Markov Equivalence and Consistency in Differentiable Structure Learning

Chang Deng
Booth Business School, University of Chicago

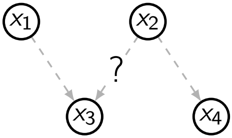
This is joint work with Kevin Bello, Pradeep Ravikumar, Bryon Aragam

https://arxiv.org/abs/2410.06163

Causal Discovery

Learning directed acyclic graph (DAGs) from data

- Inferring causal relations between variables and effects is an important task in all areas of science, e.g., genetics, finance, social science. Such causal relationship is usually represented by a graph G.
- The graph G can used to describe how the data are generating.



estimate ←

$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array}$$
$$\begin{bmatrix} 4.00 & -1.14 & 0.20 & -2.37 \\ -1.05 & 0.35 & -0.66 & -0.39 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

- The goal of causal discovery is to learning a DAG based on the observed data $\mathbf{X}$.

Score-based Structure Learning

- Score-based approaches: choosing best B to optimize the score $s(B; \mathbf{X})$.

$$\min_{B \in \{0,1\}^{p \times p}, B \in DAG} s(B; \mathbf{X})$$

  $s(B; \mathbf{X})$: how well an adjacency matrix $B \in \{0, 1\}^{p \times p}$ fits the data $\mathbf{X}$.
- Combinatorial optimization problem is generally known to be NP-complete.
- Zheng et al. [2018] has formulated such problem as a constrained continuous optimization problem, which is amendable to gradient-based optimization scheme.

Differentiable DAG Learning

- The problem is written as

$$\min_{B \in \mathbb{R}^{p \times p}} s(B; \mathbf{X}) \quad \text{subject to} \quad h(B) = 0. \tag{1}$$

- Discrete adjacency matrix $B \in \{0, 1\}^{p \times p}$ is relaxed to real matrices, i.e., $B \in \mathbb{R}^{p \times p}$
- $h : \mathbb{R}^{p \times p} \to [0, \infty)$ is a non-negative nonconvex differentiable function which penalize the circle in G. Specifically, $h(B) = 0$ if and only if B is a DAG.
- One example of $h(B)$, i.e., $h(B) = \mathrm{tr}(e^{B \circ B}) - p$.

Structural Equation Model(SEMs)

Data Generating Procedure

- Let $X = (X_1, \ldots, X_p)$
- An SEM $(X, f, P(N))$ is a collection of p structural equation

$$X_j = f_j(X, N_j), \quad \partial_k f_j = 0 \text{ if } k \notin \mathrm{PA}_j, \tag{2}$$

  1. $f = (f_j)_{j=1}^p$, $f_j : \mathbb{R}^{p+1} \to \mathbb{R}$
  2. $N = (N_1, \ldots, N_p)$ is independent noises with $P(N)$
  3. PA_j denotes parents node of j.
  4. The graphical structure implied by SEM can be represented by weighted adjacency matrix $B := B(f), B_{ij} = \|\partial_i f_j\|_2$

- In fact, essentially any distribution can be represented as an SCM of the form[Peters et al., 2017]

Parameters and the negative log-likelihood (NLL)

- Let distribution of X be $P(X, \psi, \xi)$ where $\psi \in \Psi \subseteq \mathbb{R}^m$, $\xi \in \Xi \subseteq \mathbb{R}^s$. Specifically, ψ, ξ denotes all the parameter for f, N separately. examples
- Given n i.i.d samples $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ where $\mathbf{x}_i \sim P(X; \psi, \xi)$, the negative log-likelihood and expected version

$$\ell_n(\psi, \xi) = -\frac{1}{n}\sum_{i=1}^{n} \log P(\mathbf{x}_i; \psi, \xi), \quad \ell(\psi, \xi) = -\mathbb{E}[\log P(\mathbf{x}; \psi, \xi)],$$

Identifiablity

Parameter and Structural Identifiability

Let $P(X, \psi^0, \xi^0)$ be the true distribution.

- *Parameter identifiability*: Is it possible to uniquely determine the parameters (ψ^0, ξ^0) based on observations from $P(X; \psi^0, \xi^0)$? Formally, is there any $(\widetilde{\psi}, \widetilde{\xi}) \neq (\psi^0, \xi^0)$, such that $P(X, \psi^0, \xi^0) = P(X, \widetilde{\psi}, \widetilde{\xi})$ almost surely?
- *Structural identifiability*: Is it possible to uniquely determine the DAG $G(B^0)$ based on observations from $P(X; \psi^0, \xi^0)$? In other words, is there any $(\widetilde{\psi}, \widetilde{\xi}) \neq (\psi^0, \xi^0)$ such that $P(X, \psi^0, \xi^0) = P(X, \widetilde{\psi}, \widetilde{\xi})$ but $G(B^0) \neq G(B(\widetilde{\psi}))$.

Question

What is the appropriate score $s(B; \mathbf{X})$ to ensure that the solution to (1) can recover the true G^0 (or up to an equivalent class), despite the model being unidentifiable in its parameters?

General linear Gaussian SEMs

General linear Gaussian SEMs

A nonidentifiable model

- Consider a well-known model which is nonidentifiable in term of parameters and structure.

$$
\begin{aligned}
& X = B^\top X + N, \\
& B \in \mathbb{R}^{p \times p} \\
& N \sim \mathcal{N}(0, \Omega) \qquad \Omega = \operatorname{diag}(\omega_1^2, \ldots, \omega_p^2)
\end{aligned} \tag{3}
$$

- The distribution of X

$$X \sim \mathcal{N}(0, \Theta^{-1}), \quad \Theta = \Theta_f(B, \Omega) := (I - B)\Omega^{-1}(I - B)^\top$$

  Subscript f refers to a function. In such case, Θ_f is function of (B, Ω).

- In term of general SEM (2). $\psi = B, \xi = \Omega$

Equivalence class

- It is known that model (3) is unidentifiable. This means that multiple pairs (B, Ω) can induce the same distribution $P(X)$.
- Define the equivalence class $\mathcal{E}(\Theta)$ equivalence class be the collection of all the parameters generate the

$$\mathcal{E}(\Theta) := \{(B, \Omega) : \Theta_f(B, \Omega) = \Theta\}. \quad (4)$$

- Which pair (B, Ω) to estimate? The "simplest" DAG!
- Find B that has the minimal number of nonzero entries in the equivalence class.
- Let number of edge in B, $s_B = |\{(i, j) : B_{ij} \neq 0\}|$.

Minimality

Definition (Minimality)

(B, Ω) is called a minimal-edge I-map[a] in the equivalence class $\mathcal{E}(\Theta)$ if $s_B \leq s_{\widetilde{B}}, \forall(\widetilde{B}, \widetilde{\Omega}) \in \mathcal{E}(\Theta)$. The set of all minimal-edge I-maps in the equivalence class $\mathcal{E}(\Theta)$ is referred to as the minimal equivalence class $\mathcal{E}_{\min}(\Theta)$:

$$\mathcal{E}_{\min}(\Theta) = \{(B, \Omega) : (B, \Omega) \text{ is minimal-edge I-map}, (B, \Omega) \in \mathcal{E}(\Theta)\}.$$

[a] This generalizes the classical definition for DAGs [e.g. Van de Geer and Bühlmann, 2013] to refer to the entire model with the distribution and graph encoded by the matrix B and the error variance Ω.

Regularization

- To distinguish elements in $\mathcal{E}(\Theta)$ from minimal element in $\mathcal{E}_{\min}(\Theta)$, a regularizer is needed to account the number of edges included.
- ℓ_0 is a natural choice, but its non-differentiable nature is amenable to continuous structure learning.
- ℓ_1 is not effective in precisely counting the number of edges, and also biased in parameter estimation.
- Alternatives such as smoothly clipped absolute deviation (SCAD) penalty and the minimax concave penalty (MCP) have been proposed to mitigate these shortcomings.

quasi-MCP

- A reparametrized version of MCP, termed quasi-MCP is used.

quasi-MCP: $$p_{\lambda,\delta}(t) = \lambda[(|t| - \frac{t^2}{2\delta})\mathbb{1}(|t| < \delta) + \frac{\delta}{2}\mathbb{1}(|t| > \delta)]$$

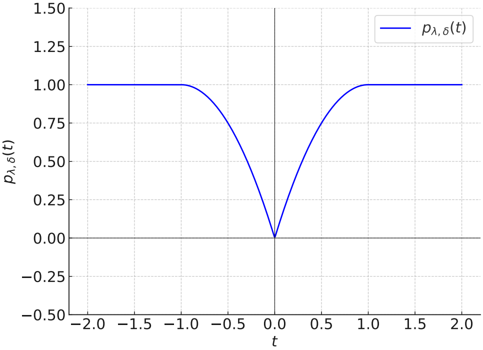


Figure: The plot $p_{\lambda,\delta}(t)$ with $\lambda = 2, \delta = 1$

Optimization

- The score function

$$s(B,\Omega;\lambda,\delta,\mathbf{X}) = \ell_n(B,\Omega) + p_{\lambda,\delta}(B)$$

  where $\ell_n(B,\Omega)$ is NLL.

- The optimization can be written as

$$\min_{B,\Omega} s(B,\Omega;\lambda,\delta,\mathbf{X}) \quad \text{subject to} \quad h(B)=0,\ \Omega > 0. \tag{5}$$

- The optimization requires minimizing $\ell_n(B,\Omega)$ and $p_{\lambda,\delta}$ simultaneously. Define the set of global minimizers

$$\mathcal{O}_{n,\lambda,\delta} = \{(B^*,\Omega^*) : (B^*,\Omega^*) \text{ is a minimizer of } (5)\}. \tag{6}$$

Provably recovering minimal models

Theorem

Let X follow model (3) with (B^0, Ω^0) and $\Theta^0 = \Theta_f(B^0, \Omega^0)$. Let $\mathbf{X}$ be n i.i.d. samples from $P(X)$. Then, for all sufficiently small $\lambda, \delta > 0$ (independent of n), it holds that $P(\mathcal{O}_{n,\lambda,\delta} = \mathcal{E}_{\min}(\Theta^0)) \to 1$ as $n \to \infty$.

The elements in $\mathcal{E}_{\min}(\Theta)$ not only represent the "simplest" DAG model for X in term of edge count, but also bears a deep connection to classical notion such as Markov equivalence.

Minimal Models and Markov Equivalence Class

Definition (Markov, faithful, Markov equivalence class)

1. $\mathcal{I}(P)$: the set of conditional independence relations implied by P
2. $\mathcal{I}(G)$ denote the set of d-separations implied by the graph G.
3. P is *markov* to G if $\mathcal{I}(G) \subset \mathcal{I}(P)$
4. P is *faithful* to G if $\mathcal{I}(P) = \mathcal{I}(G)$.
5. For any DAG G, the Markov equivalence class is $\mathcal{M}(G) = \{\widetilde{G} : \mathcal{I}(\widetilde{G}) = \mathcal{I}(G)\}$

Minimal Models in the same Markov Equivalence Class

Lemma

Let X follow model (3) with (B^0, Ω^0) and $\Theta^0 = \Theta_f(B^0, \Omega^0)$. Assume that $P(X)$ is faithful to $G^0 := G(B^0)$. Then $\mathcal{M}(G^0) = \mathcal{G}(\mathcal{E}_{\min}(\Theta^0))$.

where $\mathcal{G}(\mathcal{E}_{\min}(\Theta)) := \{G(B) : (B, \Omega) \in \mathcal{E}_{\min}(\Theta)\}$.

Theorem

Consider the setup in Theorem above and assume additionally that $P(X)$ is faithful to $G^0 := G(B^0)$. Then, for all sufficiently small $\lambda, \delta > 0$ (independent of n), it holds that $P(\mathcal{G}(\mathcal{O}_{n,\lambda,\delta}) = \mathcal{M}(G^0)) \to 1$ as $n \to \infty$.

Scale invariance and standardization

Standardization of data would make causal structural learning algorithms utilizing least square loss fail [Reisach et al., 2021]. But it turns out that NLL is scale-invariant.

Theorem (Scale invariance)

Under the same setting as previous Theorem, the solutions to (5) are scale-invariant. That is, for any $n \geq 0$, let

$$\mathcal{O}_{n,\lambda,\delta}(\mathbf{X}) = \{(B^*, \Omega^*) : (B^*, \Omega^*) \text{ is a minimizer of } (5) \text{ with data } \mathbf{X}\},$$
$$\mathcal{O}_{n,\lambda,\delta}(\mathbf{Z}) = \{(B^*, \Omega^*) : (B^*, \Omega^*) \text{ is a minimizer of } (5) \text{ with data } \mathbf{Z}\},$$

where $\mathbf{Z}$ is the standardized version of $\mathbf{X}$. For all sufficiently small $\lambda, \delta > 0$ and all n, we have $\mathcal{G}(\mathcal{O}_{n,\lambda,\delta}(\mathbf{X})) = \mathcal{G}(\mathcal{O}_{n,\lambda,\delta}(\mathbf{Z}))$. Moreover, for all sufficiently small $\lambda, \delta > 0$ we have

$$P\left[\mathcal{G}(\mathcal{O}_{n,\lambda,\delta}(\mathbf{X})) = \mathcal{G}(\mathcal{O}_{n,\lambda,\delta}(\mathbf{Z})) = \mathcal{G}(\mathcal{E}_{\min}(\Theta_f(B^0, \Omega^0)))\right] \to 1 \quad \text{as } n \to \infty.$$

General Models

General Models and its minimal models

- Assume X follows model (2) and the induced distribution is denoted by $P(X;\psi^0,\xi^0)$.
- Define the equivalence class $\mathcal{E}(\psi^0,\xi^0)$,

$$\mathcal{E}(\psi^0,\xi^0)=\{(\psi,\xi):P(x;\psi,\xi)=P(x;\psi^0,\xi^0),\forall x\in\mathbb{R}^p\}.$$

Lemma

(ψ,ξ) is called a minimal-edge I-map in the equivalence class $\mathcal{E}(\psi^0,\xi^0)$ if $s_{B(\psi)}\le s_{B(\widetilde{\psi})},\forall(\widetilde{\psi},\widetilde{\xi})\in\mathcal{E}(\psi^0,\xi^0)$. We further define

$$\mathcal{E}_{\min}(\psi^0,\xi^0)=\{(\psi,\xi):(\psi,\xi)\text{ is minimal-edge I-map},\ (\psi,\xi)\in\mathcal{E}(\psi^0,\xi^0)\}.$$

Nonconvex regularized log-likelihood

- Similar in spirit to previous Theorem, define the following problem

$$\min_{\psi \in \Psi, \xi \in \Xi} \ell_n(\psi, \xi) + p_{\lambda,\delta}(B(\psi)) \quad \text{subject to} \quad h(B(\psi)) = 0, \tag{7}$$

- The set of global minimizers.

$$\mathcal{O}_{n,\lambda,\delta} = \{(\psi^*, \xi^*) : (\psi^*, \xi^*) \text{ is minimizer of } (7)\}.$$

Theoretical Guarantee for General Model

Assumption (A)

(1) $|\mathcal{E}(\psi^0, \xi^0)|$ *is finite. (2)* $B(\psi)$ *is L-Lipschitz w.r.t.* ψ*, i.e.* $\frac{\|B(\psi_1)-B(\psi_2)\|_2}{\|\psi_1-\psi_2\|_2} \leq L$.

Assumption (B)

For any α *such that* $\ell(\psi^0, \xi^0) < \alpha$*, the level set* $\{(\psi, \xi) : \ell(\psi, \xi) \leq \alpha\}$ *is bounded, where* $\ell(\psi, \xi)$ *is the expected NLL*

Theoretical Guarantee for General Model

Theorem

Let X follow model (2) with parameters (ψ^0, ξ^0) and let $\mathbf{X}$ be n i.i.d. samples from $P(X; \psi^0, \xi^0)$. Under Assumptions A-B, for all sufficiently small $\lambda, \delta > 0$ (independent of n), it holds that $P(\mathcal{G}(\mathcal{O}_{n,\lambda,\delta}) = \mathcal{G}(\mathcal{E}_{\min}(\psi^0, \xi^0))) \to 1$ as $n \to \infty$.

Theorem

Under the setting in Theorem above and assuming that $P(X; \xi^0, \psi^0)$ is faithful with respect to $G^0 := G(B(\psi^0))$. Then, for all sufficiently small $\lambda, \delta > 0$ (independent of n), it holds that $P(\mathcal{O}_{n,\lambda,\delta} = \mathcal{M}(G^0)) \to 1$ as $n \to \infty$.

Experiments

Experiments on raw data $\mathbf{X}$

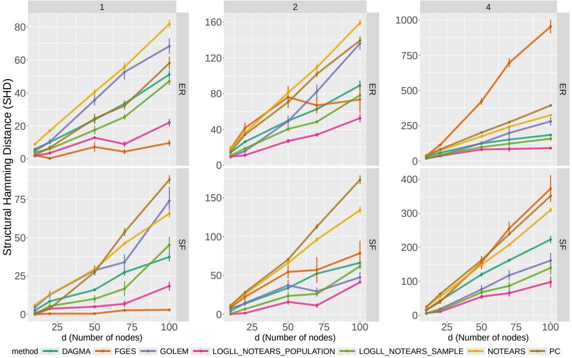


Figure: Results in terms of SHD between MECs of estimated graph and ground truth on raw data $\mathbf{X}$. Lower is better. Column: $k = \{1, 2, 4\}$. Row: random graph types. {ER,SF}-k = {Scale-Free,Erdős-Rényi } graphs with kd expected edges. Here $p = \{10, 20, 50, 70, 100\}$, $n = 1000$.

Experiments on standardized data $\mathbf{Z}$

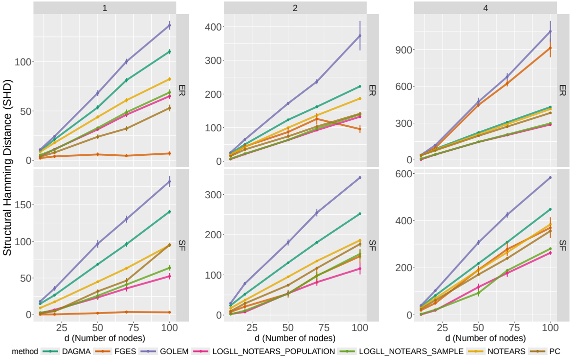
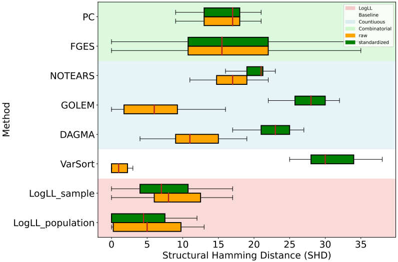
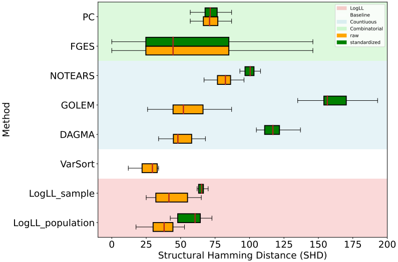


Figure: Results in terms of SHD between MECs of estimated graph and ground truth on standardized data $\mathbf{Z}$. Lower is better. Column: $k = \{1, 2, 4\}$. Row: random graph types. $\{$ER,SF$\}$-k = $\{$Scale-Free,Erdős-Rényi $\}$ graphs with kd expected edges. Here $p = \{10, 20, 50, 70, 100\}, n = 1000$.

Direct comparison



(a) $p = 10$, graph ="ER", $k = 2$



(b) $p = 50$, graph ="ER", $k = 2$

Figure: Comparison of raw (orange) vs. standardized (green) data. SHD (lower is better) between Markov equivalence classes (MEC) of recovered and ground truth graphs for ER-2 graphs with 10 (left) or 50 (right) nodes. In (b), SHD for VarSort with standardized data is omitted due to its average exceeding 300.

Solving Optimization (5) Exactly

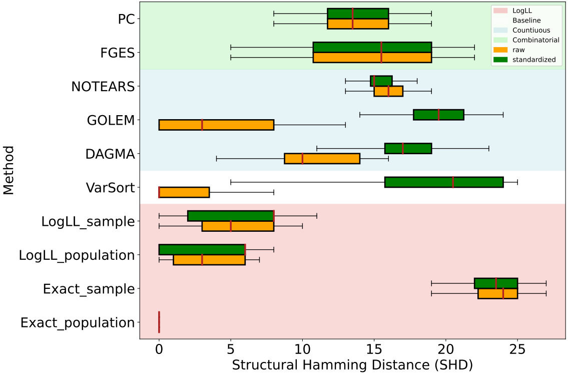


Figure: Both Exact-sample and Exact-population produce the same DAG structure for raw data $\mathbf{X}$ and standardized data $\mathbf{Z}$. When the population covariance matrix is known, $\mathcal{E}_{\min}(\Theta^0) = \mathcal{M}(G^0)$, resulting in an SHD of zero.

Neural Network

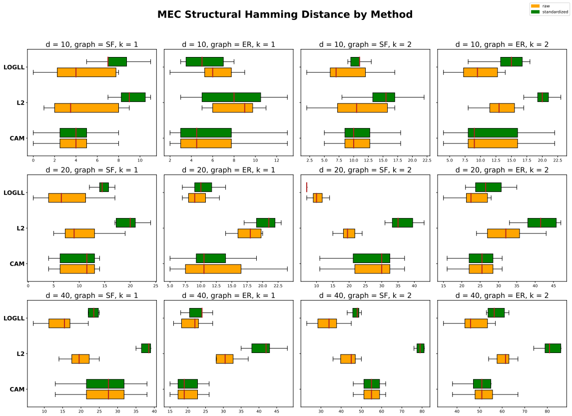


Figure: Structural Hamming distance (SHD) between Markov equivalence classes (MEC) of recovered and ground truth graphs. **LOGLL** (i.e. LOGLL-NOTEARS) stands for NOTEARS method with log-likelihood and quasi-MCP, **L2** (i.e. NOTEARS) stands for NOTEARS method with least square and ℓ_1.

General Linear Model with Binary Output

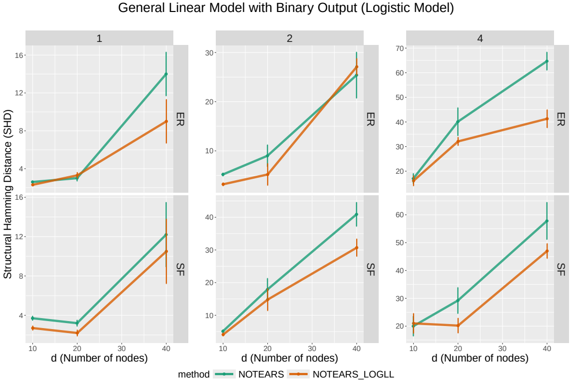


Figure: Structural Hamming distance (SHD) for Logistic Model, Row: random graph types, {SF, ER}-k= {Scale-Free,Erdős-Rényi } graphs. Columns: kd expected edges. NOTEARS_LOGLL (i.e. LOGLL-NOTEARS) uses log-likelihood with quasi-MCP, NOTEARS use log-likelihood with ℓ_1. Error bars represent standard errors over 10 simulations.

Thanks for Listening!

References I

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Sara Van de Geer and Peter Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

Appendix

Identifiablity

Definition (identifiablity)

$\mathcal{G}$ is identifiable if no other SEMs can induce the same distribution $P(X)$ with a different DAG.

Find $(\tilde{W}(\pi), \tilde{\Omega}(\pi))$

Define

$$\Theta^0 := \Theta(W^0, \Omega^0) = (I - W^0)[\Omega^0]^{-1}(I - W^0)^\top$$

$$(P_\pi A)_{ij} = A_{\pi(i),\pi(j)}$$

- Calculate $P_\pi(\Theta^0)$
- Use Cholesky decomposition: $P_\pi \Theta^0 = (I - L)D^{-1}(I - L)^\top = \Theta(L, D)$
- $\Theta^0 = (P_\pi)^{-1}\Theta^0(L, D) = P_{\pi^{-1}}\Theta^0(L, D) = \Theta^0(P_{\pi^{-1}}L, P_{\pi^{-1}}D)$
- $\tilde{B}_0(\pi) = P_{\pi^{-1}}L, \tilde{\Omega}_0(\pi) = P_{\pi^{-1}}D$

Back to Equivalence Class