Some Recent Advances and Open Problems in Post-Linkage Data Analysis

Martin Slawski[1], B. T. West[2], P. Bukke[3], E. Ben-David[4]

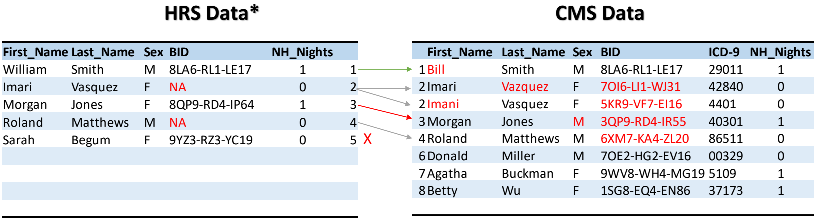
[1]University of Virginia, [2]University of Michigan,
[3]George Mason University, [4]U.S. Census Bureau

October 24, 2024
FCSM 2024

Secondary Analysis of Linked Files

Setting:

Researchers are interested in creating or enhancing a data set via record linkage (RL), but they may lack expertise, time, or access to conduct linkage themselves.



HRS Data*

First_Name	Last_Name	Sex	BID	NH_Nights	
William	Smith	M	8LA6-RL1-LE17	1	1
Imari	Vasquez	F	NA	0	2
Morgan	Jones	F	8QP9-RD4-IP64	1	3
Roland	Matthews	M	NA	0	4
Sarah	Begum	F	9YZ3-RZ3-YC19	0	5 X

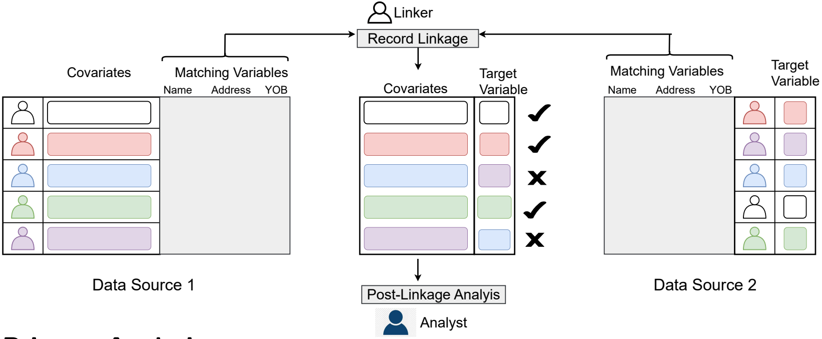
CMS Data

	First_Name	Last_Name	Sex	BID	ICD-9	NH_Nights
1	Bill	Smith	M	8LA6-RL1-LE17	29011	1
2	Imari	Vazquez	F	7OI6-LI1-WJ31	42840	0
2	Imani	Vasquez	F	5KR9-VF7-EI16	4401	0
3	Morgan	Jones	M	3QP9-RD4-IR55	40301	1
4	Roland	Matthews	M	6XM7-KA4-ZL20	86511	0
6	Donald	Miller	M	7OE2-HG2-EV16	00329	0
7	Agatha	Buckman	F	9WV8-WH4-MG19	5109	1
8	Betty	Wu	F	1SG8-EQ4-EN86	37173	1

*: Tables are fake and meant to be illustrative of matching complications.

Linkage is "outsourced" and reseachers operate on the linked file, which is taken at face value, i.e., the possibility of incorrect or missing links is not accounted for in the analysis.

RL and Post-Linkage Data Analysis (PLDA)



Primary Analysis:
Access to individual Data Sources 1 & 2. RL and subsequent data analysis can be performed jointly, with propagation of uncertainty.

Secondary Analysis (this talk):
Access only to the linked file, not the individual files. Information about underlying RL may be available, but limited.

The importance of the secondary setting is expected to increase. Data users may not be able or willing to perform linkage.

Consequences of Linkage Error & Adjustment Methods

Literature is heavily focused on false matches (mismatches); false non-matches are "argued away" using ignorability assumptions.

Mismatches tend to introduce data contamination, leading to attenuated relationships, biased parameter estimates, reduced model fit, inflated standard errors, etc.

(Neter et al., 1965; Scheuren & Winkler, 1997; Lahiri & Larsen, 2005; Wang et al., 2022; Chambers et al., 2023)

Desiderata for Adjustment Methods:

- Can be applied with no or minimum information about what happened at the RL stage. Common pieces of information are:
  - Quality score for each linked record,
  - Surrogates for such score,
  - Block indicators & blockwise error rates (or estimates thereof).
- Statistical Inference & Efficiency,
- Scalability (roughly linear in #data points).

Very brief literature review: Secondary Analysis

Approach 1 – “Weighting”:

Lahiri & Larsen (2005); Han & Lahiri (2019) consider regression analysis with $\mathbf{x}$ in File A and y in File B. The expected value $\mathbf{Q} = \mathbf{E}[\Pi^*]$ of the unknown correct linkage configuration Π^* is supposed to be known.

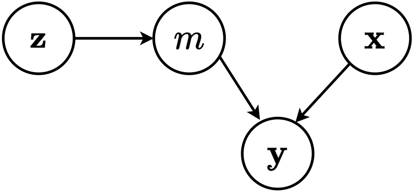
Chambers (2009); Chambers *et al.* (2023) build on Lahiri & Larsen (2005), simplifying $\mathbf{Q}$ according to an ELE (exchangeable linkage eror) model within blocks.

$$\begin{array}{cc} & \text{File B} \\ \text{File A} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \qquad \begin{pmatrix} 1-\alpha_b & \lambda_b & \cdots & \cdots & \lambda_b \\ \lambda_b & 1-\alpha_b & \ddots & & \vdots \\ \vdots & \ddots & 1-\alpha_b & \ddots & \lambda_b \\ \lambda_b & \cdots & \cdots & \lambda_b & 1-\alpha_b \end{pmatrix}$$

Latent linkage configuration Π^* ELE model (for a given block b)

Mixture model approach at a glance

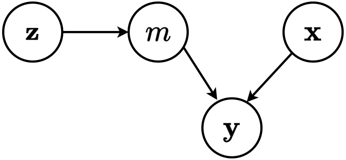
Approach 2 – “Mixture Modeling”: Hof & Zwinderman (2014), Gutman *et al.* (2016), Slawski *et al.* (2021), Slawski *et al.* (2024).



$$\mathbf{y}|\{m=1\},\mathbf{x} \sim f_{\mathbf{y}} \qquad \mathbf{y}|\{m=0\},\mathbf{x} \sim \phi(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$$

- $\mathbf{x}$ in file A, $\mathbf{y}$ in file B; (regression) parameter of interest $\boldsymbol{\theta}$,
- Latent binary mismatch indicator m, (possibly) modeled conditionally on info about RL $\mathbf{z}$,
- “Standard model” for pair $(\mathbf{x},\mathbf{y})$ if associated $m=0$ (right),
- Independence model $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$ if associated $m=1$ (left).

Mixture model approach: assumptions



$\mathbf{y}|\{m=1\},\mathbf{x} \sim f_{\mathbf{y}}$ $\qquad$ $\mathbf{y}|\{m=0\},\mathbf{x} \sim \phi(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$

Assumption 1 – Independence for mismatches: $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \,|\, m=1$

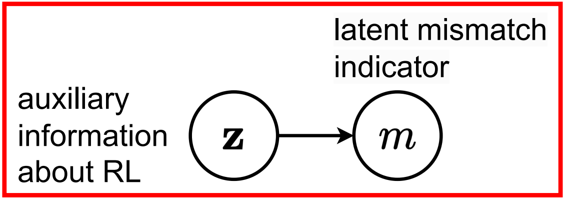
Satisfied if distinct records are independent. Can be violated if mismatches occur within correlated blocks of observations.

Assumption 2 – Mismatch error does not depend on $(\mathbf{x},\mathbf{y})$

The models for m and for $(\mathbf{x},\mathbf{y})$ are kept strictly separate.

m only depends on $\mathbf{z}$ but not on $\mathbf{x}$. This assumption is strong but renders inference more tractable. In particular, it implies that

$$f(\mathbf{y}|m=1) = f(\mathbf{y}|m=0).$$



The covariates $\mathbf{z}$ for the latent indicator m can be the following:

- … An intercept – corresponding to a constant mismatch rate model,
- … Block indicators from RL – corresponding to mismatch rates varying across blocks,
- … Output from probabilistic RL (e.g., confidence in the correctness of a match),
- … Comparison variables used during probabilistic RL.

Inference

Maximize the composite likelihood resulting from the postulated model with respect to the unknown parameters:

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \{\, \phi(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})\, \mathbf{P}(m_i = 0|\mathbf{z}_i;\boldsymbol{\gamma}) + f_{\mathbf{y}}(\mathbf{y}_i)\, \mathbf{P}(m_i = 1|\mathbf{z}_i;\boldsymbol{\gamma})\}$$

Inference (standard errors etc.) via asymptotic theory for composite maximum likelihood estimators (Varin *et al.*, 2011).

The framework can be applied to various statistical models:

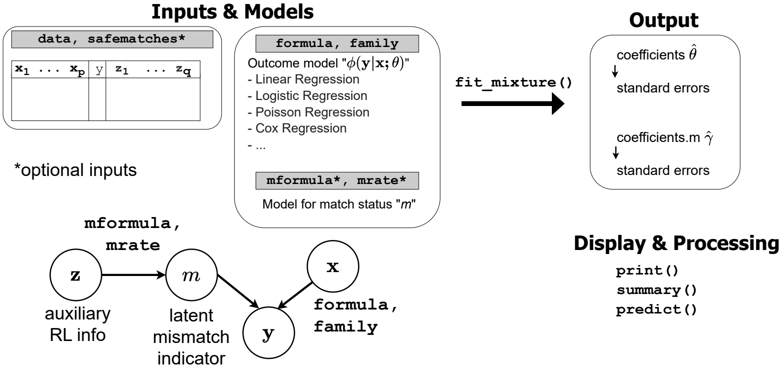
GLMs	Slawski *et al.*, 2024
Splines	
Contingency Tables	
Cox PH	Bukke *et al.*, 2024
Small Area Models	Fabrizi *et al.*, 2024
Random Forest	Ben-David *et al.*, 2024
Causal Inference	in progress
Penalized Regression	in progress

Implementation

Bukke, P., Wang, Z., Slawski, M., West, B. T., Ben-David, E. and Diao, G. (2024). `pldamixture`: Post-Linkage Data Analysis Based on Mixture Modelling. R Package. Version 0.1.1.

https://CRAN.R-project.org/package=pldamixture

https://github.com/bpriy/pldamixture



Inputs & Models

`data, safematches*`

$\mathbf{x}_1 \ \dots \ \mathbf{x}_p$	y	$\mathbf{z}_1 \ \dots \ \mathbf{z}_q$

*optional inputs

`formula, family`

Outcome model "$\phi(\mathbf{y}|\mathbf{x};\theta)$"

- Linear Regression
- Logistic Regression
- Poisson Regression
- Cox Regression
- ...

`mformula*, mrate*`

Model for match status "*m*"

`fit_mixture()`

Output

coefficients $\hat{\theta}$

standard errors

coefficients.m $\hat{\gamma}$

standard errors

`mformula, mrate`

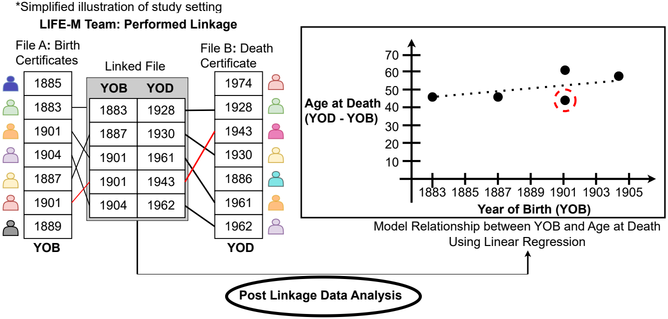
`formula, family`

Display & Processing

```
print()
summary()
predict()
```

Quick Demo

Life-M project: Longitudinal Intergenerational Family Electronic Micro-Database (`life-m.org`).



*Simplified illustration of study setting

LIFE-M Team: Performed Linkage
File A: Birth Certificates
Linked File
File B: Death Certificate
YOB
YOD
Age at Death (YOD - YOB)
70
60
50
40
30
20
10
1883
1885
1887
1889
1901
1903
1905
Year of Birth (YOB)
Post Linkage Data Analysis

File A: Birth Certificates (YOB)
1885
1883
1901
1904
1887
1901
1889

YOB	YOD
1883	1928
1887	1930
1901	1961
1901	1943
1904	1962

File B: Death Certificate (YOD)
1974
1928
1943
1930
1886
1961
1962

Model Relationship between YOB and Age at Death Using Linear Regression

The Life-M team used a hybrid of two RL procedures:

- "hand-linked" – clerically reviewed RL,
- "machine-linked" – automated probabilistic RL (anticipated mismatch rate ∼5%).

Quick Demo

*Note: R Package has demo data which is a subset of this full data.

```r
# 156,453 records w/ 6 variables
lifem <- read.csv("lifem.csv")
head(lifem, n = 4)
```

```
   yob   unit_yob age_at_death                    hndlnk commf comml
1 1905 0.95652174           83     Purely Machine-Linked  0.77  0.45
2 1883 0.00000000           79 Hand-Linked At Some Level  0.93  0.08
3 1886 0.13043478           58     Purely Machine-Linked  0.89  0.80
4 1885 0.08695652           58     Purely Machine-Linked  0.72  0.42
```

`commf`, `comml`: “commonness” scores of first and last name. It is a ratio of the log count of name in the 1940 Census and the log count of the most frequently used name in the 1940 Census.

Quick Demo

Model:

1. $y_i \mid m_i = 0, x_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \sigma^2)$,
2. $y_i \mid m_i = 1, x_i \sim N(\mu, \tau^2)$,
3. $m_i \mid \mathsf{commf}_i, \mathsf{comml}_i \sim \mathrm{Bernoulli}\left(\dfrac{\exp(\gamma_0 + \gamma_1 \mathsf{commf}_i + \gamma_2 \mathsf{comml}_i)}{1 + \exp(\gamma_0 + \gamma_1 \mathsf{commf}_i + \gamma_2 \mathsf{comml}_i)}\right)$.
4. Overall mismatch rate assumed to be $\leq 5\%$.
5. Hand-linked records are assumed to be correct links.

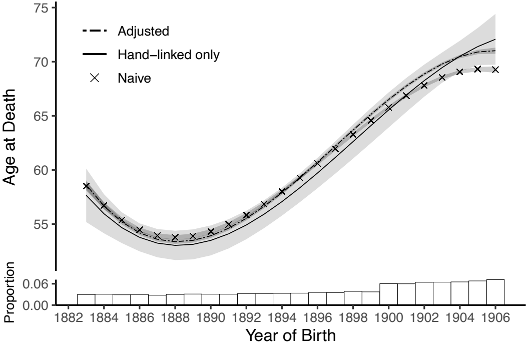
```
library(pldamixture)
fit <- fit_mixture(age_at_death ~ poly(unit_yob, 3, raw = TRUE),
                   data = lifem, family = "gaussian",
                   mformula = ~commf + comml,
                   mrate = 0.05,
                   safematches=ifelse(
                   lifem$hndlnk=="Hand-Linked At Some Level",
                                        TRUE, FALSE))
summary(fit)
```

Summary of Results:

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}$	$\widehat{\gamma}_0$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$
Naive	58.5(.2)	-46.7(1.8)	130.4(4.0)	-72.9(2.5)	21.2(.1)			
Adj‡	58.6(.1)	-51.0(1.5)	140.3(3.9)	-76.8(2.6)	20.7(.1)	-6.0(.5)	-1.5(.6)	7.2(.3)
Adj†	58.7(.2)	-52.5(1.6)	143.2(3.9)	-77.7(2.7)	20.4(.1)	-4.9(.4)	-1.4(.4)	6.1(.3)
HL*	57.7(1.3)	-44.2(11.6)	118.6(27.9)	-59.9(18.5)	19.0(.3)			

Adj‡: proposed, assuming mismatch rate $\leq$5%

Adj†: proposed, assuming mismatch rate $\leq$7.5%, HL: hand-linked only.



Open Problem I: beyond strongly non-informative mismatch error

Allowing m to depend on $\mathbf{x}$:
We are interested in eliminating the separation into two (independent) sets of covariates $\mathbf{x}$ and $\mathbf{z}$ for the outcome and mismatch indicator models, respectively.

This separation can often be limiting in applications. Its purpose is to achieve that $f(\mathbf{y}|m = 0) = f(\mathbf{y}|m = 1)$.

More generally, we can distinguish the following scenarios:

- SN: strongly non-informative linkage error – depends neither on $\mathbf{x}$ and $\mathbf{y}$.
- NL: non-informative linkage error – depends on $\mathbf{x}$ (only).
- WNL: weakly non-informative linkage error – depends on $\mathbf{x}$ and $\mathbf{y}$.
- IL: informative linkage – linkage error depends on on other possibly unobserved variables (correlated with $\mathbf{x}$ and $\mathbf{y}$).

Open Problem II: Missing Links and Mismatches

Suppose that some $\mathbf{x}$'s cannot be linked to any of the $\mathbf{y}$'s.
Let δ denote the corresponding indicator variable ($\delta = 1$ if linked).

Among the successfully linked data, we might still have mismatches. Assuming that $\delta \perp\!\!\!\perp m | \{\mathbf{x}, \mathbf{y}\}$, one possible approach is to employ the following likelihood contributions:

$$(i)\ f(\mathbf{y}, \delta = 1, m = 0 | \mathbf{x}) = \boxed{\mathbf{P}(\delta = 1 | \mathbf{x}, \mathbf{y}; \boldsymbol{\phi})} \cdot \boxed{f(\mathbf{y}, m = 0 | \mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\gamma})},$$

$$(ii)\ f(\mathbf{y}, \delta = 1, m = 1 | \mathbf{x}) = \boxed{\int \mathbf{P}(\delta = 1 | \mathbf{x}, \mathbf{y}; \boldsymbol{\phi})\, dP(\mathbf{x}) \times} \times \boxed{f(\mathbf{y}, m = 1; \boldsymbol{\theta}, \boldsymbol{\gamma})},$$

$$(iii)\ f(\delta = 0, \mathbf{x}) = \int \mathbf{P}(\delta = 0 | \mathbf{x}, \mathbf{y}; \boldsymbol{\phi}) \cdot f(\mathbf{y} | \mathbf{x})\, d\mathbf{y},$$

The terms inside $\boxed{\ \ldots\ }$ can be decomposed according to the mixture model as presented earlier.

Open Problem III: PLDA without RL?

RL may not always be feasible:

- Identifiers may not be shared,
- Respondents do not provide consent for linkage,
- Linkage is not possible for logistical reasons.

Statistical Matching (D'Orazio *et al.*, 2006) may be used, but it relies on a strong conditional independence assumption.

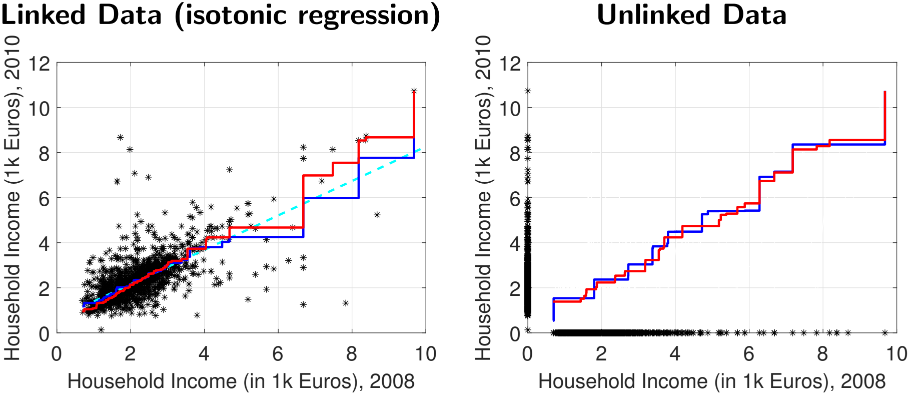
Unlinked regression is a more recent paradigm for performing regression without ever linking responses and predictors (Carpentier & Schlüter, 2016; Rigollet & Weed, 2019; Balabdaoui *et al.*, 2021; Slawski & Sen, 2022; Azadkia & Balabdaoui, 2022).

$$\begin{tikzcd}
X \sim \mu \arrow[d] & Y \overset{\mathcal{D}}{=} f^*(X) + \varepsilon \arrow[d] \\
\{X_i\}_{i=1}^{n} & \{Y_j\}_{j=1}^{m}
\end{tikzcd}$$

- X's are generated according to some distribution μ.
- Y is equal in distribution $\overset{\mathcal{D}}{=}$ to a transformation f^* of X plus noise ε.

Unlinked Regression: Illustration

Taken from the Italian Survey of Household Income & Wealth (SHIW). For unlinked regression, we use the method in Slawski & Sen, 2022.



Linked Data (isotonic regression)

Unlinked Data

blue: Assuming Gaussian noise. red: Assuming Laplacian noise.

cyan (dashed): Least squares regression line.

Acknowledgments & Supporting Materials

Acknowledgment:
Martin Slawski, Brady T. West, and Priyanjali Bukke were partially supported by NSF award SES–2120318.

Papers:
Mixture Model – `arXiv:2306.00909` JRSS-A, forthcoming.
Small Area Estimation – `arXiv:2405.20149`
Unlinked Regression – `arXiv:2201.03528`; JMLR, 25 (2024), 1–57.

References

- Slawski, Diao, Ben-David, "A Pseudo-Likelihood Approach to Linear Regression with Partially Shuffled Data", *JCGS*, 2021.
- Wang, Ben-David, Diao, Slawski, "Regression with linked data sets subject to linkage error", *WIREs Computational Statistics*, 2022.
- Wang, Ben-David, Slawski, "Regularization for Shuffled Data Problems via Exponential Family Priors on the Permutation Group", *AISTATS*, 2023.
- Neter, Maynes, Ramanathan, "The Effect of Mismatching on the Measurement of Response Errors", *JASA*, 1965.
- Scheuren & Winkler, "Regression Analyis of data files that are computer matched", *Surv Meth*, 1997.
- Lahiri & Larsen, "Regression Analysis with Linked Data", *JASA*, 2005.
- Chambers, Fabrizi, Ranalli, Salvati, Wang, "Robust regression using probabilistically linked data", *WIREs Computational Statistics*, 2023.
- Han & Lahiri, "Statistical Analysis with Linked Data", *Int Stat Rev*, 2019.
- Chambers, "Regression analysis of probability-linked data", *Technical Report, Statistics New Zealand*, 2009.
- Hof & Zwinderman, "A mixture model for the analysis of data derived from record linkage", *Stat Med.*, 2014.
- Gutman, Sammartino, Green, Montague, " Error adjustments for file linking methods using encrypted unique client identifier (euci) with application to recently released prisoners who are hiv+", *Stat Med.*, 2016.

References (Continued)

- Varin, Reid, Firth “An overview of composite likelihood estimation”, *Stat Sinica*, 2011.
- Bukke, Ben-David, Diao, Slawski, West. “Cox Proportional Hazards Regression using Linked Data: an Approach based on Mixture Modeling” *Springer IISA Series on Statistics and Data Science*, forthcoming.
- Ben-David, West, Slawski, “A Novel Methodology for Improving Applications of Modern Predictive Modeling Techniques to Linked Data Sets Subject to Mismatch Error” *IEEE Big Surv Proceedings*, 2023.
- Kim & Shao. “Statistical Methods of Handling Incomplete Data” *CRC press*, 2021.
- D’Orazio, Di Zio, Scanu. “Statistical Matching – Theory & Practice” *Wiley*, 2006.
- Carpentier & Schlüter. “Learning relationships between data obtained independently.” *AISTATS*, 2016.
- Rigollet & Weed. “Uncoupled isotonic regression via minimum Wasserstein deconvolution” *Information & Inference*, 2019.
- Balabdaoui et al. “Unlinked monotone regression” *JMLR*, 2021.
- Azadkia & Balabdaoui “Linear regression with unmatched data: a deconvolution perspective” *JMLR*, 2024.