0.626	0.682	0.698	0.712	0.760	0.994	1.833
(I-7)	0.824	0.914	0.932	0.908	0.888	0.890	0.950	0.934	0.950	0.952	0.996	5.078
(I-8)	0.888	0.912	0.934	0.884	0.856	0.914	0.916	0.932	0.932	0.958	0.996	4.017
(I-9)	0.900	0.834	0.822	0.788	0.778	0.914	0.914	0.916	0.862	0.914	0.996	2.061
(I-10)	0.954	0.956	0.904	0.866	0.836	0.956	0.948	0.962	0.964	0.942	1.000	1.725

Subsampling/bootstrap do not solve non-regular inference problems.

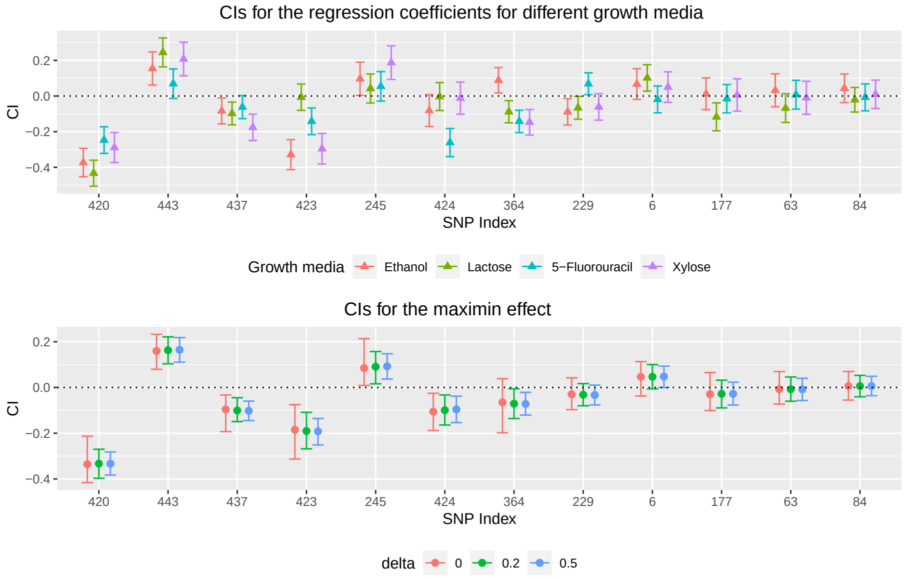
Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. Econometrica, 399–405.

Applications to GWAS

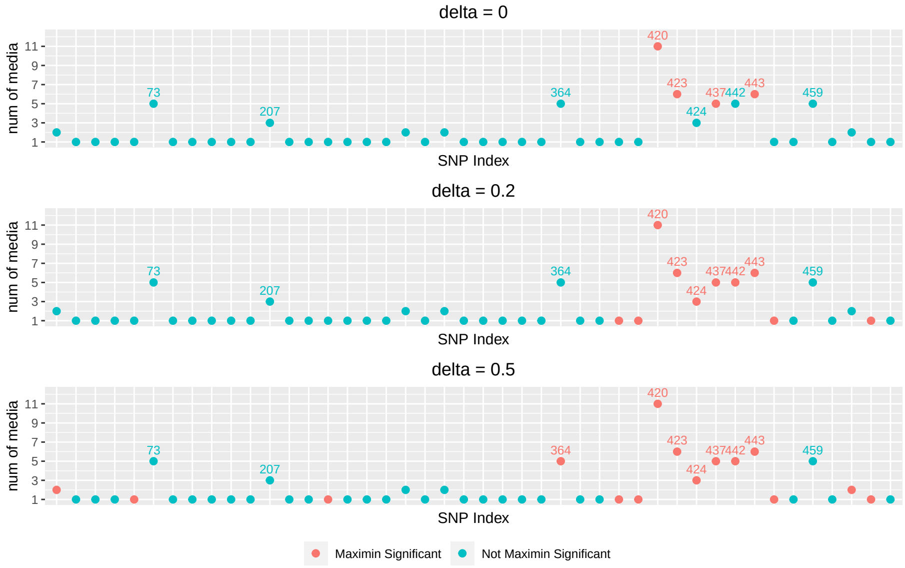
- Study the yeast colony growth under different growth media.
- Outcome: end-point colony sizes under different growth media.
- $p = 513$ genetic markers
- The study is based on $n = 1,008$ yeast segregants.
- Training media: “Ethanol", “Lactose", “5-Fluorouracil", “Xylose"

Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite, and L. Kruglyak (2013). Finding the sources of missing heritability in a yeast cross. Nature 494 (7436), 234–237.

Stable associations: causal?



Generalizability?



Surv-Maximin

- Study the COVID-19 mortality risk prediction.
- Analyze 15 participating health care centers across three countries (France, Germany, and the U.S.).
- The proposed SurvMaximin algorithm achieves higher accuracy than the existing methods.
- Privacy-preserving.

Wang, X., Zhou, H., $\cdots$, 4CE, Avillach, P., **Guo, Z.**, & Cai, Tianxi. (2022) SurvMaximin: Robust Federated Approach to Transporting Survival Risk Prediction Models. Journal of biomedical informatics 134 (2022): 104176.

Take home message

Multi-source data

- heterogeneity
- generalizability
- privacy-preserving.
- Non-standard inference: sampling
- $\cdots$

Generalizable model

- Maximin effect
- Group distributionally robust model
- $\cdots$

Acknowledgement

- Peter Bühlmann, Tianxi Cai, Nicolai Meinshausen, and Minge Xie.
- Zhenyu Wang: help with simulation studies.
- Molei Liu, Zhonghua Liu, Jue Hou, and Rong Ma: suggestions about the draft.

Acknowledgement to NSF, NIH, and ETH FIM for funding.

Guo, Z. (2020). Statistical Inference for Maximin Effects: Identifying Stable Associations across Multiple Studies. *arXiv preprint arXiv:2011.07568.*

R package `MaximinInfer`

Thank you!