Learning By Doing

CHEM Track Solution

João Bravo

Feedzai

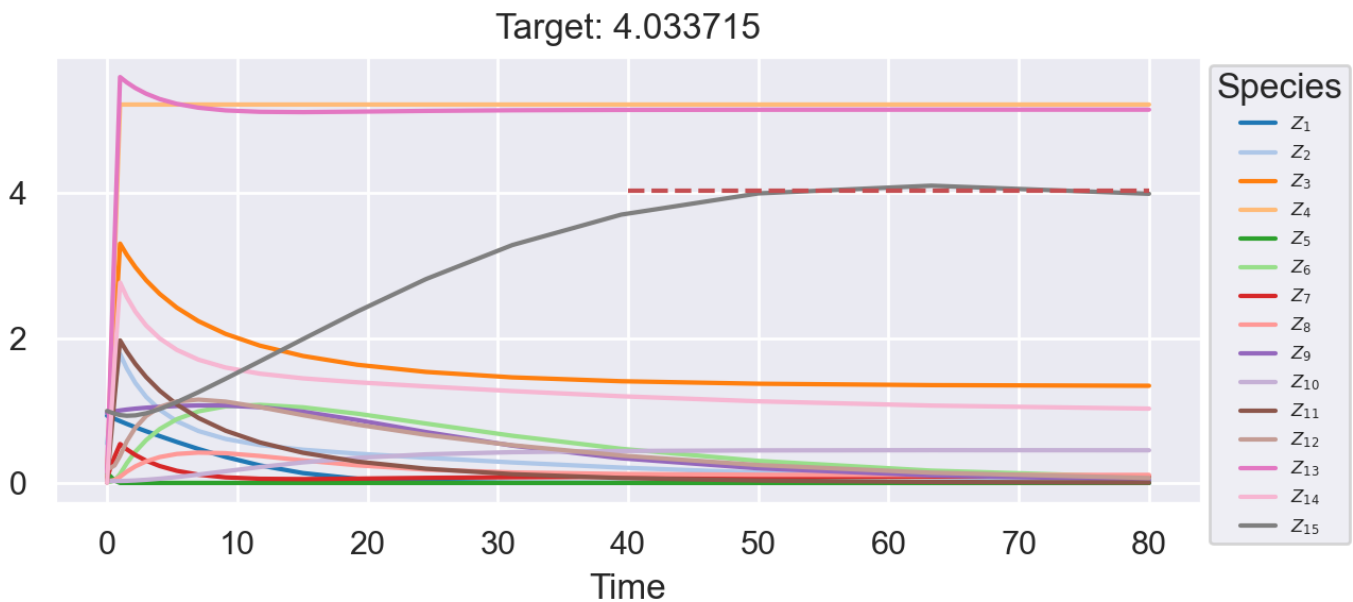
December 8, 2021

Introduction

The goal of the competition was to achieve a specified concentration for a target species in a chemical reaction network

The main challenges are:

- We can only act in the first second
- We know very little about the network
- We are provided simulated data for how 12 different networks react to control inputs



Introduction

Break down solution into 2 steps:

1. **System Identification (Inference)** Infer what is the dynamical system that gave rise to the observations
   (I) **Structure Inference** - What is the structure of the chemical reaction network
   (II) **Parameter Inference** - What are its parameters
2. **Control (Decision)** Find the control inputs to achieve the desired goal given our best estimate of the dynamics

The System Identification Problem

We have the following IVP describing all the involved species' concentrations over time:

$$Z(0) = z_0$$
$$\dot{Z}(t) = F\left(Z(t)\right)\theta + BU(t)$$

Given measurements obtained by adding zero mean white noise with an unknown distribution,

$$X_i = Z(t_i) + N_i,$$

estimate the parameters of the dynamical system

The System Identification Problem

The negative log-likelihood function for this model can be computed as:

$$L_{full}(z_0, \theta, B, \nu) = \sum_i n(X_i; \Phi(t_i; z_0, \theta, B, U), \nu)$$

Where:

- n is the negative log-likelihood for the noise model
- $\Phi(t_i; z, \theta, B, U)$ is the solution of the IVP

And the following parameters

- z : The initial condition
- θ : The parameters of the autonomous dynamics (shared across realizations)
- B : The input matrix (shared across systems)
- ν : The noise model parameters (shared across systems)

The System Identification Problem

We can forego estimating the noise model and solve the following nonlinear least squares problem instead:

$$L(z_0, \theta, B) = \sum_i \|X_i - \Phi(t_i; z_0, \theta, B, U)\|^2$$

Where:

- n is the negative log-likelihood for the noise model
- $\Phi(t_i; z, \theta, B, U)$ is the solution of the IVP

And the following parameters

- z : The initial condition
- θ : The parameters of the autonomous dynamics (shared across realizations)
- B : The input matrix (shared across systems)
- ν : The noise model parameters (shared across systems)

The System Identification Problem

For $t \in [t_3, \infty)$, U is 0 and the dynamics are autonomous. Focusing on this period we can ignore $U(t)$ and the unknown B matrix:

$$L_a(z_3, \theta) = \sum_{i=3} \|X_i - \Phi(t_i - t_3; z_3, \theta, 0, 0)\|^2$$

Where:

- n is the negative log-likelihood for the noise model
- $\Phi(t_i; z, \theta, B, U)$ is the solution of the IVP

And the following parameters

- z : The initial condition
- θ : The parameters of the autonomous dynamics (shared across realizations)
- B : The input matrix (shared across systems)
- ν : The noise model parameters (shared across systems)

A Naive Approach

Approximating the rates with finite differences gives as a crude method for estimating θ:

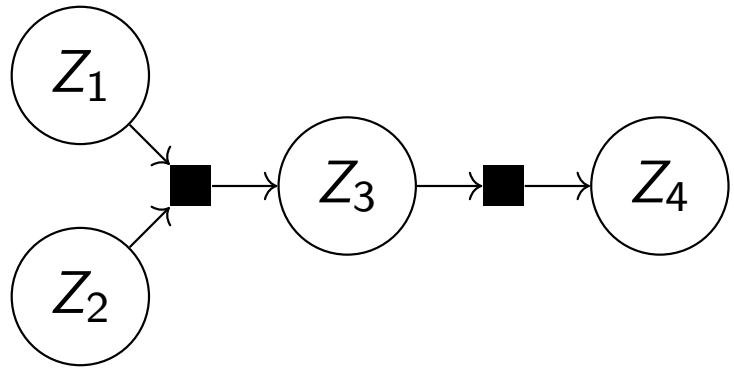
$$\hat{R}_i := \mathrm{fdiff}(X_{i-1:i+1}, t_{i-1:i+1}) \approx \dot{Z}(t_i) \approx F(X_i)\,\theta$$

- Minimizing $L_{fd}(\theta) = \sum_i \left\| \hat{R}_i - F(X_i)\,\theta \right\|^2$ is a linear least squares problem
- Residual is a messy stochastic process and not mean 0 in general
  - Error from finite differences
  - Error from replacing Z with X
- Bad estimator but still useful

The Autonomous Dynamics

The dynamics describe species concentrations for a chemical reaction network such as:

$$Z_1 + Z_2 \longrightarrow \gamma_{1,2;3} Z_3$$
$$Z_3 \longrightarrow \gamma_{3;4} Z_4$$



This leads to the following rate equations for the 4 species involved:

$$\dot{Z}_1 = -k_{1,2} Z_1 Z_2$$
$$\dot{Z}_2 = -k_{1,2} Z_1 Z_2$$
$$\dot{Z}_3 = \gamma_{1,2;3} k_{1,2} Z_1 Z_2 - k_3 Z_3$$
$$\dot{Z}_4 = \gamma_{3;4} k_3 Z_3$$

The Main Challenge

If we knew the reactions comprising the chemical network we could just estimate the reaction rates k

- We know that they only involve one or two reactants
  This gives $15 + 120$ possible reactions (times 15 potential products for each reaction)

$$F_l(Z)\theta = \sum_{j=1}^{15} \theta_j^l Z_j + \sum_{j=1,k=j}^{15} \theta_{j,k}^l Z_j Z_k, \quad l = 1, \ldots, 15$$

- Only 20 simulations per system

We are also told:

- There are only a few reactions (θ is sparse)
- The signs of the coefficients are consistent across systems (the reactions are the same)

The Main Challenge

If we knew the reactions comprising the chemical network we could just estimate the reaction rates k

- We know that they only involve one or two reactants
  This gives $15 + 120$ possible reactions (times 15 potential products for each reaction)

$$F_l(Z)\theta = \sum_{j=1}^{15} \theta_j^l Z_j + \sum_{j=1,k=j}^{15} \theta_{j,k}^l Z_j Z_k, \quad l = 1, \ldots, 15$$

- Only 20 simulations per system

We are also told:

- There are only a few reactions (θ is sparse)
- The signs of the coefficients are consistent across systems (the reactions are the same)

Structure Inference

A simple heuristic for inferring the set of reactions is then:

- Find θ by solving the LASSO problem

$$\min_{\theta} \sum_{i} \left\| \hat{R}_i - F\left(X_i\right)\theta \right\|_2^2 + \lambda \|\theta\|_1$$

  in order to encourage sparseness in the coefficients
- Select λ by 4-fold cross-validation over the 20 simulations for each system
- Check consistency of the signs of estimated coefficients across systems to narrow down the structure

Structure Inference

As an example, consider the reaction:

$$Z_j + Z_k \longrightarrow \gamma_{j,k;m} Z_m + \gamma_{j,k;n} Z_n$$

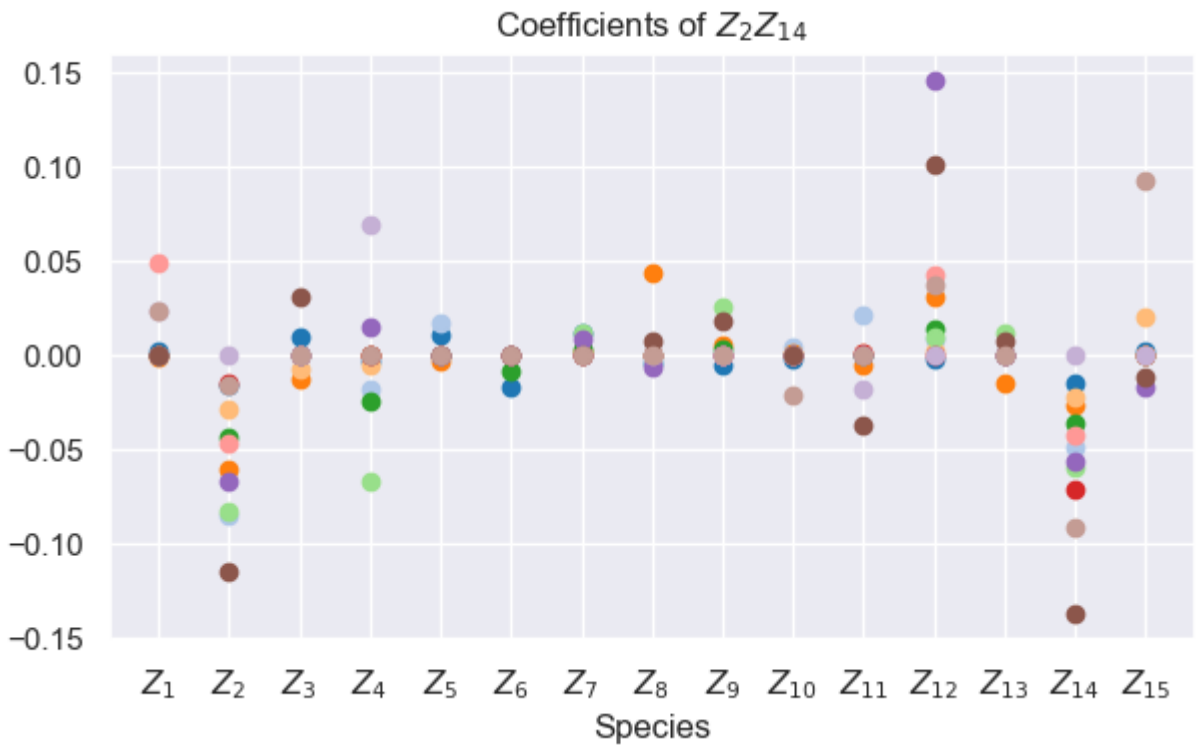
The coefficients $\theta^l_{j,k}$ of the monomial $Z_j Z_k$ should be:

- Negative for the reactants (unless one of them is also a product): $\theta^j_{j,k} = \theta^k_{j,k} < 0$
- Positive for the products: $\theta^m_{j,k} = -\gamma_{j,k;m} \theta^j_{j,k}$
- 0 otherwise

Structure Inference

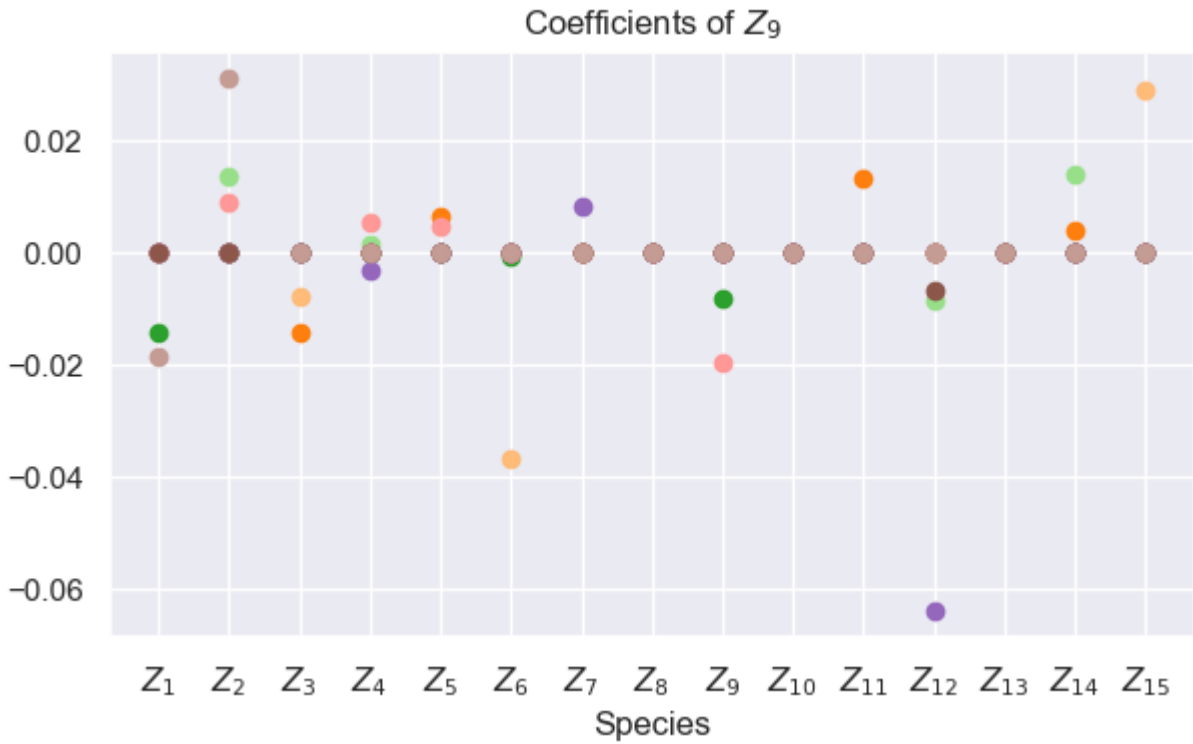
For example, looking at the coefficients of the monomial $Z_2 Z_{14}$ already suggests the reaction:

$$Z_2 + Z_{14} \longrightarrow Z_{12}$$



Structure Inference

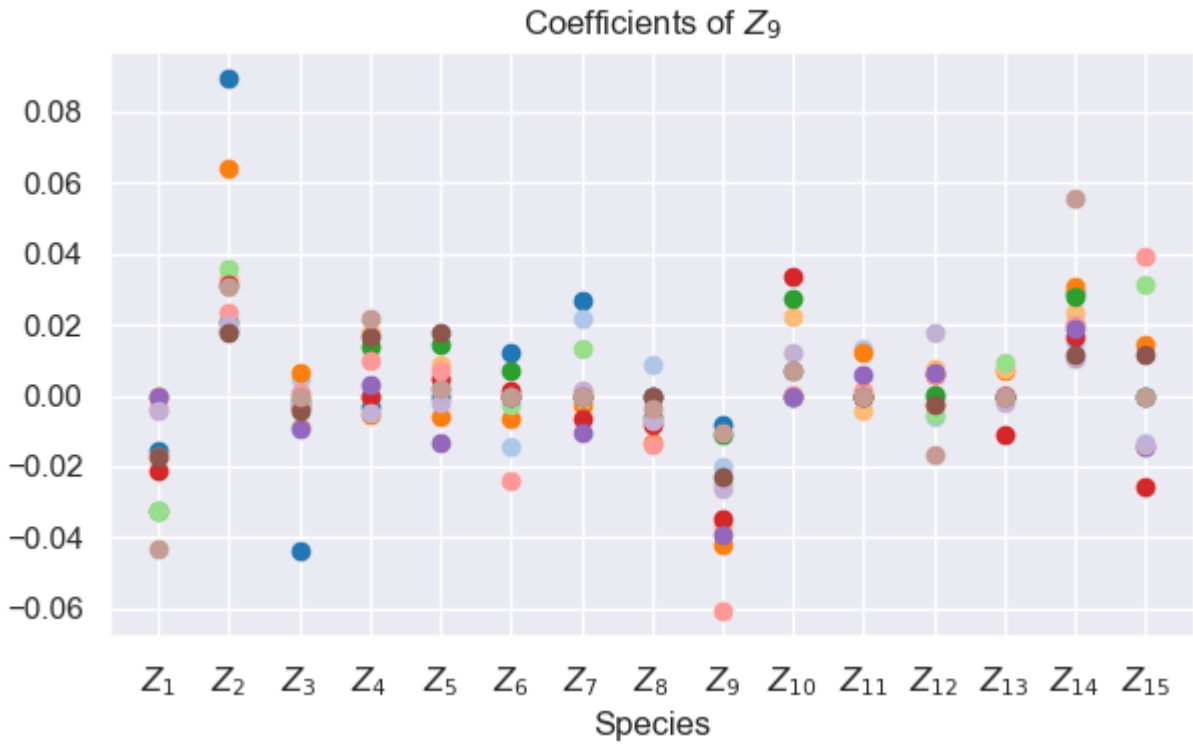
But not all reactions are this obvious at this point:



So we start by removing all monomials for which the signs of the reactants are not sufficiently consistent

Structure Inference

This removes some of the noise and the picture becomes clearer:



Structure Inference

Applying the same heuristic again we get the final set of 10 monomials to consider:

One Reactant:

$$Z_6 \longrightarrow ?$$
$$Z_9 \longrightarrow ?$$
$$Z_{10} \longrightarrow ?$$
$$Z_{12} \longrightarrow ?$$

Two Reactants:

$$Z_1 + Z_{15} \longrightarrow ?$$
$$Z_2 + Z_{14} \longrightarrow ?$$
$$Z_3 + Z_{11} \longrightarrow ?$$
$$Z_4 + Z_5 \longrightarrow ?$$
$$Z_7 + Z_{13} \longrightarrow ?$$
$$Z_8 + Z_{15} \longrightarrow ?$$

Structure Inference

We can now apply a similar heuristic and iterate to determine the products of each reaction:

One Reactant:

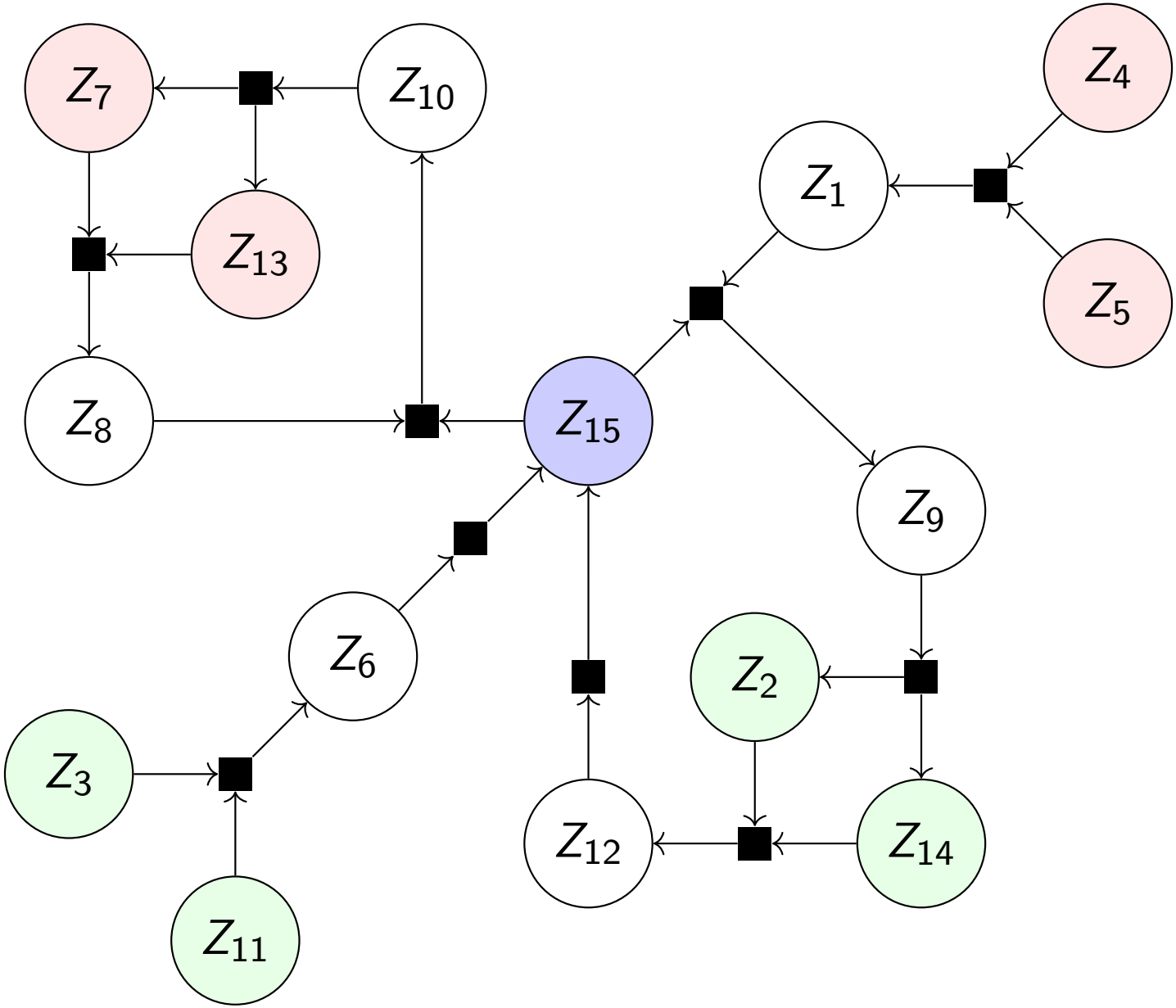
$$Z_6 \longrightarrow Z_{15}$$
$$Z_9 \longrightarrow Z_2 + Z_{14}$$
$$Z_{10} \longrightarrow Z_7 + Z_{13}$$
$$Z_{12} \longrightarrow Z_{15}$$

Two Reactants:

$$Z_1 + Z_{15} \longrightarrow Z_9$$
$$Z_2 + Z_{14} \longrightarrow Z_{12}$$
$$Z_3 + Z_{11} \longrightarrow Z_6$$
$$Z_4 + Z_5 \longrightarrow Z_1$$
$$Z_7 + Z_{13} \longrightarrow Z_8$$
$$Z_8 + Z_{15} \longrightarrow Z_{10}$$

- The products and stochiometries were in general more difficult to determine using this naive approach
- I checked a few discrete possibilities for the products structure in the next phase of parameter inference

Structure Inference



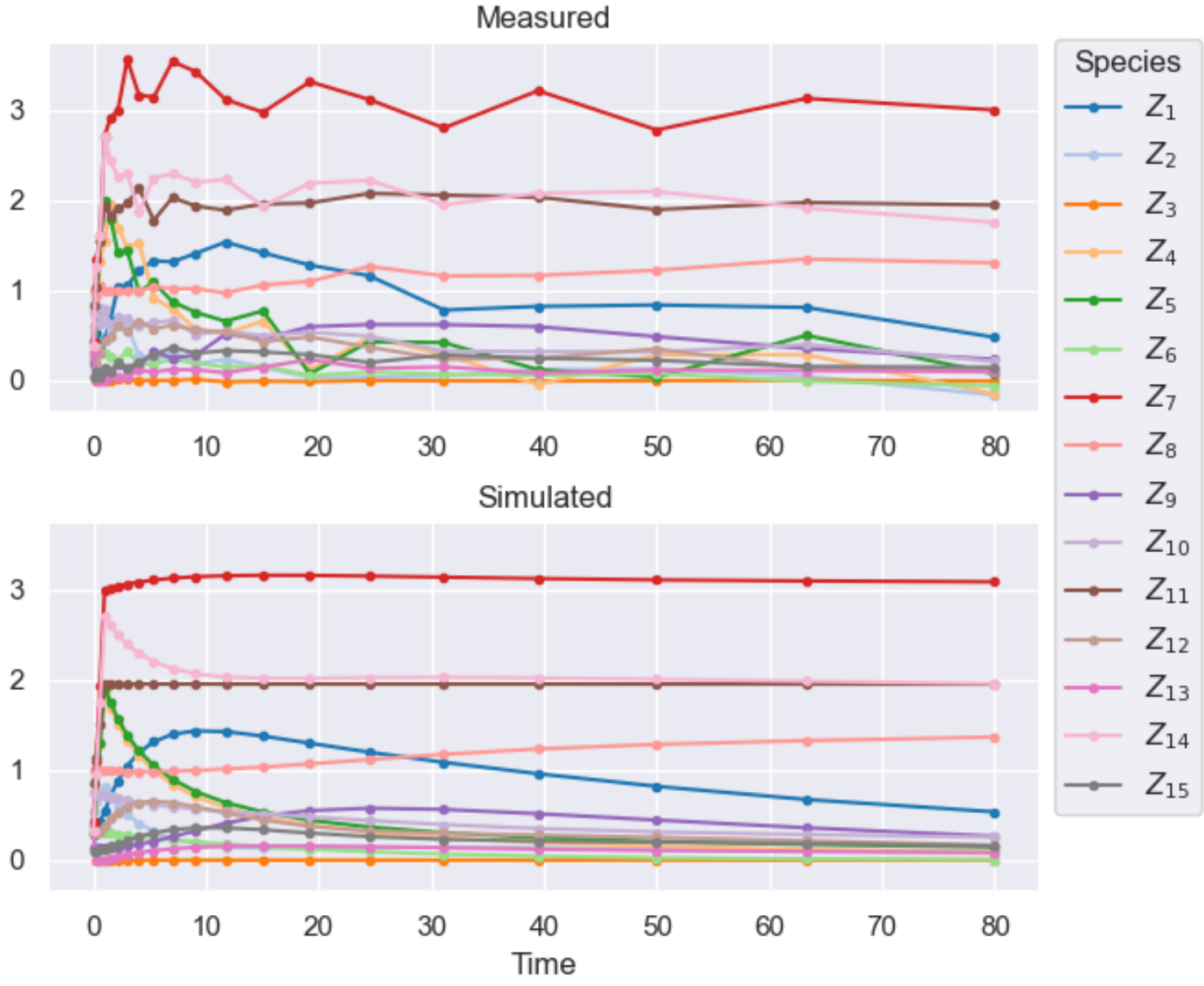
Parameter Inference

Having settled on the set of chemical reactions we estimate the parameters of the system:

- By solving the non-linear least squares problem
- Because the starting point matters we proceed in small steps

1. Solve the autonomous problem (i.e., for $t \geq t_3$):
   - I For k, fixing the initial condition $z_3 = (X_3)_+$
   - II For (z_3, k) to get a better estimate of the reaction rates
2. Solve the forced problem (i.e., for $t \geq t_0$):
   - I For B fixing k and initial condition $z_0 = (X_0)_+$ and pooling all systems together
   - II For (z_0, k) keeping B fixed to get final estimate

Parameter Inference



Control

The problem to solve for the control stage is similar to the parameter inference problem:

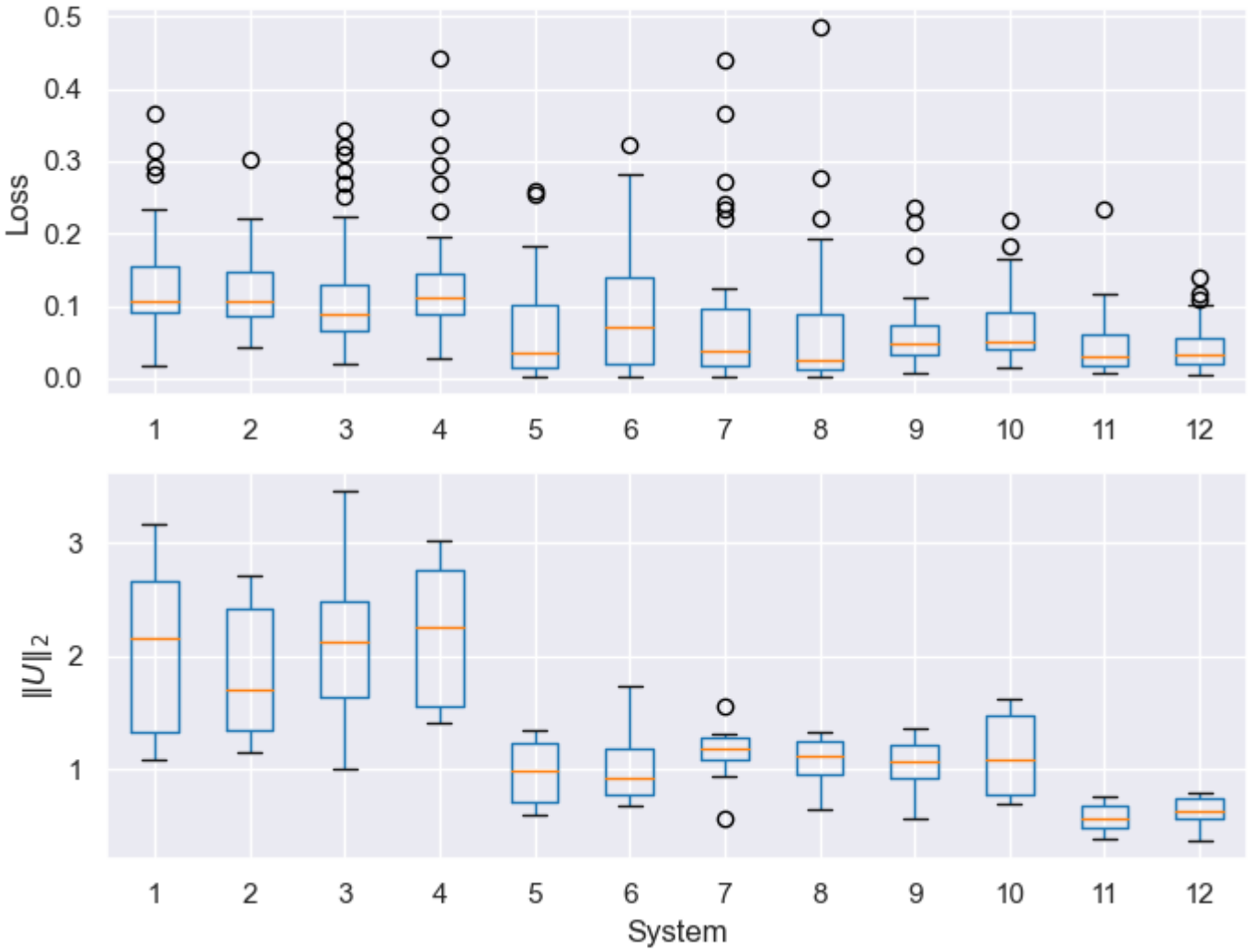
- For a set of initial conditions, z and target concentrations, y_*, select open loop controls that minimize

$$J(U) = \sqrt{\frac{1}{40}\int_{40}^{80} (\Phi_{15}(t; z, \theta, B, U) - y_*)^2 \, dt} + \frac{\sqrt{2}}{80}\|U\|_2$$

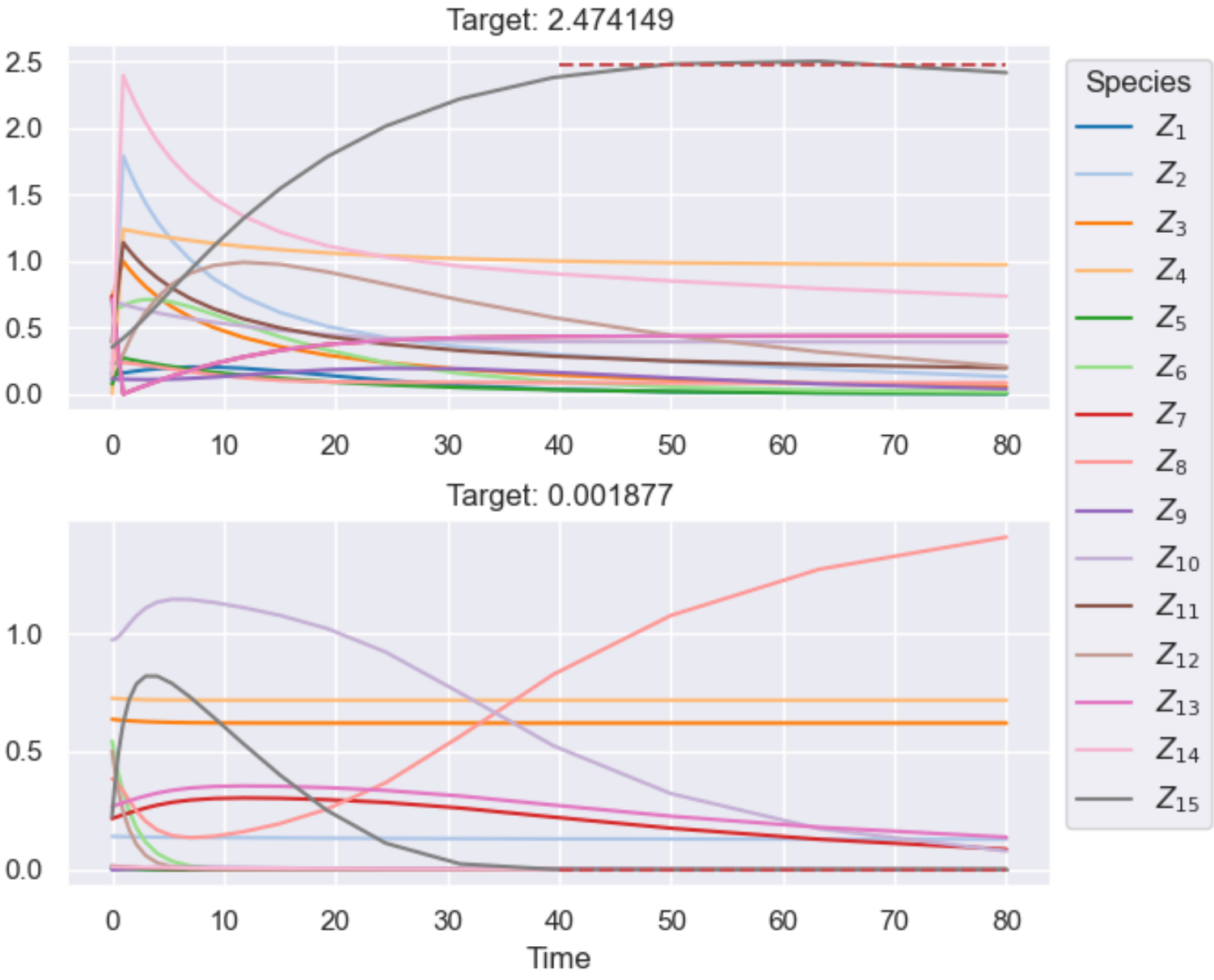
- This gives a non-convex optimization problem
- We solve it with `L-BFGS-B` from 10 different starting points sampled from $[-10, 10]^8$ with a Latin Hypercube Sampler

Control

We get somewhat consistent losses per system:



Control


Target: 2.474149
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70
80
Target: 0.001877
1.0
0.5
0.0
0
10
20
30
40
50
60
70
80
Time
Species
Z_1
Z_2
Z_3
Z_4
Z_5
Z_6
Z_7
Z_8
Z_9
Z_{10}
Z_{11}
Z_{12}
Z_{13}
Z_{14}
Z_{15}

Conclusion

- I was able to infer a plausible structure for the chemical reaction network with simple heuristics
- This allowed me to get a good predictor for the concentration of the target species over time making use of the dynamics we know
- In simulated data this seems to work very well but in reality we need to contend with model misspecification, stochastic exogenous inputs, ...