PhD Defense

Aleksei Sholokhov

Thursday 18th May, 2023

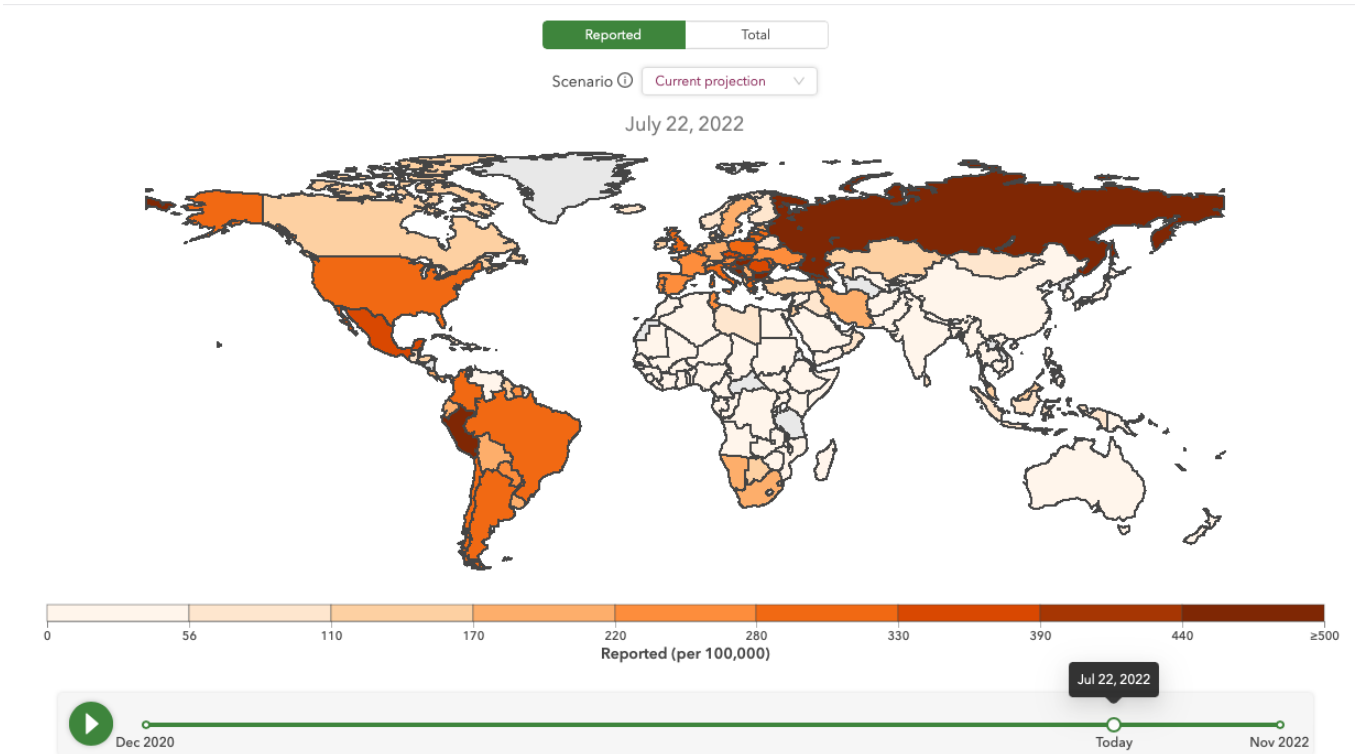
Plan of the Defense

Show topics and published papers. Mention covid

Feature Selection for Mixed-Effect Models

Mixed-effect models

- Used for analyzing **combined data** across a range of **groups**.
- Use covariates to separate the **population variability** from the **group variability**.
- **Borrow strength** across groups to estimate key statistics.



[1]

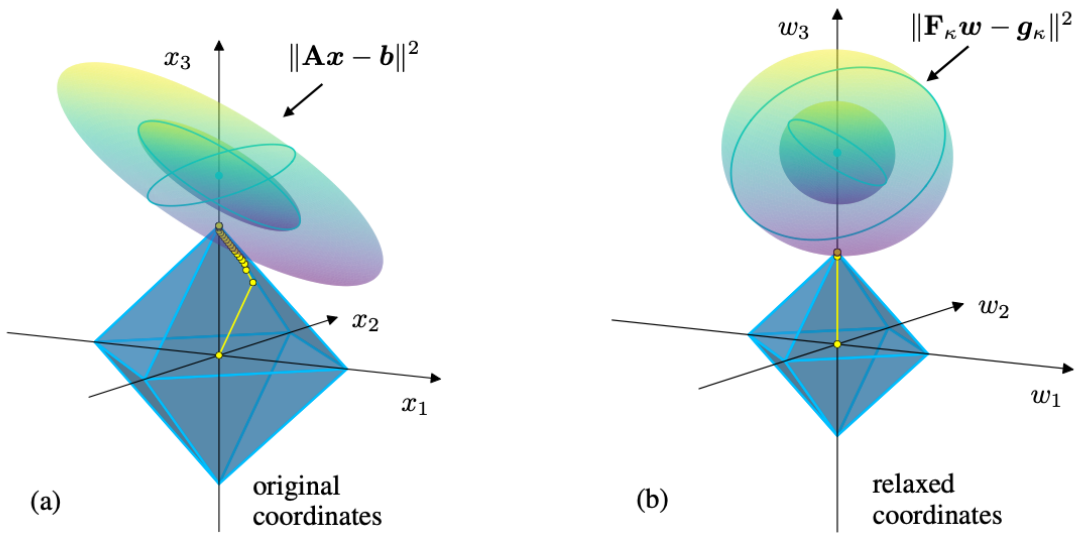
[1] Picture is taken from covid19.healthdata.org

Feature Selection for Mixed-Effect Models

Practitioners:

- Often seek **sparse models** that select **most informative** covariates.
- Want to be **flexible but efficient** in using various sparsity-promoting terms.
- Want a library to be **universal and compatible** with e.g. `sklearn`.

Sparse Relaxed Regularized Regression ($\mathcal{SR}3$) [9] showed great results for t linear models:



Goal: *create a feature selection library that uses a relaxation approach for feature-selection in mixed-effect models.*

Linear Mixed-Effect (LME) Models

Dataset: m groups (X_i, Z_i, y_i), $\quad i = 1, \ldots m$, each has n_i observations

- $X_i \in \mathbb{R}^{n_i \times p}$ – group i design matrix for fixed features
- $Z_i \in \mathbb{R}^{n_i \times q}$ – group i design matrix for random features
- $y_i \in \mathbb{R}^{n_i}$ – group i observations

Model:

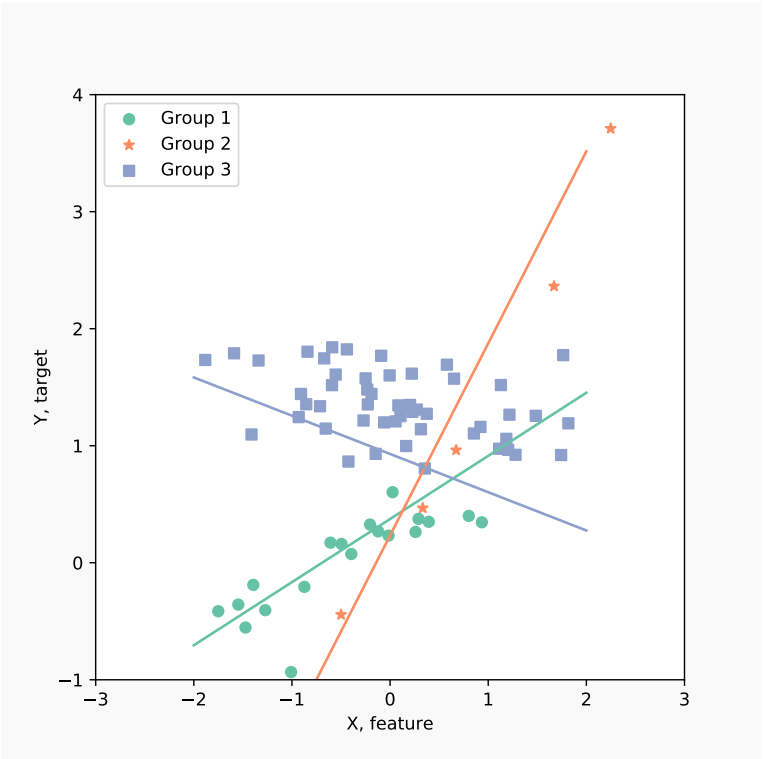
$$y_i = X_i\beta + Z_i u_i + \varepsilon_i$$
$$\varepsilon_i \sim \mathcal{N}(0, \Lambda_i)$$
$$u_i \sim \mathcal{N}(0, \Gamma)$$

Equivalently:

$$y_i = X_i\beta + \omega_i$$
$$\omega_i \sim \mathcal{N}(0, Z_i \Gamma Z_i^T + \Lambda_i)$$

Simplifying assumption:

$$\Gamma = \mathrm{diag}\,((\,)\,\gamma)$$



Notation

$$
\begin{aligned}
y_i &= X_i\beta + Z_i u_i + \varepsilon_i \quad i = 1 \dots m \\
\varepsilon_i &\sim \mathcal{N}(0, \Lambda_i) \\
u_i &\sim \mathcal{N}(0, \Gamma)
\end{aligned} \tag{1}
$$

- p – number of fixed features, q – number of random effects.
- $\beta \in \mathbb{R}^p$ – fixed effects, or mean effects
- $u_i \in \mathbb{R}^q$ – random effects
- $\Gamma \in \mathbb{R}^{q \times q}$ – covariance matrix of random effects, often $\Gamma = \operatorname{diag}((\gamma))$
- $\varepsilon_i \in \mathbb{R}^{n_i}$ – observation noise
- $\Lambda_i \in R^{n_i \times n_i}$ – covariance matrix for noise

Unknowns: β, u_i, γ, sometimes Λ_i.

Likelihood for Mixed Models

Optimization problem:

$$\mathcal{FS} - \mathcal{LME} \qquad \min_{\beta \in \mathbb{R}^p,\, \gamma \in \mathbb{R}^q_+} \mathcal{L}(\beta, \gamma) + R(\beta, \gamma) \tag{2}$$

Where $\mathcal{L}$:

$$\begin{aligned} \mathcal{L}(\beta, \gamma) = \sum_{i=1}^{m} \frac{1}{2} (y_i - X_i \beta)^T (Z_i \Gamma Z_i^T + \Lambda_i)^{-1} (y_i - X_i \beta) + \\ + \frac{1}{2} \log\det (Z_i \Gamma Z_i^T + \Lambda_i), \quad \Gamma = \operatorname{diag}((\gamma)) \end{aligned} \tag{3}$$

- $\mathcal{L}(\beta, \gamma)$ is smooth on its domain, quadratic w.r.t. β and $\bar{\eta}$-weakly-convex w.r.t. γ.
- $R(\beta, \gamma)$ is closed, proper, with easily computed *prox operator*

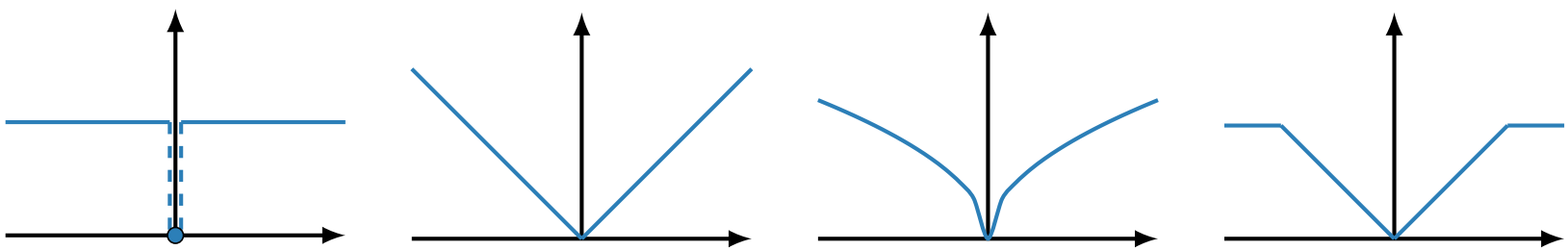
Regularization

- $R(\beta, \gamma)$ is closed, proper, with easily computed *prox operator*

$$\text{prox}_{\alpha R + \delta_{\mathcal{C}}}(\tilde{\beta}, \tilde{\gamma}) := \underset{(\beta,\gamma)\in\mathcal{C}}{\text{argmin}}\, R(\beta, \gamma) + \frac{1}{2\alpha}\|(\beta, \gamma) - (\tilde{\beta}, \tilde{\gamma})\|_2^2, \quad \text{where } \mathcal{C} := \mathbb{R}^p \times R_+^q \tag{4}$$

Examples:

- $R(x) = \lambda \sum_{j=1}^{p} w_j \|x_j\|_1$ – LASSO and Adaptive LASSO penalties [1, 6]
- $R(x) = \lambda \|x\|_0$ – ℓ_0 penalty [8, 5]
- $R(x)$ – SCAD penalty ([2, 3])



(a) ℓ_0 (b) ℓ_1 (c) ℓ_p, $p = 1/2$ (d) SCAD

Figure: Four commonly-used regularizers which promote sparsity

SR3-Relaxation for Mixed-Effect Models ($\mathcal{MSR}3$)

Original problem $\mathcal{FS}-\mathcal{LME}$:

$$\min_{\beta\in\mathbb{R}^p,\,\gamma\in\mathbb{R}^q_+} \mathcal{L}(\beta,\gamma)+R(\beta,\gamma) \tag{5}$$

Relaxed problem $\mathcal{MSR}3$:

$$\min_{\beta,\tilde{\beta}\in\mathbb{R}^p,\,\gamma,\tilde{\gamma}\in\mathbb{R}^q_+} \mathcal{L}(\beta,\gamma)+\phi_\mu(\gamma)+\kappa_\eta(\beta-\tilde{\beta},\gamma-\tilde{\gamma})+R(\tilde{\beta},\tilde{\gamma}) \tag{6}$$

where the *relaxation* κ_η decouples the likelihood and the regularizer

$$\kappa_\eta(\beta-\tilde{\beta},\gamma-\tilde{\gamma}) := \frac{\eta}{2}\|\beta-\tilde{\beta}\|_2^2+\frac{\eta}{2}\|\gamma-\tilde{\gamma}\|_2^2, \quad \eta>\bar{\eta} \tag{7}$$

and the *perspective mapping* ϕ_μ replaces $\gamma\geq 0$ with a log-barrier

$$\phi_\mu(\gamma) := \begin{cases} -\mu\sum_{i=1}^q \ln(\gamma_i/\mu), & \mu>0 \\ \delta_{\mathbb{R}^q_+}(\gamma), & \mu=0 \\ +\infty, & \mu<0 \end{cases} \tag{8}$$

Value Function Reformulation

$\mathcal{MSR}3$-relaxation replaces the original likelihood $\mathcal{L}$ with a *value function* $u_{\eta,\mu}$:

$$\begin{aligned} v_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) &:= \min_{(\beta,\gamma)} \mathcal{L}_{\eta,\mu}((\beta,\gamma),(\tilde{\beta},\tilde{\gamma})) \\ &:= \min_{(\beta,\gamma)} \mathcal{L}(\beta,\gamma) + \phi_\mu(\gamma) + \kappa_\eta(\beta-\tilde{\beta},\gamma-\tilde{\gamma}) \end{aligned} \tag{9}$$

so $\mathcal{MSR}3$-formulation (6) becomes

$$\min_{\beta\in\mathbb{R}^p,\,\gamma\in\mathbb{R}^q_+} v_{\eta,\mu}(\tilde{\beta},\tilde{\gamma}) + R(\tilde{\beta},\tilde{\gamma}) \tag{10}$$

When η is larger than the weak-convexity constant

- $v_{\eta,\mu}$ is well-defined and continuously differentiable.
- As $\mu \to 0$ and $\eta \to \infty$, cluster points of solutions to $\mathcal{MSR}3$ are first-order stationary points for $\mathcal{FS}-\mathcal{LME}$
- $v_{\eta,\mu}$ don't need to be evaluated precisely.

Value Function Reformulation

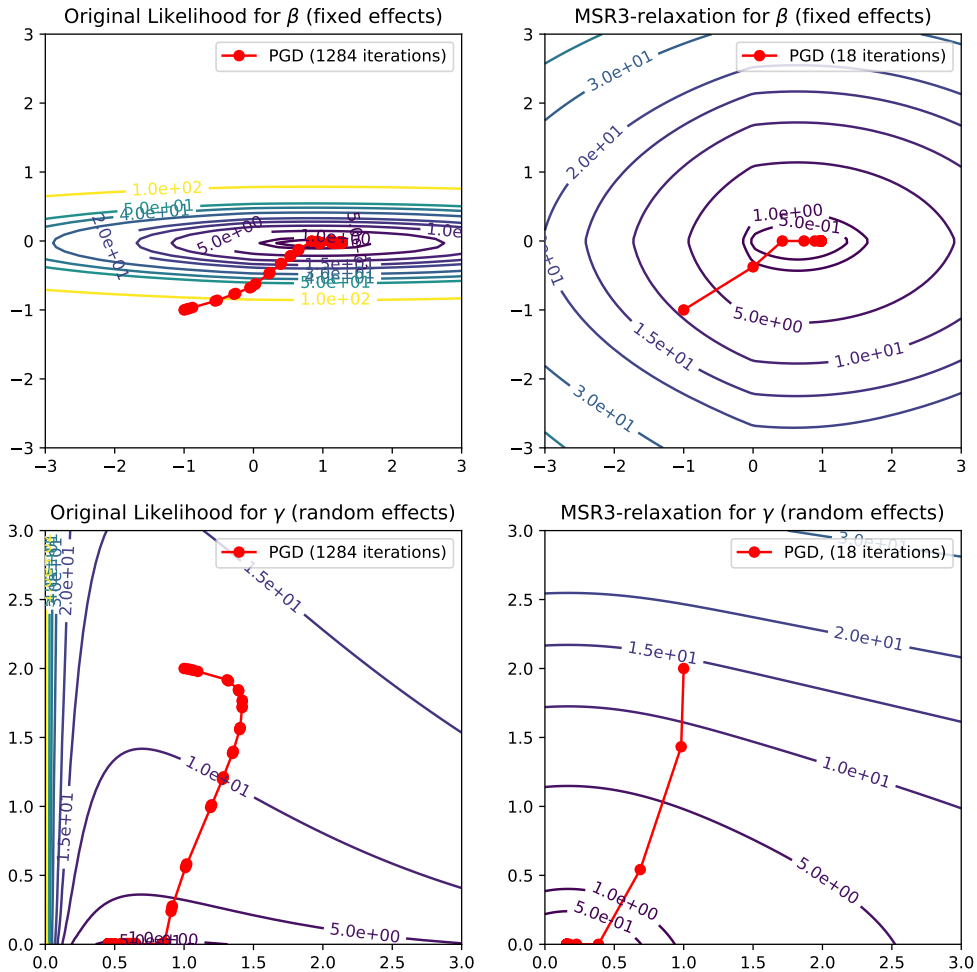


Figure: Comparison of the level-sets for the original likelihood (left) and $\mathcal{MSR}3$-likelihood (right), for fixed (top) and random (bottom) effects.

Designing an Algorithm

$G_{\nu,\eta}$ encodes both gradient of a Lagrangian (lines 1-2) and the complementarity condition (line 3):

$$G_{\nu,\eta}((\beta,\gamma,v),(\tilde{\beta},\tilde{\gamma})) := \begin{bmatrix} \nabla_\beta \mathcal{L}(\beta,\gamma) + \eta(\beta-\tilde{\beta}) \\ \nabla_\gamma \mathcal{L}(\beta,\gamma) + \eta(\gamma-\tilde{\gamma}) - v \\ v \bigodot \gamma - \mu \mathbf{1} \end{bmatrix} \tag{11}$$

We apply Newton method to G while geometrically decreasing μ.
Lemma: For every $(\mu,\eta) \in \mathbb{R}_+ \times \mathbb{R}_{++}$,

$$\begin{aligned} &(\hat{\beta},\hat{\gamma}) = \underset{(\beta,\gamma)}{\operatorname{argmin}}\, \mathcal{L}_{\eta,\mu}((\beta,\gamma),(\tilde{\beta},\tilde{\gamma})) \\ &\iff \\ &\exists \hat{v} \in \mathbb{R}_+^q \text{ s.t. } G_{\nu,\eta}((\beta,\gamma,\hat{v}),(\tilde{\beta},\tilde{\gamma})) = 0 \end{aligned} \tag{12}$$

If $\mu > 0$, then $\hat{v} = -\nabla\phi_\mu(\hat{\gamma})$, and if $\mu = 0$, then $\hat{v}$ is the unique KKT multiplier associated with the constraint $0 \leq \gamma$.

$\mathcal{MSR}3$-fast Algorithm

progress $\leftarrow$ **True**; iter = 0;

$\beta^+, \tilde{\beta}^+ \leftarrow \beta_0$; $\gamma^+, \tilde{\gamma}^+ \leftarrow \gamma_0$; $v^+ \leftarrow 1 \in \mathbb{R}^q$; $\mu \leftarrow \frac{{v^+}^T \gamma^+}{10q}$

while iter $<$ max_iter and $\|G_\mu(\beta^+, \gamma^+, v^+)\| >$ tol and progress **do**

  $\beta \leftarrow \beta^+$; $\gamma \leftarrow \gamma^+$; $\tilde{\beta} \leftarrow \tilde{\beta}^+$; $\tilde{\gamma} \leftarrow \tilde{\gamma}^+$

  $[dv, d\beta, d\gamma] \leftarrow \nabla G_\mu((\beta, \gamma, v), (\tilde{\beta}, \tilde{\gamma}))^{-1} G_\mu((\beta, \gamma, v), (\tilde{\beta}, \tilde{\gamma}))$

  $\alpha \leftarrow 0.99 \times \min\left(1, -\frac{\gamma_i}{d\gamma_i}, \forall i : \ d\gamma_i < 0\right)$

  $\beta^+ \leftarrow \beta + \alpha d\beta$; $\gamma^+ = \gamma + \alpha d\gamma$; $v^+ \leftarrow v + \alpha dv$

  if $\|\gamma^+ \odot v^+ - q^{-1}{\gamma^+}^T v^+ \mathbf{1}\| > 0.5 q^{-1} {v^+}^T \gamma^+$ **then** continue;

  else

    $\tilde{\beta}^+ = \text{prox}_{\alpha R}(\beta^+)$; $\tilde{\gamma}^+ = \text{prox}_{\alpha R + \delta_{\mathbb{R}_+}}(\gamma^+)$; $\mu = \frac{1}{10} \frac{{v^+}^T \gamma^+}{q}$

  end

  progress $= (\|\beta^+ - \beta\| \geq \text{tol}$ or $\|\gamma^+ - \gamma\| \geq \text{tol}$ or $\|\tilde{\beta}^+ - \tilde{\beta}\| \geq \text{tol}$ or $\|\tilde{\gamma}^+ - \tilde{\gamma}\| \geq \text{tol})$

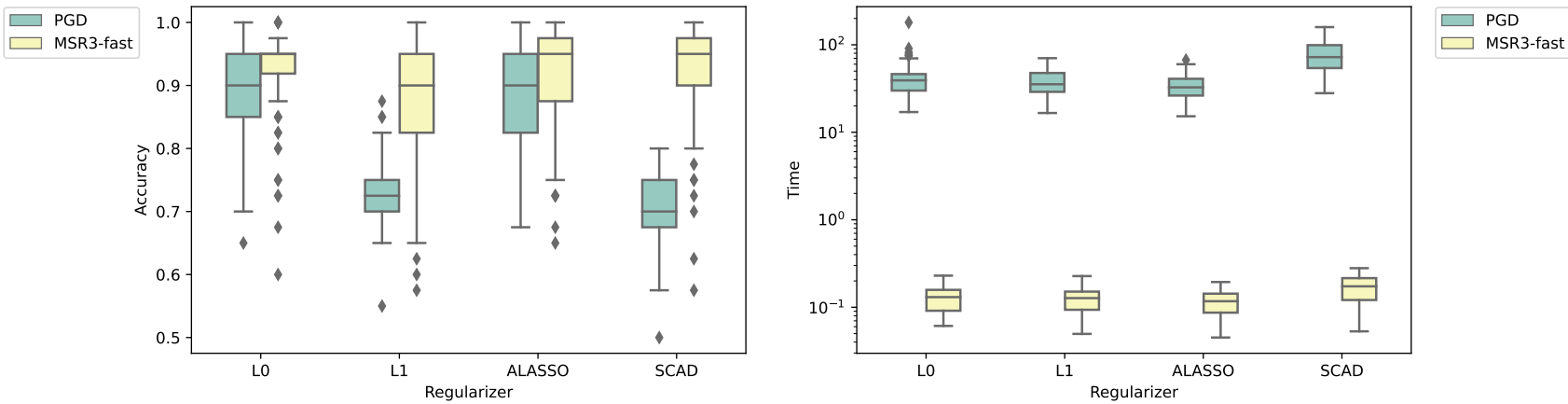
  iter += 1

end

return $\tilde{\beta}^+, \tilde{\gamma}^+$

Application to Synthetic Problems

- The number of fixed effects p and random effects q is 20.
- $\beta = \gamma = \frac{1}{2}[1, 2, 3, \ldots, 10, 0 \ldots, 0]$
- 9 groups with sizes [10, 15, 4, 8, 3, 5, 18, 9, 6]
- $X_i \sim \mathcal{N}(0, I)^p$, $Z_i = X_i$, $\varepsilon_i \sim \mathcal{N}(0, 0.3^2 I)$
- Each experiment is repeated 100 times.
- Grid-search for $\eta \in [10^{-4}, 10^2]$, golden search for $\lambda \in [0, 10^5]$
- Final model is chosen to maximize BIC



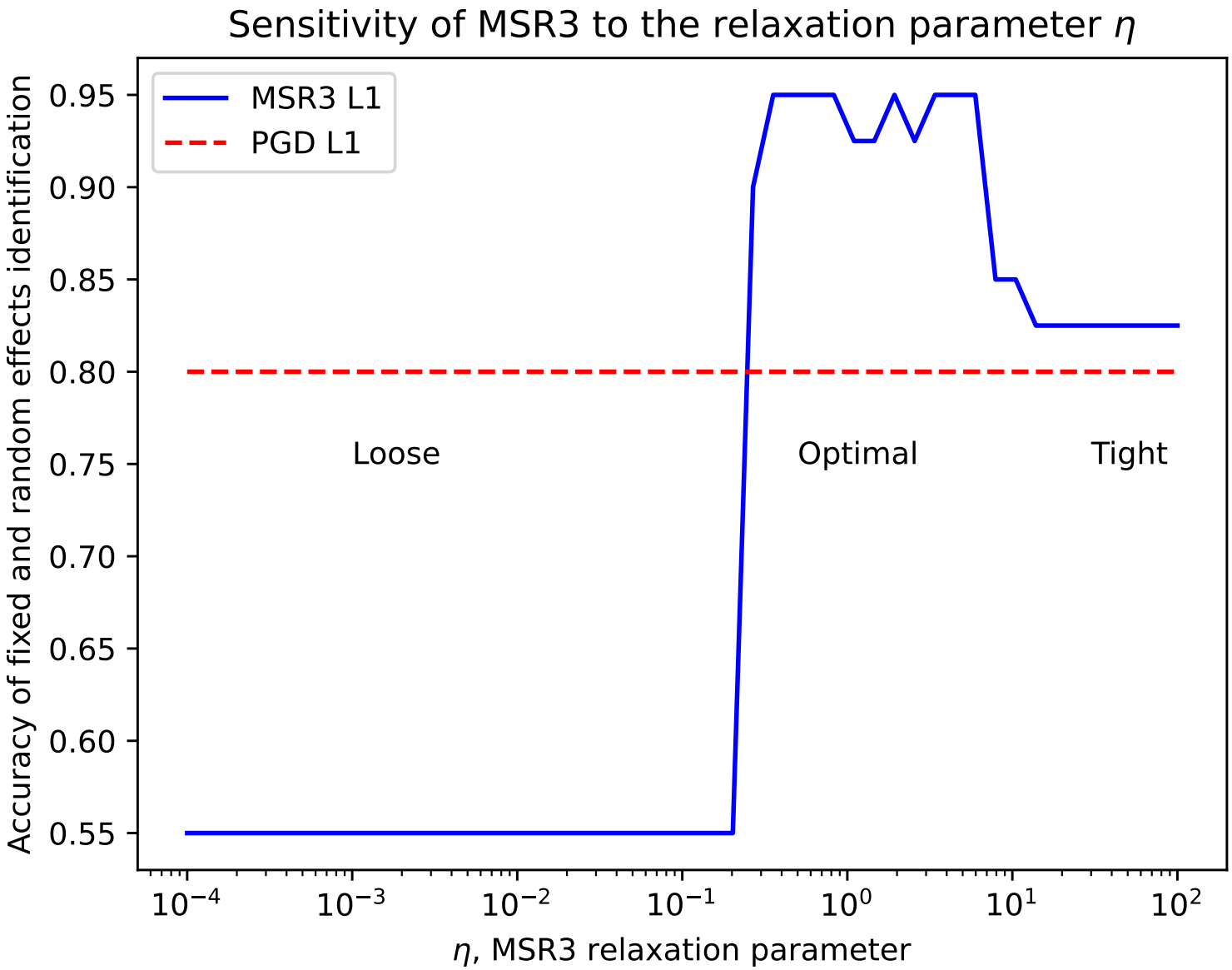
+ $\mathcal{MSR}3$-relaxation improves feature selection performance of the original likelihood.
+ $\mathcal{MSR}3$-fast optimization accelerates the compute time by $\sim 10^2$.
– Initialization of η is problem-specific

Comparison to Other Libraries

Algorithm	MSR3-Fast (ℓ_1)	`glmmLasso`[2] [4]	`lmmLasso`[3][7]	PGD (ℓ_1)
Accuracy, %	**88**	48	66	73
FE Accuracy, %	**86**	52	47	56
RE Accuracy, %	**91**	45	84	**91**
Time, sec	**0.19**	1.37	11.51	38.39
Iterations, num	34	50	-	7693

[2] https://rdrr.io/cran/glmmLasso/man/glmmLasso.html
[3] https://rdrr.io/cran/lmmlasso/

Choice of η



ℓ_0-based Covariate Selection for Bullying Study from GBD

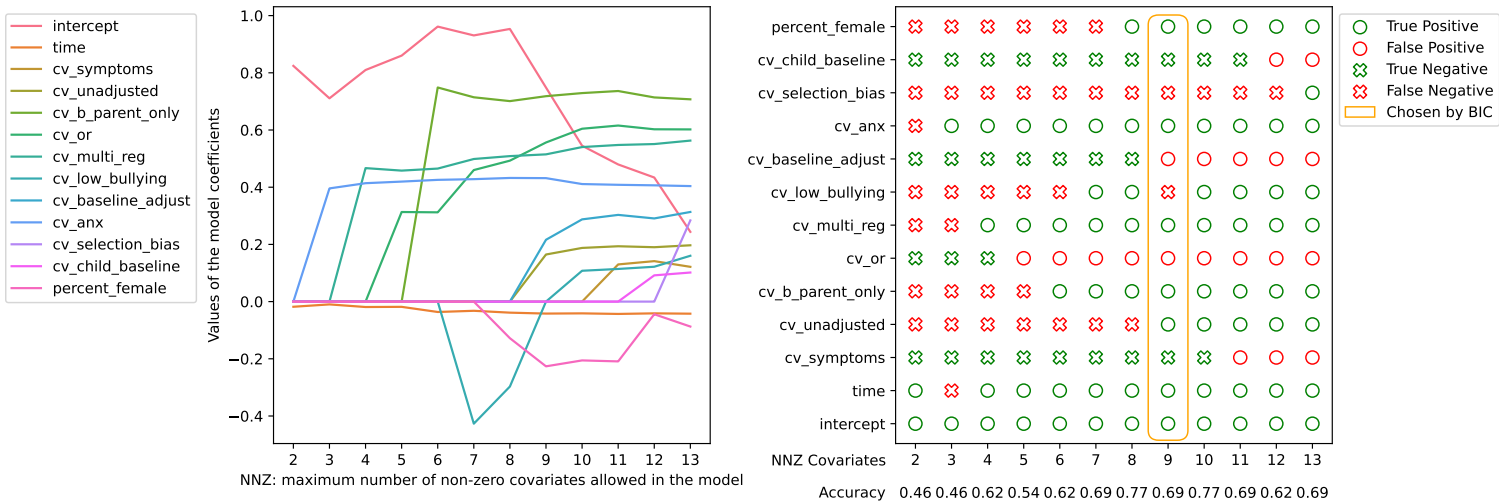


Figure: Fixed and random covariate selection for Bullying dataset[4]. The model selected 9 covariates, 7 of which were historically significant, and did not select 4 covariates, 1 of which was historically significant.

[4] Institute for Health Metrics and Evaluation (IHME). Bullying Victimization Relative Risk Bundle GBD 2020. Seattle, United States of America (USA), 2021.

Software

The code is available on GitHub: https://github.com/aksholokhov/pysr3

- All estimators are fully compatible to `sklearn` library.
- Implements SR3 for linear, generalized-linear, and linear mixed-effect models.
- Has tutorials, tests, and documentation.

Data-Driven Modeling of Physical Systems

1) People used to model physical systems with first-principle knowledge 2) Data-Driven modelling of dynamical systems became a big thing 3) However, it requires a lot of data 4) Incorporating prior knowledge is a big recent trend, so history does a spiral

Incorporating Knowledge into Models

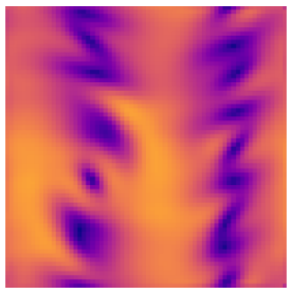
1) There are multiple ways of incorporating knowledge into system 4) The overall umbrella term for it is physics-informed machine learning 2) Some use the equations that model phenomena 3) Some take aspects of it, e.g. symmetries and preservation laws, and forces A network to respect those 5) Our work falls into the first category of approaches

Reduced-Order Models (ROMs)

$$x \in \mathbb{R}^n$$

$$\frac{dx}{dt} = f(x)$$

$$x_0$$



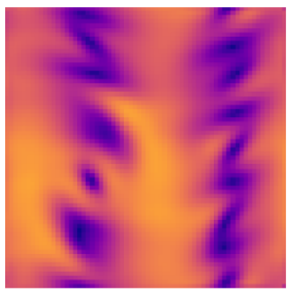
Reduced-Order Models (ROMs)

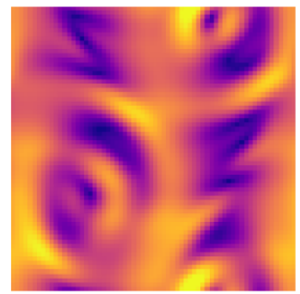
$x \in \mathbb{R}^n$

$$\frac{dx}{dt} = f(x)$$

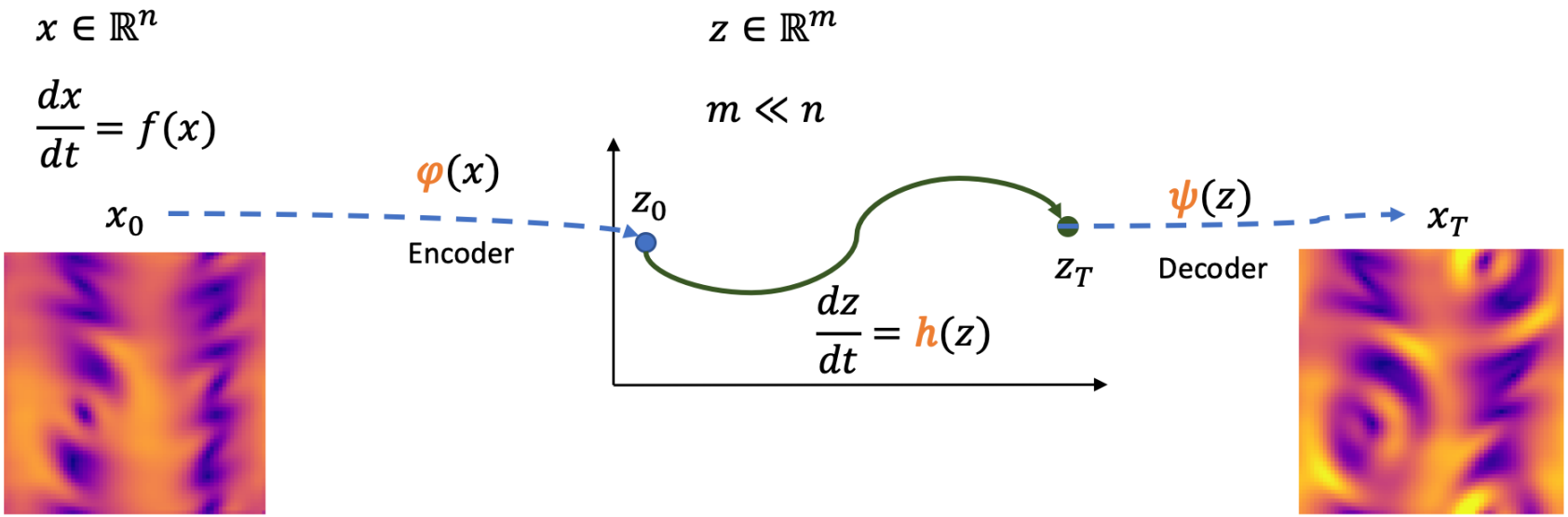
$$x_T = x_0 + \int_0^T f(x)dt$$

x_0 → x_T

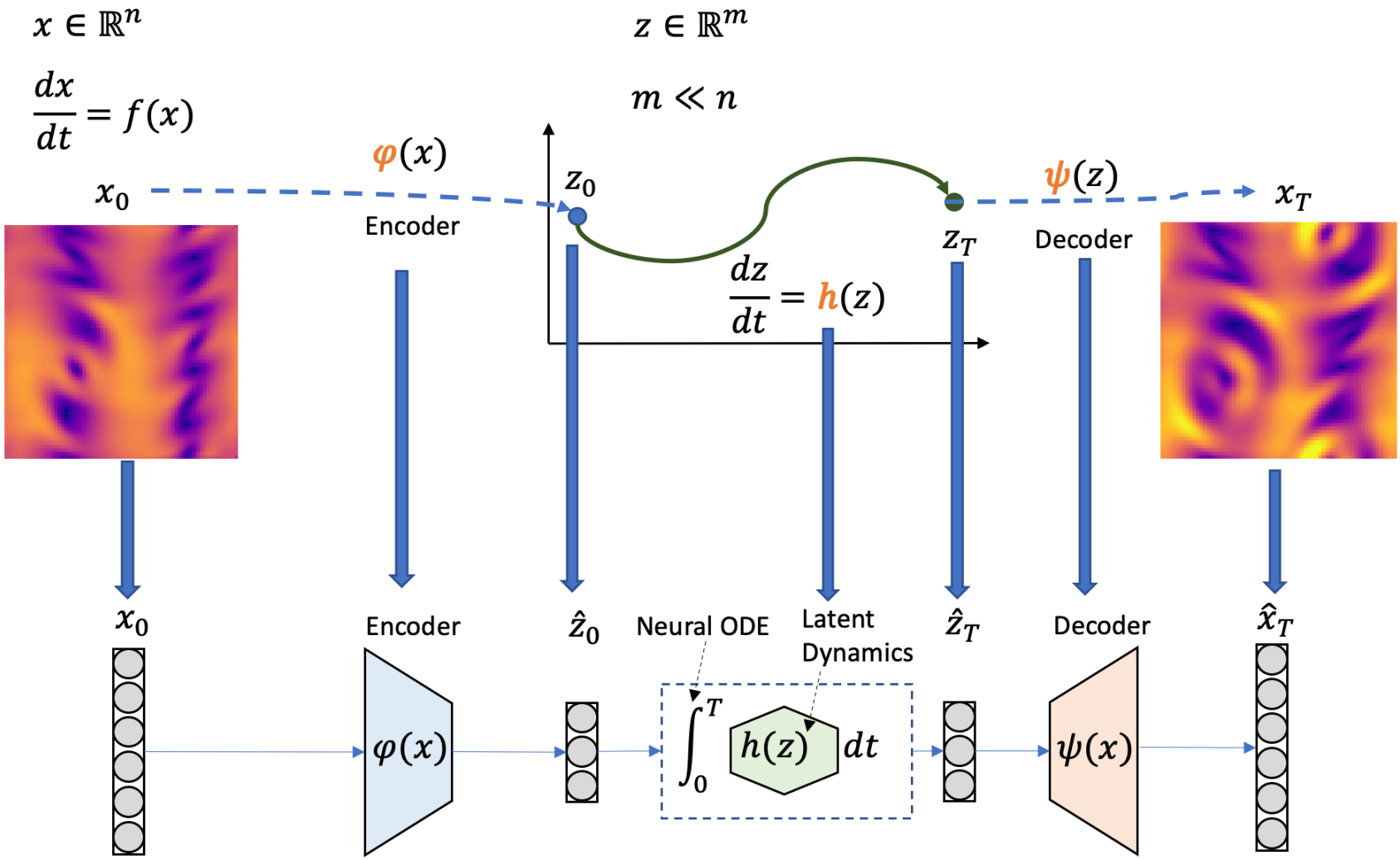




Reduced-Order Models (ROMs)



Reduced-Order Models (ROMs)



Physics-Informed Loss

$$\frac{dz}{dt} = \frac{dz}{dx}\frac{dx}{dt} = \nabla\varphi(x)^T f(x)$$

Physics-Informed Loss

$$\frac{dz}{dt} = \frac{dz}{dx}\frac{dx}{dt} = \nabla\varphi(x)^T f(x) \qquad \frac{dz}{dt} = h(\varphi(x))$$

Physics-Informed Loss

$$\frac{dz}{dt} = \frac{dz}{dx}\frac{dx}{dt} = \nabla\varphi(x)^T f(x) \qquad \frac{dz}{dt} = h(\varphi(x))$$

$$\mathcal{L}^{physics}(\tilde{x}) = \|\nabla\varphi(\tilde{x})^T f(\tilde{x}) - h(\varphi(\tilde{x}))\|_2^2 + \|\tilde{x} - \psi(\varphi(\tilde{x}))\|_2^2$$

Physics-Informed Loss = Latent Gradient Loss + Collocation Reconstruction Loss

Physics-Informed Loss

$$\frac{dz}{dt} = \frac{dz}{dx}\frac{dx}{dt} = \nabla\varphi(x)^T f(x) \qquad \frac{dz}{dt} = h(\varphi(x))$$

1 2 3

$$\mathcal{L}^{physics}(\tilde{x}) = \|\nabla\varphi(\tilde{x})^T f(\tilde{x}) - h(\varphi(\tilde{x}))\|_2^2 + \|\tilde{x} - \psi(\varphi(\tilde{x}))\|_2^2$$

Physics-Informed Loss = Latent Gradient Loss + Collocation Reconstruction Loss

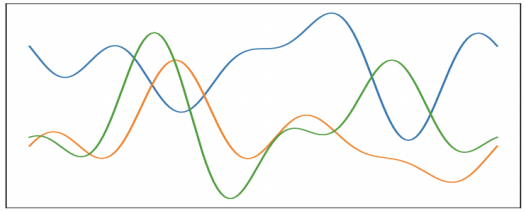
1

$$-u_t = u_{xx} + u_{xxxx} + \frac{1}{2}u_x^2 \Rightarrow \dot{x} = f(x)$$

$$u(x) = \frac{a}{1 + e^{-k(x-x_0)}} - \frac{a}{1 + e^{-k(x-x_1)}}, \quad x_0 < x_1$$

$$u(x) = \sum_{w=1}^{30} a(w)\sin(2\pi x) + b(w)\cos(2\pi x)$$

$$u(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-x_0)^2}{2\sigma^2}}$$



Physics-Informed Loss

$$\frac{dz}{dt} = \frac{dz}{dx}\frac{dx}{dt} = \nabla\varphi(x)^T f(x) \qquad \frac{dz}{dt} = h(\varphi(x))$$

$$\mathcal{L}^{physics}(\tilde{x}) = \|\nabla\varphi(\tilde{x})^T f(\tilde{x}) - h(\varphi(\tilde{x}))\|_2^2 + \|\tilde{x} - \psi(\varphi(\tilde{x}))\|_2^2$$

Physics-Informed Loss = Latent Gradient Loss + Collocation Reconstruction Loss

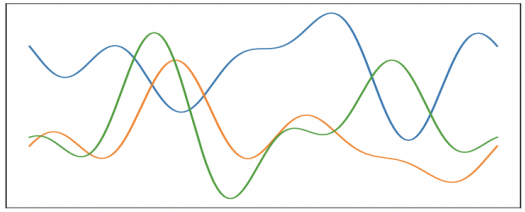
1

$$-u_t = u_{xx} + u_{xxxx} + \frac{1}{2}u_x^2 \Rightarrow \dot{x} = f(x)$$

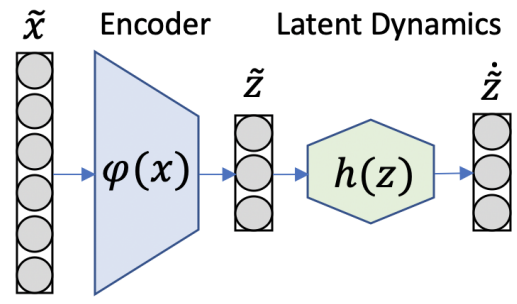
$$u(x) = \frac{a}{1 + e^{-k(x-x_0)}} - \frac{a}{1 + e^{-k(x-x_1)}}, \quad x_0 < x_1$$

$$u(x) = \sum_{w=1}^{30} a(w)\sin(2\pi x) + b(w)\cos(2\pi x)$$

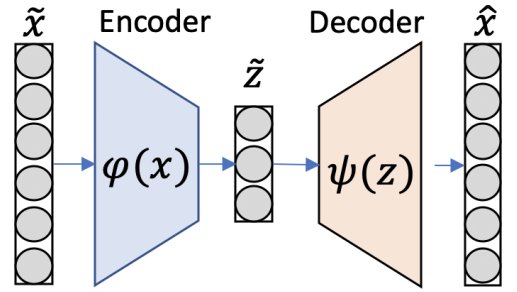
$$u(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-x_0)^2}{2\sigma^2}}$$



2



3



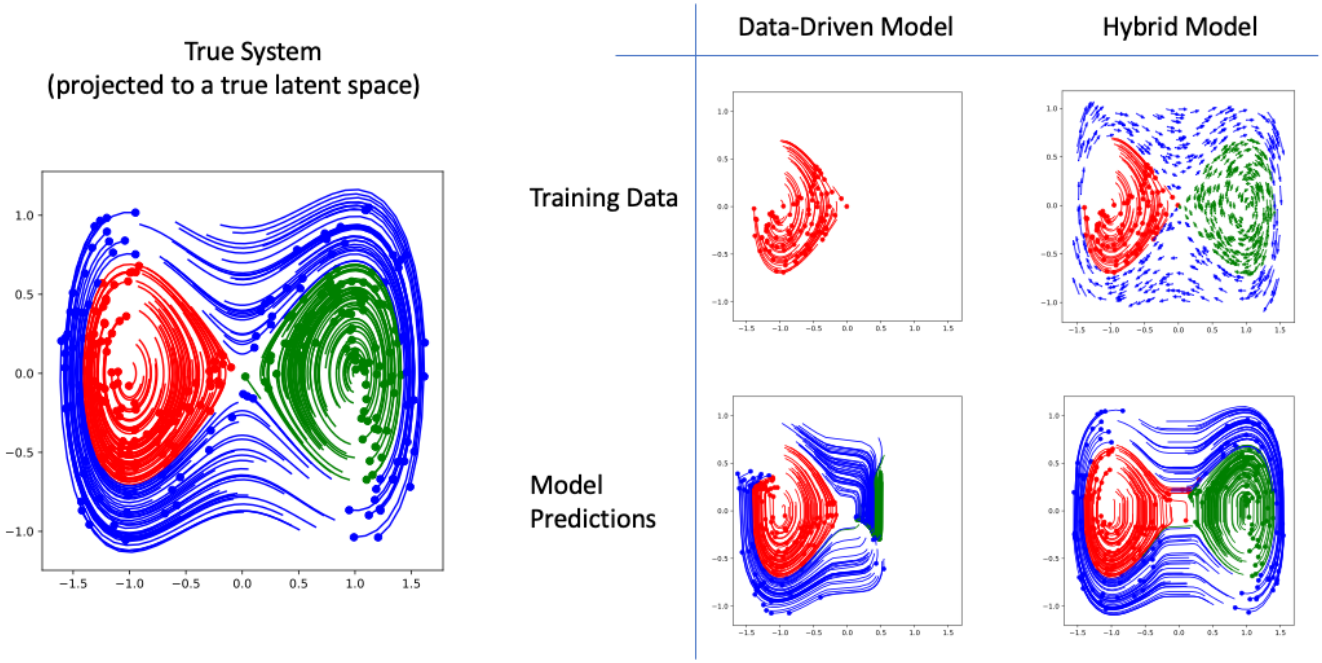
Results: Extrapolation to Unknown Regions

Duffing Oscillator on a low-dimensional (2D) manifold:

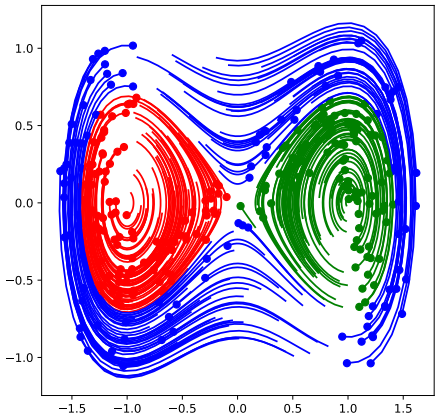
$$\begin{aligned} \frac{dz_1}{dt} &= z_2 \\ \frac{dz_2}{dt} &= z_1 - z_1^3 \end{aligned} \tag{13}$$

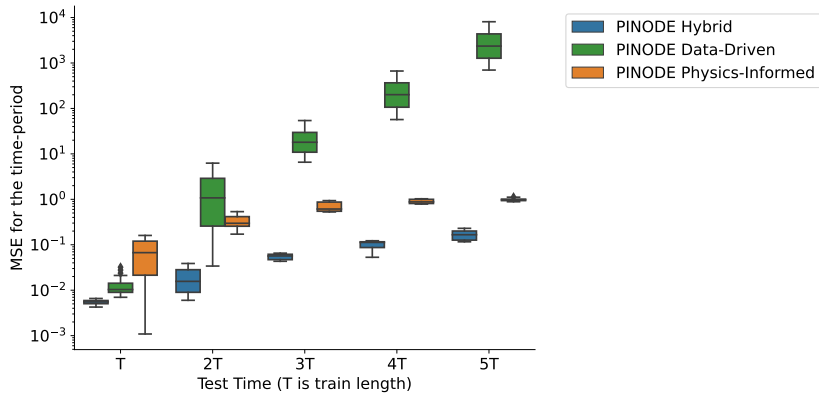
Projection to a high-dimensional (128) space:

$$\boldsymbol{x} := \mathcal{A}(\boldsymbol{z}) = A\boldsymbol{z}^3, \quad A \in \mathbb{R}^{128\times 2}, \quad A_{ij} \sim_{i.i.d.} \mathcal{N}(0,1) \tag{14}$$

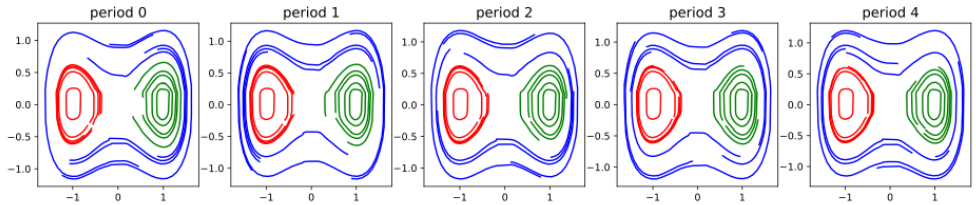


Results: Stable Long-Term Predictions

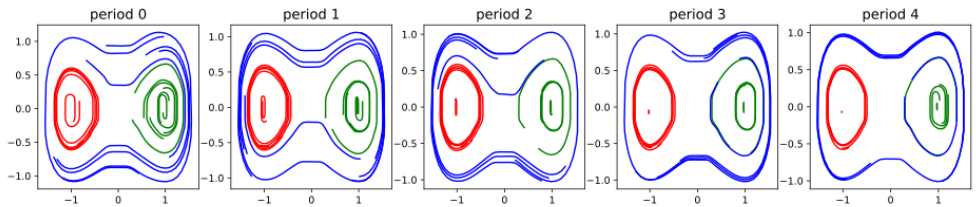




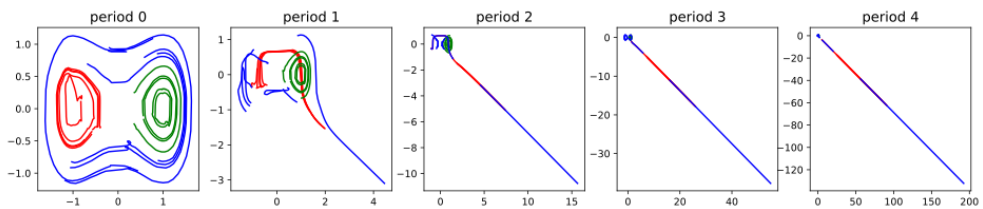
Hybrid Model



Physics-Informed Model



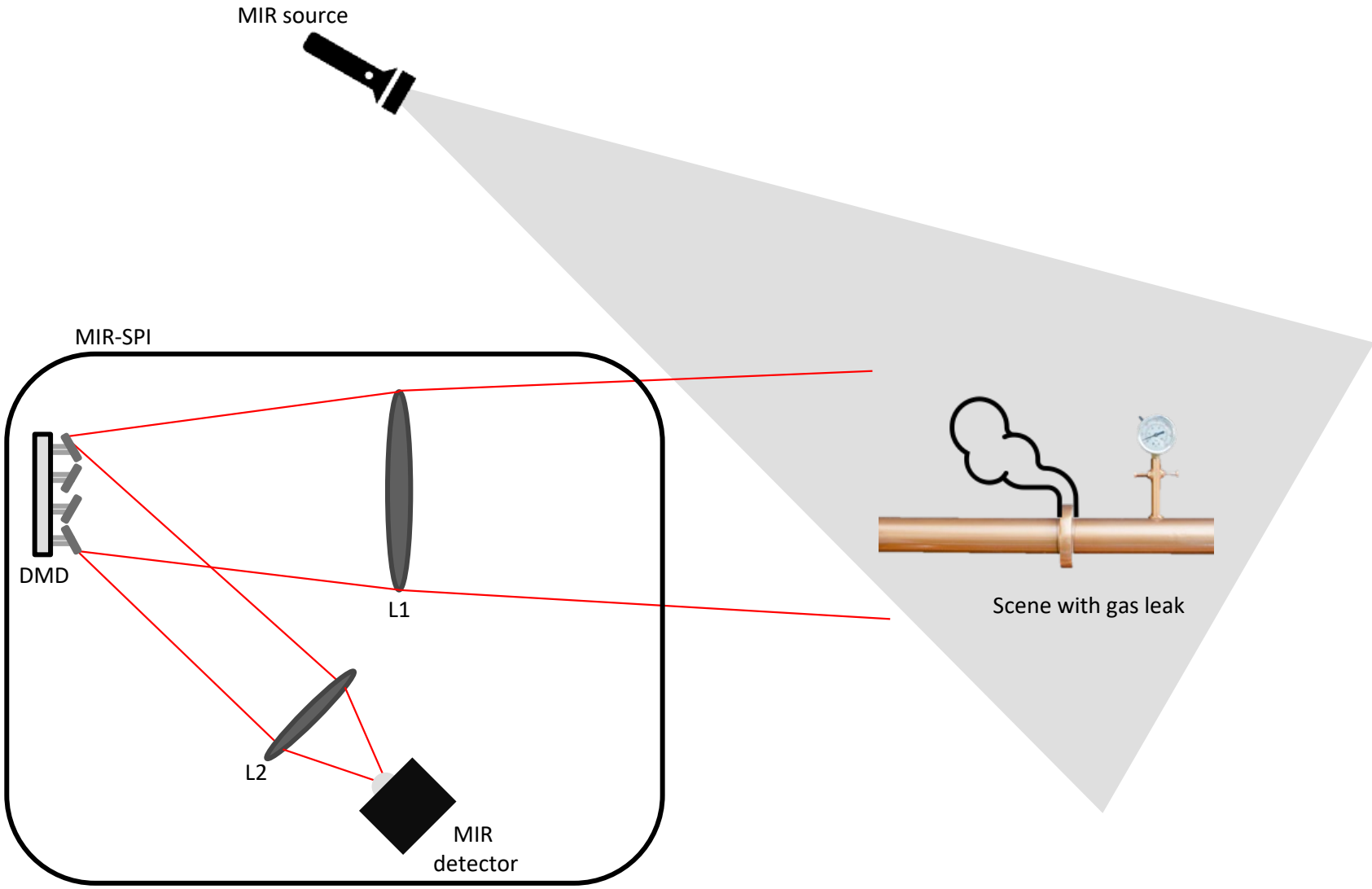
Data-Driven Model



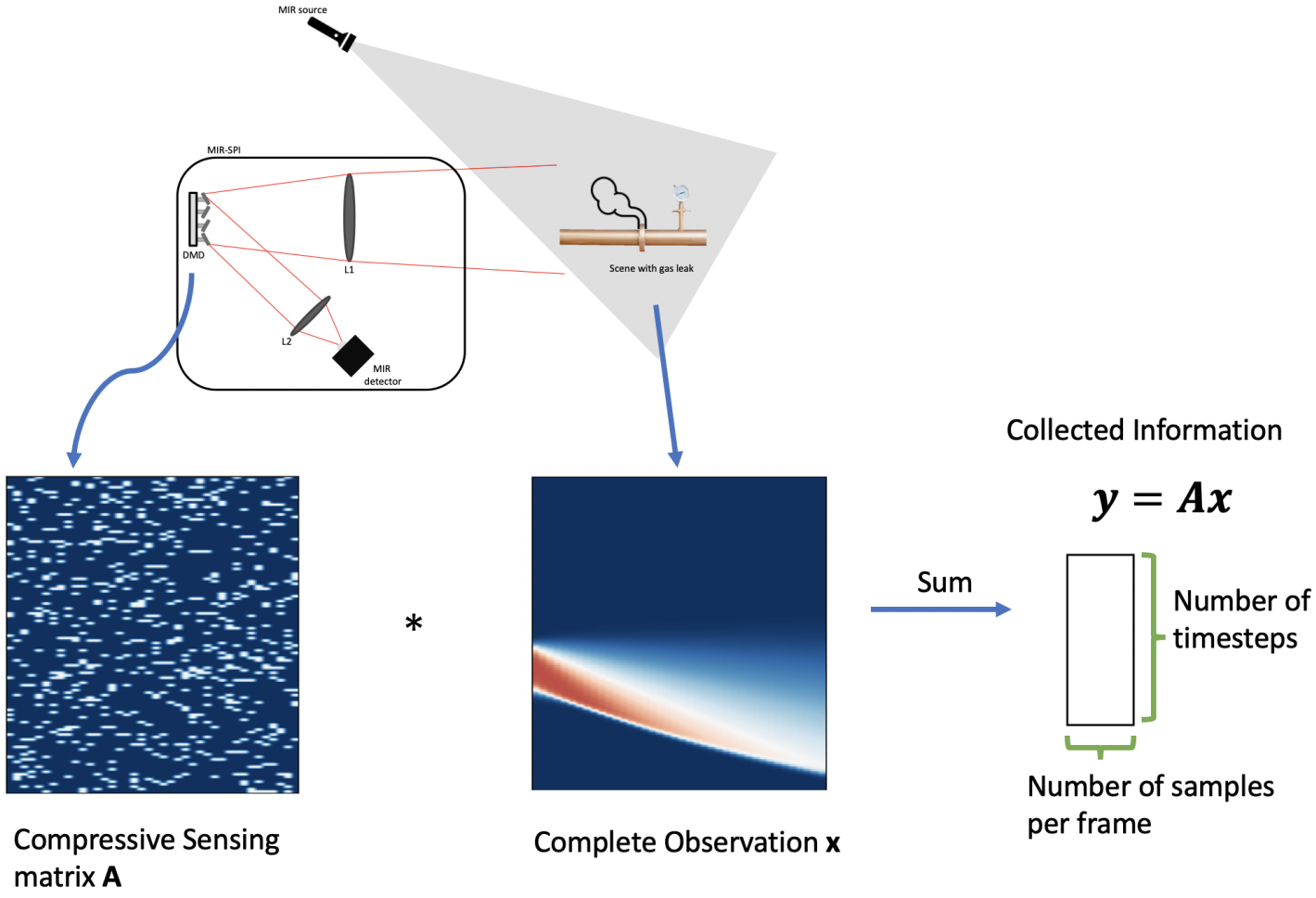
Results: Learning From Collocations

1) Finally we show that collocations can be even more useful than the data itself. 2) The difference is especially prominent in low-data regime. 3) It shows that collocations are powerful source of information and that the network can indeed interpolate between them.

Single-Pixel Imaging

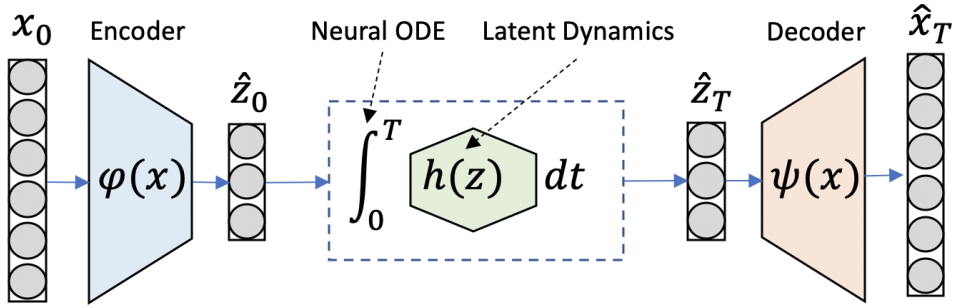


Single-Pixel Imaging



Compressive Sensing with Reduced-Order Models

Offline Step: Train a Data-Driven Reduced-Order Model



Online Step: Reconstruct Complete Observations by Optimizing in Latent Space

Reconstruction Loss

Compressive Sensing Loss

Loss for Prediction in Latent space

$$\mathcal{L}^{recon.}(z) = ||y - A\psi(z)|| + \lambda \left|\left| z - (z_0 + \int_0^T h(z)dz) \right|\right|$$

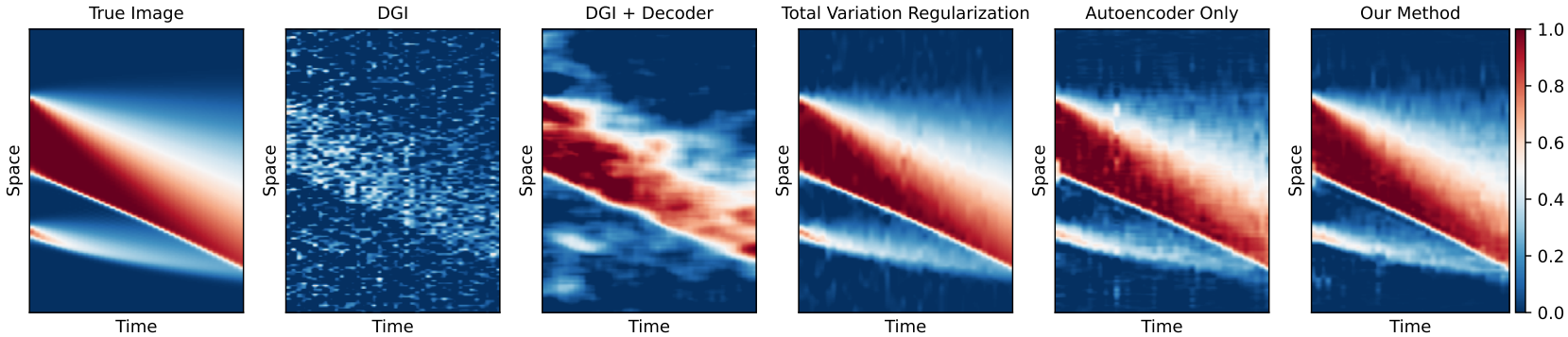
Latent-space representation of the trajectory

"What the data tells us the trajectory should be"

"What the model thinks the trajectory should be"

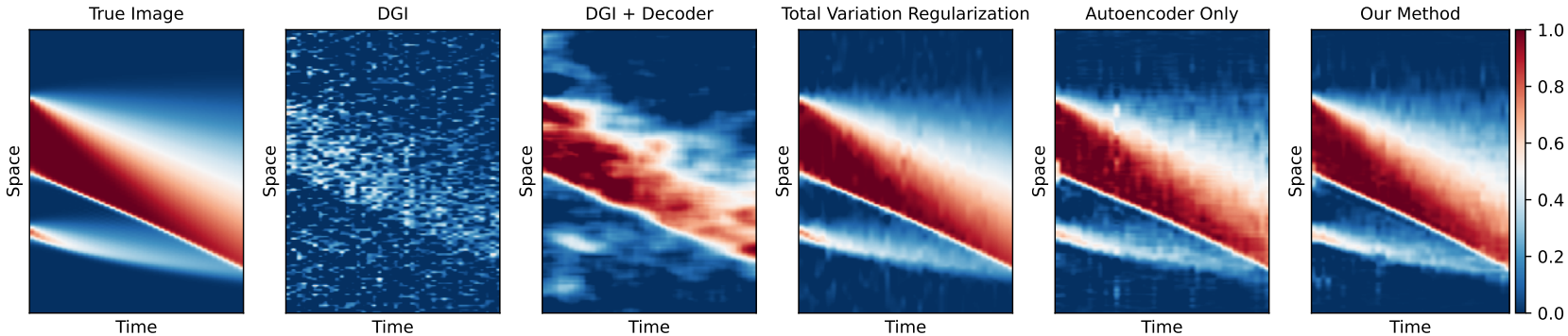
Results: Burger's Equation

When we capture 32 samples per frame:

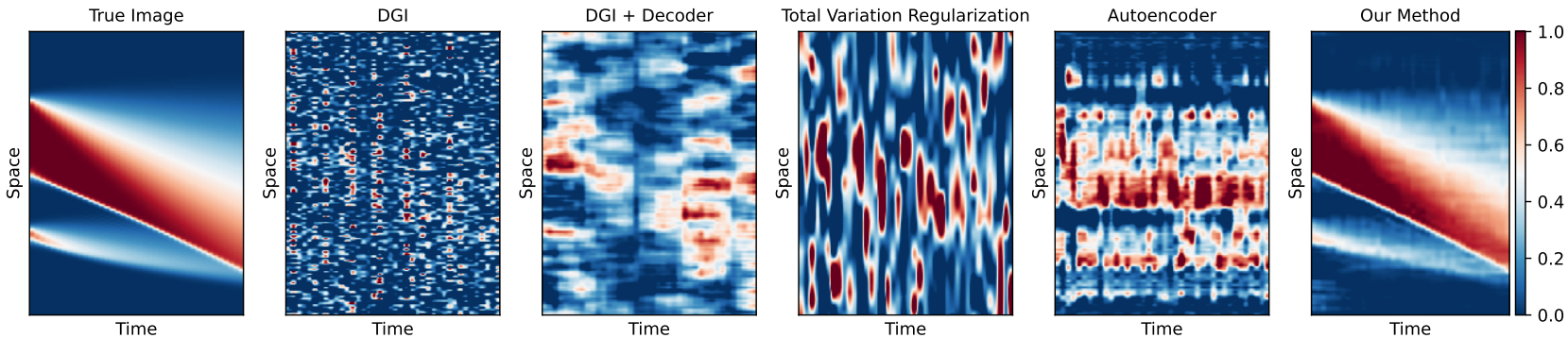


Results: Burger's Equation

When we capture 32 samples per frame:



When we capture 2 samples per frame:



Results: Burger's Equation

Aggregated results

Results: Interpretation

Results: Kolmogorov Flow OR Real Example

Conclusion

Results on Burgers Maybe results on a harder problem

References

References:

[1] Howard D. Bondell, Arun Krishna, and Sujit K. Ghosh. Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. Biometrics, 66(4):1069–1077, dec 2010.

[2] Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association, 96(456):1348–1360, dec 2001.

[3] Yingying Fan and Runze Li. Variable selection in linear mixed effects models. The Annals of Statistics, 40(4):2043–2068, aug 2012.

[4] Andreas Groll and Gerhard Tutz. Variable selection for generalized linear mixed models by l 1-penalized estimation. Statistics and Computing, 24(2):137–154, 2014.

[5] Richard H. Jones. Bayesian information criterion for longitudinal and clustered data. Statistics in Medicine, 30(25):3050–3056, nov 2011.

[6] Bingqing Lin, Zhen Pang, and Jiming Jiang. Fixed and random effects selection by REML and pathwise coordinate optimization. Journal of Computational and Graphical Statistics, 22(2):341–355, 2013.

[7] Jürg Schelldorfer, Peter Bühlmann, and SARA VAN DE GEER. Estimation for high-dimensional linear mixed-effects models using l1-penalization. Scandinavian Journal of Statistics, 38(2):197–214, 2011.

[8] Florin Vaida and Suzette Blanchard. Conditional Akaike information for mixed-effects models. Biometrika, 92(2):351–370, jun 2005.

[9] Peng Zheng and Aleksandr Aravkin. Relax-and-split method for nonconvex inverse problems. Inverse Problems, 36(9), 2020.