Simulating counterfactuals: Online Appendices

Juha Karvanen, Santtu Tikka, Matti Vihola
Department of Mathematics and Statistics
University of Jyvaskyla, Finland

Online Appendix 1: Simple Illustration of Counterfactual Inference

Consider the following SCM $\mathcal{M}$:

$$\begin{aligned} U_Z &\sim \mathrm{N}(0,1), \\ U_X &\sim \mathrm{N}(0,1), \\ U_Y &\sim \mathrm{N}(0,1), \\ Z &= U_Z, \\ X &= Z + U_X, \\ Y &= X + Z + U_Y, \end{aligned}$$

where $\mathbf{U} = \{U_Z, U_X, U_Y\}$ are unobserved background variables and Z, X and Y are observed variables.

The aim is to derive the counterfactual distribution

$$p(Y_{\mathrm{do}(X=-1)} = y \,|\, Y = 1).$$

The condition $Y = 1$ is fulfilled if $U_Y + U_X + 2U_Z = 1$. This equation defines a plane in the space of (U_Z, U_X, U_Y). In the general case, it might not be possible to present the surface of interest in a closed form but for the normal distribution this is doable.

Following Pearl (2009), the solution can be obtained in three steps:

1. Find the distribution $p(U_Z, U_X, U_Y \,|\, Y = 1)$.
2. Modify $\mathcal{M}$ by the intervention $\mathrm{do}(X = -1)$ to obtain the submodel $\mathcal{M}_{\mathrm{do}(X=-1)}$.
3. Use the submodel $\mathcal{M}_{\mathrm{do}(X=-1)}$ and the updated distribution $p(U_Z, U_X, U_Y \,|\, Y = 1)$ to compute the probability distribution of $Y_{\mathrm{do}(X=-1)}$.

The unconditional joint distribution of (U_Z, U_Y, Y) is a multivariate normal distribution with zero expectations and the covariance matrix

$$\begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \\ 2 & 1 & 6 \end{pmatrix}.$$

Because X is determined by the intervention $\mathrm{do}(X = -1)$ and U_X affects only X, it suffices to consider only background variables U_Z and U_Y. The distribution of (U_Z, U_Y) conditional on $Y = 1$ is a multivariate normal distribution with the expectation

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \frac{1}{6} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{6} \end{pmatrix}$$

and the covariance matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 2 \\ 1 \end{pmatrix} \frac{1}{6} \begin{pmatrix} 2 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{5}{6} \end{pmatrix}.$$

The submodel $\mathcal{M}_{\mathrm{do}(X=-1)}$ is the following:

$$\begin{aligned} U_Z &\sim \mathrm{N}(0,1), \\ U_X &\sim \mathrm{N}(0,1), \\ U_Y &\sim \mathrm{N}(0,1), \\ Z &= U_Z, \\ X &= -1, \\ Y &= X + Z + U_Y = -1 + U_Z + U_Y. \end{aligned}$$

The counterfactual distribution of interest, i.e., the distribution of $Y = -1 + U_Z + U_Y$ on the condition of $Y = 1$ is a normal distribution with the expectation

$$-1 + \frac{1}{3} + \frac{1}{6} = -\frac{1}{2},$$

and the variance

$$\frac{1}{3} + \frac{5}{6} - 2 \cdot \frac{1}{3} = \frac{1}{2}.$$

In other words, $Y_{\mathrm{do}(X=-1)} \,|\, (Y = 1) \sim \mathrm{N}(-\frac{1}{2}, \frac{1}{2})$.

Online Appendix 2: Counterfactual Inference for Linear Gaussian SCMs

Consider a linear Gaussian SCM with observed variables $\mathbf{V} = (V_1, \ldots, V_J)$ and mutually independent background variables $\mathbf{U} = (U_1, \ldots, U_H)$ that follow the standard normal distribution. The model is written as

$$\mathbf{V} = \mathbf{b}_0 + \mathbf{B}_1 \mathbf{V} + \mathbf{B}_2 \mathbf{U},$$

where $\mathbf{b}_0$ is a constant vector, $\mathbf{B}_1$ is a $J \times J$ strictly triangular matrix, and $\mathbf{B}_2$ is a $H \times H$ matrix. The observed variables can be expressed in terms of the background variables as follows

$$\mathbf{V} = (\mathbf{I} - \mathbf{B}_1)^{-1}(\mathbf{b}_0 + \mathbf{B}_2 \mathbf{U}),$$

and because $\mathbf{b}_0$ is a constant and $\mathbf{U} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ we have $\mathbf{V} \sim \mathrm{N}(\boldsymbol{\mu}_\mathbf{V}, \boldsymbol{\Sigma}_{\mathbf{VV}})$ where

$$\begin{aligned} \boldsymbol{\mu}_\mathbf{V} &= (\mathbf{I} - \mathbf{B}_1)^{-1}\mathbf{c}, \\ \boldsymbol{\Sigma}_{\mathbf{VV}} &= (\mathbf{I} - \mathbf{B}_1)^{-1}(\mathbf{B}_2 \mathbf{B}_2^T)((\mathbf{I} - \mathbf{B}_1)^{-1})^T. \end{aligned}$$

The joint distribution of $\mathbf{V}$ and $\mathbf{U}$ is

$$\begin{pmatrix} \mathbf{V} \\ \mathbf{U} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_\mathbf{V} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{VV}} & \boldsymbol{\Sigma}_{\mathbf{VU}} \\ \boldsymbol{\Sigma}_{\mathbf{VU}}^T & \mathbf{I} \end{pmatrix} \right),$$

where $\Sigma_{\mathbf{VU}} = (\mathbf{I} - \mathbf{B}_1)^{-1}\mathbf{B}_2$.

Next, we consider the conditional distribution when the values of some observed variables are fixed. We partition the observed variables as $\mathbf{V} = \mathbf{V}_1 \cup \mathbf{V}_2$ such that the values of $\mathbf{V}_2$ are fixed and write

$$\begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \mathbf{U} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{V}_1} \\ \boldsymbol{\mu}_{\mathbf{V}_2} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{V}_1\mathbf{V}_1} & \boldsymbol{\Sigma}_{\mathbf{V}_1\mathbf{V}_2} & \boldsymbol{\Sigma}_{\mathbf{V}_1\mathbf{U}} \\ \boldsymbol{\Sigma}_{\mathbf{V}_2\mathbf{V}_1} & \boldsymbol{\Sigma}_{\mathbf{V}_2\mathbf{V}_2} & \boldsymbol{\Sigma}_{\mathbf{V}_2\mathbf{U}} \\ \boldsymbol{\Sigma}^T_{\mathbf{V}_1\mathbf{U}} & \boldsymbol{\Sigma}^T_{\mathbf{V}_2\mathbf{U}} & \mathbf{I} \end{pmatrix} \right).$$

The distribution of $(\mathbf{V}_1, \mathbf{U})^T$ conditional on $\mathbf{V}_2 = \mathbf{c}$ is a normal distribution with the mean vector

$$\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{V}_1} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{V}_1\mathbf{V}_2} \\ \boldsymbol{\Sigma}^T_{\mathbf{V}_2\mathbf{U}} \end{pmatrix} \boldsymbol{\Sigma}^{-1}_{\mathbf{V}_2\mathbf{V}_2}(\mathbf{c} - \boldsymbol{\mu}_{\mathbf{V}_2}), \tag{1}$$

and the covariance matrix

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{V}_1\mathbf{V}_1} & \boldsymbol{\Sigma}_{\mathbf{V}_1\mathbf{U}} \\ \boldsymbol{\Sigma}^T_{\mathbf{V}_1\mathbf{U}} & \mathbf{I} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{V}_1\mathbf{V}_2} \\ \boldsymbol{\Sigma}^T_{\mathbf{V}_2\mathbf{U}} \end{pmatrix} \boldsymbol{\Sigma}^{-1}_{\mathbf{V}_2\mathbf{V}_2} \begin{pmatrix} \boldsymbol{\Sigma}^T_{\mathbf{V}_1\mathbf{V}_2} & \boldsymbol{\Sigma}_{\mathbf{V}_2\mathbf{U}} \end{pmatrix}. \tag{2}$$

Finally, we derive the distribution after the intervention. The observed variables after the intervention can be expressed as

$$\mathbf{V}^\circ = (\mathbf{I} - \mathbf{B}^\circ_1)^{-1}(\mathbf{b}^\circ_0 + \mathbf{B}^\circ_2(\mathbf{U} \,|\, \mathbf{V}_2 = \mathbf{c})),$$

where $\mathbf{b}^\circ_0$ is obtained from $\mathbf{b}_0$ by setting the constants of intervened variables to the value of the intervention, $\mathbf{B}^\circ_1$ and $\mathbf{B}^\circ_2$ are obtained from $\mathbf{B}_1$ and $\mathbf{B}_2$, respectively, by setting the values in the row vectors of the intervened variables to zero, and the distribution of $\mathbf{U} \,|\, \mathbf{V}_2 = \mathbf{c}$ is defined by equations (1) and (2).

Online Appendix 3: Details of the Simulation Experiment

Here we describe additional details of the simulation experiment presented in Section 4 of the main paper . The parameters of the simulation experiment can be divided into the SCM parameters, the counterfactual parameters, and the parameters of Algorithm 3. The SCM parameters include the number of observed variables, the average number of neighbouring observed variables per observed variable, the average number of unobserved confounders per observed variable, and the probability distributions from where the coefficients in $\mathbf{B}_1$, $\mathbf{B}_2$, and $\mathbf{b}_0$ (defined in Online Appendix 2) will be drawn. The counterfactual parameters include the number of conditioning variables. The only free parameter of Algorithm 3 is the sample size n.

The performance measures of the simulation experiment evaluate the proportion of unique observations in the sample, univariate statistics of observed variables, and the covariance structure between variables. The proportion of unique observations is calculated by dividing the number of unique observations by the sample size.

The univariate statistics are calculated for an arbitrarily chosen observed variable. The chosen variable is standardized by subtracting the true mean and dividing by the true standard deviation. The mean and standard deviation of the standardized variable z should be

0 and 1, respectively. In addition, the Kolmogorov-Smirnov statistic measuring the largest difference between the true cumulative distribution function and the empirical cumulative distribution function is reported. The correlation coefficient between two arbitrarily chosen observed variables was compared to the true correlation coefficient.

Online Appendix 4: Details of the Credit-Scoring Causal Model

The SCM depicted in Figure 1 of the main paper is explained here variable by variable in a topological order starting from the bottom of the graph. Ethnicity, age and gender do not have parents in the graph and are assumed to be independent from each other. Ethnicity is a categorical variable with with probabilities 0.75, 0.15 and 0.10 for the classes 1, 2 and 3, respectively. Age is uniformly distributed between 18 and 78 years. Gender has two values, 0 and 1, with equal probabilities. We have deliberately chosen not to label ethnicity and gender classes.

Education is an ordered factor with four levels (1 primary, 2 secondary, 3 tertiary, 4 doctorate) and is affected by ethnicity, age, gender, and the unobserved confounders U_2 and U_4. On average, education is higher for ethnicity class 1, for gender 0, and for older individuals. Marital status has three classes 1 (single), 2 (married or cohabiting) and 3 (divorced or widowed) and is affected by age and unobserved confounder U_2. The odds of class 2 compared to class 1 and the odds of class 3 compared to class 2 increase as a function of age. The number of children follows a Poisson distribution whose expectation is affected by ethnicity, age, marital status and education. On average, the number of children is higher for ethnicity classes 2 and 3, higher age (up to age 45), marital status 2 and 3 and higher education.

Job is a categorical variable with three classes (1 not working and not retired, 2 working, 3 retired) and is affected by ethnicity, age, gender, the number of children, education, and the unobserved confounders U_2 and U_4. Ethnicity classes 2 and 3, gender 0, young age and low education increase the probability that a person to the job class 1. The same features increase the odds of job class 2 compared to job class 3. The length of employment (in years) is a continuous variable that is affected by age and education that also determine a technical upper limit for the length of employment. Income is a continuous variable that describes the annual income (in euros) and is affected by job, education, age, the length of employment and unobserved confounder U_4. On average, income is higher for working individuals (job class 2), higher education classes and longer length of employment. For job class 2, income has its peak at age 58.

Address is an ordered factor with classes $1, 2, \ldots, 10$ and is affected by marital status, ethnicity, age, the number of children, income and the unobserved confounder U_5. The address variable is thought to be derived from the street address in such a way that higher classes of address are associated with higher income. In addition, ethnicity classes 2 and 3 have higher odds for living in an address of class 1 compared to ethnicity class 1. Higher age, higher number of children and marital status 2 and 3 are associated with higher classes of address. Housing has two possible values depending whether the home is rented (value 1) or owned (value 2). Housing is affected by age, marital status, the number of children, education, income and the unobserved confounders U_3 and U_5. Higher age, marital status

2 and 3, higher number of children, higher education and higher income increase the odds of home ownership.

Savings (in euros) is a continuous variable that is affected by income, age, ethnicity, education, marital status, and the unobserved confounders U_1 and U_3. Age and income have a joint effect on savings and it is assumed that on average 5% of income has been saved every year starting from age 18. In addition, there is an age-specific reduction to savings that has its peak at age 27. Higher education, ethnicity class 3 and marital status 2 (married or cohabiting) increase the amount of savings. The number of children decrease the amount of savings. Finally, there is a Gamma-distributed multiplier for the savings that depends on the background variables. This multiplier reflects the success of investments and inherited property.

Credit amount (in euros) is a continuous variable that is affected by age, income, job, housing, marital status, the number children, savings and the unobserved confounder U_1. The credit amount is the highest at age 40 and increases as a function of income. On average, the credit amount is higher for individuals who are working, have rented their house, have marital status 2 or 3 and have a high number of children. The credit amount decreases as a function of savings. The minimum of credit amount is 5000 euros.

Default is a binary variable that is affected by ethnicity, age, education, job, the length of employment, income, housing, savings, credit amount and the unobserved confounder U_1. Being a member of minority group (ethnicity classes 2 and 3) reduces the risk of default. Higher age, higher education, having a job, longer length of employment, higher income, home ownership and high amount of savings also reduce the risk of default. The risk of default increases as a function of credit amount.

The R code for the credit-scoring example is available in the repository `https://github.com/JuhaKarvanen/simulating_counterfactuals` in the file `fairness_example.R`.

References

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd edition). Cambridge University Press.