LUCIDus: an R package to implement integrated clustering analysis

Yinqi Zhao

Advisor: Dr. David V. Conti

Division of Biostatistics
Department of Population and Public Health Sciences
University of Southern California

Outline

Part 1: Motivation and introduction of the LUCID model

Part 2: Overview of the *LUCIDus* R package

Part 3: An application of LUCID on liver injury

Outline

Part 1: Motivation and introduction of the LUCID model

Part 2: Overview of the *LUCIDus* R package

Part 3: An application of LUCID on liver injury

Data with more and more complex structure

Multi-view data:

1. Definition:

- A collection of datasets measured from multiple sources with complementary and consistent information
- Generated by large research consortium

2. Example:

The Human Early Life Exposome (HELIX) Study – to investigate environmental exposures (exposome) during early life and associate these exposures with molecular omics signatures and child health outcome

Human Early Life Exposome (HELIX) Study

Pregnancy

Child: 6-10

Exposome

- Outdoor exposures (GIS)
- Exposoure to chemicals (biomarker)
- Lifestyles (questionnaire)

Child: 6-10

Health outcome

Molecular omics data

Whole blood cells	DNA methylation data
	Transcriptome data
Plasma and serum	Metabolomics data
	Proteomics data
Urine	Metabolomics data

Covariates

Maternal:
cohort,
age,
education,
parity
...

Child:
sex,
gestational age,
weight,
height,
...

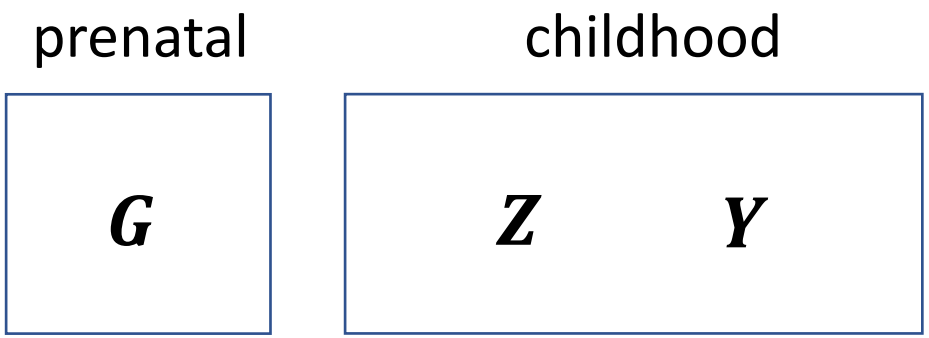
Inspired by ISGlobal 2021 data challenge

Maitre et al. *BMJ Open* (2018)

What questions can we ask?

Simplify multi-view data into notations:

1. $\boldsymbol{G}$ (exposure): environmental exposure
2. $\boldsymbol{Z}$ (omics data): metabolomics, proteomics
3. $\boldsymbol{Y}$ (outcome): liver injury

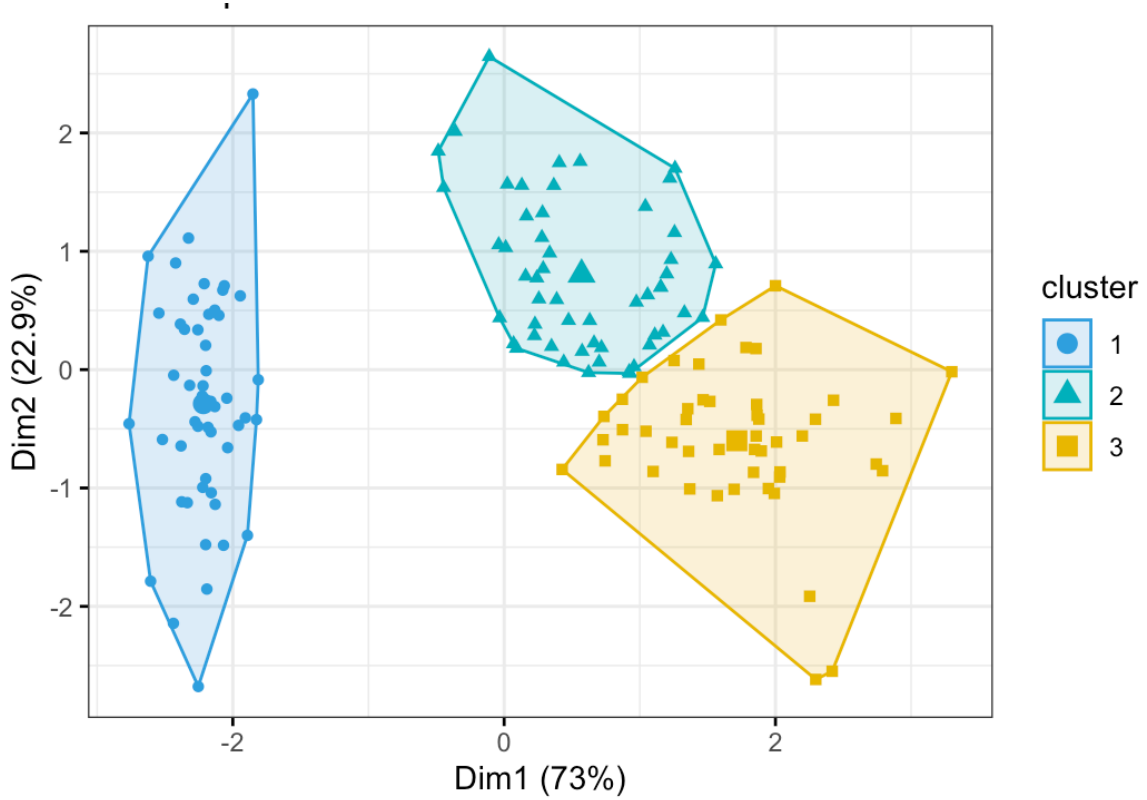


How to build an integrated clustering model to analyze the multi-view data $\boldsymbol{G}$, $\boldsymbol{Z}$, $\boldsymbol{Y}$ ($\boldsymbol{Cov}$), while adjusting for temporal sequence of measurements?

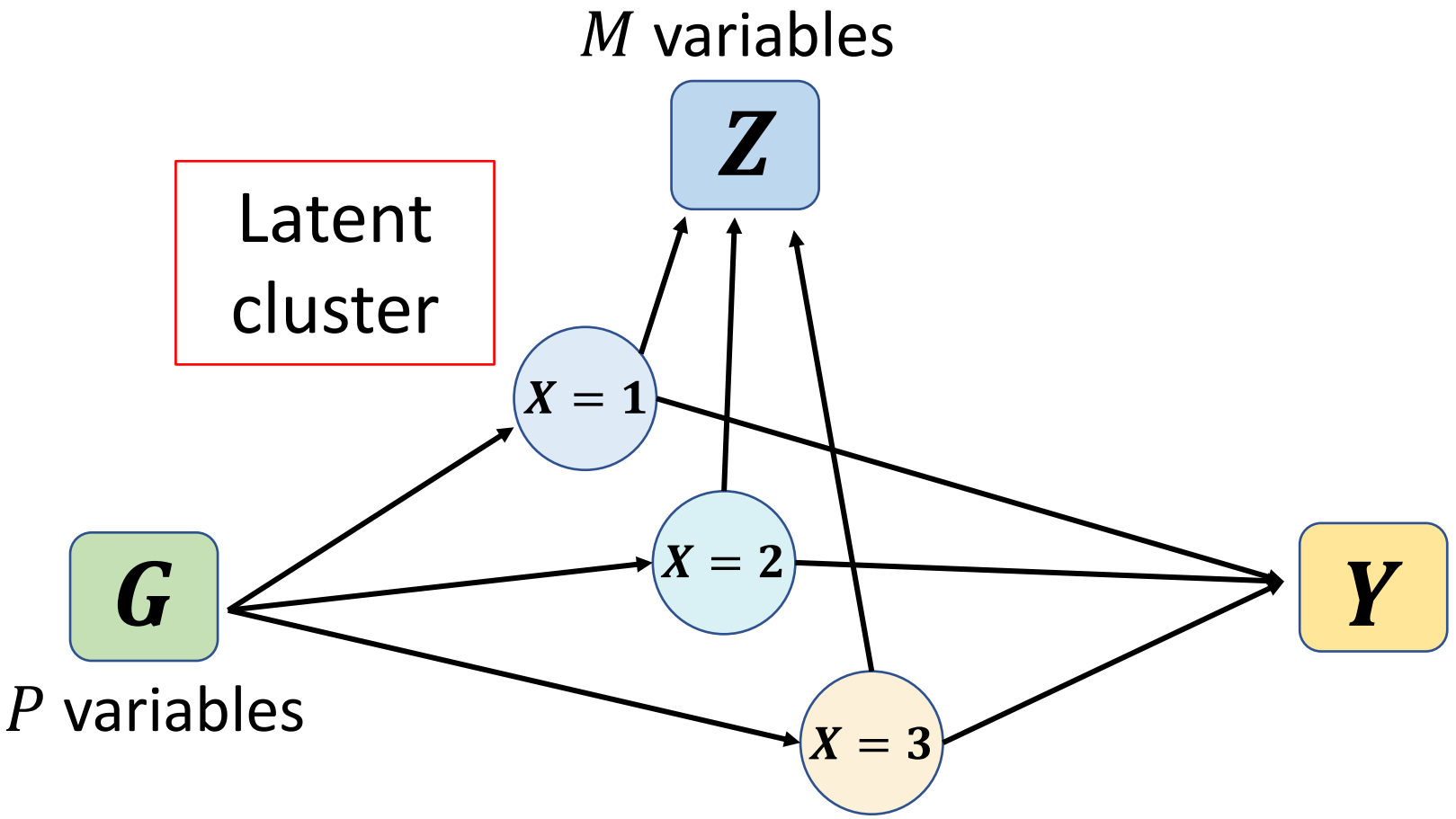
Clustering analysis

Aim:

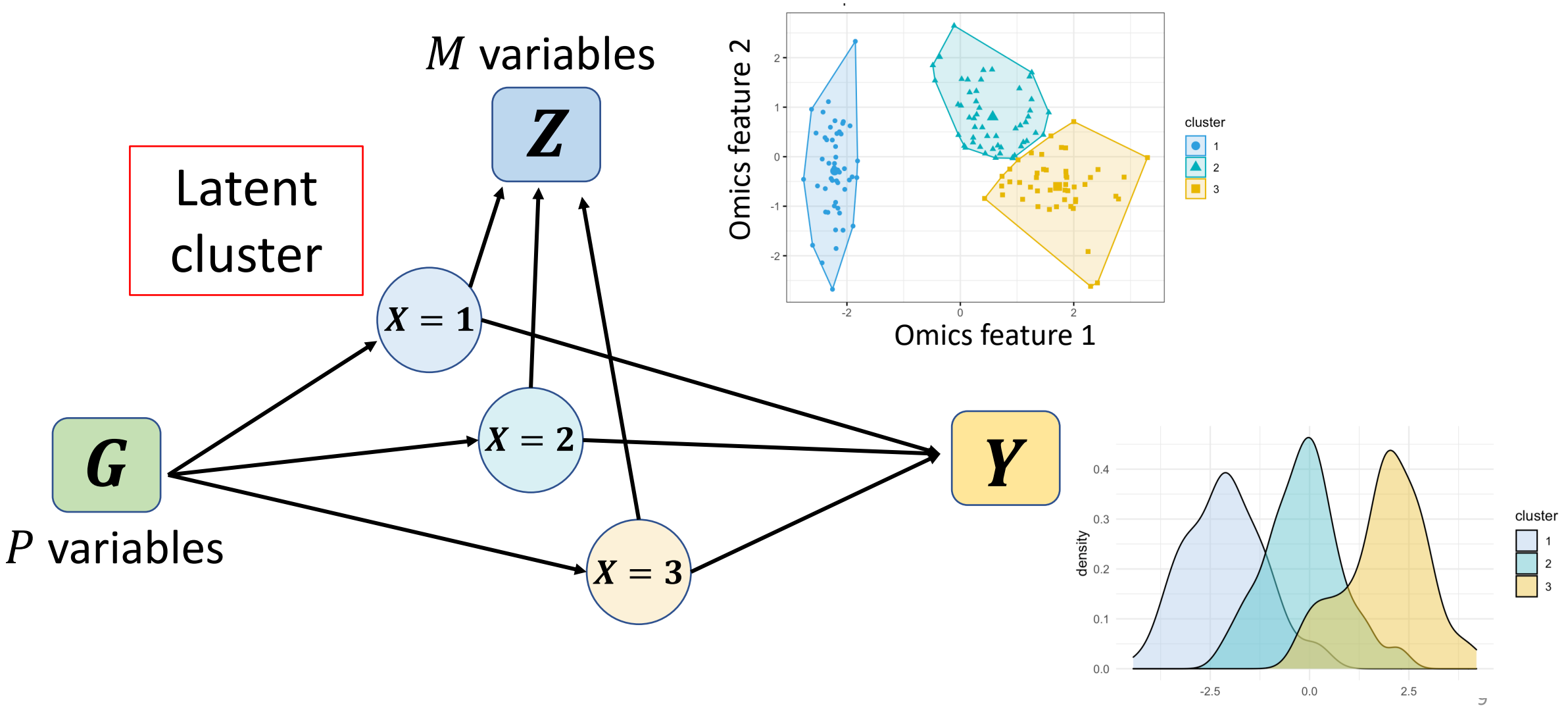
To divide samples into several groups, such that samples within a group are similar, and samples in different groups are dissimilar.



Integrated clustering



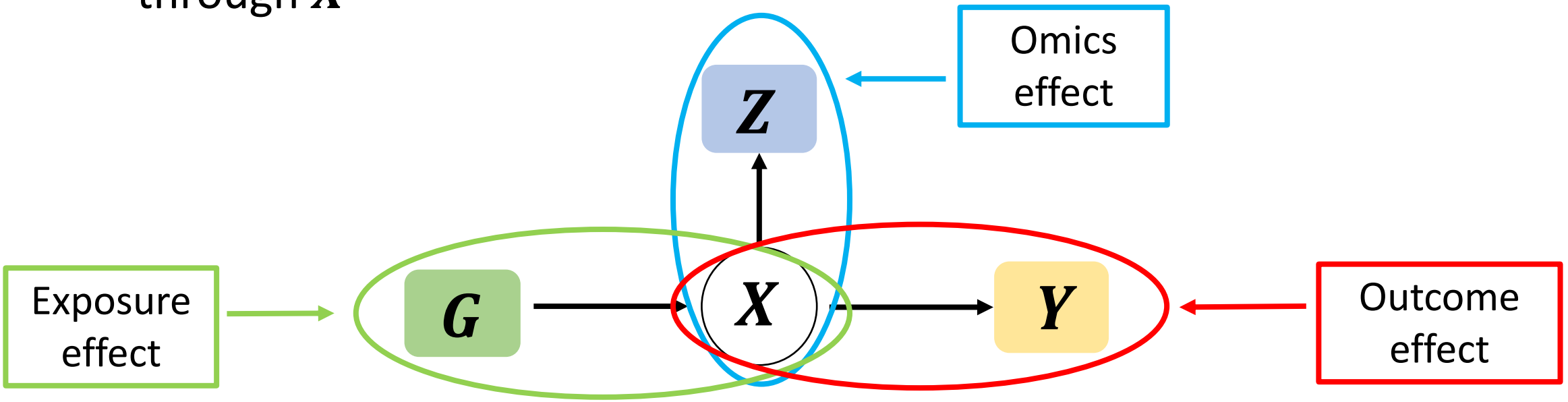
Integrated clustering



Latent Unknown Clustering Integrating omics Data (LUCID)

Aim:

1. Identify clusters ($\boldsymbol{X}$) characterized by $\boldsymbol{G}, \boldsymbol{Z}, \boldsymbol{Y}$
2. Estimate the association between exposure ($\boldsymbol{G}$) and outcome ($\boldsymbol{Y}$) through $\boldsymbol{X}$



Peng et al., *Bioinformatics* (2019)

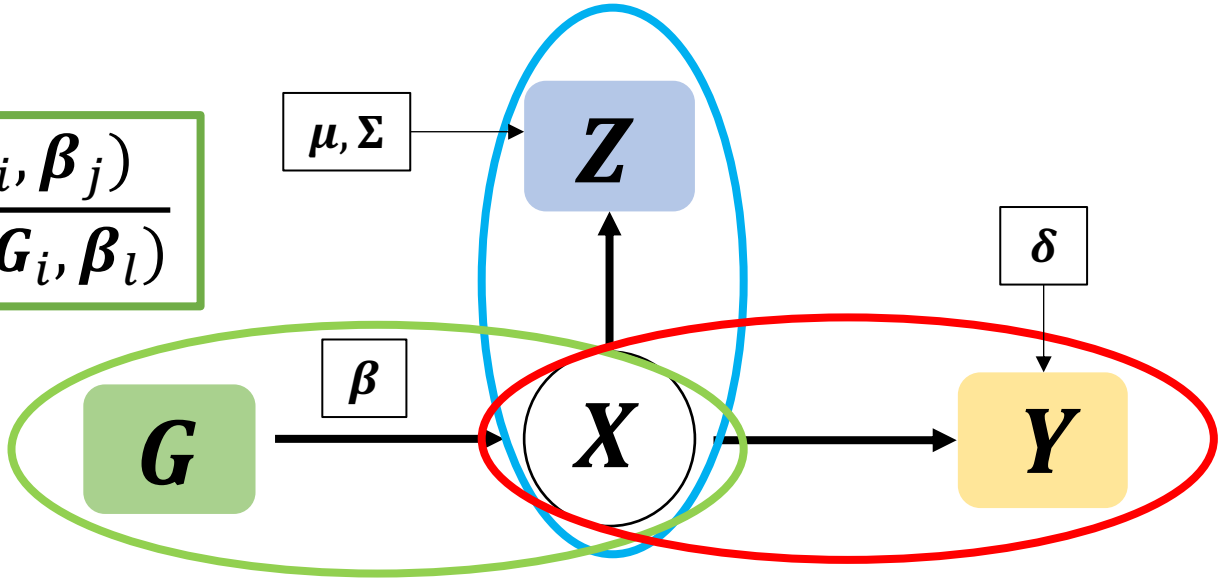
LUCID: Jointly model multi-view data via a latent variable

$$f(X_i = j|\boldsymbol{G}_i, \boldsymbol{\beta}) = S(X_i = j|\boldsymbol{G}_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{G}_i, \boldsymbol{\beta}_j)}{\sum_l \exp(\boldsymbol{G}_i, \boldsymbol{\beta}_l)}$$

$$f(\boldsymbol{Z}_i|X_i = j) \sim MVN(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

$$f(Y_i|X_i = j) \sim N(\delta_j, \sigma_j^2) \quad \textbf{or}$$

$$f(Y_i|X_i = j) = \frac{\exp(\delta_j)}{1+\exp(\delta_j)}$$



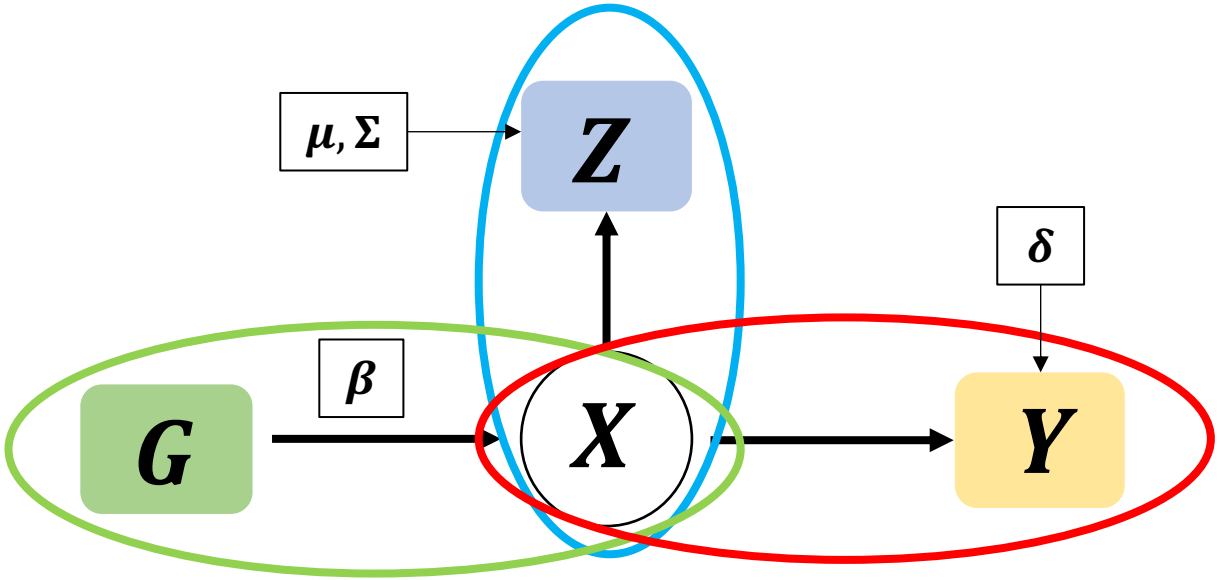
Assumption:

Conditional independence among $f(X|G)$, $f(Z|X)$ and $f(Y|X)$

LUCID: Jointly model multi-view data via a latent variable

The joint likelihood of the LUCID model

$$
\begin{aligned}
& l(\boldsymbol{\Theta}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} I(X_i = j) \log S\left(X_i = j | \boldsymbol{E}_{\mathrm{i}}, \boldsymbol{\beta}_j\right) \\
&+ \sum_{i=1}^{n} \sum_{j=1}^{k} I(X_i = j) \log \phi\left(\boldsymbol{M}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \\
&+ \sum_{i=1}^{n} \sum_{j=1}^{k} I(X_i = j) \log \phi\left(Y_i | \delta_j, \sigma_j^2\right)
\end{aligned}
$$



Assumption:

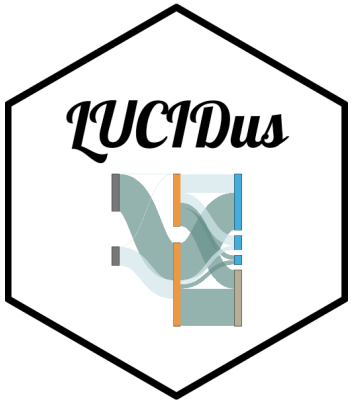
Conditional independence among $f(X|G), f(Z|X)$ and $f(Y|X)$

Outline

Part 1: Motivation and introduction of the LUCID model

Part 2: Overview of the *LUCIDus* R package

Part 3: An application of LUCID on liver injury



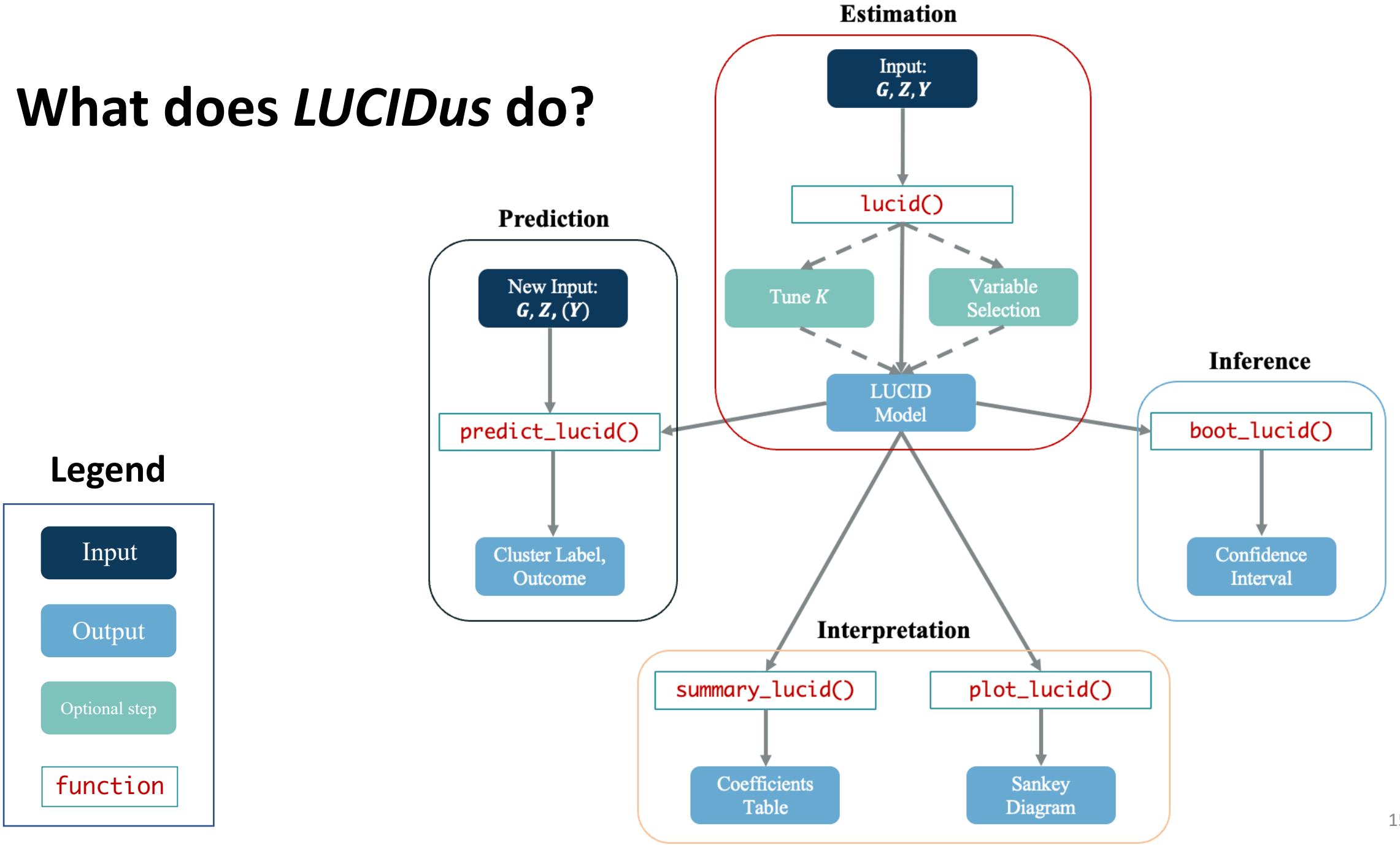
LUCIDus: an R package implementing LUCID

- Currently, version 2.2.0 is available on CRAN with **~ 19k downloads**
- Developer version on Github: [USCbiostats/LUCIDus: the new version of LUCID (github.com)]

Biostats IMAGE

USC Biostats P01: Integrative Methods of Analysis for Genetic Epidemiology

What does *LUCIDus* do?



Example: data

- A publicly available dataset: HELIX data challenge

```
library(LUCIDus)
# load data
data("helix_data")
exposome <- helix_data$exposure
proteomics <- helix_data$omics
zBMI <- helix_data$outcome["zBMI"]
```

`exposome`: a 100 x 8 data frame
`proteomics`: a 100 x 10 data frame
`zBMI`: a vector of length 100

Example: estimation `lucid()`

1. Fit LUCID model

```
> fit1 <- lucid(G = exposome, Z = proteomics, Y = zBMI,
family = "normal", K = 2)
```

2. Fit LUCID model, choose optimal number of clusters, K

```
> fit2 <- lucid(G = exposome, Z = proteomics, Y = zBMI,
K = 2:6)
```

3. Fit LUCID model, select informative exposures and omics variables

```
> fit3 <- lucid(G = exposome, Z = proteomics, Y = zBMI,
K = 2, Rho_G = 0.05, Rho_Z_Mu = 5, Rho_Z_Cov = 0.5)
Fitting LUCID model

3/8 exposures are selected

4/10 omics variables are selected
```

Example: interpretation `summary_lucid()`

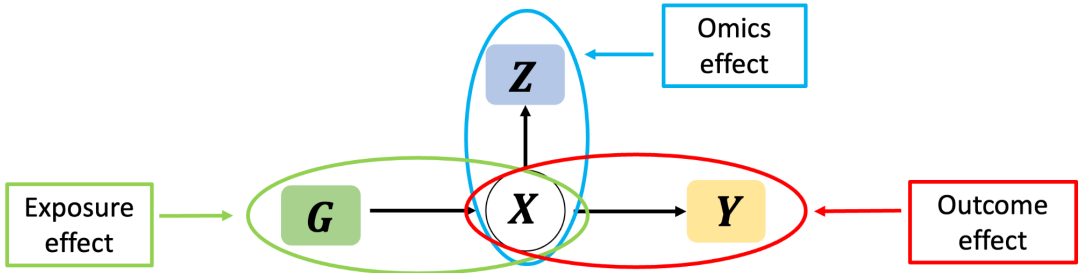
```
> summary_lucid(fit1)
----------Summary of the LUCID model----------

K =  2 , log likelihood = -1071.893 , BIC =  3216.791
```

```
(1) Y (continuous outcome): mean of Y for each latent cluster (and effect of
covariates if included)
                beta
cluster1 0.08771338
cluster2 0.86691317
```

```
(2) Z: mean of omics data for each latent cluster
        mu_cluster1 mu_cluster2
IL1beta -0.21282280  0.39620999
IL6     -0.25313089  0.47125114
INSULIN -0.31598417  0.58826444
IFNalfa  0.15370117 -0.28614387
IL1RA   -0.05274367  0.09819234
IL2      0.17884736 -0.33295828
IP10    -0.14934740  0.27803850
IL2R     0.13363312 -0.24878339
MIG      0.11645372 -0.21680069
IL4      0.15979421 -0.29748722
```

```
(3) E: odds ratio of being assigned to each latent cluster for each exposure
                        beta        OR
DDE_c.cluster2 -1.787901798 0.1673109
DDE_m.cluster2  0.367653805 1.4443419
DDT_c.cluster2  0.211258288 1.2352314
DDT_m.cluster2  0.004310885 1.0043202
HCB_c.cluster2  1.847153233 6.3417403
HCB_m.cluster2  0.174522198 1.1906772
PCB_c.cluster2 -0.751327088 0.4717401
PCB_m.cluster2 -0.027843969 0.9725401
```

- Summarize the LUCID model in a table of statistics:



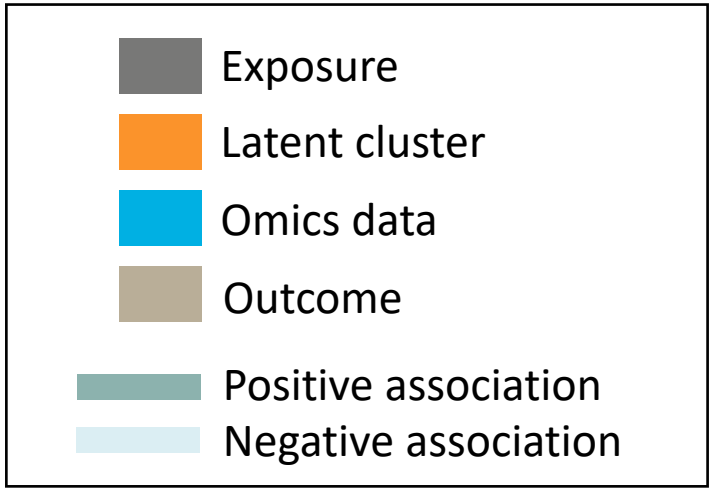
Example: interpretation `plot_lucid()`

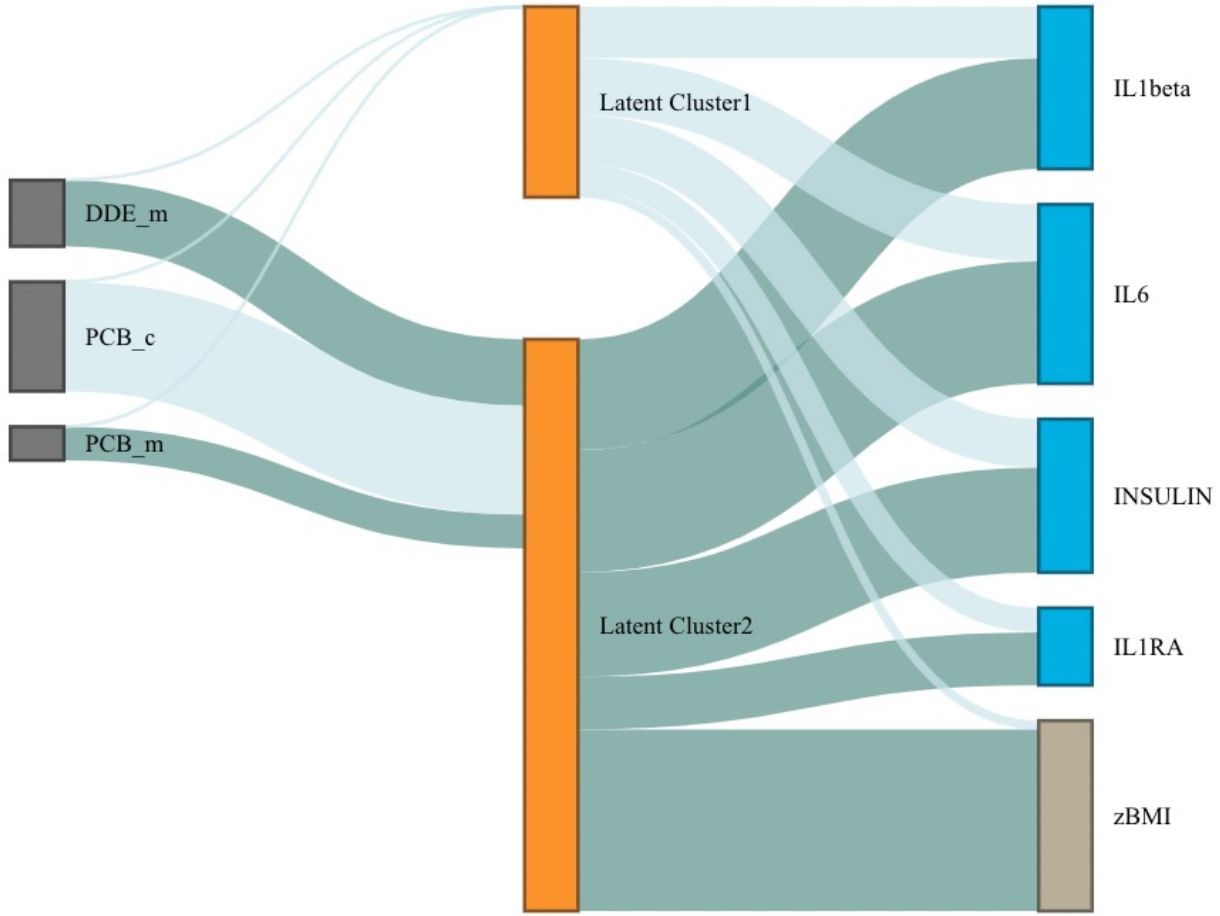
- Visualize the LUCID model

```
> plot_lucid(fit3)
```

Legend

(The colors are customizable!)





Example: prediction `pred_lucid()`

- Predicted cluster assignment (X) and outcome (Y)

```
> # predicted cluster label
> table(pred1$pred.x)
 1  2
66 34
> # predicted outcome
> pred1$pred.y[1:5]
[1] 0.86691245 0.21046383 0.08780161 0.09210515 0.10272194
```

Example: inference `boot_lucid()`

- Derive confidence intervals (CIs) given a confidence level

```
> boot1 <- boot_lucid(G = exposome, Z = proteomics, Y = zBMI,
model = fit1, R = 200)
Use Bootstrap resampling to derive 95% CI for LUCID
[==========>-----------------------------------------]  20%
```

Outline

Part 1: Motivation and introduction of the LUCID model

Part 2: Overview of the *LUCIDus* R package

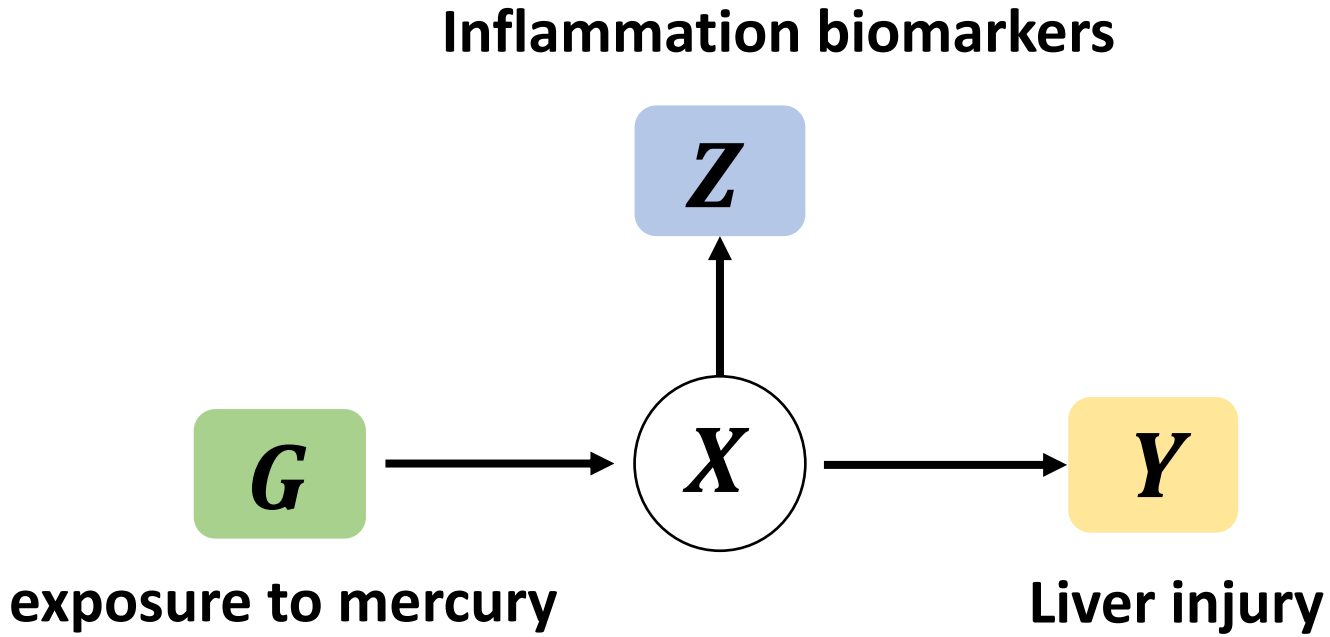
Part 3: An application of LUCID on liver injury

Study Background

- Mercury (Hg) is a ubiquitous toxic metal.
- Animal studies have show that Hg exposures increase liver enzyme levels including ALT
- There is no studies examining the hepatotocix effect of prenatal exposure to Hg.
- **Study population:** 872 mothers and their children
- **Exposure:** Hg concentration in maternal blood during pregnancy
- **Omics**: plasma concentrations of inflammation-related cytokines
- **Outcome:** elevated ALT (binary outcome, an indicator of livery injury)

Stratakis et al. *Hepatology* (2021)

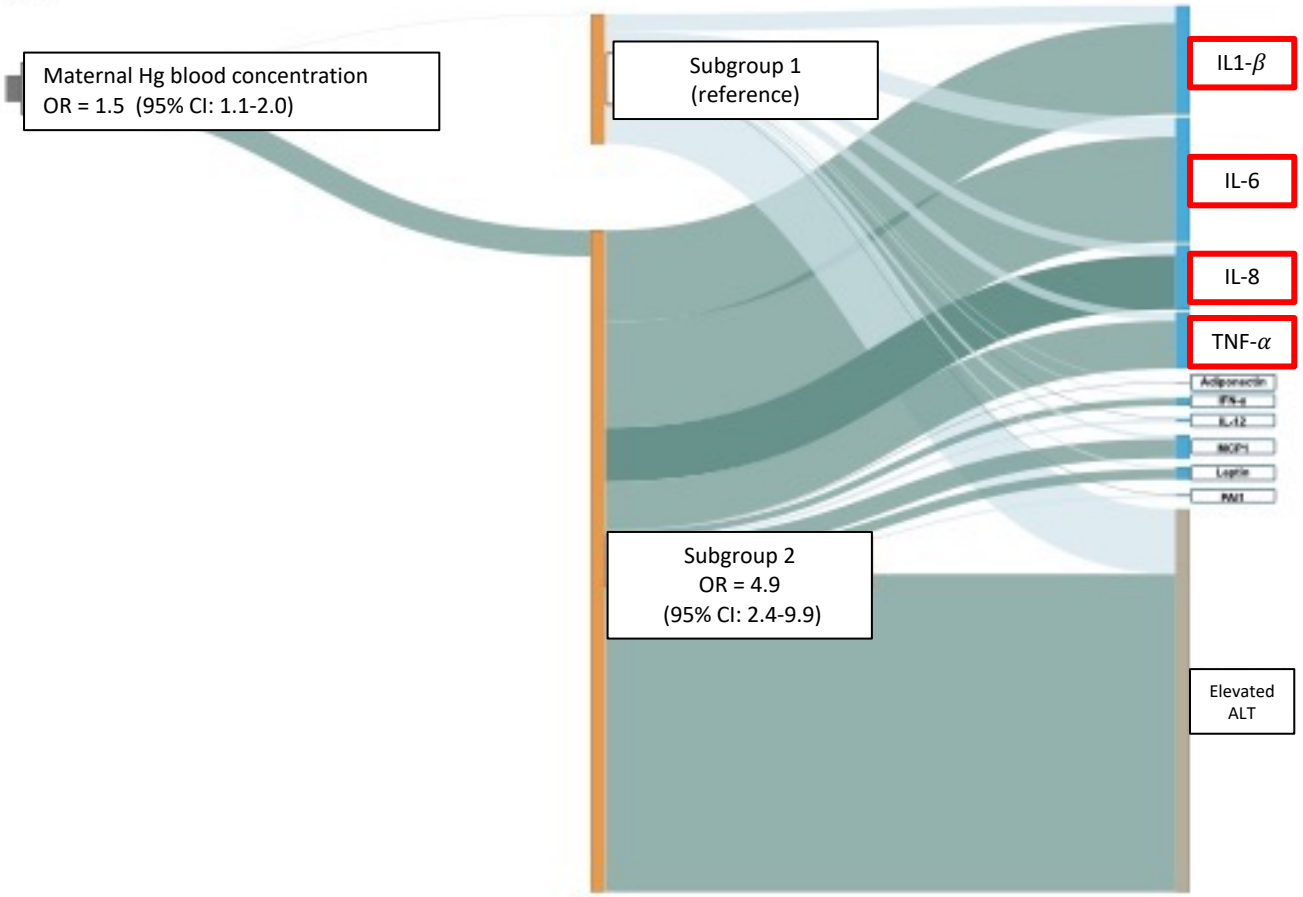
How exposure to mercury associates with inflammation and liver injury in children



Stratakis et al. *Hepatology* (2021)

Exposure to mercury associates with inflammation and liver injury in children

A



B

	Subgroup 1 (N=669)	Subgroup 2 (N=123)
Maternal blood Hg concentration, µg/L	1.8 (1.0, 3.4)	2.7 (1.4, 4.4)
≥ 5.8 µg/L	55 (8.2)	22 (17.9)
ALT, U/L	14 (11.3, 17.3)	18 (13.6, 23.6)
Elevated ALT	37 (5.5)	30 (24.4)
Inflammatory cytokines		
IL-1β, pg/mL	9.7 (7.3, 13.6)	36.9 (17.6, 54.8)
IL-6, pg/mL	9.6 (7.8, 12.3)	26.4 (20.3, 32.9)
IL-8, pg/mL	87.7 (80.5, 95.7)	101.3 (92.8, 114.6)
TNF-α, pg/mL	30.2 (26.2, 31.9)	33.6 (30.3, 38.8)

Stratakis et al. *Hepatology* (2021)

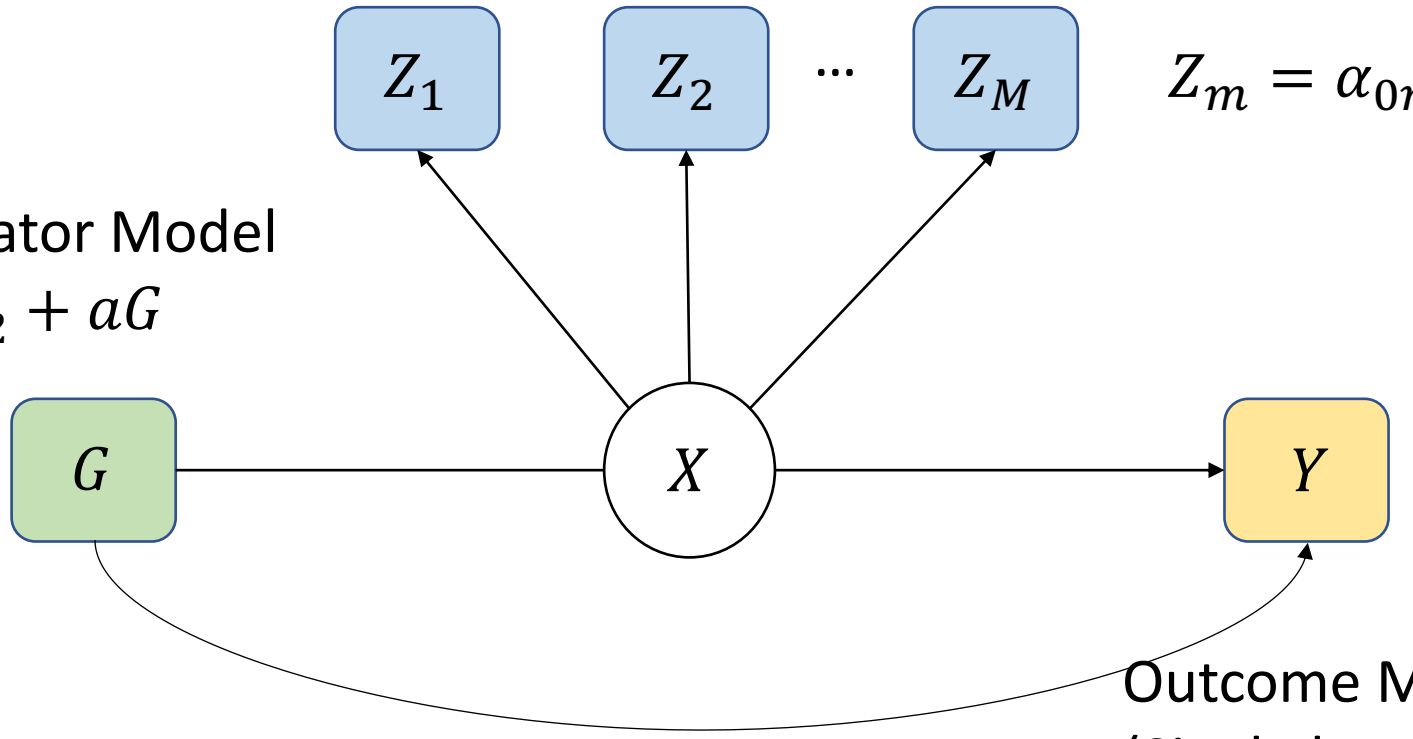
Q&A

Thanks!

References

1. Maitre, Léa, et al. "Human Early Life Exposome (HELIX) study: a European population-based exposome cohort." *BMJ open* 8.9 (2018): e021311.
2. Peng, Cheng, et al. "A latent unknown clustering integrating multi-omics data (LUCID) with phenotypic traits." *Bioinformatics* 36.3 (2020): 842-850.
3. Stratakis, Nikos, et al. "In utero exposure to mercury is associated with increased susceptibility to liver injury and inflammation in childhood." *Hepatology* 74.3 (2021): 1546-1559.

Mediation analysis with one latent mediator



Albert et al., *Biometrical Journal* (2016)

EM algorithm for MLE

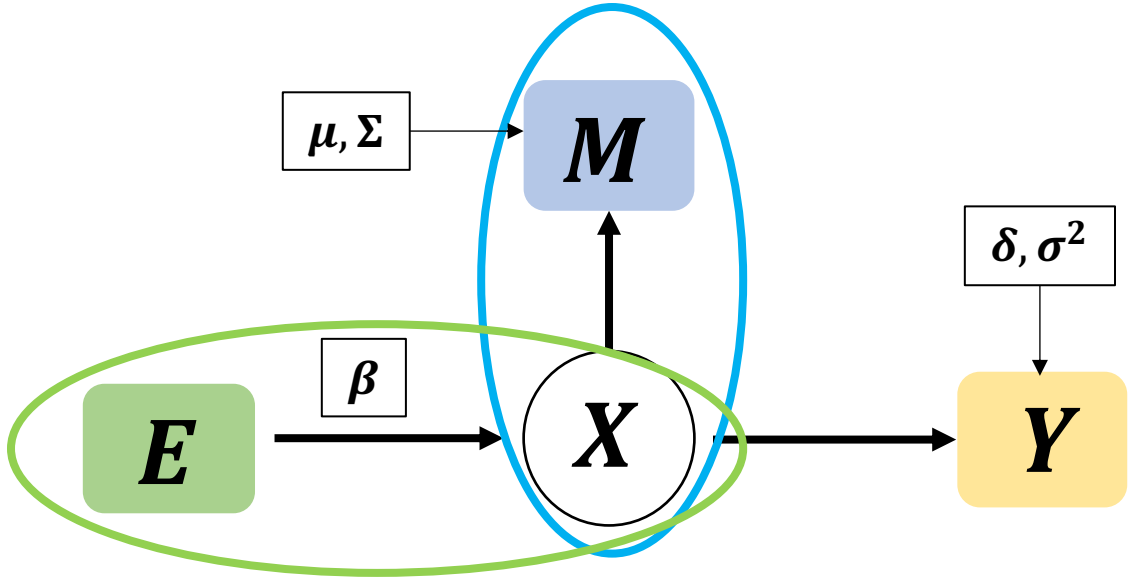
$$\mathcal{D} = \{\boldsymbol{E}, \boldsymbol{M}, \boldsymbol{Y}\}$$

$$Q(\boldsymbol{\Theta}) = \mathrm{E}_{\boldsymbol{X}}(l(\boldsymbol{\Theta}))$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} P(X_i = j|\mathcal{D};\boldsymbol{\Theta}) \log S\left(X_i = j|\boldsymbol{E}_i, \boldsymbol{\beta}_j\right)$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{k} P(X_i = j|\mathcal{D};\boldsymbol{\Theta}) \log \phi\left(\boldsymbol{M}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{k} P(X_i = j|\mathcal{D};\boldsymbol{\Theta}) \log \phi\left(Y_i|\delta_j, \sigma_j^2\right)$$



At iteration $t+1$,

E-step: Evaluate $Q(\boldsymbol{\Theta})$ at $P\left(X_i = j|\mathcal{D}, \boldsymbol{\Theta}^{(t)}\right)$

M-step: $\boldsymbol{\Theta}^{(t+1)} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})$

$$r_{ij}^{(t+1)} = P(X_i = j|\mathcal{D}, \boldsymbol{\Theta}^{(t)})$$

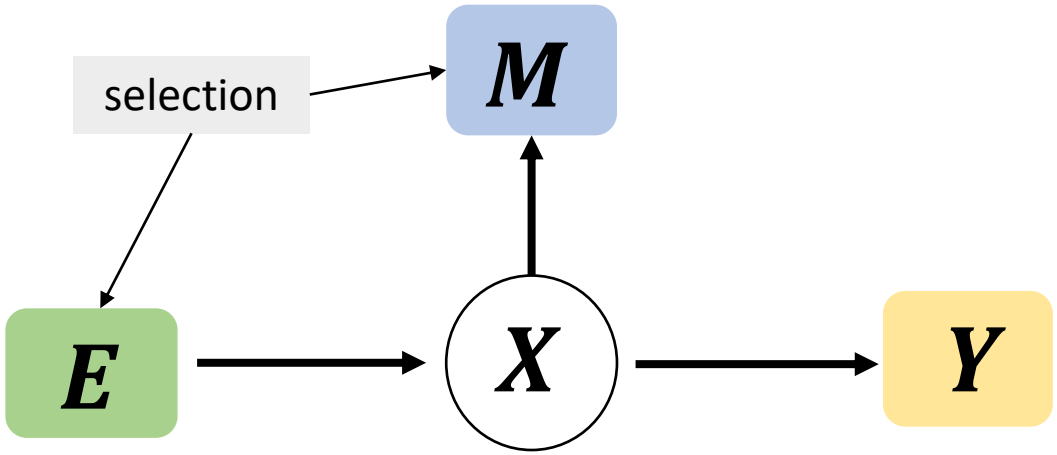
$$= \frac{S\left(X_i = j\middle|\boldsymbol{E}_i, \boldsymbol{\beta}^{(t)}\right)\phi\left(\boldsymbol{M}_i\middle|X_i = j, \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)\phi\left(Y_i\middle|X_i = j, \delta_j^{(t)}, \sigma_j^{2(t)}\right)}{\sum_{j=1}^{k} S\left(X_i = j\middle|\boldsymbol{E}_i, \boldsymbol{\beta}^{(t)}\right)\phi\left(\boldsymbol{M}_i\middle|X_i = j, \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)\phi\left(Y_i\middle|X_i = j, \delta_j^{(t)}, \sigma_j^{2(t)}\right)}$$

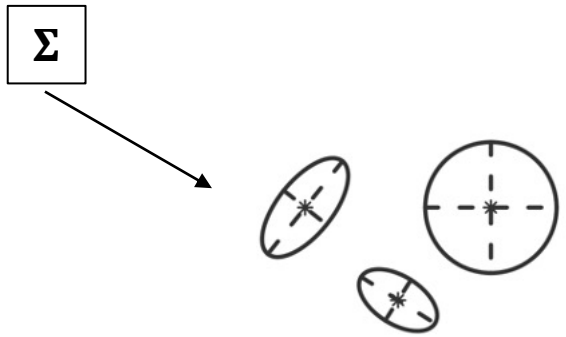
Integrated variable selection

Penalized Likelihood function:

$$Q_p(\boldsymbol{\Theta}) = \mathrm{E}_{\boldsymbol{X}}\left(l_p(\boldsymbol{\Theta})\right) - p_\lambda(\boldsymbol{\Theta})$$

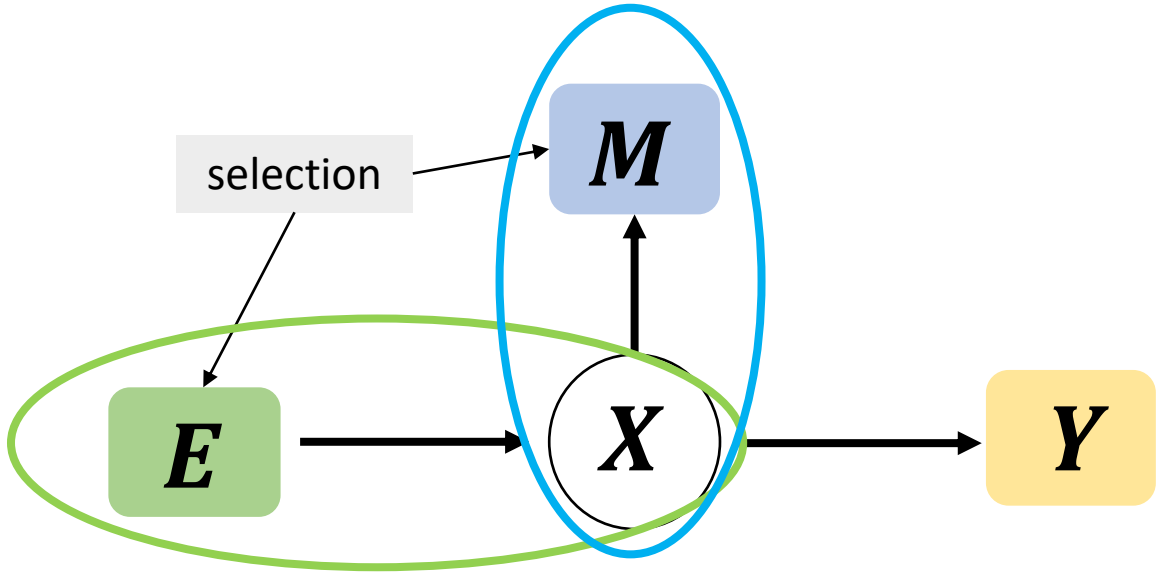
$$p_\lambda(\boldsymbol{\Theta}) = \lambda_E \sum_{q=1}^{p+1}\sum_{j=1}^{k}|\beta_{qj}| + \lambda_{\boldsymbol{W}} \sum_{l=1}^{m}\sum_{s\neq l}^{m}|w_{jls}| + \lambda_{\boldsymbol{\mu}} \sum_{j=1}^{k}\sum_{l=1}^{m}|\mu_{jl}|$$





Integrated variable selection

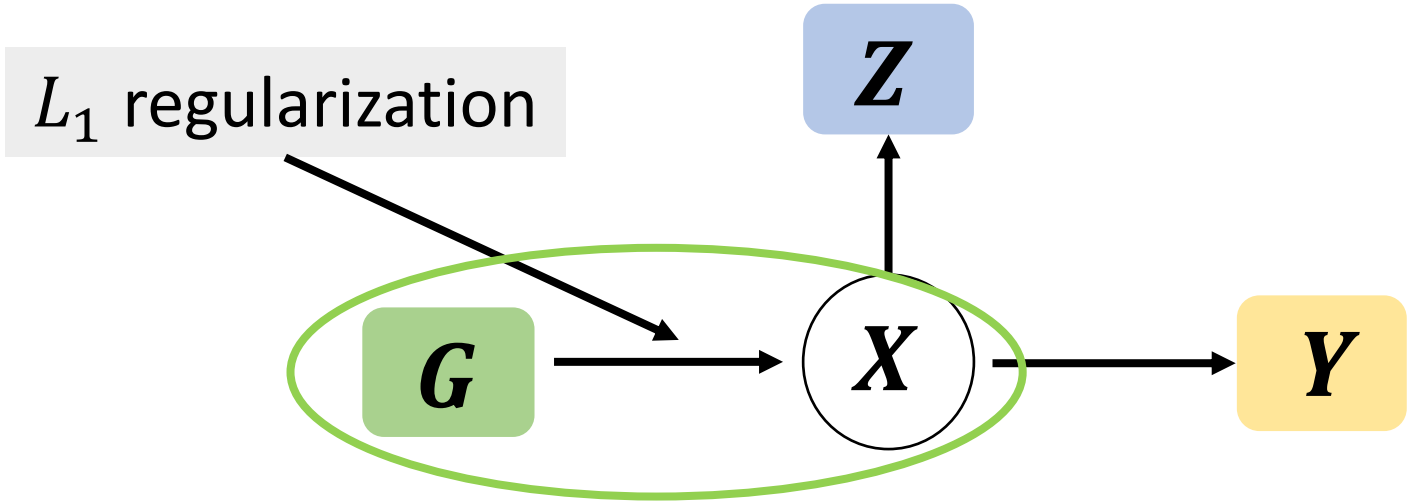
$$Q(\boldsymbol{\Theta}) = \mathrm{E}_{\boldsymbol{X}}(l(\boldsymbol{\Theta}))$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{k} P(X_i = j|\boldsymbol{\mathcal{D}};\boldsymbol{\Theta}) \log S\left(X_i = j|\boldsymbol{E}_{\mathrm{i}},\boldsymbol{\beta}_j\right)$$
$$+ \sum_{i=1}^{n}\sum_{j=1}^{k} P(X_i = j|\boldsymbol{\mathcal{D}};\boldsymbol{\Theta}) \log \phi\left(\boldsymbol{M}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)$$
$$+ \sum_{i=1}^{n}\sum_{j=1}^{k} P(X_i = j|\boldsymbol{\mathcal{D}};\boldsymbol{\Theta}) \log \phi\left(Y_i|\delta_j,\sigma_j^2\right)$$



Penalize each likelihood component with corresponding L_1 penalty

$$p_\lambda(\boldsymbol{\Theta}) = \lambda_E \sum_{q=1}^{p+1}\sum_{j=1}^{k}|\beta_{qj}| + \lambda_{\boldsymbol{W}} \sum_{l=1}^{m}\sum_{s\neq l}^{m}|w_{jls}| + \lambda_{\boldsymbol{\mu}} \sum_{j=1}^{k}\sum_{l=1}^{m}|\mu_{jl}|$$

Variable selection for exposures



LASSO problem for multinomial logistic regression

$$\boldsymbol{\beta}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij}^{(t+1)} \log S\left(X_i = j | \boldsymbol{E}_{\mathrm{i}}, \boldsymbol{\beta}_j\right) - \lambda_E \sum_{p=1}^{P+1} \sum_{j=1}^{K} |\beta_{pj}|$$

Variable selection for omics data

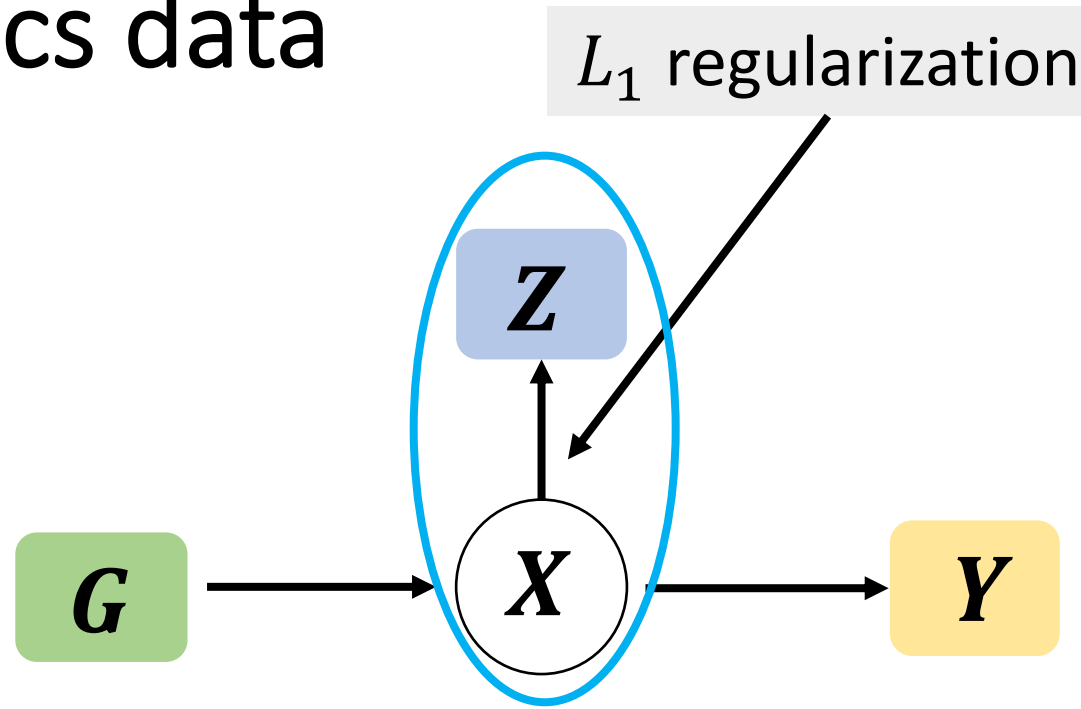
$$f(\boldsymbol{Z}_i|X_i = j) \sim MVN(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

+

$$p_\lambda(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \lambda_{\boldsymbol{\mu}} \sum_{j=1}^{K} \sum_{l=1}^{M} |\mu_{jl}| + \lambda_{\boldsymbol{\Sigma}^{-1}} \sum_{l=1}^{M} \sum_{s \neq l}^{M} |w_{jls}|$$

LASSO problem for Gaussian mixture model

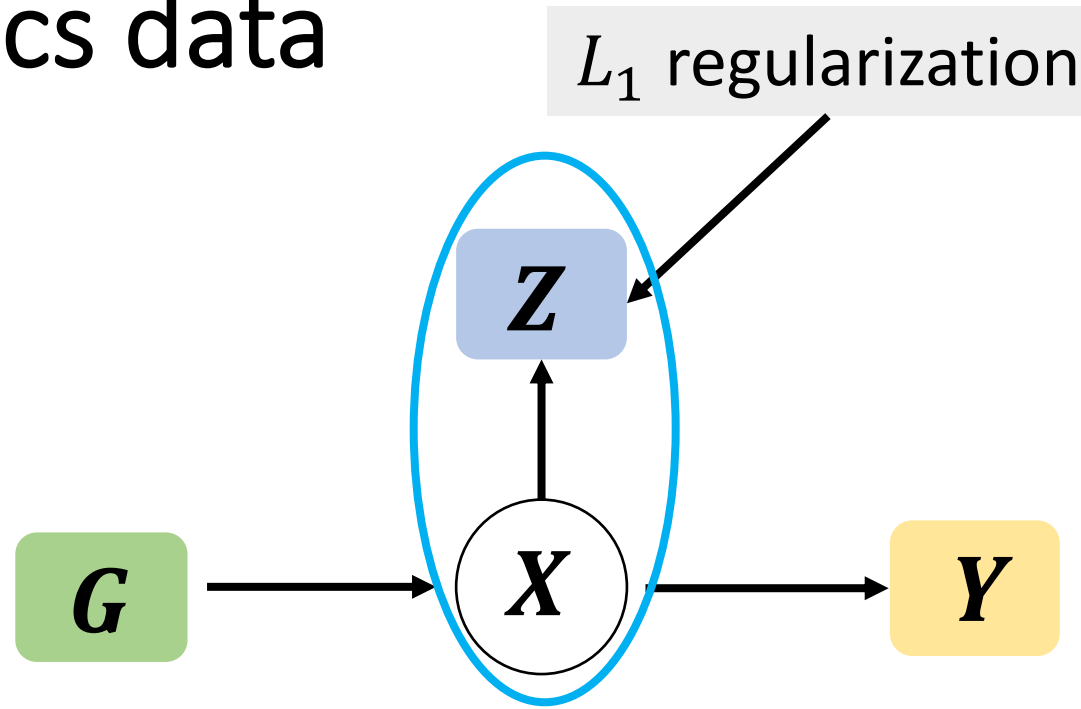
Graphical LASSO problem



Variable selection for omics data

$$f(\boldsymbol{Z}_i|X_i = j) \sim MVN(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

Step 1: Obtain sparse solution for $\boldsymbol{\mu_j}$



$$\boldsymbol{\mu}^{(t+1)} = \underset{\boldsymbol{\mu}}{\arg\max} \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}^{(t+1)} \log \phi\left(\boldsymbol{M}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) - \lambda_{\boldsymbol{\mu}} \sum_{j=1}^{k}\sum_{l=1}^{m} |\mu_{jl}|$$

LASSO problem for Gaussian Mixture Model

Variable selection for omics data

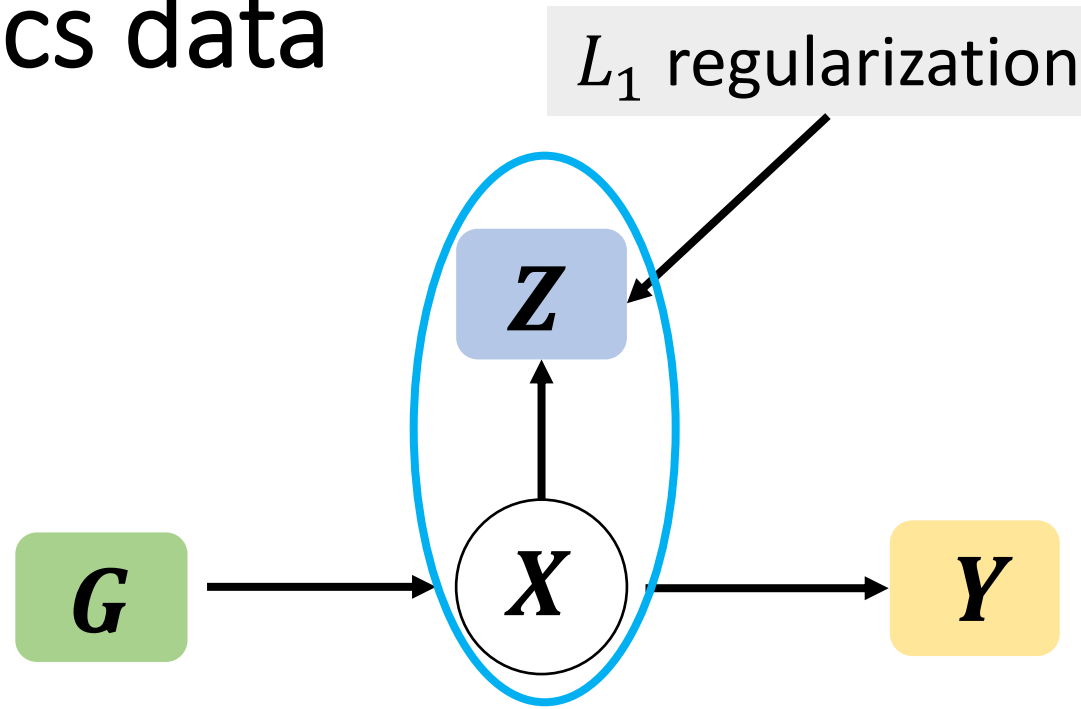
$$f(\boldsymbol{Z}_i|X_i = j) \sim MVN(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

Step 2: Obtain sparse solution for $\boldsymbol{\Sigma_j}$

$$\boldsymbol{S}_j = \frac{\sum_{i=1}^{n} r_{ij}^{(t+1)} (\boldsymbol{M}_i - \boldsymbol{\mu}_j)(\boldsymbol{M}_i - \boldsymbol{\mu}_j)^T}{\sum_{i=1}^{n} r_{ij}^{(t+1)}}$$

$$\boldsymbol{\Sigma}_j^{(t+1)} = \underset{\boldsymbol{W}_j = \boldsymbol{\Sigma}_j^{-1}}{\arg\max} \sum_{j=1}^{k} \left(\frac{1}{2} \sum_{i=1}^{n} r_{ij}^{(t+1)} \log\det(W_j) - \frac{1}{2} \sum_{i=1}^{n} r_{ij}^{(t+1)} \mathrm{trace}(\boldsymbol{S}_j \boldsymbol{W}_j) - \lambda_{\boldsymbol{W}} \sum_{l=1}^{m} \sum_{s \neq l}^{m} |w_{jls}| \right)$$

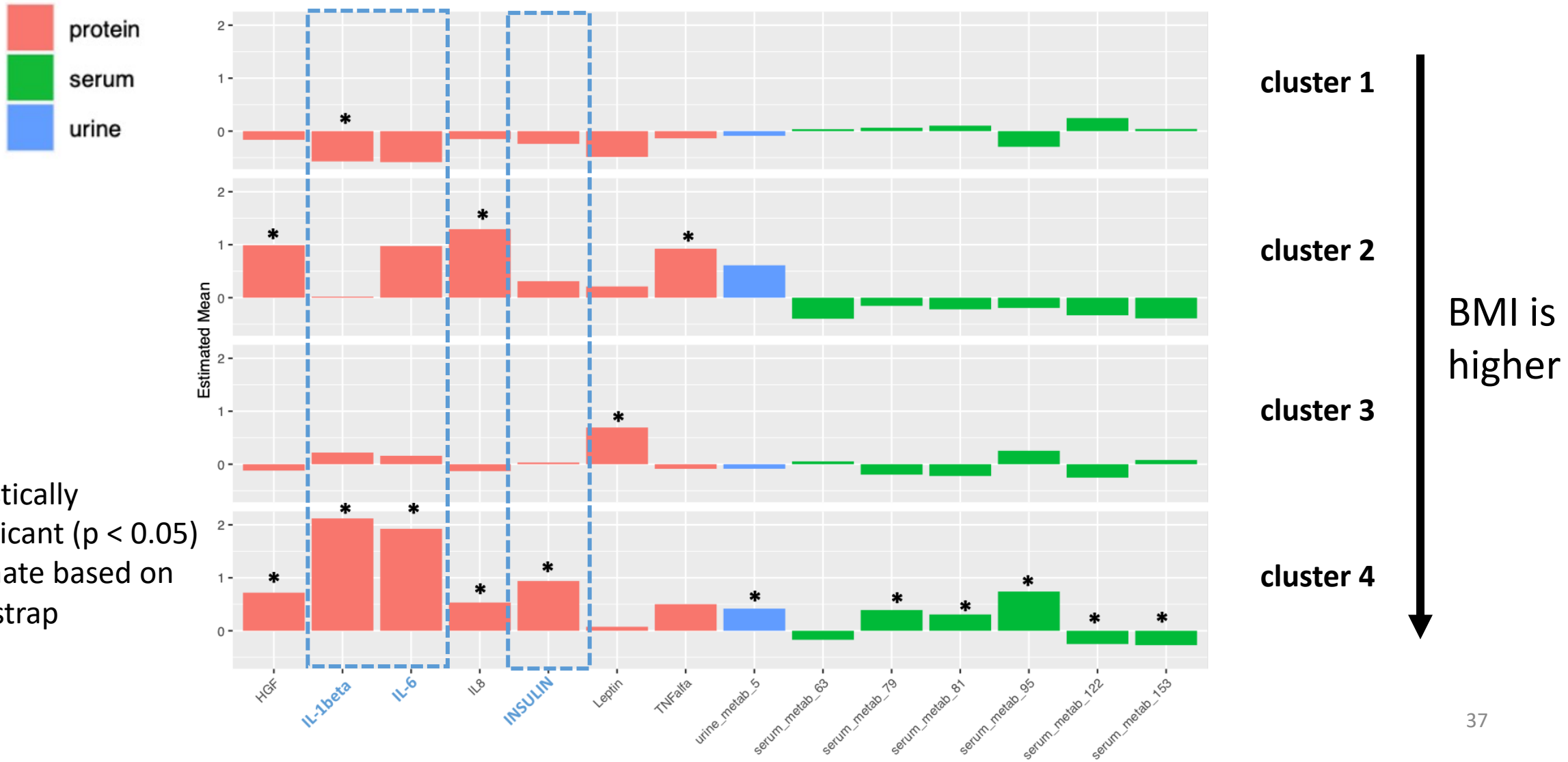
Graphical LASSO problem



Application on HELIX

- Selected presentation at ISGlobal Data challenge (2021)
- We conducted integrated clustering analysis using exposure to organochlorines, omics data (proteomics, serum and urine metabolites), and associates the estimated clusters to children's BMI.
- We used the integrated variable selection based on the LUCID model and picked 14 out of 257 omics features.

Omics profile



*: Statistically significant ($p < 0.05$) estimate based on bootstrap

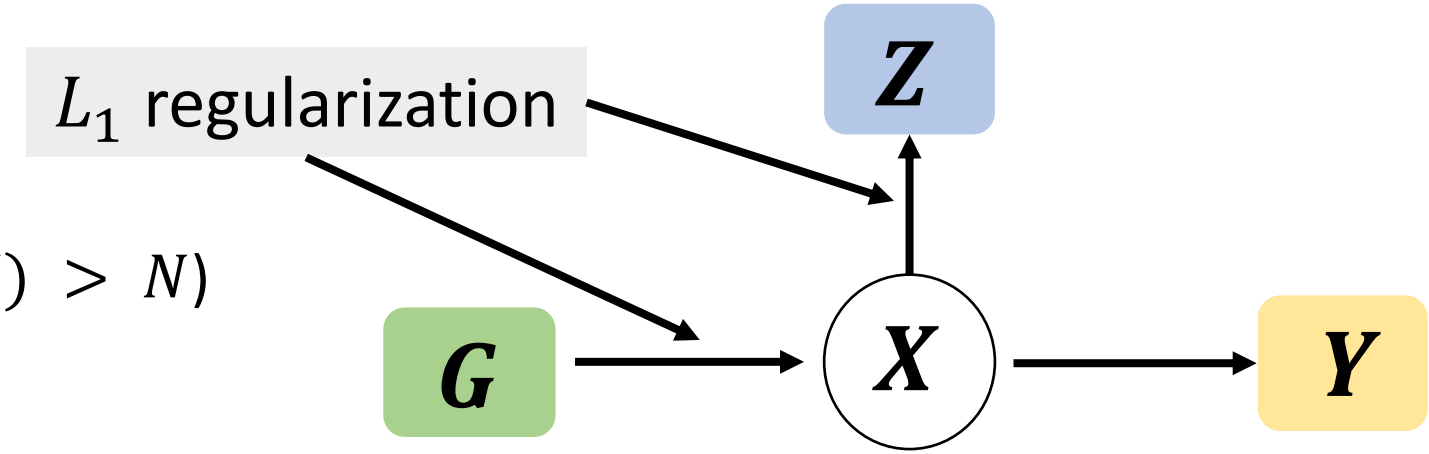
Features of the LUCID model

1. Regularization and variable selection
2. A more flexible geometric model for omics data
3. Incorporation of missingness in omics data
4. Incorporation of multiple omics data by using multiple latent variables
5. A powerful R package, *LUCIDus* to conduct a comprehensive analysis framework by using the LUCID model
6. Inclusion of covariates
7. Use BIC to determine number of clusters
8. Supervised and unsupervised analysis

Integrated variable selection

Motivation:

1. High dimensionality ($(P + M) > N$)
2. Reduce noises
3. Parsimony and interpretability
4. Stabilize the EM algorithm

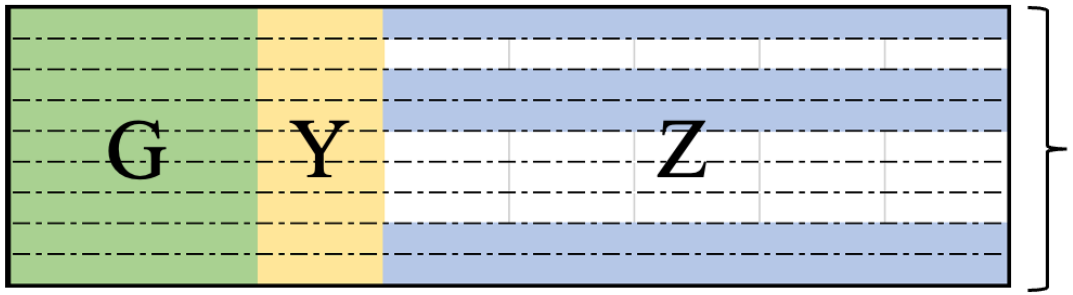


Incorporating missingness in omics data

Motivation:
In large cohort studies, some omics data are not available for all participants for various reasons, such as budgetary constraints, low sample availability, or lack of consent for future use of biospecimens.
We refer to this type of missingness as a list-wise missing pattern

Example:
Multi-Ethnic Cohort (MEC) study, 4346 African American men have genotype data (genetic exposure) and status of prostate cancer (outcome), but only 672 out of 4346 have metabolomics data.

Incorporating missingness in omics data



$\{i = 1, 2, \ldots, n\}$

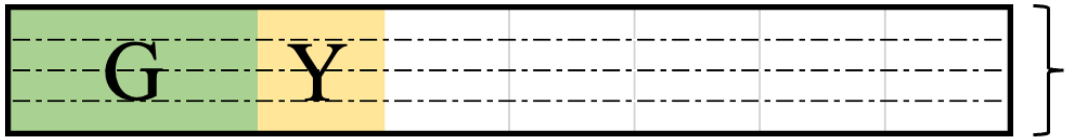
$l(\boldsymbol{\Theta})$

Likelihood partition

$\{i_{(\text{obs})}; Z_i \text{ is observed}\}$

$l_{\text{obs}}(\boldsymbol{\Theta}) = l_G + l_Y + l_Z$

+

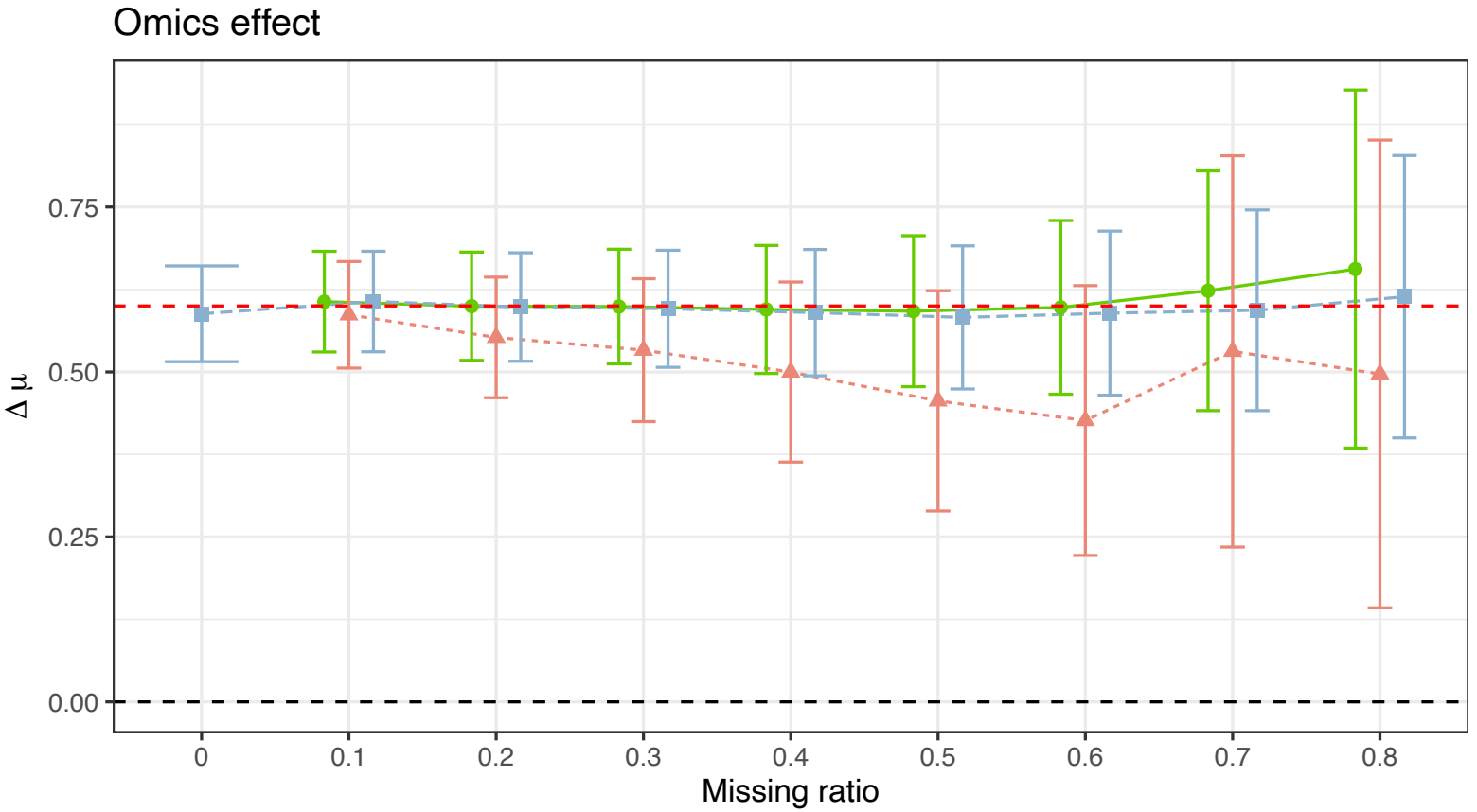


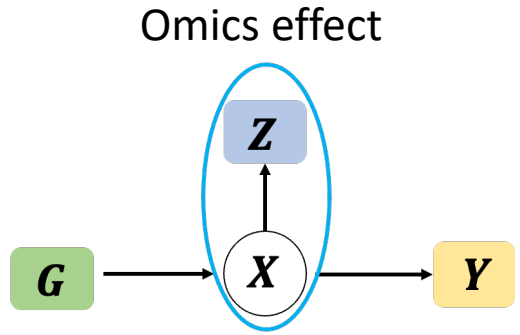
$\{i_{(\text{miss})}; Z_i \text{ is missing}\}$

$l_{\text{miss}}(\boldsymbol{\Theta}) = l_G + l_Y$

Incorporating missingness in omics data

Simulation example





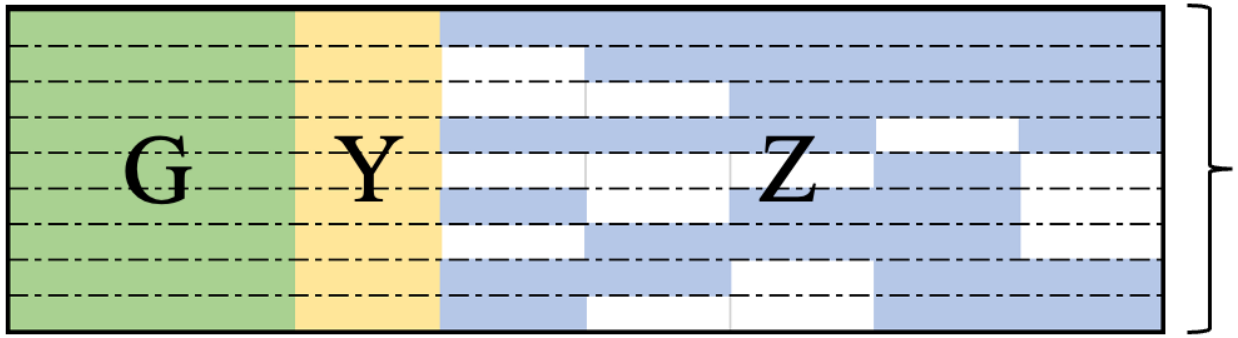
Complete case analysis

Imputation

Likelihood partition

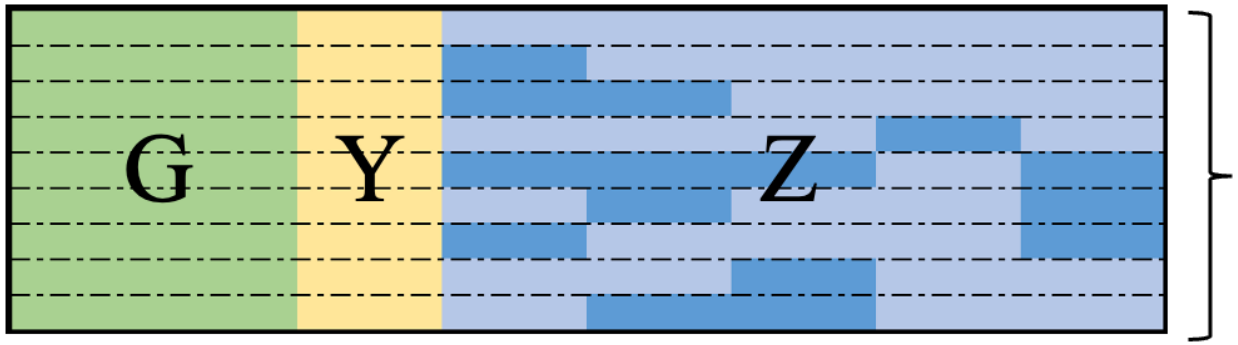
Incorporating missingness in omics data

Some omics features are randomly missing due to the measurement process.
(referred to as sporadic missing pattern)



$\{i = 1, 2, \ldots, n\}$

Impute missing values
(optimize $\boldsymbol{Z}$ given fixed $\boldsymbol{\Theta}$ in the EM algorithm)



$\{i = 1, 2, \ldots, n\}$

Visualize the LUCID model