On the Relationship between Equivariant Predictive Models and Structural Causal Model Identification

Grace Yin *[1]

[1]Department of Statistics, University of British Columbia

April 21, 2022

Abstract

Structural Causal Models (SCMs) are a key component in causal inference and they have been used for a long time in many fields. We proposed an approach to identify the structure of an SCM based on Constant Risk Theorem. In particular, we proved that for an equivariant model and predictor, the risk function is constant across interventions described by the action of a group. These theories give a straightforward understanding of certain types of causal model identification. We also explored the risks on a specific SCM for linear regression predictive models with different types of interventions through simulation experiments.

1 Introduction

Structural causal models (SCMs) are a critical component of causal inference and has had broad implications in many fields for a long while [PJS17][MJ22]. The definition of an SCM is

Definition 1. *[PJS17]A structural causal model (SCM)* $\mathfrak{C} := (G, S, P_N)$ *consists of a collection* S *of* d *(structural) assignments*

$$X_j := f_j(PA_j, N_j), j = 1, \cdots, d$$

where $PA_j \subset \{X_1, \cdots, X_d\} \backslash \{X_j\}$ *are parents of* X_j *and* $P_N = P_{N_1, \cdots, N_d}$ *is the joint (product) distribution over the noise variables which are assumed to be jointly independent.* $G = (V, E)$ *is the graph contains vertices and edges.*

An intervention is a forcing term that pulls a variable toward a desired value [PJS17], denoted by $do(\cdots)$. An intervention distribution can be denoted by

$$P_X^{\tilde{\mathfrak{C}}} =: P_X^{\mathfrak{C}; do(X_k := \tilde{f}(\widetilde{PA})_k, \tilde{N}_k)}$$

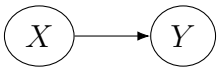

Figure 1: Example of a certain type SCM

It is important to identify the structure of an SCM to reveal the causal direction between the variables, and one common method is to perform conditional independence tests. However, numerous existing methods of conditional independence tests have high computational costs[PBM15][BK17], which precludes the utility. We observed that for some specific SCMs, there is a constant risk under certain shifts. For example, as Figure 1 shows, X is the direct parent of Y, and the functional assignment is a simple linear regression

$$f(X, E) = \beta_0 + \beta_1 X + E,$$

where $E \sim \mathcal{N}(0, \sigma^2)$. Then the risk of the predictive model $\hat{Y} = \hat{\beta}_1 X + \hat{\beta}_0$ remains constant if we intervene X by shifting $X' = X + a, a \in \mathbb{R}$. This gave us the intuition for an alternative approach with constant risk. Therefore, we proposed and proved the Constant Risk Theorem, and proposed an alternative approach to identify the structure of certain SCMs by exploiting constant risk theorem. This new approach based on constant risk theorem, is one of the main contributions of this project. We applied constant risk theorem to identify the structure of SCMs across interventions described by group actions.

In our alternative approach, we built the relationship between predictive models and SCMs from the perspective of invariant and equivariant mapping. Invariance is a property of a mathematical object which remains unchanged after a certain type group action applied to the objects whereas equvariance is a property of a mathematical object that has the same change after being applied by group action. Lehamm and Casella (1983) proposed the definition of an equivariant estimator.

Definition 2. *[LC98] If $\delta(X)$ is the estimator of ξ, and it satisfies that*

$$\delta(X_1 + a, \cdots, X_n + a) = \delta(X_1, \cdots, X_n) + a, \forall a \in \mathbb{R}$$

then the estimator $\delta(X)$ is equivariant under the transformation

$$X_i' = X_i + a,\ \xi' = \xi + a,\ d' = d + a$$

where d' is the estimated value of ξ'.

In our project, we used an equivariant predictor for constant risk theorem and SCM identification.

This project includes two main parts. First, we extended several decision theories and proposed the constant risk theorem. Particularly, we demonstrated that for an equivariant decision procedure, the risk function is constant under conditional shifts with an equivariant

*grace.yin@stat.ubc.ca

kernel and invariant loss function. More details can be found in Constant Risk Theorem section. The second part illustrated how to identify the structure of a specific SCM via the constant risk theorem. We also conducted several simulation experiments on a linear regression model and a particular SCM to demonstrate the framework of SCM identification.

2 Background

In probabilistic modelling, one basic problem can be expressed to use a random variable $X \in (\mathbf{X}, \mathcal{X})$ to predict a random variable $Y \in (\mathbf{Y}, \mathcal{Y})$ where $\mathbf{X}$ and $\mathbf{Y}$ are measurable spaces and $\mathcal{X}$ and $\mathcal{Y}$ are $\sigma-$algebras. The joint distribution of (X, Y) can be decomposed as

$$\tilde{P} = P \otimes Q_x,$$

where P is a marginal distribution on $(\mathbf{X}, \mathcal{X})$ and Q_x is a regular conditional distribution (or Markov kernel) from $(\mathbf{X}, \mathcal{X})$ into $(\mathbf{Y}, \mathcal{Y})$. We denote ρ to be the decision procedure, which can be thought as a predictor, $\mathbf{X} \times \mathcal{Z} \to [0, 1]$ where $(\mathbf{Z}, \mathcal{Z})$ is the decision space, L is the loss function $\mathcal{Y} \times \mathcal{Z} \to [-\infty, \infty)$. Then the risk function is defined as

$$R(\tilde{P}, \rho) = \int_{\mathcal{X}\times\mathcal{Y}\times Z} P(dx)Q_x(dy)\rho_x(dz)L(y, z).$$

Define $\mathcal{G}$ to be a group acting measurably on $\mathbf{X}$ and $\mathbf{Y}$, then for each $x \in \mathbf{X}$ and $y \in \mathbf{Y}$, we define a **conditional shift** g_x on joint distribution $\tilde{P}$ by

$$\forall g \in \mathcal{G},\ g_x\tilde{P} = (P \circ g^{-1}) \otimes Q_x,$$

and a **joint shift** g on joint distribution $\tilde{P}$ is

$$\forall g \in \mathcal{G},\ g\tilde{P} = \tilde{P} \circ g^{-1} = (P \circ g^{-1}) \otimes Q_{g^{-1}x} \circ g^{-1}.$$

The process of disintegration [Blo21] in invariant models under joint shifts can be found in **Appendix**. We say a model $\mathcal{P} = \{\tilde{P} : \theta \in \Theta\}$ is $\mathcal{G}$-invariant if for any group action $g \in \mathcal{G}$,

$$g\tilde{P} = \tilde{P} \circ g^{-1} \in \mathcal{P}$$

Define the kernel to be equivariant if

$$\forall g \in \mathcal{G},\ Q_{gx} = Q_x \circ g^{-1}.$$

We also define an invariant loss function and equivariant decision rule as below:

Definition 3. *[LC98] A loss function is invariant under g if it satisfies that for all $g \in \mathcal{G}$*

$$L(gy, gz) = L(y, z), \text{ and } y \in \mathcal{Y}, z \in \mathcal{Z}$$

When $gy = y + g$ for all $g \in \mathbb{R}$ and the equation holds, the problem is said to be location invariant.

Definition 4. *The function ρ: $\boldsymbol{X} \times \boldsymbol{Z} \to [0, 1]$, is equivariant if $\rho_{gx} = \rho_x \circ g^{-1}$; that is $\forall g \in \mathcal{G}, x \in \boldsymbol{x}$*

$$\int \rho_{gx}(dz)f(z) = \int \rho_x(dz)f(gz)$$

3 Decision Theorem Extension

In this section, we proposed and proved constant risk theorem for both conditional shifts and joint shifts. We also extended the theorem with unbiased minimum risk equivariant (MRE) estimator.

3.1 Constant Risk Theorem

Theorem 1. *For an invariant loss function, the risk of a decision procedure ρ is constant under conditional shifts $g_x\tilde{P}$ for any group action $g_x \in \mathcal{G}$, if ρ is equivariant and Q_x is equivariant. That is, for any $\rho \in$ equivariant predictors,*

$$\forall g_x \in \mathcal{G}, R(\rho, g_x\tilde{P}) = R(\rho, \tilde{P})$$

Proof.

$$\begin{aligned}
R(\rho, g_x\tilde{P}) &= g_x\tilde{P}\rho L && (1)\\
&= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} g_x P(dx)Q_x(dy)\rho_x(dz)L(y,z) && (2)\\
&= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P\circ g_x^{-1}(dx)Q_x(dy)\rho_x(dz)L(y,z) && (3)\\
&= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx)Q_{gx}(dy)\rho_{gx}(dz)L(y,z) && (4)\\
&= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx)Q_x\circ g_x^{-1}(dy)\rho_x\circ g_x^{-1}(dz)L(y,z) && (5)\\
&= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx)Q_x(dy)\rho_x(dz)L_(g_xy, g_xz) && (6)\\
&= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx)Q_x(dy)\rho_x(dz)L(y,z) && (7)\\
&= R(\rho, \tilde{P}) && (8)
\end{aligned}$$

□

Theorem 1 is the crucial result in our project, as it provides the theoretical foundation for our SCM identification approach. We will then apply Theorem 1 to simulation experiments in the following sections to explain the framework of SCM identification in details. We also proposed constant risk theorem under joint shifts, which is denoted by Theorem 2.

Theorem 2. *For an invariant predictive model and invariant loss function, the risk of a decision procedure ρ is constant under joint shifts $g\tilde{P}$ for group action $g \in \mathcal{G}$, if ρ is equivariant. That is,*

$$\forall \rho, R(\rho, g\tilde{P}) = R(\rho, \tilde{P})$$

Proof. Proof of Theorem 2:

$$R(g\tilde{P}\rho) = g\tilde{P}\rho L \tag{9}$$
$$= \int P \circ g^{-1}(dx) Q_{g^{-1}x} \circ g^{-1}(dy) \rho_x(dz) L(y,z) \tag{10}$$
$$= \int P(dx) Q_x(dy) \rho_{gx}(dz) L(gy,z) \tag{11}$$
$$= \int P(dx) Q_x(dy) \rho_x \circ g^{-1}(dz) L(gy,z) \tag{12}$$
$$= \int P(dx) Q_x(dy) \rho_x(dz) L(gy,gz) \tag{13}$$
$$= \int P(dx) Q_x(dy) \rho_x(dz) L(y,z) \tag{14}$$
$$= R(\rho,\tilde{P}) \tag{15}$$

□

3.2 Unbiased Risk Theorem

Using the technical tools in the proof of constant risk theorem, we can extend the risk-unbiased theorem. We showed that for a transitive group, the minimum risk equivariant estimator is risk-unbiased under both conditional shifts and joint shifts.

Definition 5. *[LC98] In an invariant estimation problem, if an equivariant estimator exists which minimizes the constant risk, it is called the minimum risk equivariant (MRE) estimator.*

Definition 6. *[LC98] The decision procedure ρ is risk-unbiased if*

$$\forall g \in \mathcal{G},\ R(\rho,\tilde{P}) \leq R(\rho, g\tilde{P})$$

Theorem 3. *If $\mathcal{G}$ is transitive, ρ is an MRE estimator, the kernel Q_x is equivaraint, then it is risk-unbiased under conditional shifts.*

Proof. Since $\mathcal{G}$ is transitive, then ρ_x is equivariant $\implies$ $\tilde{\rho} = \rho_{g_x x} \circ g_x$ is equivariant for $g_x \in \mathcal{G}$. Thus,

$$R(\rho, g_x\tilde{P}) = \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P \circ g_x^{-1}(dx) Q_x(dy) \rho_x(dz) L(y,z) \tag{16}$$
$$= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx) Q_{g_x x}(dy) \rho_{g_x x}(dz) L(y,z) \tag{17}$$
$$= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx) Q_x(dy) \rho_{g_x x} \circ g_x(dz) L(g_x y, g_x z) \tag{18}$$
$$= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx) Q_x(dy) \tilde{\rho}_x(dz) L(y,z) \tag{19}$$
$$\geq \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx) Q_x(dy) \rho_x(dz) L(y,z) \tag{20}$$
$$= R(\rho,\tilde{P}) \tag{21}$$

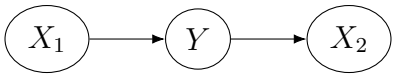

Figure 2: SCM 1

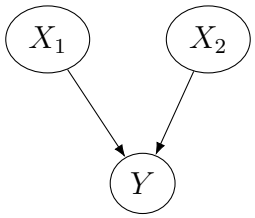

Figure 3: SCM 2

$$\begin{cases} X_1 \sim \mathcal{N}(0, \sigma^2) \\ Y := \beta_1 X_1 + \varepsilon_y \\ X_2 := \beta_2 Y + \varepsilon_2 \end{cases}$$

where $\varepsilon_y \sim \mathcal{N}(0, \sigma^2), \varepsilon_2 \sim \mathcal{N}(0, 1)$.

$$\begin{cases} X_1 \sim \mathcal{N}(0, \sigma^2) \\ X_2 := \varepsilon_2 \\ Y := \beta_1 X_1 + \beta_2 X_2 + \varepsilon_y \end{cases}$$

where $\varepsilon_y \sim \mathcal{N}(0, \sigma^2), \varepsilon_2 \sim \mathcal{N}(0, 1)$.

□

Theorem 4. *If $\mathcal{G}$ is transitive, ρ is an MRE estimator, then it is risk-unbiased under joint shifts.*

Proof. Since $\mathcal{G}$ is transtive, then ρ_x is equivariant $\implies \tilde{\rho} = \rho_{gx} \circ g$ is equivariant for $g \in \mathcal{G}$.

$$\begin{aligned} R(\rho, g\tilde{P}) &= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P \circ g^{-1}(dx) Q_{g^{-1}x} \circ g^{-1}(dy) \rho_x(dz) L(y, z) && (22) \\ &= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx) Q_x \circ g^{-1}(dy) \rho_{gx}(dz) L(y, z) && (23) \\ &= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx) Q_x(dy) \rho_{gx} \circ g(dz) L(gy, gz) && (24) \\ &= \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx) Q_x(dy) \tilde{\rho}_x(dz) L(y, z) && (25) \\ &\geq \int_{\mathbf{X}\times\mathbf{Y}\times\mathbf{Z}} P(dx) Q_x(dy) \rho_x(dz) L(y, z) && (26) \\ &= R(\rho, \tilde{P}) && (27) \end{aligned}$$

□

4 SCM Identification

In Section 3, we proved that for an invariant loss function and equivariant kernel, the risk of decision procedure is constant under conditional shifts if the decision procedure is equivariant. Going back to our original problems, suppose there are two candidate SCMs for a simulated data, and their graphical models and structural equations are displayed by Figure 2 and Figure 3. We want to identify the structure of the true SCM.

In both SCM 1 and SCM 2, Y is equivariant with X_1. Therefore, by constant risk theorem, under conditional shifts, the risk function of the equivariant predictive model using X_1 to predict Y should remain constant with an invariant loss function. However, Y is equivariant with X_2 in SCM 2 whereas in in SCM 1, Y is invariant with X_1. Then with an invariant loss function, the risk function of the predictive model using X_2 to predict Y is not constant under conditional shifts. Thus, we can generate the framework of identifying the structure of this certain SCM connected with constant risk theorem across shifting by group actions. There are different types of interventions and only the right intervention can yield expected results for identification. Assume that $\mathcal{G}_1 = \mathcal{G}_1 = (\mathbb{R}, +)$, and the group action is just shifting. We can conduct the interventions on X_1 or/and X_2 by assigning new values to X_1 or/and X_2.

$$\begin{cases} X_1' \leftarrow X_1 + g_1 \\ X_2' \leftarrow X_2 + g_2 \end{cases}$$

where $g_1, g_2 \in \mathcal{G} = (\mathbb{R}, +)$. Note that in an shifting environment, the conditional shifts are respected with the "soft" interventions in SCMs. We conducted several simulation experiments to demonstrate the framework of the identification of this certain SCM.

5 Simulation Experiments and Results

The goal of conducting simulation experiments is to verify constant risk theorem on identifiable SCMs by applying different group actions through conditional shifts on different causal variables. In our simulation experiment, we simulated the data from SCM 1. The predictive model we used is a linear regression model. The proof of an equivariant linear predictive model can be found in **Appendix**. We first simulated estimated coefficients of the following three linear regression models in shifting environment 0 ($g_1 = g_1^0,\ g_2 = g_2^0$).

$$\begin{cases} \ell_1 : \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ \ell_2 : \hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_2 \\ \ell_3 : \hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_2 \end{cases}$$

The estimated coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ only depend on X_1; $\hat{\alpha}_0$, $\hat{\alpha}_1$ only depend on X_2, and similarly, $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ depend on both X_1 and X_2. Then we set these coefficients for predictive models and computed the risks of these three predictive models with different shifting values on different variables. The invariant loss function we used is least square error function where

$$L(y, \hat{y}) = \frac{1}{2} \sum_{i=1}^{n} (y - \hat{y})^2.$$

The risk function is

$$R(\rho, P) = \int_{\mathcal{X} \times \mathcal{Y}} L(y, f(x)) dP(x, y), \text{ where } \rho \text{ is a Dirac measure} = \delta_{f(x)}$$

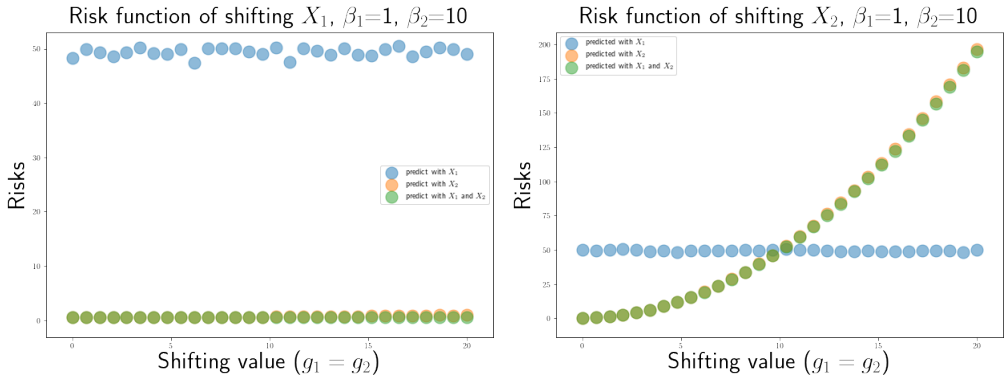

Figure 4: Verification of constant risk theorem by shifting X_1 and X_2

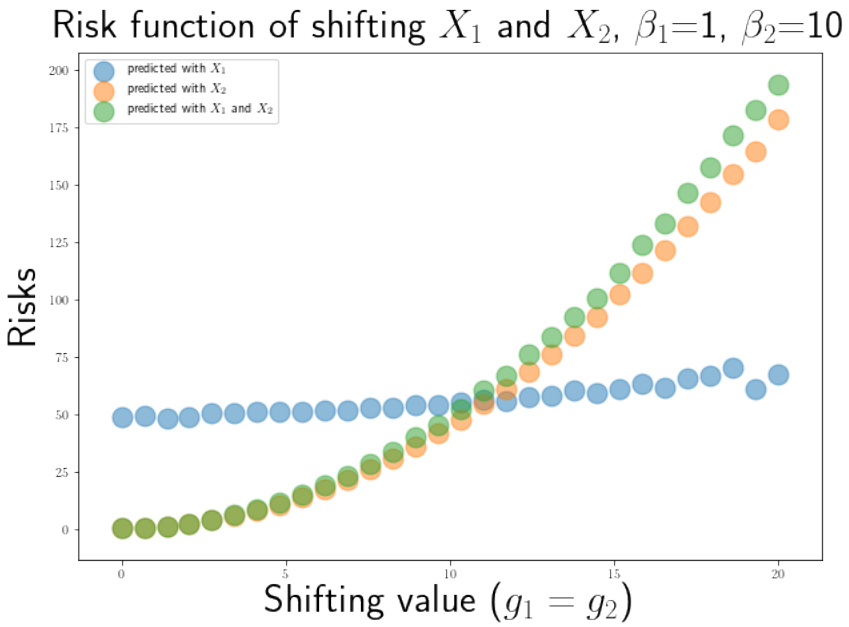

Figure 5: Crossover among ℓ_1, ℓ_2 and ℓ_3

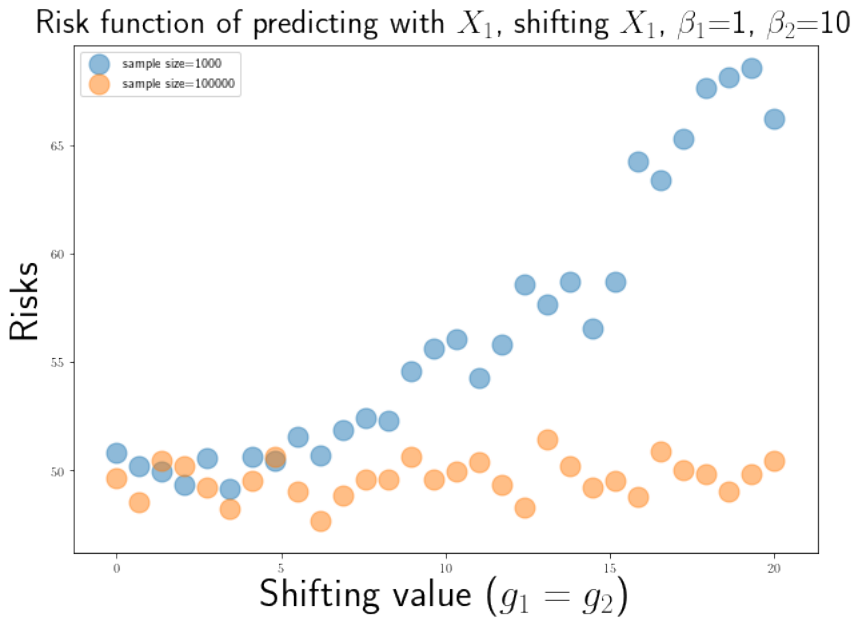

Figure 6: Risk of ℓ_1 across group action on X_1 with sample size $n = 1000$ and $n = 100000$

Our simulation results verified the constant risk theorem, that is the risk of the predictive model with X_1 is constant no matter what the shifting circumstances, since Y is equivariant with X_1. As Figure 4 shows,the risk of all the three predictive models remain constant shifting X_1, but only the risk of ℓ_1 is constant shifting X_2. This verified our framework of SCM identification that SCM 1 is the correct model for our simulated data points.

The simulation results also illustrated the main difference between our approach and the risk minimization approach. As Figure 5 shows, when we shift both X_1 and X_2, there is a crossover among these three predictive models. To be detailed, when the shifting value is less than the intersection, the risk of ℓ_1 is larger than the other two predictive models. Reversely, its risk is the lowest when the shifting value is greater than the intersection. The result implied that the "correct" model does not always hold the lowest risk.

Additionally, the influence of the sample size of simulated coefficients on the constant risk theorem is also a noteworthy finding. As Figure 6 shows, when simulating sample size for estimated coefficients is 1000, shifting X_1, the empirical risks of the predictive model with X_1 have a quadratic growth with the shifting values g; however, if we increase the sample size to 100000, the trend of the empirical risks is close to a constant trend. This can be explained by the following mathematical equation:

$$\hat{R}(g) \propto (\beta_0 - \hat{\beta}_0)^2 + (x_{1,i} + g)^2(\beta_1 - \hat{\beta}_1)^2,$$

Increasing the sample size of simulating coefficients can reduce the estimated errors $(\beta_0 - \hat{\beta}_0)$ and $(\beta_1 - \hat{\beta}_1)$ so that the term

$$(\beta_1 - \hat{\beta}_1)g^2 < \varepsilon, \ \forall \varepsilon > 0.$$

Hence, in the future work with constant risk theorem, it is important to set the simulating sample size to be large enough as in real life dataset is finite. The derivation of the empirical risk and the overall plots of simulated results can be found in **Appendix**.

6 Discussion and Conclusion

The simulation results showed that the constant risk theorem under conditional shifts can shed new light on the identification of structural features of a causal model given the framework across appropriate group actions. Therefore, we can apply constant risk theorem for certain types of SCMs under certain interventions. There are several directions for modification and improvement in the future. First, we can apply group actions on other examples of SCMs to verify the constant risk theorem. Identifying the theoretical interconnection between the risk function among several linear regression models is also important, as it lays the foundation for algorithmic computation. Implementing the respective algorithm and employing it in the context of learning from both observational and experimental data remain a problem worth investigating. Finally, we might also fit non-linear predictive models to verify the constant risk theorem and improve SCM identification.

7 Acknowledgement

I would like to sincerely appreciate my supervisor, Professor Benjamin Bloem-Reddy for his valuable guidance and support. I would also like to acknowledge the various members of the UBC Department of Statistics, particularly Johnny, Kenny and Sherry, who have played a key role during my master program. Last but not least, I would like to thank my parents for their unconditional love and support for my M.Sc. I also want to thank my friend, Anqi for helping me with scientific writing.

References

[Blo21] B. Bloem-Reddy. *Notes on predictive statistical decision theory*. July 2021.

[BK17] S. Burkart and F. J. Király. *Predictive Independence Testing, Predictive Conditional Independence Testing, and Predictive Graphical Modelling*. 2017. DOI: 10.48550/ARXIV.1711.05869. URL: https://arxiv.org/abs/1711.05869.

[LC98] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Second. New York, NY, USA: Springer-Verlag, 1998.

[MJ22] H. Miguel and R. James. *Causal inference: What if (the book)*. Jan. 2022. URL: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

[PBM15] J. Peters, P. Bühlmann, and N. Meinshausen. "Causal inference using invariant prediction: identification and confidence intervals". In: (2015). DOI: 10.48550/ARXIV.1501.01332. URL: https://arxiv.org/abs/1501.01332.

[PJS17] J. Peters, D. Janzing, and B. Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN: 0262037319.

A Appendix

A.1 Disintegration for Joint Shifts

Proof.

$$
\begin{aligned}
g\tilde{P}f &= (\tilde{P} \circ g^{-1})f && (28)\\
&= \tilde{P}(f \circ g) && (29)\\
&= (P \otimes Q_x)(f \circ g) && (30)\\
&= \int P(dx) \int Q_x(dy) f(gx, gy) && (31)\\
&= \int P(dx) \int (Q_x \circ g^{-1})(dy) f(gx, y) && (32)\\
&= \int (P \circ g^{-1})(dx) \int (Q_{g^{-1}x} \circ g^{-1})(dy) f(x, y) && (33)\\
&= ((P \circ g^{-1}) \otimes (Q_{g^{-1}x} \circ g^{-1})) && (34)
\end{aligned}
$$

Since the marginal model P on X is equivariant, by the uniquesness of disintegration,

$$gQ = Q_{g^{-1}x} \circ g^{-1}$$

□

A.2 Conditional Shifts and Joint Shifts on SCM 1

There are two types of interventions: conditional shifts and joint distributions shifts on the SCM4.1.

A.2.1 Conditional shifts

Conditional shifts are applied to the conditional distribution only.

Example 1. *Apply $g_1 \in \mathcal{G}$ to X_1:*

$$
\begin{cases}
g_1 X_1 := g_1 \varepsilon_1 \\
Y := g_1 X_1 + \varepsilon_y \\
X_2 := Y + \varepsilon_2
\end{cases}
$$

And the joint distribution is

$$\int (P_1 \circ g_1^{-1})(dx_1)(Q_1(x_1, dy))Q_2(y, dx_2) f(x_1, x_2, y) = \int P_1(dx_1)(Q_1(g_1 x_1, dy))Q_2(y, dx_2) f(g_1 x_1, x_2, y)$$

Example 2. *Apply $g_y \in \mathcal{G}_y$ to Y so that the SCM becomes*

$$
\begin{cases}
X_1 := \varepsilon_1 \\
g_Y Y := g_Y (X_1 + \varepsilon_y) \\
X_2 := g_Y Y + \varepsilon_2
\end{cases}
$$

So the joint distribution obeys

$$\int P_1(dx_1)(Q_y \circ g_y^{-1}(x_1, dy))Q_2(y, dx_2)f(x_1, x_2, y) = \int P_1(dx_1)(Q_y(x_1, dy))Q_2(g_y y, dx_2)f(x_1, x_2, g_y y)$$

Example 3. *Apply g_1 to X_1 *and* g_Y to Y*

$$\begin{cases} g_1 X_1 := g_1 \varepsilon_1 \\ g_y Y := g_y(g_1 X1 + \varepsilon_y) \\ X_2 := g_y Y + \varepsilon_2 \end{cases}$$

Then

$$\int (P_1^\theta \circ g_1^{-1})(dx_1)(Q_y^\theta \circ g_y^{-1})(x_1, dy)Q_2^\theta(y, dx_2)f(x_1, x_2, y)$$
$$= \int P_1^\theta(dx_1)Q_y^\theta(g_1 x_1, dy)Q_2^\theta(g_y y, dx_2)f(g_1 x_1, x_2, g_y y)$$

A.2.2 Joint shifts

The joint distribution with joint shifts is

$$P(g_1 X_1, g_y Y.g_2 X_2) = gP(X_1, Y, X_2)$$

In the SCM, it is

$$\begin{cases} g_1 X_1 := g_1 \varepsilon_1 \\ g_y Y := g_y(g_1^{-1}(g_1 X_1) + \varepsilon_y) = g_y(X_1 + \varepsilon_y) \\ g_2 X_2 := g_2(g_y^{-1}(g_y Y) + \varepsilon_2) = g_2(Y + \varepsilon_2) \end{cases}$$

This implies that

$$\int (P_1 \circ g_1^{-1})(dx_1)(Q_1 \circ g_y^{-1})(g_1^{-1} x_1, dy)(Q_2 \circ g_2^{-1})(g_y^{-1} y, dx_1)f(x_1, x_2, y)$$

$$= \int P_1(dX_1)Q_1(x_1, dy)Q_2(y, dx_2)f(g_1 x_1, g_y y, g_2 x_2)$$

In our simulation experiments, we will only focus on **conditional shifts**.

Now assume that $\mathcal{G}_1 = \mathcal{G}_2 = \mathcal{G}_y = (\mathbb{R}, +)$ and the group action is just a shift, i.e.,

$$\begin{cases} g_1 X_1 := X_1 + g_1 \\ g_y Y := Y + g_y \\ g_2 X_2 := X_2 + g_2 \end{cases}$$

Then we have the following lemma:

Lemma 1. *1. For any joint shifts $(g_1, g_y, g_2)_{joint}$, there exists a unique conditional shift $(\tilde{g}_1, \tilde{g}_y, \tilde{g}_2)_{cond}$ such that the shifted joint distributions are equal.i.e.,*

$$P^\theta \circ (g_1, g_y, g_2)_{joint}^{-1} = P^\theta \circ (\tilde{g}_1, \tilde{g}_y, \tilde{g}_2)_{cond}^{-1}$$

2. *For any conditional shifts $(g_1, g_y, g_2)_{joint}$, there exists a unique joint shift $(g_1, g_y, g_2)_{joint}$ such that the shifted joint distributions are equal.*

Proof. 1. **Lemma 1.1** $\forall g_1, g_2, g_y \in \mathbb{R}^+$,

$$\begin{cases} g_1 X_1 := X_1 + g_1 \\ g_y Y := Y + g_y \\ g_2 X_2 := X_2 + g_2 \end{cases}$$

whereas the conditional shift $(\tilde{g}_1, \tilde{g}_2, \tilde{g}_3)$ is

$$\begin{cases} \tilde{g}_1 X_1 := X_1 + \tilde{g}_1 \\ \tilde{g}_y Y := \tilde{g}_y(\tilde{g}_1 X_1 + \varepsilon_y) = X_1 + \tilde{g}_1 + \tilde{g}_y + \varepsilon_y \\ \tilde{g}_2 X_2 := \tilde{g}_2(g_y Y + \varepsilon_2) = \tilde{g}_2 + X_1 + \tilde{g}_1 + \tilde{g}_y + \varepsilon_2 + \varepsilon_y \end{cases}$$

For any joint shifts $(g_1, g_y, g_2)_{\text{joint}}$, there exists a (unique) conditional shift $(\tilde{g}_1, \tilde{g}_y, \tilde{g}_2)_{\text{cond}}$ such that the shifted joint distributions are equal as long as the conditional shift satisfies

$$\begin{cases} \tilde{g}_1 = g_1 \\ \tilde{g}_y = g_y - g_1 \\ \tilde{g}_2 = g_2 - g_y \end{cases}$$

2. The proof of **Lemma 1.2** is similar. For any conditional shifts,

$$\begin{cases} \tilde{g}_1 X_1 := X_1 + \tilde{g}_1 \\ \tilde{g}_y Y := X_1 + \tilde{g}_1 + \tilde{g}_y + \varepsilon_y \\ \tilde{g}_2 X_2 := \tilde{g}_2 + X_1 + \tilde{g}_1 + \tilde{g}_y + \varepsilon_2 + \varepsilon_y \end{cases}$$

there exists a (unique) joint distribution (g_1, g_2, g_3) such that the shifted joint ddistributions are equal as long as the joint shift satisfies

$$\begin{cases} g_1 = \tilde{g}_1 \\ g_y = \tilde{g}_y + \tilde{g}_1 \\ g_2 = \tilde{g}_1 + \tilde{g}_2 + \tilde{g}_y \end{cases}$$

□

A.3 Equivariant Prediction of Linear Regression

Lemma 2. *Shifting the predictors would not change the slope of the linear regression. In other words, the OLS estimator for $\hat{\beta}_1$ is **invariant**.*

Proof. When the constant vector lies in the span off the columns of the design matrix,

$$\hat{\beta} = b^{-1} Cov(X, Y)$$

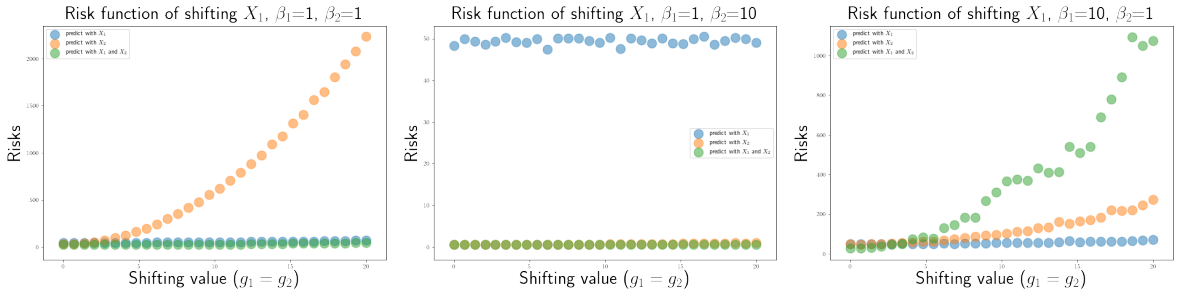

Figure 7: Prediction risks of shifting by X_1. In the first plot, $\beta_1 = \beta_2 = 1$; in the second plot, $\beta_1 = 10, \beta_2 = 1$; in the third plot, $\beta_1 = 1, \beta_2 = 10$.

where

$$b_{i,j} = Cov[X_i, X_j]$$

For a simple linear regression model,

$$\hat{\beta}_1(x_1) = \frac{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(y - \bar{y})}{\sum_{i=1}^{n}(x_{1,i} - \bar{x})^2}$$

Therefore, shifting the predictors would not change the slope of the linear regression. □

Lemma 3. *The linear regression model for prediction is equivariant.*

Proof. Suppose $\bar{x}_g \mapsto \bar{x} + g$, then

$$\bar{y}_g = \beta_1 \bar{x}_g + \beta_0 + \bar{\varepsilon}_{1,g} = \beta_1(\bar{x} + g) + \beta_0 + \bar{\varepsilon}_{1,g}$$

Therefore,

$$\hat{\beta}_0^{(g)} = \beta_1(\bar{x} + g) + \beta_0 + \bar{\varepsilon}_{1,g} - \hat{\beta}_1(\bar{x} + g) = (\beta_1 - \hat{\beta}_1)(\bar{x} + g) + \beta_0 + \bar{\varepsilon}_{1,g}$$

Thus, the prediction is

$$\begin{aligned}
\hat{y}_g &= \hat{\beta}_1(x + g) + \hat{\beta}_0^{(g)} \\
&= \hat{\beta}_1(x + g) + (\beta_1 - \hat{\beta}_1)(\bar{x} + g) + \beta_0 + \bar{\varepsilon}_{1,g} \\
&= \beta_1 \bar{x} + \beta_0 + \bar{\varepsilon}_{1,g} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x + \beta_1 g \\
&= \hat{\beta}_0 + \hat{\beta}_1 x + \beta_1 g \\
&= \hat{y} + \beta_1 g
\end{aligned}$$

Therefore, the predictive model is equivariant. □

A.4 Overall Results of Simulation Experiments

Figure 7, 8, and 9 respectively demonstrated the risks of the linear models predicted by X_1, X_2, as well as both X_1, X_2 under different conditional shifts with different shifting values.

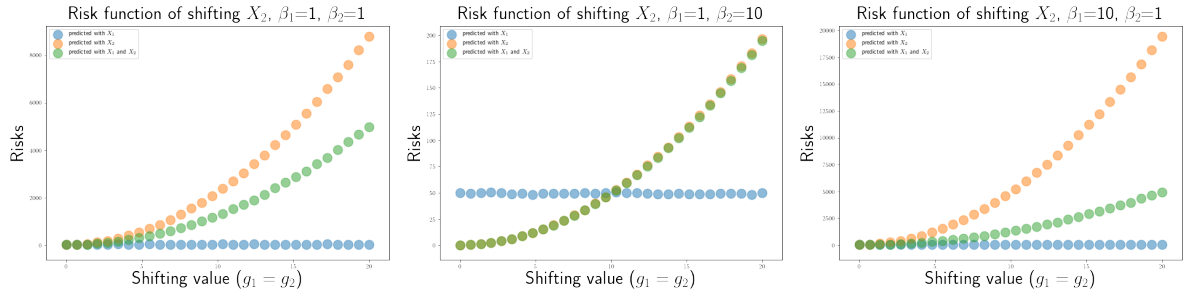

Figure 8: Prediction risks of shifting by X_2. In the first plot, $\beta_1 = \beta_2 = 1$; in the second plot, $\beta_1 = 10, \beta_2 = 1$; in the third plot, $\beta_1 = 1, \beta_2 = 10$.

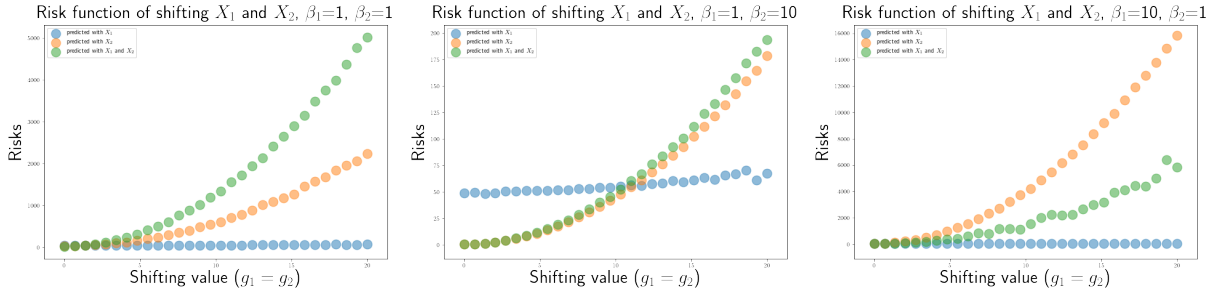

Figure 9: Prediction risks of shifting by X_1 and X_2. In the first plot, $\beta_1 = \beta_2 = 1$; in the second plot, $\beta_1 = 10, \beta_2 = 1$; in the third plot, $\beta_1 = 1, \beta_2 = 10$.

A.5 Lemmas with Risks of Linear Regression

Lemma 4. *The empirical variance of the prediction error has a quadratic growth with shifting values for finite data.*

Proof.

$$y - \hat{y} = (\beta_1 - \hat{\beta}_1)(x + g) + (\beta_0 - \hat{\beta}_0) + \varepsilon$$

Therefore,

$$Var(y - \hat{y}) \propto (x + g)^2 Var(\beta_1 - \hat{\beta}_1)$$

This implies that the random prediction error grows quadratically with the shifting value g. □

Lemma 5. *The empirical risk of linear predictive models under conditional shifts has a quadratic growth with shifting values for finite data.*

Proof.

$$\mathbb{E}(y - \hat{y})^2 = \mathbb{E}((\beta_1 - \hat{\beta}_1)(x + g) + (\beta_0 - \hat{\beta}_0) + \varepsilon)^2 \propto (\beta_0 - \hat{\beta}_0)^2 + (x_i + g)^2(\beta_1 - \hat{\beta}_1)^2$$

□

Therefore, for finite dataset, the empirical risk has a quadratic growth with g.

A.6 Github Link

The Python codes of this project can be found here.