Nonparametric Partial Disentanglement via Mechanism Sparsity

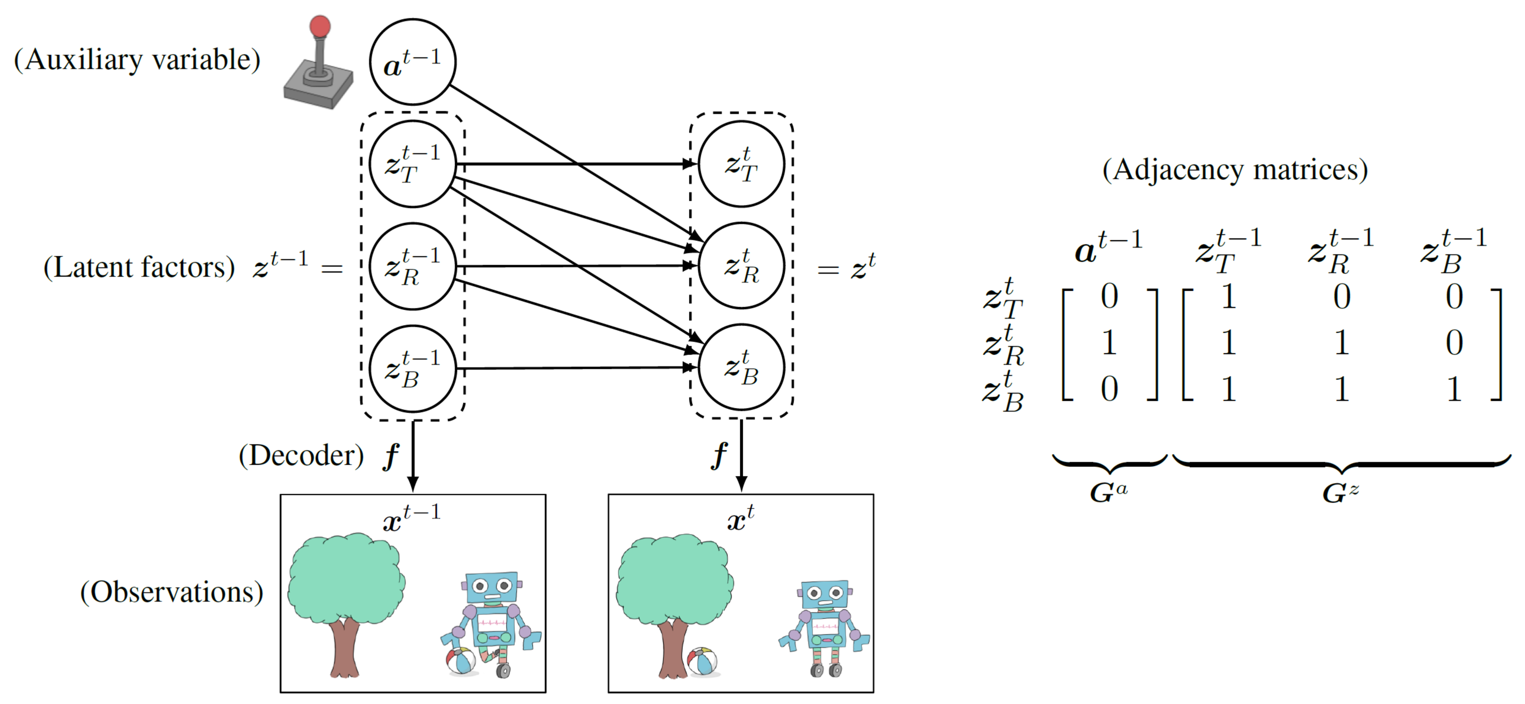
Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, Simon Lacoste-Julien

SAMSUNG Samsung Advanced Institute of Technology AI Lab Montreal · Mila · Université de Montréal

Contributions

- New principle for disentanglement based on mechanism sparsity regularization motivated by novel identifiability guarantees
- Extending [2] to nonparametric and partial disentanglement results
- Given a latent ground-truth graph, our theory describes how entangled the learned representation is expected to be
- Algorithm based on VAEs and constrained optimization to enforce sparsity
- Many examples to show the scope of our theory

An identifiable model with latent dynamics



- **Observation (e.g. image):** $\boldsymbol{x}^t \in \mathbb{R}^{d_x}$ for all $t \in [T]$
- **Latent factors:** $\boldsymbol{z}^t \in \mathbb{R}^{d_z}$ for all $t \in [T]$, with $d_z \leq d_x$
- **Auxiliary variables (e.g. action or intervention index):** $\boldsymbol{a}^t \in \mathbb{R}^{d_a}$ for all $t \in [T]$
- $\boldsymbol{x}^t = \boldsymbol{f}(\boldsymbol{z}^t) + \boldsymbol{n}^t$, where $\boldsymbol{n}^t \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{f}$ is a diffeomorphism onto its image
- **Latent dynamical system:** $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \prod_{i=1}^{d_z} p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}_{\mathbf{Pa}_i^z}, \boldsymbol{a}^{<t}_{\mathbf{Pa}_i^a})$ where $\mathbf{Pa}_i^z$ and $\mathbf{Pa}_i^a$ are the parents of $\boldsymbol{z}_i^t$ in graphs $\boldsymbol{G}^z$ and $\boldsymbol{G}^a$.

Terminology & Notation

- **Ground-truth parameter:** $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$
- **Learned parameter:** $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{p}, \hat{\boldsymbol{G}})$
- **Entanglement map:** $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \hat{\boldsymbol{f}}$, assuming $\boldsymbol{f}(\mathbb{R}^{d_z}) = \hat{\boldsymbol{f}}(\mathbb{R}^{d_z})$
- **Entanglement graph:** $\boldsymbol{V}_{i,j} = 0 \iff \forall \boldsymbol{z} \in \mathbb{R}^{d_z}, \frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{z}_j}(\boldsymbol{z}) = 0$
- **Complete disentanglement:** Graph $\boldsymbol{V}$ is a permutation, i.e. $\boldsymbol{v} = \boldsymbol{d} \circ \boldsymbol{P}^\top$ where $\boldsymbol{d}$ is element-wise
- **Partial disentanglement:** Graph $\boldsymbol{V}$ is not complete nor a permutation
- $\mathbb{R}^{m \times n}_{\boldsymbol{B}} := \{\boldsymbol{M} \in \mathbb{R}^{m \times n} \mid \boldsymbol{B}_{i,j} = 0 \implies \boldsymbol{M}_{i,j} = 0\}$ (it's a vector space!)
- Abuse of notation: $\boldsymbol{M} \subseteq \boldsymbol{B} \iff \boldsymbol{M} \in \mathbb{R}^{m \times n}_{\boldsymbol{B}}$

Constrained VAE approach

- **Approximate posterior:** $q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T}) := \prod_{t=1}^T q(\boldsymbol{z}^t \mid \boldsymbol{x}^t)$
- **Transition model:** $\hat{p}(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ is a Gaussian distribution with mean $\hat{\boldsymbol{\mu}}_i(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ (theory allows for more flexibility)
- **Evidence lower bound:**

$$\log \hat{p}(\boldsymbol{x}^{\leq T} | \boldsymbol{a}^{<T}) \geq \text{ELBO}(\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{G}}, q; \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T}) :=$$

$$\sum_{t=1}^T \mathop{\mathbb{E}}_{q(\boldsymbol{z}^t | \boldsymbol{x}^t)} [\log \hat{p}(\boldsymbol{x}^t \mid \boldsymbol{z}^t)] - \mathop{\mathbb{E}}_{q(\boldsymbol{z}^{<t} | \boldsymbol{x}^{<t})} KL(q(\boldsymbol{z}^t \mid \boldsymbol{x}^t) || \hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}))$$

- **Adding sparsity constraint:**

$$\max_{\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \boldsymbol{\gamma}, q} \mathbb{E}_{\hat{\boldsymbol{G}} \sim \sigma(\boldsymbol{\gamma})} \text{ELBO}(\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{G}}, q) \text{ subject to } \mathbb{E}_{\hat{\boldsymbol{G}} \sim \sigma(\boldsymbol{\gamma})} ||\hat{\boldsymbol{G}}||_0 \leq \beta\,.$$

- Using Gumbel-sigmoid trick to estimate gradient w.r.t. $\boldsymbol{\gamma}$.
- Constrained optimization is done by doing **gradient ascent-descent on the Lagrangian**. We are using the python library `cooper` [1].

[1] J. Gallego-Posada and J. Ramirez. Cooper: a toolkit for lagrangian-based constrained optimization. https://github.com/cooper-org/cooper, 2022.

[2] S. Lachapelle, Rodriguez Lopez, P., Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In First Conference on Causal Learning and Reasoning, 2022.

Proof sketch & G-preserving matrices

Proposition: If $p(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T}) = \hat{p}(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T})$ everywhere, then $\boldsymbol{f}(\mathbb{R}^{d_z}) = \hat{\boldsymbol{f}}(\mathbb{R}^{d_z})$ and $\hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) |\det D\boldsymbol{v}(\boldsymbol{z}^t)|\,.$

- By taking the log on both sides and computing derivative w.r.t. both $\boldsymbol{z}^t$ and $\boldsymbol{a}^\tau$ with $\tau < t$ we get

$$\underbrace{H^{t,\tau}_{z,a} \log \hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})}_{\subseteq \hat{\boldsymbol{G}}^a} = D\boldsymbol{v}(\boldsymbol{z}^t)^\top \underbrace{H^{t,\tau}_{z,a} \log p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})}_{\subseteq \boldsymbol{G}^a}\,.$$

- The Hessian of the log-conditional-densities have the same sparsity as $\boldsymbol{G}^a$!
- If we assume that $\hat{\boldsymbol{G}}^a = \boldsymbol{G}^a$, we have that $D\boldsymbol{v}(\boldsymbol{z}^t)$ preserves the graph $\boldsymbol{G}^a$, which motivates the following definition:

  $\boldsymbol{G}$-preserving matrix: $\boldsymbol{C}^\top \mathbb{R}^{m \times n}_{\boldsymbol{G}} \subseteq \mathbb{R}^{m \times n}_{\boldsymbol{G}}$ (forms a group when $\boldsymbol{C}$ are invertible!)

  Proposition: A matrix $\boldsymbol{C}$ is $\boldsymbol{G}$-preserving if and only if

$$\text{for all } i, j,\ \boldsymbol{G}_{i,\cdot} \not\subseteq \boldsymbol{G}_{j,\cdot} \implies \boldsymbol{C}_{i,j} = 0\,. \tag{1}$$

- In other words, $\boldsymbol{G}$-preserving matrices are sparse!
- Thus, if for all $\boldsymbol{z}^t$, $H^{t,\tau}_{z,a} \log p$ spans $\mathbb{R}^{d_z \times d_a}_{\boldsymbol{G}^a}$, then $D\boldsymbol{v}(\boldsymbol{z}^t)$ is $\boldsymbol{G}^a$-preserving!
- Result assumes only $\|\hat{\boldsymbol{G}}^a\|_0 \leq \|\boldsymbol{G}^a\|_0$, so additional permutation indeterminacy.
- Similar argument works for sparsity of $\boldsymbol{G}^z$:

$$\underbrace{H^{t,\tau}_{z,z} \log \hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})}_{\subseteq \hat{\boldsymbol{G}}^z} = D\boldsymbol{v}(\boldsymbol{z}^t)^\top \underbrace{H^{t,\tau}_{z,z} \log p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})}_{\subseteq \boldsymbol{G}^z} D\boldsymbol{v}(\boldsymbol{z}^\tau)$$

Identifiability results

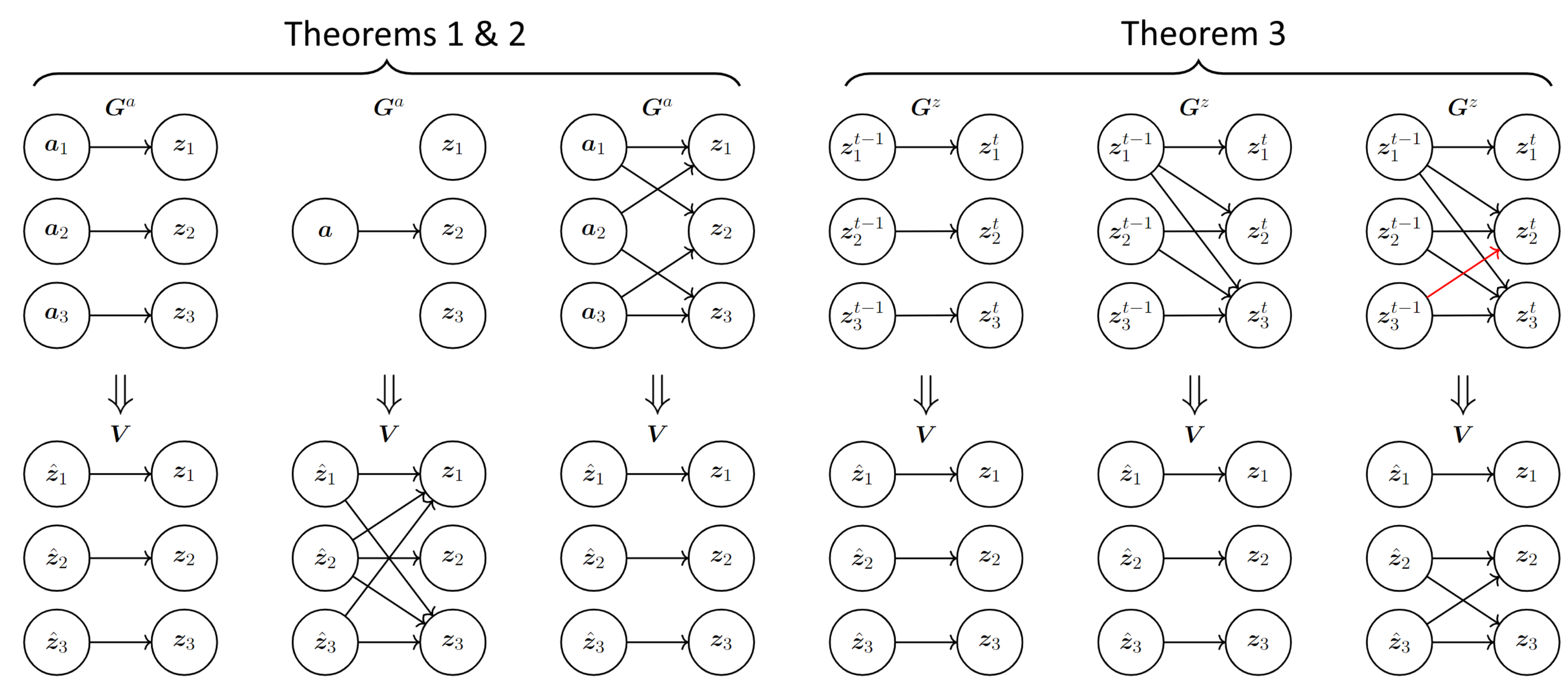
Nonparametric identifiability results

Assume $p(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T}) = \hat{p}(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T})$.

Theorem 1: Partial disentanglement via sparse $\boldsymbol{G}^a$ - continuous a
If "$H^{t,\tau}_{z,a} \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ spans $\mathbb{R}^{d_z \times d_a}_{\boldsymbol{G}^a}$" and $||\hat{\boldsymbol{G}}^a||_0 \leq ||\boldsymbol{G}^a||_0$, then $\boldsymbol{V} = \boldsymbol{C}\boldsymbol{P}^\top$ where $\boldsymbol{C}$ is $\boldsymbol{G}^a$-preserving.

Theorem 2: Partial disentanglement via sparse $\boldsymbol{G}^a$ - discrete a (important for **interventions**!)
If "$\Delta^{\tau,\vec{\epsilon}}_a D^t_z \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ spans $\mathbb{R}^{d_z \times d_a}_{\boldsymbol{G}^a}$" and $||\hat{\boldsymbol{G}}^a||_0 \leq ||\boldsymbol{G}^a||_0$, then $\boldsymbol{V} = \boldsymbol{C}\boldsymbol{P}^\top$ where $\boldsymbol{C}$ is $\boldsymbol{G}^a$-preserving.

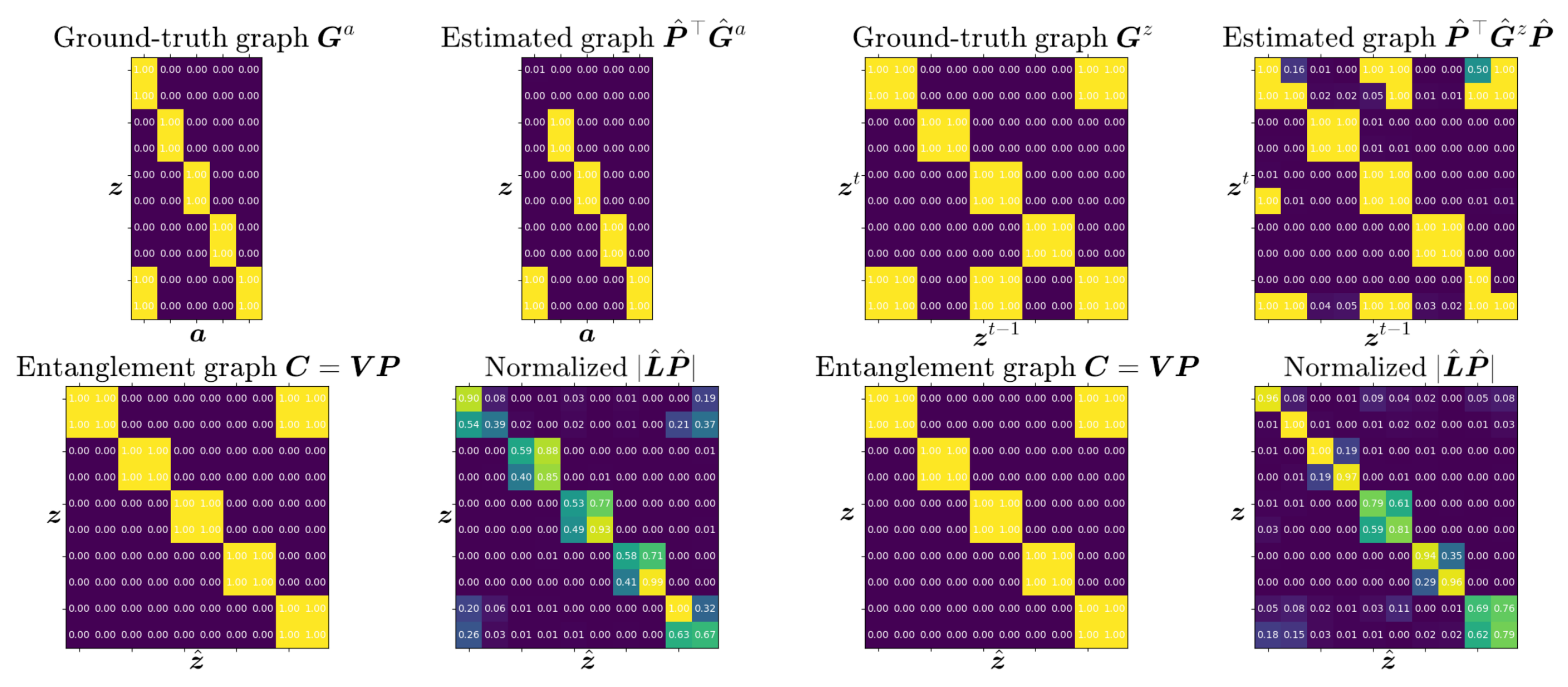
Theorem 3: Partial disentanglement via sparse $\boldsymbol{G}^z$
If "$H^{t,\tau}_{z,z} \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ spans $\mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z}$" and $||\hat{\boldsymbol{G}}^z||_0 \leq ||\boldsymbol{G}^z||_0$, then $\boldsymbol{V} = \boldsymbol{C}\boldsymbol{P}^\top$ where $\boldsymbol{C}$ is $\boldsymbol{G}^z$-preserving and $(\boldsymbol{G}^z)^\top$- preserving.



(a) Example 2 (b) Example 3 (c) Example 4 (d) Example 5 (e) Example 6 (f) Example 7

- Since invertible $\boldsymbol{G}$-preserving matrices form a group, the dependency graph of $\boldsymbol{z} = \boldsymbol{v}(\hat{\boldsymbol{z}})$ is the same as $\hat{\boldsymbol{z}} = \boldsymbol{v}^{-1}(\boldsymbol{z})$ (modulo permutation)
- Graphical criterion of [2] implies complete disentanglement!

Experiments



(a) ActionBlockNonDiag dataset, $\beta = 10$ (b) TimeBlockNonDiag dataset, $\beta = 30$