LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Statistical Matching of Discrete Data by Bayesian Networks

Eva Endres, Thomas Augustin
Ludwig-Maximilians-Universität München, Germany

PGM '16, Lugano

Statistical matching

A

denomination	gender	age
no rel. community	m	[17.9, 21.9]
Roman Catholic	f	(45, 48.8]
Roman Catholic	m	(45, 48.8]
Roman Catholic	f	(41.1, 45]
Protestant	m	(48.8, 52.6]

⋮

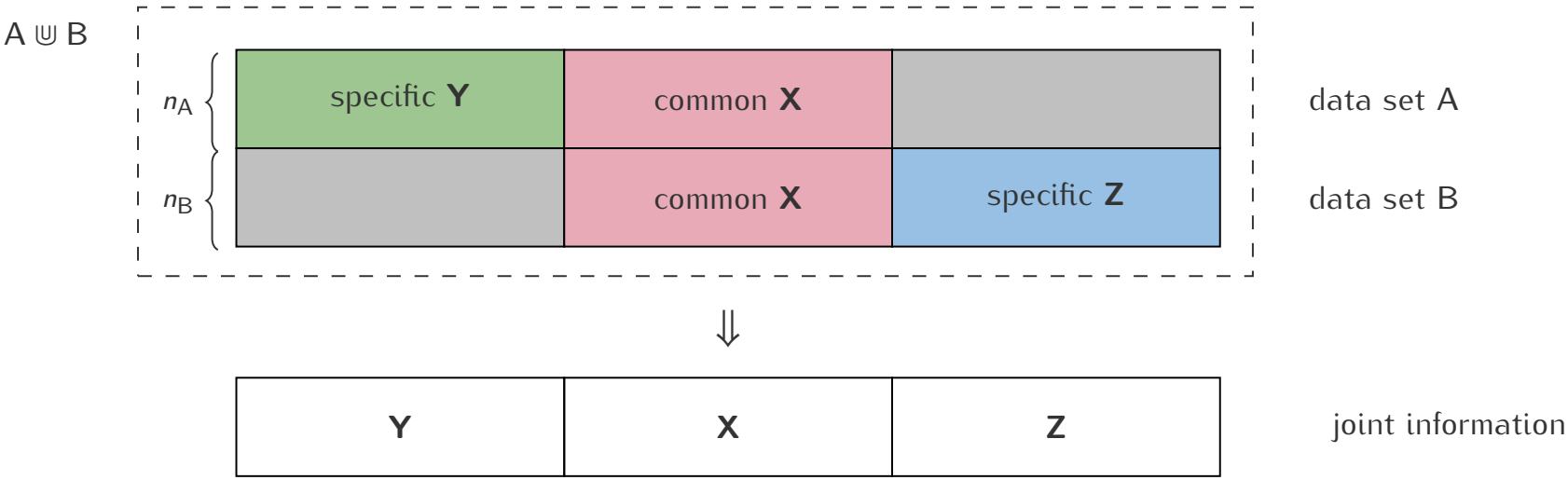
gender	age	alcohol consumption
m	(33.4,37.2]	none
m	(79.6,83.5]	several times/month
m	(33.4,37.2]	several times/month
f	(37.2,41.1]	none

⋮

B

Statistical matching

(e.g. D'Orazio et al. (2006))



- Assumption: conditional independence of **Y** and **Z** given **X**
- Parametric macro approach:

$$L(\mathbf{p}^{\mathrm{A}\uplus\mathrm{B}}|\mathrm{A}\uplus\mathrm{B}) = \prod_{i\in\mathcal{I}_\mathrm{A}} p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_i|\mathbf{x}_i) \prod_{i\in\mathcal{I}_\mathrm{B}} p_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}_i|\mathbf{x}_i) \prod_{i\in\mathcal{I}_\mathrm{A}\cup\mathcal{I}_\mathrm{B}} p_{\mathbf{X}}(\mathbf{x}_i)$$

Bayesian networks – basic concepts and notations

(e.g. Koller and Friedman (2009))

- Discrete random variables $\mathbf{W} = (W_1, \ldots, W_s)'$
- Global probability distribution $P(\mathbf{W} = \mathbf{w}) = P(W_1 = w_1, \ldots, W_s = w_s)$
- Directed acyclic graph (DAG) $\mathcal{G}_{\mathbf{W}}$, where
  - each random variable W_m, $m = 1, \ldots, s$, is represented by an eponymous node
  - the dependencies among the random variables are represented by a set of directed edges between pairs of nodes
- *Chain rule* of Bayesian networks

$$P(\mathbf{W} = \mathbf{w}) = \prod_{m=1}^{s} P(W_m = w_m | \mathbf{Pa}(W_m) = \mathbf{pa}(W_m)) =: \prod_{m=1}^{s} p(w_m | \mathbf{pa}(W_m))$$

The German General Social Survey

(GESIS – Leibniz Institute for the Social Sciences (2013))

- Survey from 2012 covers 3480 observations on 752 variables, inter alia, regarding to demography, religiousness and physical health of the respondents
- Idea: pretend that variables measured in the GGSS data came from two different sources
- Extraction of 17 variables for illustration
  - **common**: *sex*, *age*, *graduation*, professional *activity*, *marital* status, and net *income* of the respondents
  - **specific in** A: *denomination*, frequency of *church*goings, frequency of experiencing the presence of *God* through faith, frequency of experiences that can only be explained through the intervention of *supernatural* powers, any experience with miracle *healers*/spirit healers, and frequency of *pray*ing
  - **specific in** B: frequency of visiting a *doctor*, *hospital* stay in the last 12 month, number of *cigarettes* per day, *alcoholic* beverages per day, and general *health*

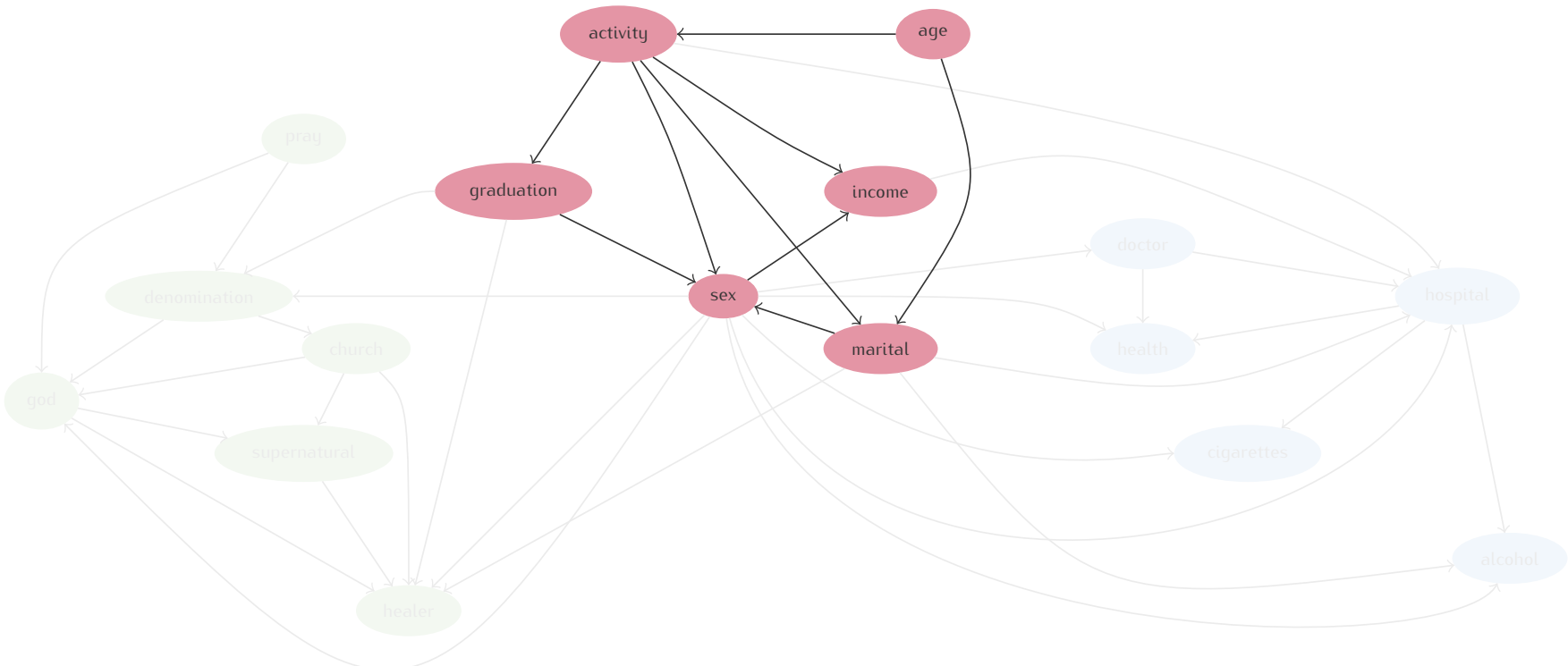
$\Rightarrow$ A and B can be matched as if they came from different surveys

Statistical matching by Bayesian networks

- Representation of $\mathbf{Y} \perp \mathbf{Z} \mid \mathbf{X}$ by Bayesian networks and incorporation of further information on the conditional independence structure
- Restriction of the DAG to basic form $\mathbf{Y} \leftarrow \mathbf{X} \rightarrow \mathbf{Z}$
- Basic procedure:
  1. Estimate two DAGs on A and on B and combine them to a joint DAG
     - 1.1 fix graph structure for the common variables
     - 1.2 individual graph structures for the common variables
  2. Estimate the corresponding local parameters and the global probability distribution
  3. Generate synthetic data set by imputation

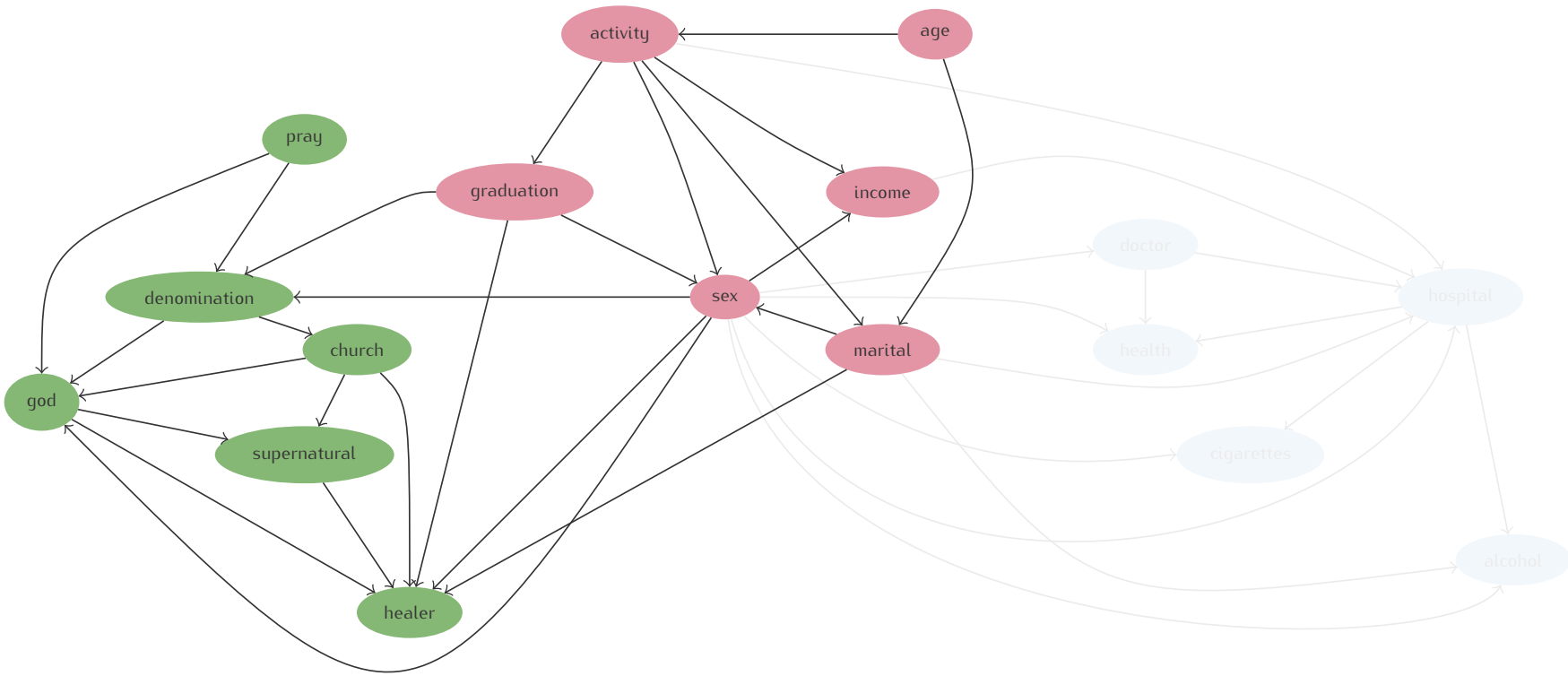
Step 1: estimation and combination of the graph structures

fix graph structure for the common variables



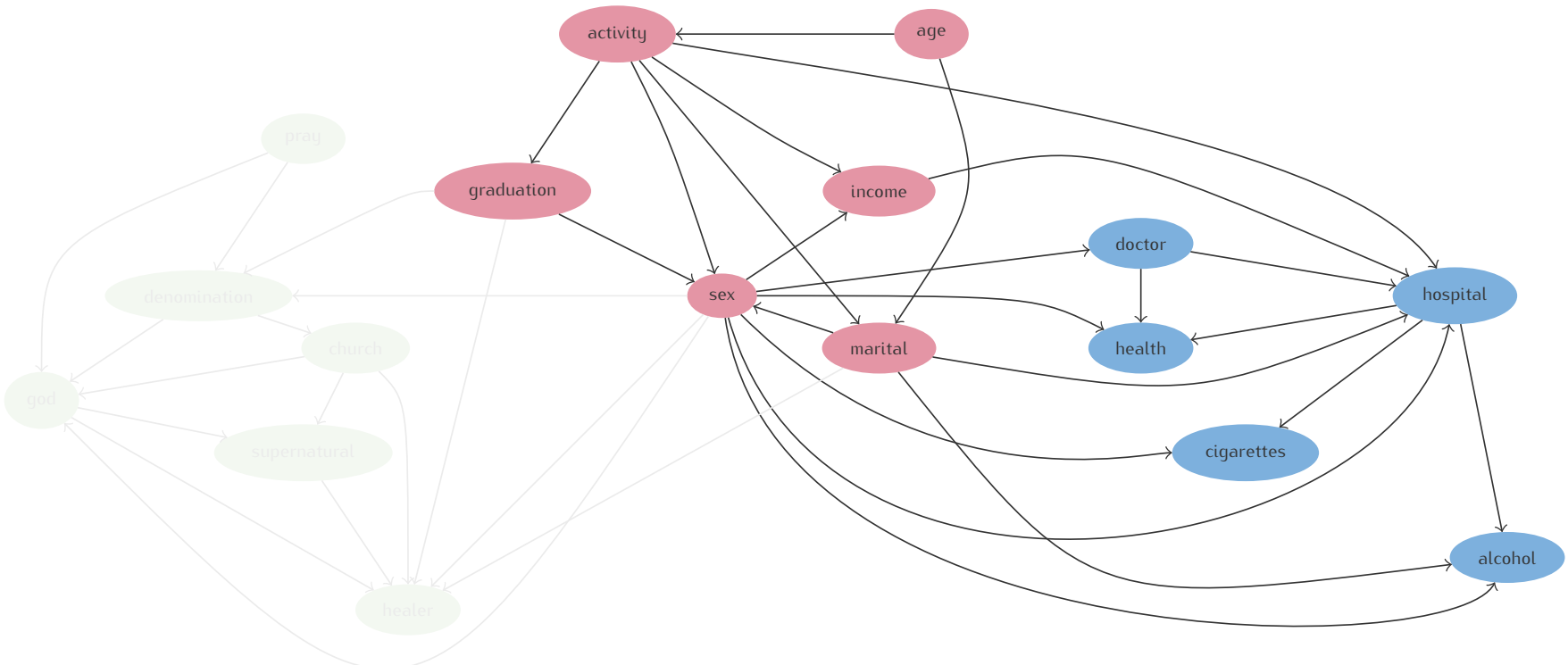
Step 1: estimation and combination of the graph structures

fix graph structure for the common variables



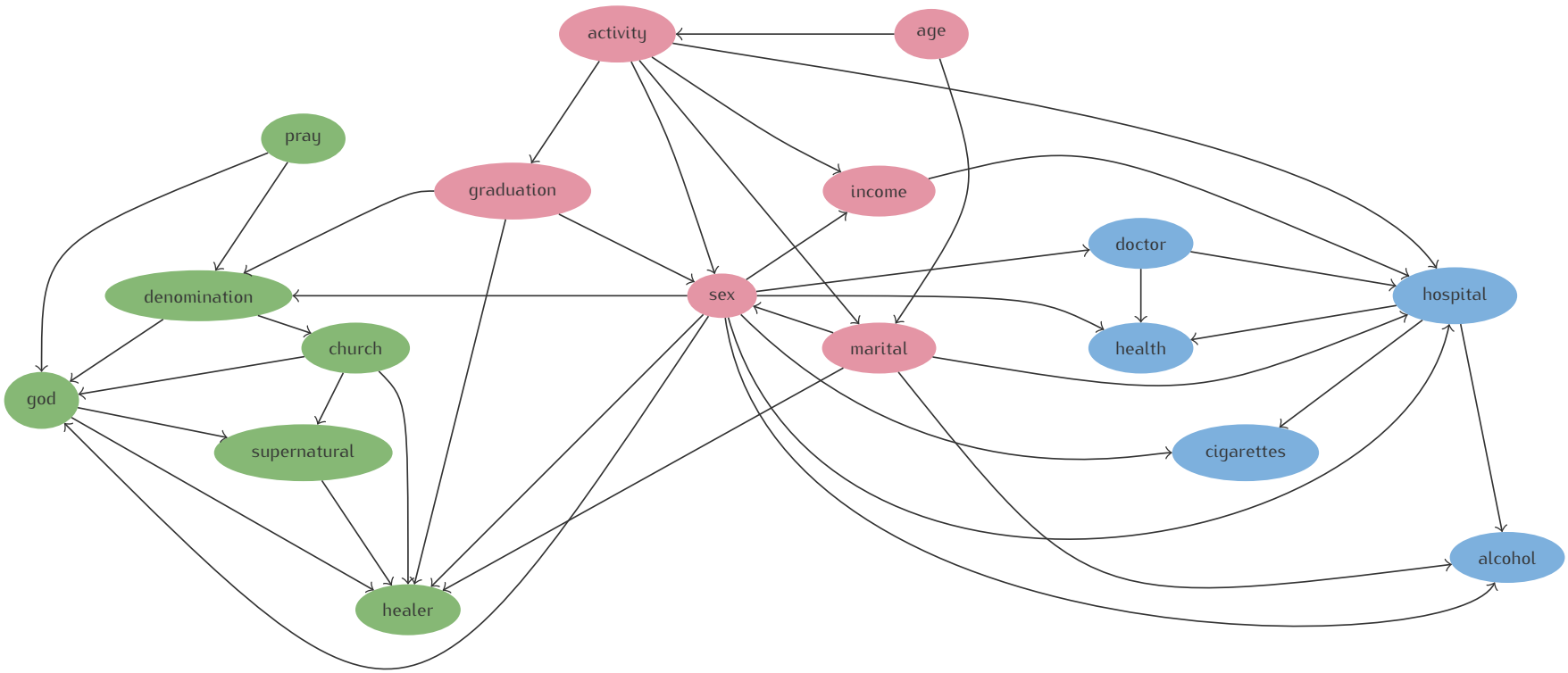
Step 1: estimation and combination of the graph structures

fix graph structure for the common variables



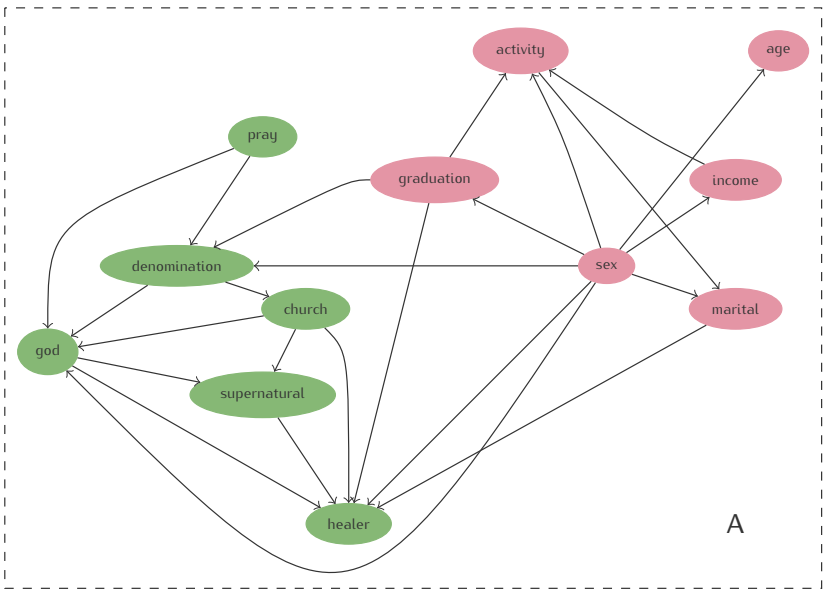
Step 1: estimation and combination of the graph structures

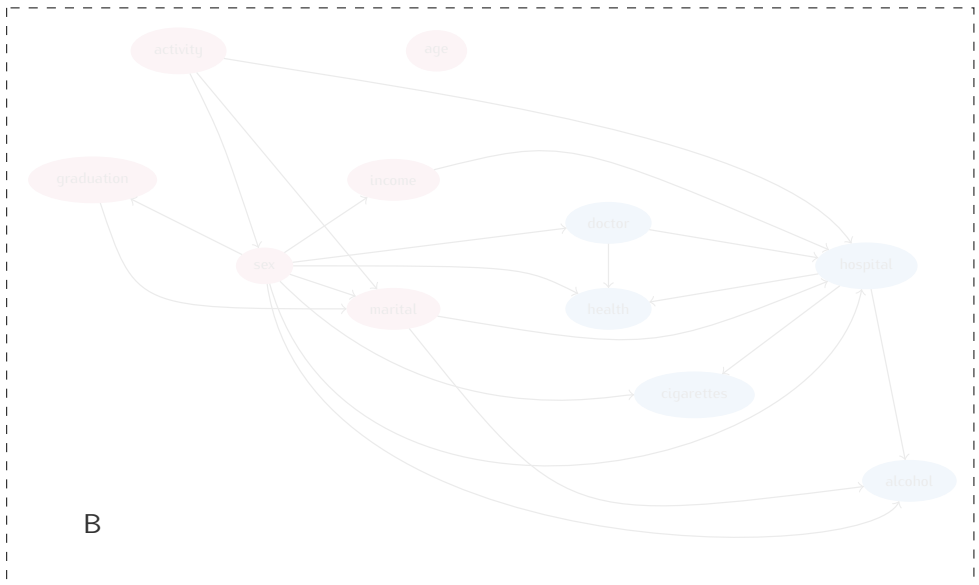
fix graph structure for the common variables



Step 1: estimation and combination of the graph structures

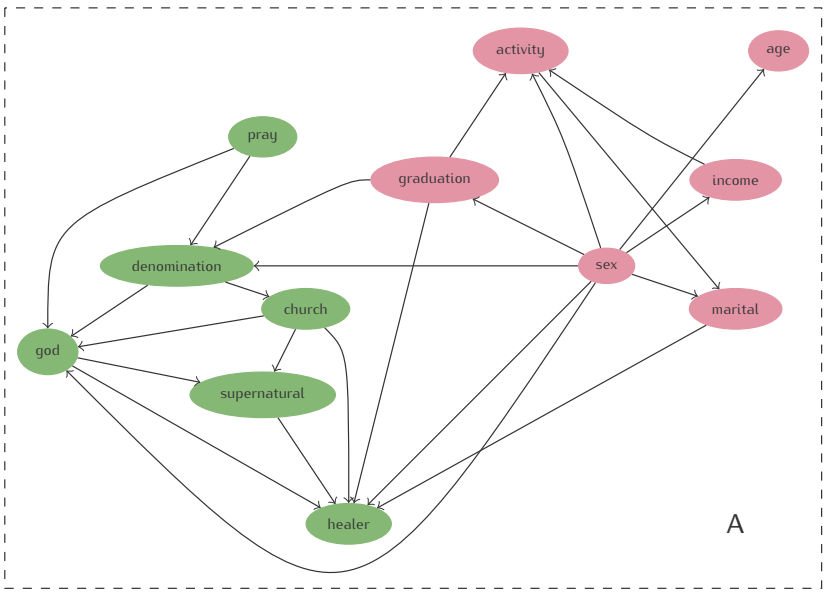
individual graph structures for the common variables

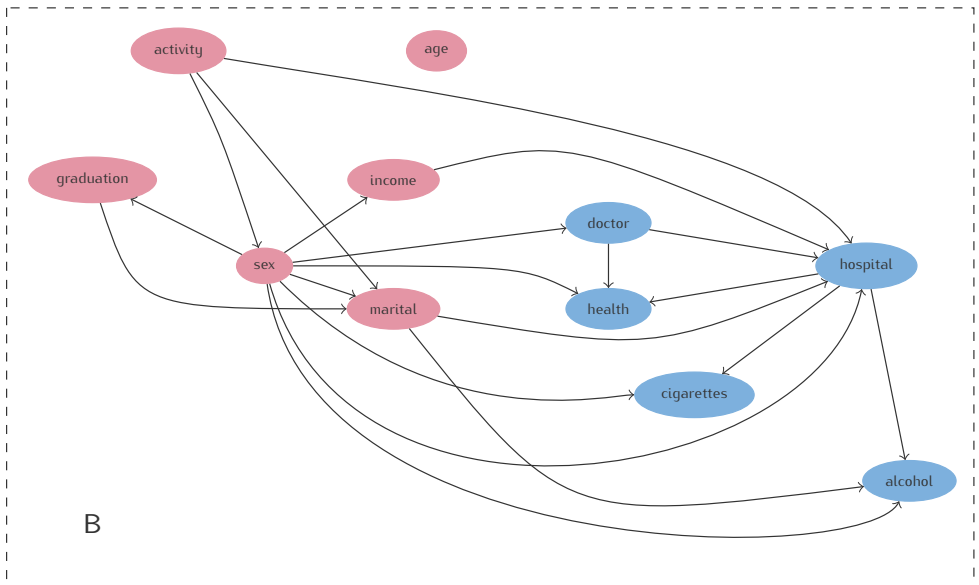




Step 1: estimation and combination of the graph structures

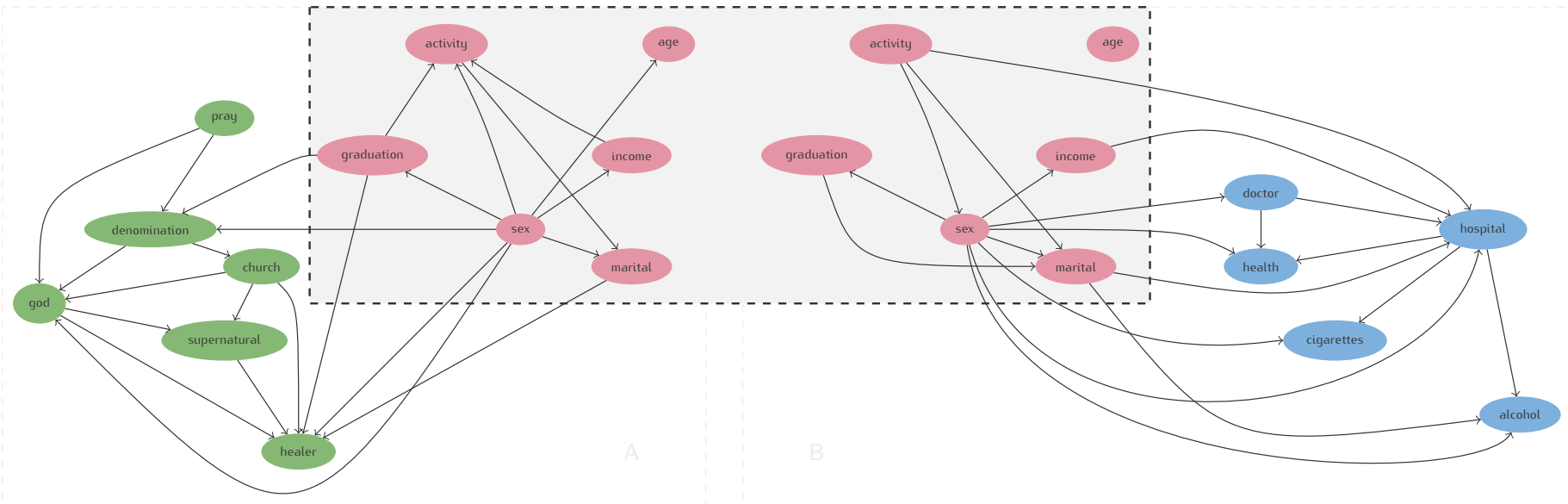
individual graph structures for the common variables





Step 1: estimation and combination of the graph structures

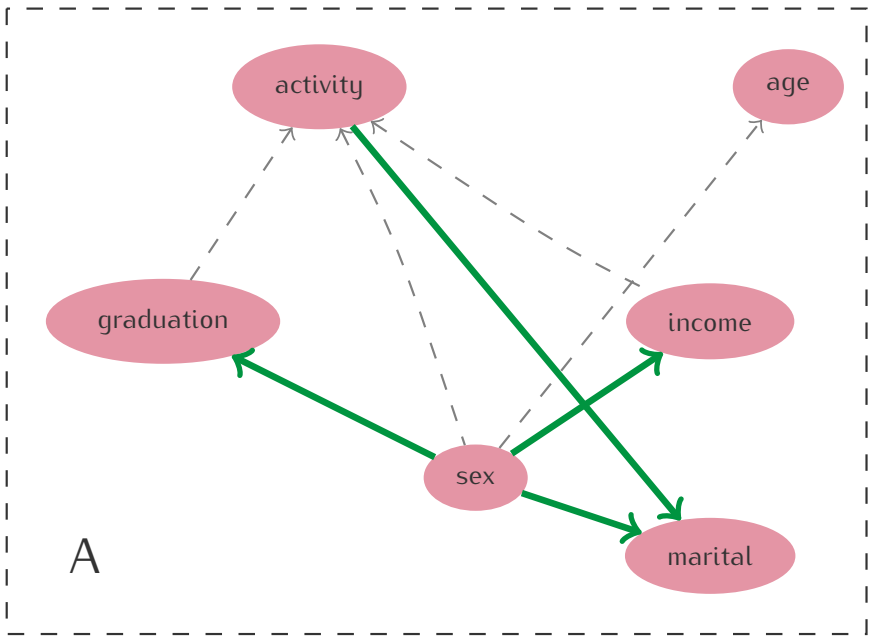
individual graph structures for the common variables

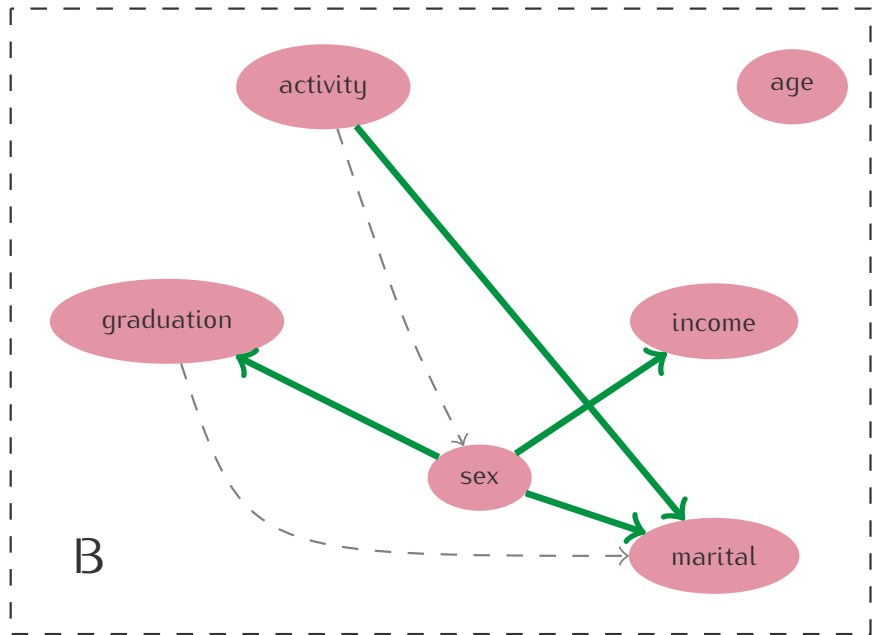


Step 1: estimation and combination of the graph structures

individual graph structures for the common variables

edge intersection

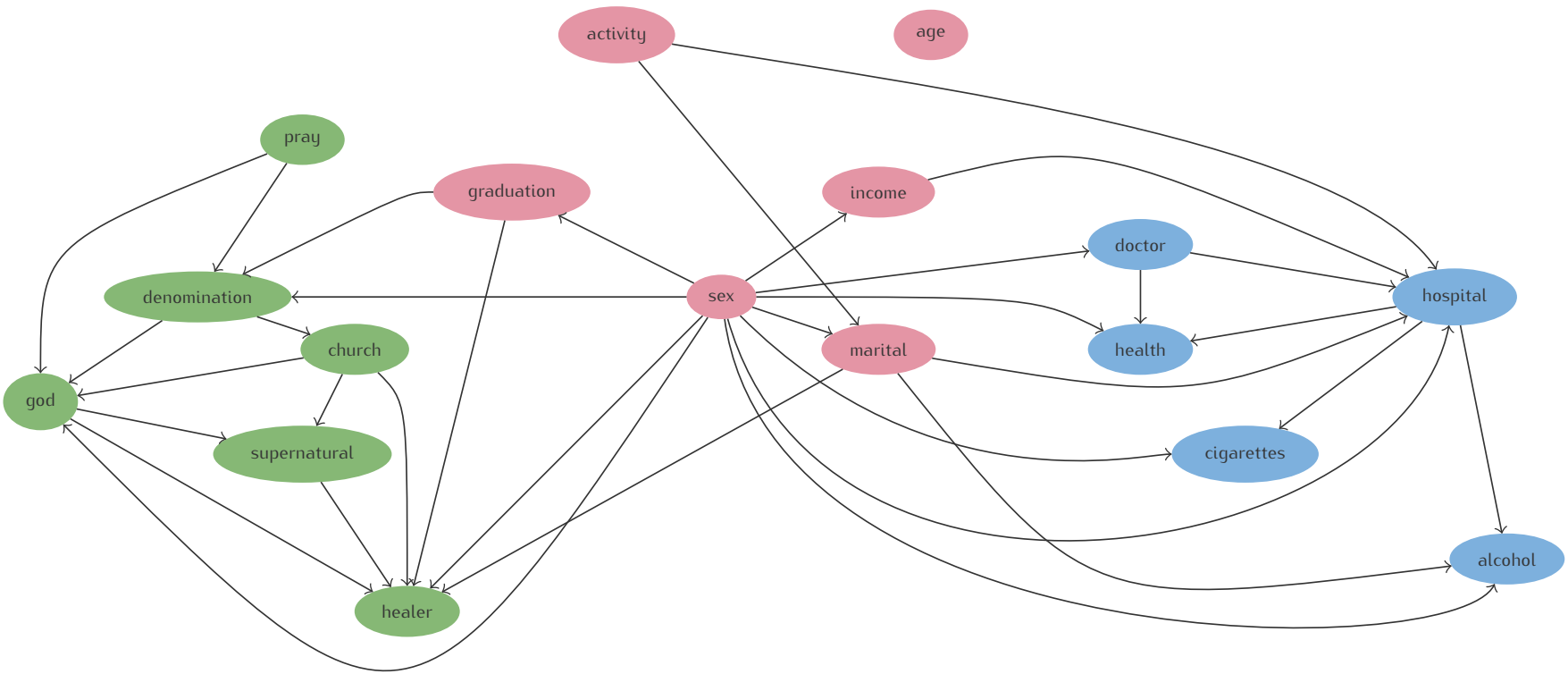




Step 1: estimation and combination of the graph structures

individual graph structures for the common variables

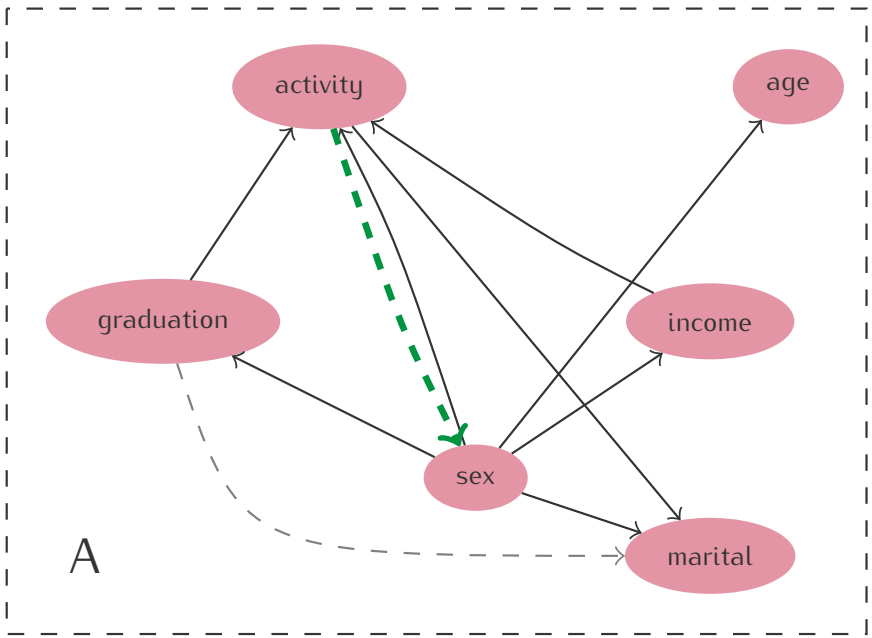
edge intersection

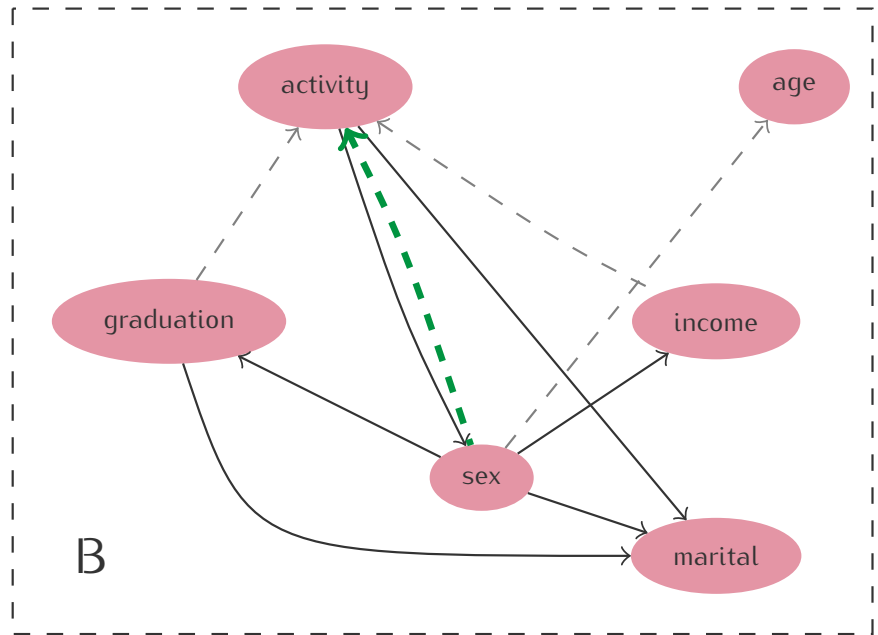


Step 1: estimation and combination of the graph structures

individual graph structures for the common variables

edge union

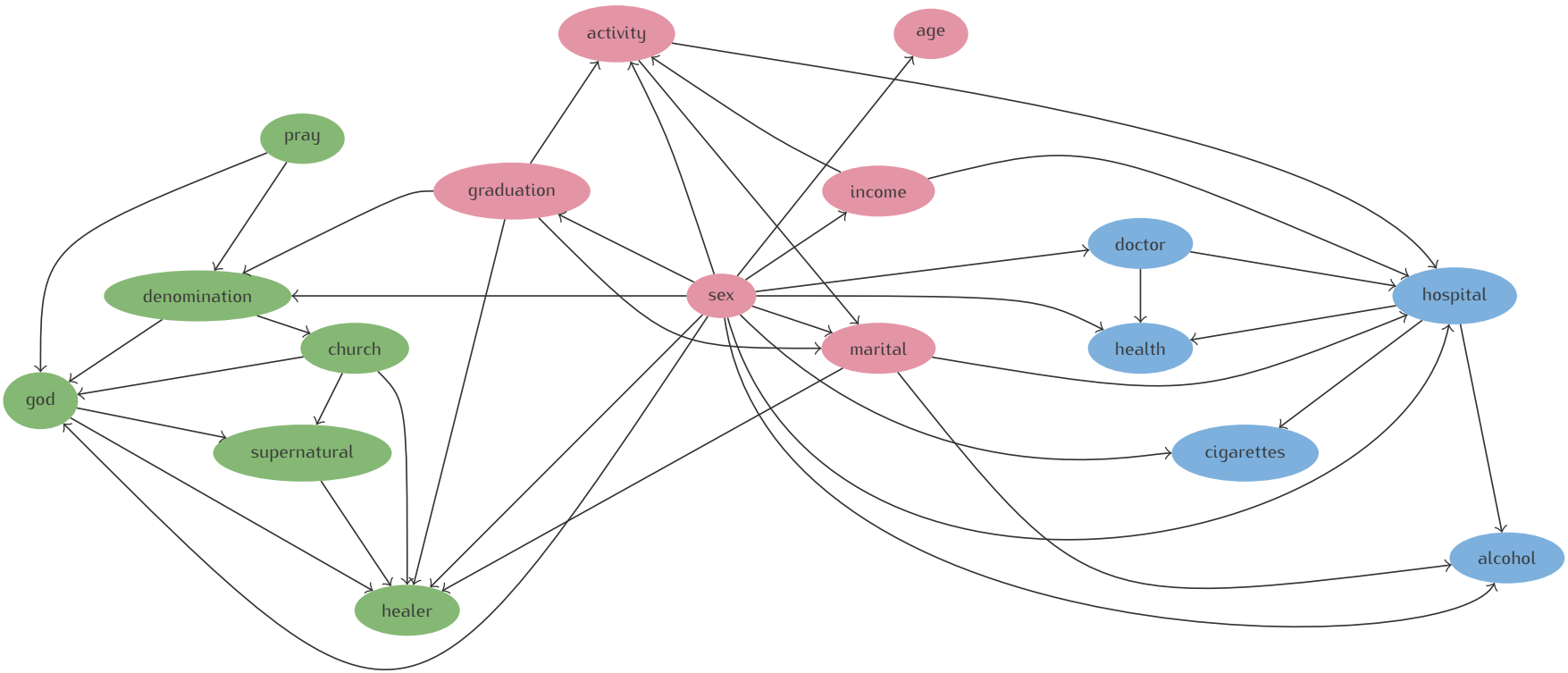




Step 1: estimation and combination of the graph structures

individual graph structures for the common variables

edge union



Step 2: estimation of the local parameters and the joint probability distribution

$$\hat{P}^{\mathrm{A}\cup\mathrm{B}}(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z}) = \prod_{k=1}^{q} \hat{p}_{\hat{\mathcal{G}}^{\mathrm{A}}_{\mathbf{X},\mathbf{Y}}}(y_k|\mathbf{pa}(Y_k)) \cdot \prod_{\ell=1}^{r} \hat{p}_{\hat{\mathcal{G}}^{\mathrm{B}}_{\mathbf{X},\mathbf{Z}}}(z_\ell|\mathbf{pa}(Z_\ell)) \cdot \prod_{j=1}^{p} \hat{p}_{\hat{\mathcal{G}}^{\mathrm{A}\cup\mathrm{B}}_{\mathbf{X}}}(x_j|\mathbf{pa}(X_j))$$

Step 3: imputation of the missing values

- data set A: draw synthetic values for Z_ℓ, $\ell = 1,\ldots,r$, given the realizations of $\mathbf{X}$ for every $i \in \mathcal{I}_A$ from the estimated posterior distribution $\hat{P}^{\mathrm{A}\cup\mathrm{B}}(\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x})$
- data set B: draw synthetic values for Y_k, $k = 1,\ldots,q$ given the realizations of $\mathbf{X}$ for every $i \in \mathcal{I}_B$ from the estimated posterior distribution $\hat{P}^{\mathrm{A}\cup\mathrm{B}}(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x})$

Representativeness of the synthetic file

(Rässler (2002))

- Four quality levels
  1. preserving marginal distributions
  2. preserving correlation/association structures
  3. preserving the joint distribution
  4. preserving the individual values

Jenson-Shannon divergence

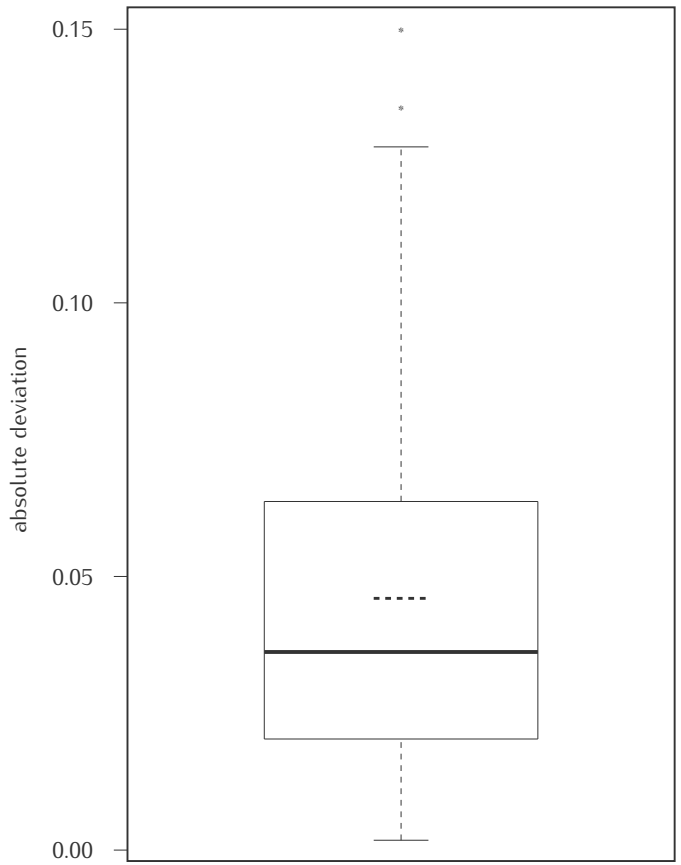
denomination	church	god	supernatural	healer	pray
0.000	0.001	0.001	0.004	0.021	0.002

doctor	hospital	cigarettes	alcohol	health
0.001	0.003	0.004	0.001	0.000

Representativeness of the synthetic file

(Rässler (2002))

- Four quality levels
  1. preserving marginal distributions
  2. preserving correlation/association structures
  3. preserving the joint distribution
  4. preserving the individual values



corrected contingency coefficient

Conclusion and outlook

- Representation of conditional independence assumption in the framework of statistical matching by probabilistic graphical models
- Incorporation of further information about the conditional independence structure achieved by the Bayesian network approach
- Marginal distributions and association structure among variables are well preserved

- Simulation studies
- Extension for continuous and mixed discrete/continuous data
- Undirected probabilistic graphical models
- Inclusion of auxiliary information
- Imprecise probabilistic graphical models

References (selection)

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*, Wiley, Chichester, United Kingdom.

GESIS – Leibniz Institute for the Social Sciences (2013). Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2012/German General Social Survey GGSS 2012. ZA4614 Data file Version 1.1.1, doi:10.4232/1.12209.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, Cambridge, MA.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer, New York, NY.