Limitations of using LASSO for feature selection in causal discovery

Aaron Do

Supervisor: Dr. Wei Huang

Introduction to causal discovery

Causal Inference is an area of statistics that determines the direct causal effect of a treatment variable on the target variable. Using causal discovery algorithms like PC or FGES, a directed graph (Fig. 1) can be generated to show the causal relationships between variables.

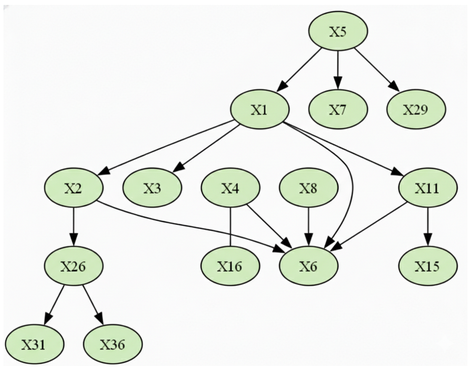

Figure 1. An example of a causal graph

Definitions of some key terms:

- For a causal relationship, $A \to B$, A is the **parent** of the **child** B, thus A is a direct cause of B.
- An **adjacent node** is any node that is connect by an edge. This includes undirected edges where the algorithm cannot determine the causal direction.
- **Siblings** share one or more parents.

These algorithms run the conditional independence test between all possible connections, which results in a time complexity of $\mathcal{O}(n^p)$ [1] and a memory complexity of $\mathcal{O}(2^p)$, for n observations and p features. Thus, alternative algorithms are required for high-dimensional, often dense, datasets, which often require a method of feature selection.

Introduction to LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) is a method of variable selection for linear regression. It finds the set of indicators which return a good linear fit, by reducing the square of the residuals, whilst penalising larger models. This is done by finding β to minimise the cost function, ϕ (eq. 1), for n observations and p coefficients in β. [3]

$$\phi = \underbrace{\frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2}_{\text{mean squared residual}} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{\text{penalty term}} \tag{1}$$

Using LASSO in feature selection for causal discovery has a time complexity of $\mathcal{O}(p^3)$ [2] and a memory capacity of $\mathcal{O}(p^3)$ [4] for n observations and p features. However, LASSO has some challenges:

- **Nonlinear Relationships:** LASSO assumes linearity between variables.
- **Dense Data:** LASSO assumes and tries to enforce sparsity, which means that variables with small impacts may not be included.

Research objectives

Determine which conditions LASSO can identify adjacent nodes in a causal discovery graph

Methodology

PyTetrad and scikit-learn libraries were used in Python in order to run LASSO regression and causal discovery on the simulated data. The simulated data explored various relationships between potential indicators and the target variable X_1. LASSO was then conducted to see if adjacent nodes would be considered indicators.

1. **Linear Relationships between Variables**
2. **Non-Linear Relationships between Variables**
   - Centred Parabola
   - Flat Line of Best Fit

It was assumed that there was a well defined equation between variables with a normally distributed error, ϵ, with mean 0.

Results: linear relationships

When LASSO is conducted, parents and children are found. Consider the following causal relationship,

$$\begin{aligned} A &\to X \to B \\ X &= m_1 A + \epsilon_1 \\ B &= m_2 X + \epsilon_2 \end{aligned}$$

where ϵ_1 and ϵ_2 are normally distributed with mean 0. For the predicted value, $\hat{X}$,

$$\begin{aligned} \hat{X} &= \gamma m_1 A + (1-\gamma)\frac{1}{m_2}B \\ &= \gamma X - \gamma\epsilon_1 + (1-\gamma)X + (1-\gamma)\epsilon_2 \\ &= X - \gamma\epsilon_1 + (1-\gamma)\epsilon_2 \end{aligned}$$

is an unbiased estimator with $E[\hat{X}_1] = X_1$. By Law of Large Numbers, as the sample size approaches infinity, $\hat{X} \sim N(X, |-\gamma\epsilon_1 + (1-\gamma)\epsilon_2|)$. This is more accurate than using only A or only B as an indicator, hence LASSO recovers both A and B as indicators of X.

Simulations show that assuming the underlying relationships are linear, **LASSO can identify the adjacent nodes** in various instances, including

- Presence of confounders, including Simpson's Paradox
- Presence of siblings

However, LASSO often **did not identify all adjacent nodes** if they were **proportional** to each other.

Results: non-linear relationships

Centred Parabola

For the causal relationship defined by

$$\begin{aligned} &A \to B \\ &B = \alpha A^2 + c + \epsilon \end{aligned}$$

where $\alpha \in \mathbb{R}\backslash\{0\}$, $c \in \mathbb{R}$, and the distribution of A is symmetric about 0, no relationship will be found by LASSO. This is because the line of best fit will have a gradient of 0.

Flat Line of Best Fit

A similar logic can be applied to non-parabolic equations where the gradient of the line of best fit is 0. For example, for a simple graph with only one connection, (Fig. 2)

$$\begin{aligned} &A \to B \\ &B = A^3 + A^2 - 2.4A + \epsilon \end{aligned}$$

for $A \sim U[-2, 2]$, no relationship is found. However if another non-linear connection is added (even if it has a zero gradient), such as a sibling or child of B,

$$\begin{aligned} &A \to B \quad \text{and} \quad A \to C \\ &B = A^3 + A^2 - 2.4A + \epsilon_1 \\ &C = A^2 + \epsilon_2 \end{aligned}$$

LASSO regression on B returns a linear combination of A and C as indicator variables. (Fig. 3)

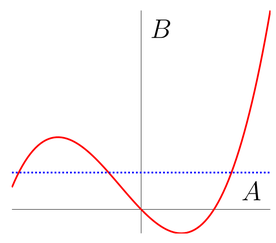

Figure 2. True causal relationship between A and B (red), which has a line of best fit with gradient 0 (blue).

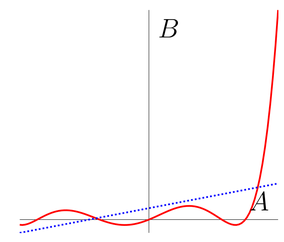

Figure 3. True causal relationship (red) shown by the data after taking a linear combination of A and C, written in terms of A, with a line of best fit (blue).

Equation for non-identifiability of a single connection

For the target variable, Y, with only one parent, X, and no children, it is not identifiable when the line of best fit has a gradient of 0. Thus the condition where it is non-identifiable is given by

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &\approx 0 \\ \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &\approx 0 \end{aligned}$$

Thus, by altering the formula to a continuous case which includes the distribution of X, we obtain the formula for non-identifiability

$$0 \approx \int f(x)\left[x - \int x f(x)\,dx\right]\left[g(x) - \int f(x)g(x)\,dx\right]dx \tag{2}$$

where $f(x)$ is the probability density function of X and $y = g(x) + \epsilon$. This equation is the same as showing orthogonality (zero covariance).

Key findings

- LASSO **almost always identified adjacent nodes**, including in instances with confounders. However, LASSO failed when adjacent nodes were proportional to each other, which is supported by how LASSO **enforces sparsity**.
- LASSO **failed to identify nodes** where the line of best fit had a **zero gradient**, unless other variables with a non-linear relationship was present. This is the same conditions as orthogonality. This method is still better than only using correlation as it can identify adjacent nodes using a linear combination of orthogonal variables.
- **Transforming variables** non-linearly could result in LASSO effectively identifying adjacent nodes even when there is orthogonality.

Limitations and future study

- The **bounds for identifiability** using the orthogonality equation need to be explored.
- Simulations were used to test different circumstances. The **robustness** needs to be analysed for high dimensionality for different sample sizes.
- The impact of **variable weights and distribution of errors** on the identifiability of adjacent nodes needs to be explored.

Conclusions

- LASSO is a fast method for variable selection that maintains low memory requirements.
- A potential algorithm: LASSO could be used for variable selection, followed by the PC algorithm to determine the causal direction. This could be repeated over all variables, whilst concatenating graphs, in order to construct a full causal graph.
- Extra care should be taken for non-linear relationships, transforming variables when necessary.

References

[1] Rahul Biswas and Somabha Mukherjee.
Consistent causal inference from time series with pc algorithm and its time-aware extension.
Statistics and Computing, 34(1):14, 2024.

[2] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani.
Least angle regression.
The Annals of Statistics, 32(2):407–499, 2004.

[3] Robert Tibshirani.
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

[4] Yaohui Zeng and Patrick Breheny.
The biglasso package: A memory- and computation-efficient solver for lasso model fitting with big data in r.
The R Journal, 12(2):6–24, 2020.