Rodent: Relevance determination in ODE

Niklas Heim, Václav Šmídl, Tomáš Pevný*
Artificial Intelligence Center
Czech Technical University
Prague, CZ 120 00
{niklas.heim, vasek.smidl, tomas.pevny}@aic.fel.cvut.cz

Abstract

From a set of observed trajectories of a partially observed system, we aim to learn its underlying (physical) process without having to make too many assumptions about the generating model. We start with a very general, over-parameterized *ordinary differential equation* (ODE) of order N and learn the minimal complexity of the model, by which we mean both the order of the ODE as well as the minimum number of non-zero parameters that are needed to solve the problem. The minimal complexity is found by combining the *Variational Auto-Encoder* (VAE) with *Automatic Relevance Determination* (ARD) to the problem of learning the parameters of an ODE which we call *Rodent*. We show that it is possible to learn not only one specific model for a single process, but a manifold of models representing harmonic signals in general.

1 Problem definition

We are concerned with models of time series $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_K]$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ that are generated by discrete-time noisy observations of a continuous-time process

$$\boldsymbol{x}_k = H(\boldsymbol{\xi}(\Delta t k)) + e_k, \tag{1}$$

where $k = 1 \dots K$, $e_k \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I})$. The partial observation operator H has fixed sampling intervals Δt, and the temporal evolution of the state variable $\boldsymbol{\xi}(t) \in \mathbb{R}^N$ is governed by a dynamical system described by an ODE:

$$\frac{\partial \boldsymbol{\xi}}{\partial t} = f(\boldsymbol{\theta}, t) \approx \boldsymbol{W}\boldsymbol{\xi} + \boldsymbol{b}. \tag{2}$$

For simplicity, we use linear ODEs with unknown parameters $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}\}$, however, the method is applicable to any differentiable dynamical system. The order of the ODE is given by $\dim(\boldsymbol{\xi}) = N$. The solution of the ODE for given parameters $\boldsymbol{\theta}$ and initial conditions $\boldsymbol{\xi}(0)$, is $\boldsymbol{\xi}(t) = \psi(\boldsymbol{\theta}, \boldsymbol{\xi}(0), t)$.

We aim to learn the structure of the ODE model from a training set of L trajectories $\{\boldsymbol{X}_i\}_{i=1}^L$ generated by the same generative process but with different parameters and different initial conditions, for each trajectory, i.e.

$$\boldsymbol{X}_i = H(\psi(\boldsymbol{\theta}_i, \boldsymbol{\xi}_i(0), \boldsymbol{t})) + \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{t} = [0, \Delta t, \dots, K\Delta t]$. Assuming we observe a system with expected order M and unknown structure, we choose $N \geq M$. Eq. 3 thus defines a generative model of sequences $\boldsymbol{X}_i$ from the latent space of parameters $\boldsymbol{\theta}$ and initial conditions $\boldsymbol{\xi}(0)$. Following the variational autoencoder approach [1], we define the latent space $\boldsymbol{z} = [\boldsymbol{\theta}, \boldsymbol{\xi}(0)]$ with the ODE (2) playing the role of the decoder. We seek an encoder in the form of a distribution $q(\boldsymbol{z}|\boldsymbol{x})$, parametrized by a deep neural network. Moreover, we use a constrictive prior $p(\boldsymbol{z})$ to promote sparsity on $\boldsymbol{z}$ to obtain a simple model, which is important for the explainability of the learned model. We will use the normal-gamma prior proposed in [2] and recently used e.g. in Bayesian compression of Neural Networks [3, 4].

*Tomáš Pevný is also with Avast Software s.r.o

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

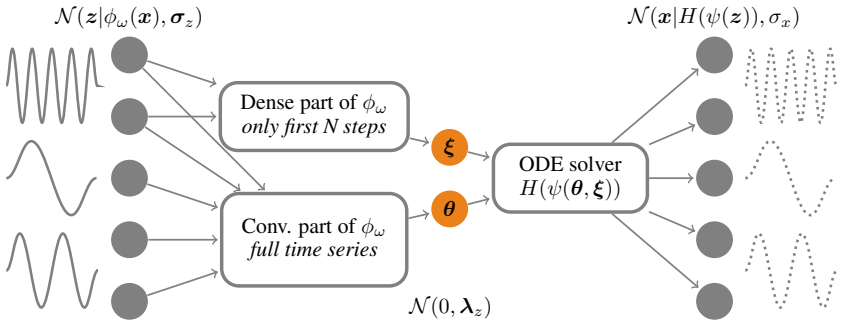


Figure 1: Schematic of the Rodent. Inputs and outputs in grey, latent dimension $\boldsymbol{z} = [\boldsymbol{\theta}, \boldsymbol{\xi}(0)]$ in the middle in brown. The encoder network ϕ_ω on the left represents the mean of the posterior. It consists of a dense part to estimate the initial conditions and a convolutional part that is responsible for the ODE parameters. By using convolutions for the parameter estimation we enable the network to process time series of variable lengths. The decoder on the right is a combination of an ODE solver and the observation operator H. The latent variable has the ARD prior $\mathcal{N}(0, \boldsymbol{\lambda}_z)$.

2 Rodent – Relevant ODE identifier

By combining ODE state and parameters in a latent variable $\boldsymbol{z} = [\boldsymbol{\theta}, \boldsymbol{\xi}(0)]$ we can define the data likelihood as

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|H(\psi(\boldsymbol{z})), \sigma_x^2). \tag{4}$$

With the decoder being an ODE solver (not a neural network), the resulting structure of the latent space allows for an interpretation e.g. in terms of physical properties of the model.

To determine the structure of the ODE, we employ the *Automatic Relevance Determination* (ARD, developed by [5, 2]) prior on the latent layer:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|0, \mathrm{diag}(\boldsymbol{\lambda}_z^2)) \qquad p(\boldsymbol{\lambda}_z) = 1/\boldsymbol{\lambda}_z. \tag{5}$$

where a new vector variable $\boldsymbol{\lambda}_z > 0$ of the same dimenson as $\boldsymbol{z}$ has been introduced. We will treat $\boldsymbol{\lambda}_z$ as an unknown variable and we will seek its point estimate.

The approximated posterior distribution of the latent variable given the observed sequence is prescribed by

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\phi_\omega(\boldsymbol{x}), \boldsymbol{\sigma}_z^2) \tag{6}$$

where mean $\boldsymbol{\mu}_z = \phi_\omega(\boldsymbol{x})$ is a deep neural network with parameters ω, and both standard deviations $\boldsymbol{\lambda}_z$, and $\boldsymbol{\sigma_z}$ are shared for all $\boldsymbol{x}$.

The parameters of the posterior are obtained by maximization of the *Evidence Lower Bound* (ELBO):

$$\mathrm{ELBO} = \mathrm{E}_{p(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z})\right] - \mathrm{KL}(p(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})). \tag{7}$$

By applying the reparametrization trick and Monte-Carlo sampling inside the expectation we can perform stochastic gradient descent on the ELBO:

$$\begin{aligned}\mathrm{ELBO} = &\sum_{i=1}^{n} \mathrm{E}_{p(z|x)}\left[\frac{(\boldsymbol{x}_i - \psi(\phi_\omega(\boldsymbol{x}_i) + \boldsymbol{\sigma}_z \odot \boldsymbol{\epsilon}))^2}{2\sigma_e^2}\right] + \frac{nd}{2}\log(\sigma_e) \\ &+ \sum_{i=1}^{n}\left(\log\left(\frac{\boldsymbol{\lambda}_z^2}{\boldsymbol{\sigma}_{\boldsymbol{z}}^2}\right) - m + \frac{\boldsymbol{\sigma}_{\boldsymbol{z}}^2}{\boldsymbol{\lambda}_z^2} + \frac{\phi_\omega(\boldsymbol{x}_i)^2}{\boldsymbol{\lambda}_z^2}\right),\end{aligned} \tag{8}$$

with noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, reconstruction $H(\psi(\boldsymbol{\theta}, \boldsymbol{\xi}(0))) \equiv \phi(\boldsymbol{z})$, and $\dim(\boldsymbol{z}) = m$. The ELBO is maximized with respect to the parameters of the encoder, and the variances $\boldsymbol{\lambda}_z, \sigma_e$. The resulting algorithm is a relevance determination of ODEs (*Rodent*). A schematic of the Rodent is shown in Fig. 1. The encoder network consists of two parts. A dense network that receives only few steps of the beginning of the time series. It is responsible for predicting the initial state $\boldsymbol{\xi}(0)$. The second part

of the encoder is a convolutional neural network (CNN) and predicts the ODE parameters $\boldsymbol{\theta}$. The CNN averages over the time dimension after the convolutions, which makes it possible to use samples of different length. Out approach unifies the following four aspects, which extend the conventional Variational Autoencoder:

- **Explainability.** The parameters of the latent vector $\boldsymbol{z}$ are decoded through an ODE solver, which makes the latent codes physically interpretable.
- **Sparsity.** The automatic relevance determination prior on the latent vector $\boldsymbol{z}$ encourages the simplest solution with fewest non-zero parameters.
- **Time series.** The convolutional part of the encoder is agnostic to different time series lengths.
- **Partial observations.** Rodent allows learning of an ODE without the knowledge of trajectories of all state variables. The observation operator is assumed to be known.

3 Related work

Bayesian compression Recent work has shown that adopting a Bayesian point of view to reduce parameters can significantly reduce computational cost while still achieving competitive accuracy [3, 4]. These approaches focus on pruning network weight or complete neurons in order to decrease computational cost and reduce overfitting. Our approach, however, enforces sparsity *only on the latent variable* and the main goal is interpretability of the latent variable.

Learning differential equations Identifying the parameters of differential equations purely from data has been addressed by a range of different approaches. Some using linear methods such as [6], others with Gaussian Processes or other kernel methods [7, 8], or Koopman theory [9, 10] which relies on finding a transformation into a higher dimensional space in which the ODE becomes linear. Combining Koopman theory and autoencoders has produced promising results, such as learning the ODE for a non-linear pendulum and the Lorenz system [11]. While the Koopman approach provides very accurate predictions once its nonlinear transformation is found, it lacks explainable results which we provide by learning the ODEs directly.

Discovering physical concepts purely from the data is a new field that neural networks are being applied to. Learning physically relevant concepts of the two body problem was recently solved both by *Hamiltonian Neural Networks* [12, 13] and by very problem specific VAE architectures [14].

Hyper networks are neural nets that generate weights for another network [15]. Learning all the parameters of an ODE through an encoder is a similar approach because the ODE can be regarded as a simple neural net. It is of course possible to represent the ODE with more complicated architectures, but this makes the learned parameters harder to interpret.

4 Identification of harmonic oscillator

We demonstrate the performance of Rodent on the problem of identification of a generative model of harmonic signals. For a frequency ω, the signal can be generated by a second order ODE $\ddot{\xi} = -\omega^2 \xi$ which can be written as a system of first order equations

$$\begin{bmatrix} \dot{\xi} \\ \ddot{\xi} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \dot{\xi} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{9}$$

The RHS of Eq. 9 is a 2×2 dense layer with weights $\boldsymbol{W}$, bias $\boldsymbol{b}$ and input $\boldsymbol{\xi}$. We want to learn both the network parameters and the initial conditions which means we have eight parameters of which only four are relevant: "1", "ω^2", and the two initial conditions $\xi(0)$ and $\dot{\xi}(0)$. To show that we can learn the simplest manifold of harmonic signals we over-parameterize the problem with a third order ODE (equivalent to a 3×3 dense layer). The third order ODE has 15 parameters in total, of which still only four are relevant. The observation operator H is defined such that only the first component of the state enters into the ELBO:

$$\phi(\boldsymbol{z}) = H \cdot \psi(\boldsymbol{\theta}, \boldsymbol{\xi}(0)) = [1, 0, 0] \cdot \psi(\boldsymbol{\theta}, \boldsymbol{\xi}(0)). \tag{10}$$

To demonstrate the advantage of the ARD prior, we train a Rodent and its variant lacking the ARD prior (called *Odent*) to reconstruct harmonic signals sampled from

$$p(\boldsymbol{x}|\omega, \alpha_0, \sigma_e) = \mathcal{N}(\boldsymbol{x}|\sin(\omega \boldsymbol{t} + \alpha_0), \sigma_e), \tag{11}$$

where $p(\alpha_0) = \mathcal{U}(0, 2\pi)$ and $p(\omega) = \mathcal{U}(0.5, 3)$. For both Rodent and Odent the dense part of the encoder (which predicts $\boldsymbol{\xi}(0)$) has two relu layers with 50 neurons. The convolutional part (responsible for $\boldsymbol{W}$ and $\boldsymbol{b}$) has four convolutional layers with 16 channels and a filter size of 3×1. An example of a single reconstruction and its generating latent code is shown in Fig. 2.

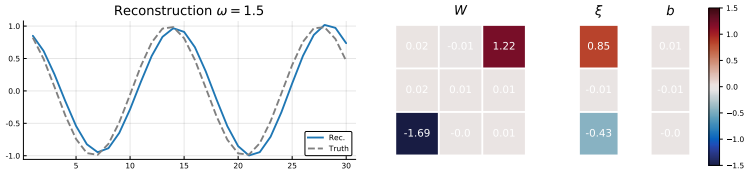


Figure 2: Rodent reconstruction compared to true data on the left. The heatmaps on the right show the corresponding encodings for the weights $\boldsymbol{W}$, biases $\boldsymbol{b}$, and initial conditions $\boldsymbol{\xi}$. It is clearly visible that the Rodent reduced the latent space to the four truly relevant parameters (see Eq. 9).

The learnt structures of the latent codes for samples from Eq. 11 are shown in Fig. 4. Both Rodent and Odent learn a reduced structure of the latent space, but we can clearly see that the pruning is much more effective in the Rodent. It keeps only the four relevant parameters with the rest being almost exactly zero, while without ARD five parameters remain and the rest being not as close to zero.

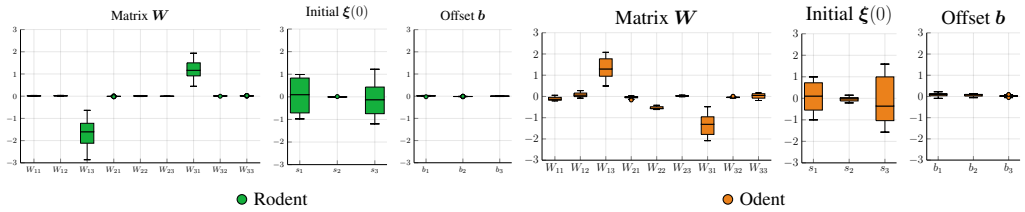


Figure 3: Latent codes of the Rodent compared to Odent. All redundant parameters are pushed to zero by Rodent while Odent keeps one irrelevant parameter (W_{22}). Additionally we can see that the Rodent squeezes irrelevant parameters closer to zero than Odent.

To show that the Rodent has learnt the manifold of harmonic signals, Fig. 4 shows the anomaly score (i.e. reconstruction error) of harmonic samples with frequencies $\{2.8, 2.9, \ldots, 3.9, 4\}$ (recall that the training set contained only frequencies up to 3), random Gaussian noise (denoted by "r"), and a square signal of a frequency within the trained range (denoted by "s"). Brown boxes represent the score of samples where the latent variables were taken directly from the encoder. Their score quickly increases as the frequency increases beyond the trained frequency range (i.e. over 3). If we *reidentify* the parameters, the Rodent can extrapolate much further than Odent while still detecting noise and square signals as anomalies (not being compatible with the generative model).

Reidentification During reidentification we sample a batch of latent codes from the encoder for each input sample. The latent samples are used as starting points for another optimization of the reconstruction error R

$$R = \text{MSE}(\phi(\boldsymbol{z}) - \boldsymbol{x}), \tag{12}$$

while keeping fixed all parameters that were identified as irrelevant (note that only the decoder ϕ enters in R). In case of the Rodent only four parameters, namely W_{13}, W_{31}, ξ_1, and ξ_3, were found to be relevant, so only those are allowed to change during the optimization with respect to R. This means that we stay in the identified model manifold, but are able to extrapolate far beyond the training range. Signals that fall outside of the model manifold such as the square signal are still identified as anomalous as shown in Fig. 4.

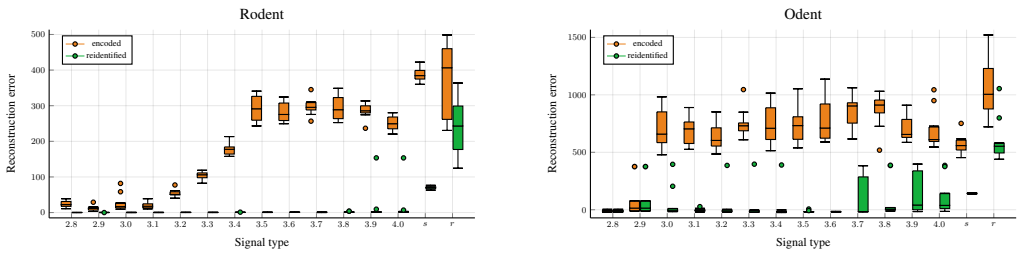


Figure 4: Reconstruction error of time series that are generated by different models: sine waves of frequency $\omega = \{2.8, \ldots, 4.0\}$, (denoted by the frequency value on x-axis), square signal ("s") and random noise ("r"). Reconstruction error based on parameters from the encoder in brown, error based on reidentified parameters in green.

5 Identification of superimposed harmonic oscillators

Many interesting signals are superpositions of different processes, so in this section we show that the Rodent can learn the superposition of harmonic signals as well. We sample time series from

$$p(\boldsymbol{x}|\alpha_0, \sigma_e) = \mathcal{N}(\boldsymbol{x}|\sin(\omega_1(\boldsymbol{t} + \Delta t)) + \sin(\omega_2(\boldsymbol{t} + \Delta t)), \sigma_e), \tag{13}$$

where $\omega_1 = 0.5$, $\omega_2 = 0.8$, and $p(\Delta t) = \mathcal{U}(0, 2\pi/\omega_1)$. For this problem the Rodent would optimally learn that the signal is the superposition of two harmonic oscillators, so we need a state size of four (two times the state size of a single harmonic oscillator). In order to make a super position of the two oscillators possible while still keeping the assumptions on the observation operator minimal ($H = [1, 0, 0, 0, 0]$), we increase the state size to five. The optimal solution of the problem is then given by:

$$\begin{bmatrix} \dot{\xi} \\ \dot{\xi}_1 \\ \ddot{\xi}_1 \\ \dot{\xi}_2 \\ \ddot{\xi}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -\omega_1^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -\omega_2^2 & 0 & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \xi_1 \\ \dot{\xi}_1 \\ \xi_2 \\ \dot{\xi}_2 \end{bmatrix}. \tag{14}$$

An exemplary reconstruction and the corresponding latent space is shown in Fig. 5. We observe that the Rodent actually learns a representation that is even more compressed than the optimal solution. The learned latent space is missing one component in the initial state, which means that we have not learned a superposition of two harmonic oscillators but a different ODE that can reconstruct Eq. 13. This becomes clearly visible when plotting the full state trajectory that is visible in Fig. 7A.

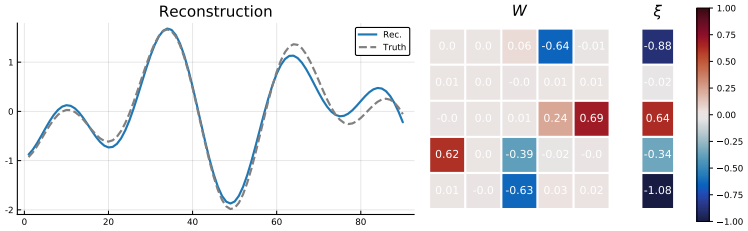


Figure 5: Reconstruction of a double harmonic signal on the left and corresponding latent variables on the right. The Rodent observes only the first state variable. It is clearly visible that one of the remaining state variables is compressed to zero, which means that we did not learn a superposition of harmonic signals.

The easiest way to obtain a decoupled state is to prescribe the state size to four and define the observation operator as $H = [1, 0, 1, 0]$. This is still quite a soft assumption, as we are only forcing the Rodent to find a solution which is an addition of two internal state variables. The reconstructions and the latent structure for this case are shown in Fig. 6. The corresponding, *decoupled* state trajectories are shown in Fig. 7B.

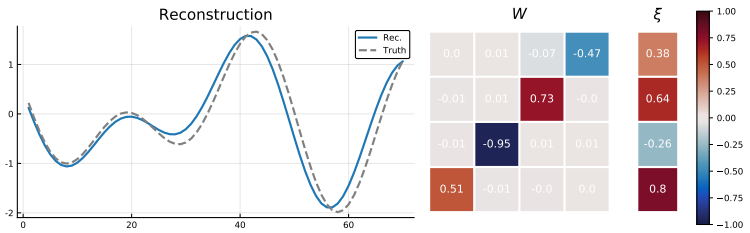
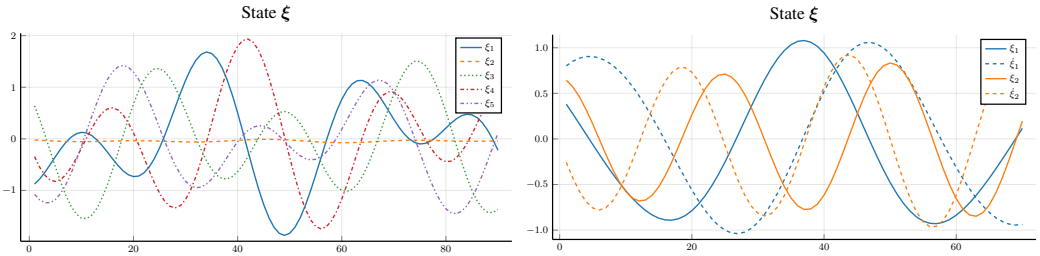


Figure 6: Reconstruction of double harmonic signal and corresponding latent variables. The observation operator $H = [1, 0, 1, 0]$ replaces the fifth state variable, which is why the only non-zero values are on the counter-diagonal.



(A) (B)

Figure 7: State trajectories corresponding to Fig. 5. The first state ξ_1 is the reconstruction (A). State trajectories to Fig. 6 (B). The states on the right show two clearly decoupled harmonic oscillators with the expected phase shift of $\alpha_0 = \pi/2$ between ξ_i and its time derivative $\dot{\xi}_i$.

6 Conclusion

This work aimed to identify the structure of a partially observed dynamical system modeled by an ODE. We used ARD to promote sparsity, which (i) identifies a model that is as simple as possible including the order of the ODE; (ii) improves the explainability of the learnt model; (iii) improves the performance on samples not present in the training set (extrapolation). All these properties and advantages were demonstrated on dynamics of harmonic oscillators. We plan to apply the Rodent to real problems in future work.

7 Acknowledgments

Research presented in this work has been supported by the Grant agency of Czech Republic no. 18-21409S. The authors also acknowledge the support of the OP VVV MEYS funded project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics”.

References

[1] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. December 2013. arXiv: 1312.6114v10.

[2] Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1996.

[3] Christos Louizos, Karen Ullrich, and Max Welling. Bayesian Compression for Deep Learning. page 11, 2017.

[4] Christos Louizos, Max Welling, and Diederik P. Kingma. Learning Sparse Neural Networks through L0 Regularization. *arXiv:1712.01312 [cs, stat]*, December 2017. arXiv: 1712.01312.

[5] David J C MacKay. Bayesian Non-linear Modeling for the Prediction Competition. page 14, 1994.

[6] W. Pan, Y. Yuan, J. Gonçalves, and G. Stan. A Sparse Bayesian Approach to the Identification of Nonlinear State-Space Systems. *IEEE Transactions on Automatic Control*, 61(1):182–187, January 2016.

[7] Maziar Raissi and George Em Karniadakis. Hidden physics models: Machine learning of nonlinear partial differential equations. *Journal of Computational Physics*, 357:125–141, March 2018.

[8] M. Bocquet, J. Brajard, A. Carrassi, and L. Bertino. Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models. *Nonlinear Processes in Geophysics*, 26(3):143–162, 2019.

[9] Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, April 2017.

[10] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, April 2016.

[11] Kathleen Champion, Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of coordinates and governing equations. page 27, 2019.

[12] Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks. *arXiv:1906.01563 [cs]*, September 2019. arXiv: 1906.01563.

[13] Peter Toth, Danilo Jimenez Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. Hamiltonian Generative Networks. *arXiv:1909.13789 [cs, stat]*, September 2019. arXiv: 1909.13789.

[14] Raban Iten, Tony Metger, Henrik Wilming, Lidia del Rio, and Renato Renner. Discovering physical concepts with neural networks. *arXiv:1807.10300 [physics, physics:quant-ph]*, September 2018. arXiv: 1807.10300.

[15] David Ha, Andrew Dai, and Quoc V. Le. HyperNetworks. *arXiv:1609.09106 [cs]*, December 2016. arXiv: 1609.09106.

Appendix: Reidentification with Rodent

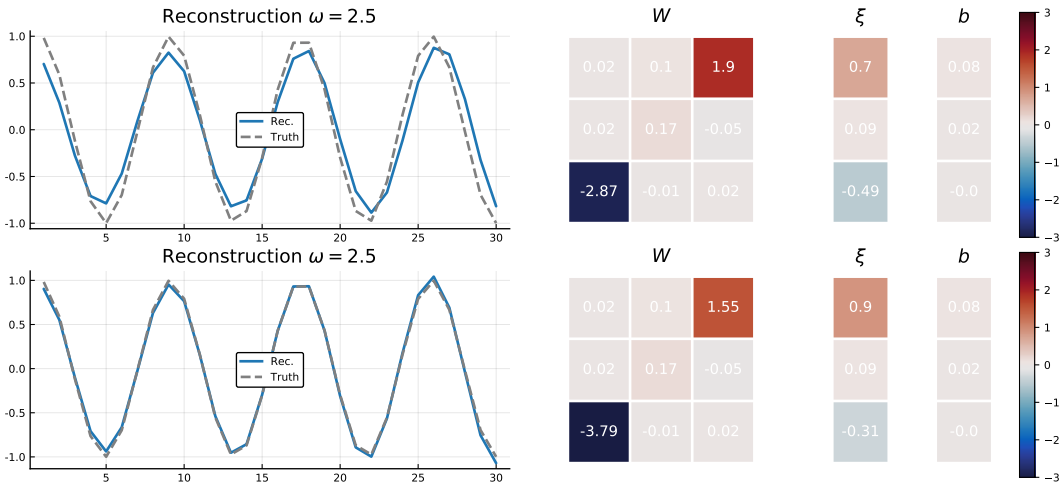


Figure 8: Rodent reconstructions and encodings of a harmonic signal in the upper plots. Reidentified reconstruction and encodings in the bottom. The values of the latent codes only change were the variables were detected as relevant.

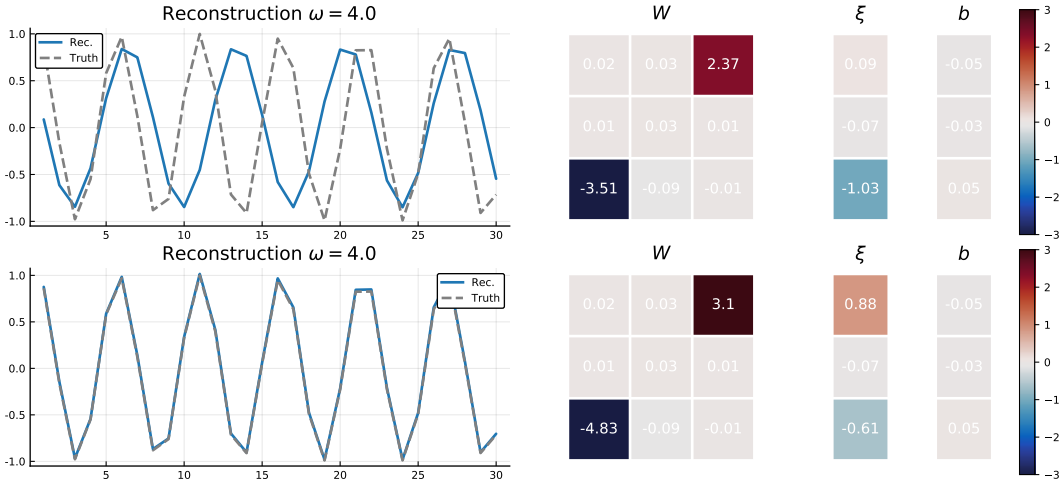


Figure 9: Rodent reconstructions and encodings of a harmonic signal in the upper plots. Reidentified reconstruction and encodings in the bottom. The values of the latent codes only change were the variables were detected as relevant. Recall that the Rodent was only trained with frequencies in the range of $\omega = (0.5, 3)$.

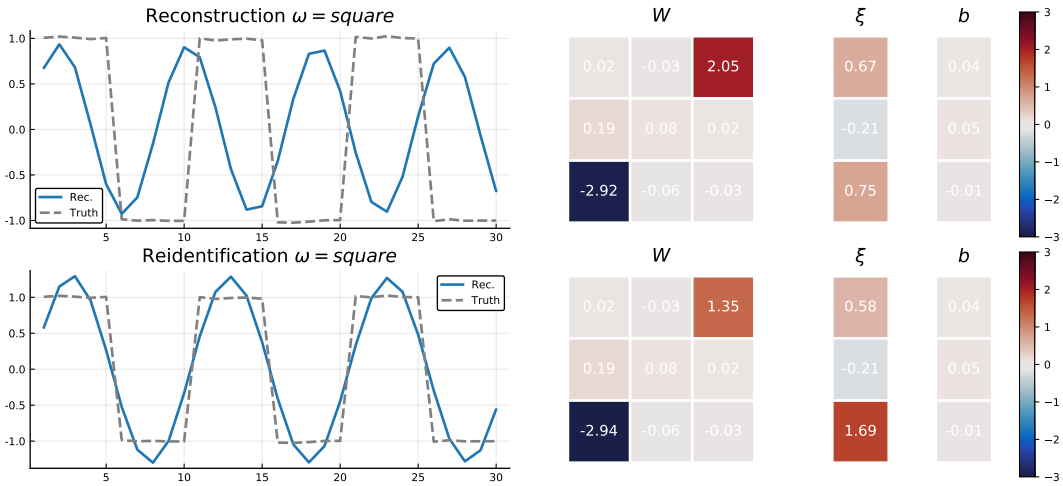


Figure 10: Rodent reconstructions and encodings of a harmonic signal in the upper plots. Detailed description in the caption of Fig. 9.

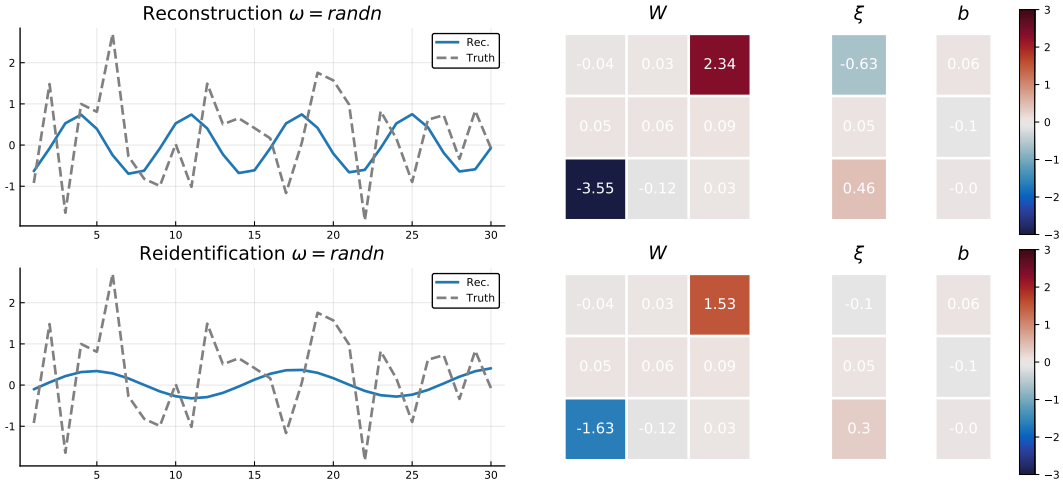


Figure 11: Rodent reconstructions and encodings of a harmonic signal in the upper plots. Detailed description in the caption of Fig. 9.