Bayesian Transfer Learning to Improve Predictive Performance of an ODE-Based Kinetic Model

Loïc Iapteff[1,2] **Benoit Celse**[1] Julien Jacques[2] Victor Costa[1]

[1]IFP Energies nouvelles
[2]Laboratoire ERIC, Université Lyon 2

2022 AIChE Spring Meeting 18th GCPS

Table of contents

The Challenge

- Objective: improve process modeling quality and robustness
- Modern industry: lot of data generated but for a new modeling problem, frequently start from zero
- IFPEN example: new catalyst = new model
- Aim: keep information from older dataset → **Transfer Learning**

Transfer Learning

Notations:

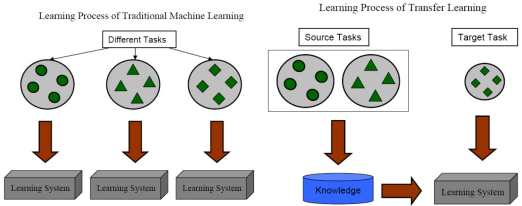
- Domain: $D = (X, P(X))$
- Task: $T = (Y, f)$
- Index "s" the source and "t" the target

Transfer Learning[1]

Improve the learning of f_t using D_s and T_s when $D_s \neq D_t$ or $T_s \neq T_t$

Transfer Learning approaches:

- Transfer knowledge of instances
- Transfer knowledge of features representation
- **Transfer knowledge of parameters**



(a) Traditional Machine Learning (b) Transfer Learning

[1] Pan and Yang, "A Survey on Transfer Learning"

Bayesian Inference

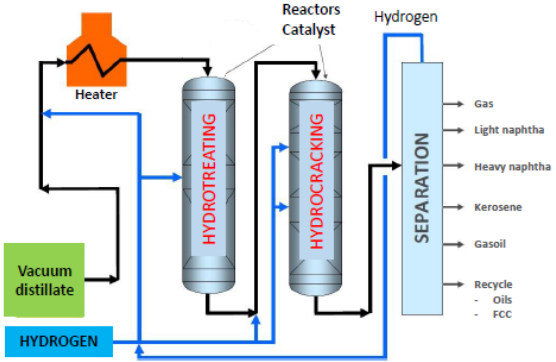
$$\underbrace{\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})}_{\text{(posterior)}} = \frac{\overbrace{\pi(\boldsymbol{\beta})}^{\text{(prior)}}\overbrace{f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{X})}^{\text{(likelihood)}}}{\underbrace{f(\boldsymbol{y}|\boldsymbol{X})}_{\text{(marginal likelihood)}}}$$

- In Frequentist statistics, $\boldsymbol{\beta}$ are optimized s.t. the likelihood is maximal.
- In Bayesian statistics, a prior on $\boldsymbol{\beta}$ is added and the posterior is maximized. It makes it useful for Transfer Learning problems.

The idea: Use as prior the distribution of the parameters learned on source

Previous work: Modeling of Hydrocracking process[1]

- Two hydrocracking industrial datasets:
  - Source dataset: from refineries using catalyst (n)
  - Target dataset: from refineries using catalyst (n+1)
- Objective: model the output Diesel Density for the catalyst (n+1)
  - Constraint: few observations for target dataset
  - **12 features** used defined by the expert
- Bayesian transfer Learning to use the knowledge from the catalyst (n) to predict (n+1)



[1] "Modeling the hydrocracking process with kriging through Bayesian Transfer Learning", 2021 AIChE Virtual Spring Meeting

Previous work: Modeling of Hydrocracking process[1]

Results

- Models used: kriging and linear model
- The prior distribution is fitted using the source dataset
- The Bayesian transfer method:
  - Reduce the number of required observations to fit model of good quality
  - Increase model predictive performance

Comparison of Bayesian transfer Learning and classical approaches:

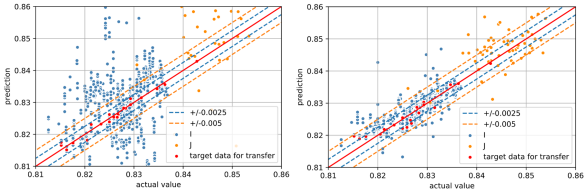


Figure: Results of Linear model for a sample of 20 target observations (left: without transfer, right: Bayesian approach

[1] "Modeling the hydrocracking process with kriging through Bayesian Transfer Learning", 2021 AIChE Virtual Spring Meeting

Work in study

- Aim: Model the Nitrogen content after Hydrotreatment stage
- Model used: ODE based Kinetic model
- Bayesian transfer method available for every parametric model:
  Method: use of Bayesian transfer to improve the model quality with few data
- Two datasets from pilot units:
  - Source: from "old" catalyst
  - Target: from "new" catalyst

Data Presentation

Source Dataset

- 144 observations
- **Aim: fit a good model to use parameters distribution as prior**

Target Dataset

- 126 observations
- **Aim: fit a good model when few observations are available**

Outlier detection:
Local outlier factor

The features:

- $LHSV$: Liquid Hourly Space Velocity, inverse of the residence time t
- T: Temperature of the hydrotreating reactor
- ppH_2: Hydrogen partial pressure
- TMP: Weighted average of the simulated distillation: $TMP = \frac{1}{7}(\text{FEED\_DS05} + 2 \times \text{FEED\_DS50} + 4 \times \text{FEED\_DS95})$
- N_0: Nitrogen content in feedstock
- S_0: Sulfur content in feedstock
- Res_0: Resines content in feedstock
- N: Nitrogen content after hydrotreating (to be predicted)

The model

- ODE based kinetic model:

$$\frac{dN}{dt} = -k_0 \frac{exp(-\frac{E_a}{R_g}(\frac{1}{T} - \frac{1}{Tref}))(\frac{ppH_2}{ppH_{2,ref}})^m N^n}{(1 + A_0 Res_0)(1 + \frac{C_0 N_0}{1+S_0})} \times (1 - u \cdot exp(-\frac{b}{R_g}(\frac{1}{T} - \frac{1}{Tref}))(\frac{ppH_2}{ppH_{2,ref}})^a (\frac{TMP}{TMP_{ref}})^v N^t).$$

  where $\boldsymbol{\theta} = (k_0, E_a, m, n, a, b, A_0, C_0, u, t, v)$ are the parameters to be optimized
- Boundary for parameters value to keep a physical sense
- Score to minimize: $\sum_{i=1}^{K} \frac{(\hat{y}_i - y_i)^2}{max(5, y_i)}$

Statistical model

In order to perform Bayesian inference, need to have a statistical model:

$$\begin{aligned} y_i &= f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + \epsilon_i, \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_i^2), \end{aligned}$$

where $f_{\boldsymbol{\theta}}(\cdot)$ is the solution of the differential equation (1)
Heteroscedastic model is considered and expression of σ_i must be chosen to fit with the cost function $\sum_{i=1}^{K} \frac{(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)-y_i)^2}{max(5,y_i)}$

Statistical model

With $\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_K^2 \end{pmatrix}$, we obtain:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) &\sim \mathcal{N}(f_{\boldsymbol{\theta}}(\boldsymbol{X}),\mathbf{\Sigma}) \\ &\propto exp(-\frac{1}{2}(\boldsymbol{y}-f_{\boldsymbol{\theta}}(\boldsymbol{X}))^T\mathbf{\Sigma}^{-1}(\boldsymbol{y}-f_{\boldsymbol{\theta}}(\boldsymbol{X}))) \\ \boldsymbol{\theta}_{ML} &= \underset{\boldsymbol{\theta}}{argmin}\frac{1}{2}(\boldsymbol{y}-f_{\boldsymbol{\theta}}(\boldsymbol{X}))^T\mathbf{\Sigma}^{-1}(\boldsymbol{y}-f_{\boldsymbol{\theta}}(\boldsymbol{X})) \\ &= \underset{\boldsymbol{\theta}}{argmin}\sum_{i=1}^{K}\frac{(y_i-f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))^2}{\sigma_i^2} \end{aligned}$$

With $\sigma_i^2 = \sigma \cdot max(5, y_i)$, σ unknown, the maximum likelihood estimator $\boldsymbol{\theta}_{ML}$ minimizes $\sum_{i=1}^{K}\frac{(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)-y_i)^2}{max(5,y_i)}$

Fitted source model

The model is fitted using source dataset and offer satisfying results:

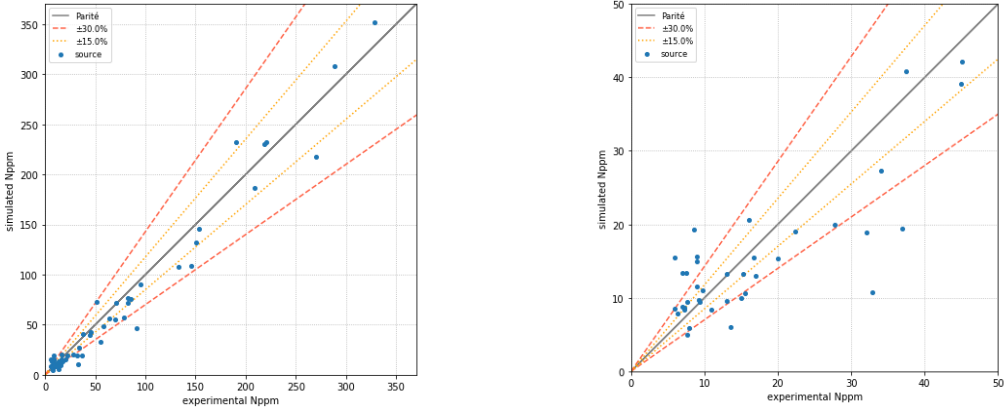


Figure: Parity plot of the fitted source model. A zoom is applied on the right.

Source model on target dataset

The source model is tested on the target dataset:

- Prediction higher than actual value: new catalyst more active
- Model readjustment needed
- Aim: use few target observations to fit the target model with the help of source knowledge

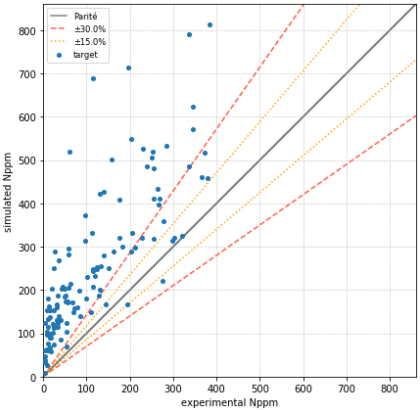


Figure: Parity plot of the source model applied to the target dataset

Prior choice

To use Bayesian inference, a prior distribution is needed:

- A Gaussian distribution is assumed for the model parameters:

$$\pi(\boldsymbol{\theta}_t) \quad \sim \quad \mathcal{N}(\hat{\boldsymbol{\theta}}_s, gVar(\hat{\boldsymbol{\theta}}_s))$$

- g is a scalar, that must be chosen, to adapt prior impact:

$$\hat{\boldsymbol{\theta}}_t \quad \underset{g \to 0}{\to} \quad \hat{\boldsymbol{\theta}}_s$$

$$\hat{\boldsymbol{\theta}}_t \quad \underset{g \to +\infty}{\to} \quad \hat{\boldsymbol{\theta}}_{t,ML}$$

- A MCMC algorithm is used to obtain source parameters distribution and estimate $\hat{\boldsymbol{\theta}}_s$ and $Var(\hat{\boldsymbol{\theta}}_s)$

Experimentation

Experiments are carried out in order to

- Find an effective method for the choice of g-value
- Compare Bayesian transfer with classical approach

Testing process

- Different target sample sizes considered: 5, 10, 15, 20
- For each size, 10 random samples tested
- For each sample, different value for g: 1, 10, 100, 1000, 10000

Experimentation

Example for a random sample of size 15:

- Training set:
  - High g-value, with and without transfer scores are similar (prior neglected)
  - When the value of g decreases, the training score decreases.
- Test set:
  - The score evolution is not monotonous
  - A well chosen g-value increases model quality and conversely a badly chosen g-value decreases it

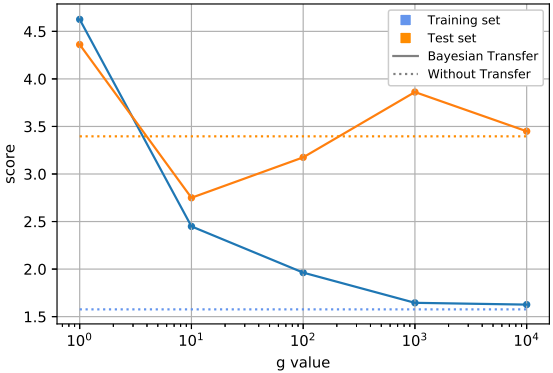


Figure: Example of g-value impact for a random sample of size 15

Example of the size 15 random sample

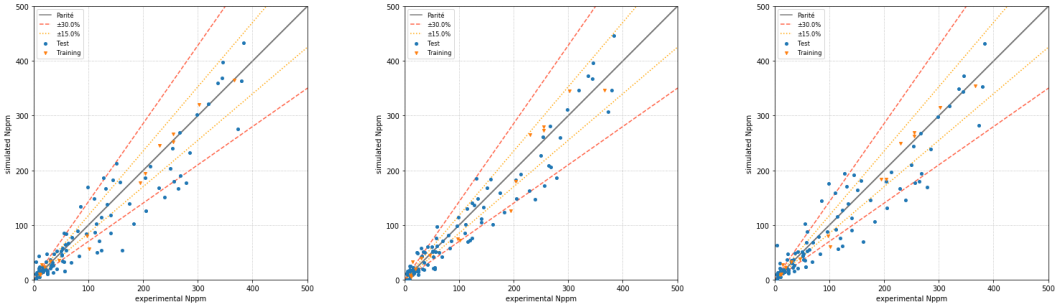


Figure: Parity plot for the size 15 random sample. Left: Without transfer, Center: Bayesian transfer with g=10, Right: Bayesian transfer with g=1000

- Similar results as 15 observations are sufficient to fit a satisfying model, but improvement with Bayesian transfer with a good g-value

Example of the size 15 random sample

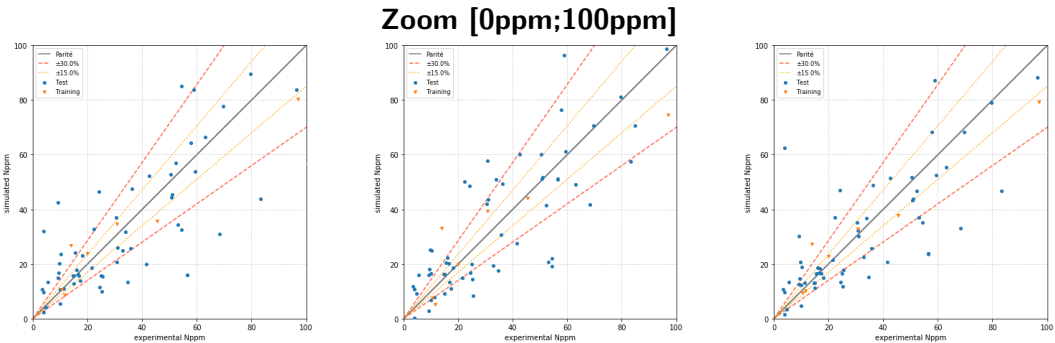


Figure: Parity plot for the size 15 random sample. Left: Without transfer, Center: Bayesian transfer with g=10, Right: Bayesian transfer with g=1000

- Similar results as 15 observations are sufficient to fit a satisfying model, but improvement with Bayesian transfer with a good g-value

Example of a size 10 random sample

- For this application and model, small designs of 15 observations offer good results
- With less observations, the classical approach can lead to really bad model
- The Bayesian transfer model still offers satisfying results and thus a great improvement

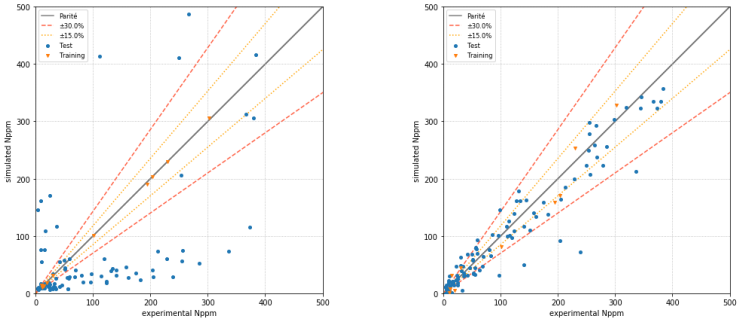


Figure: Parity plot for a random sample of size 10. Left: Without transfer, Right: Bayesian transfer with g=10

Choice of g-value

Two approaches tested:

First method: Cross Validation

For a given sample of size n_{sample}:

- For each g-value tested, leave one out cross validation is performed on the training set:
  - n_{sample} model are fitted using $n_{sample} - 1$ observations and score is evaluated on the remaining observation
  - The mean of the n_{sample} test scores is considered
- The value of g with the lowest averaged score is kept
- Many model to fit: time consuming

The chosen g-value is not the same for the different designs

Choice of g-value

Two approaches tested:

Second method: Bound on training score

Idea: The training score starts with the score of the source model with $g \approx 0$ and monotonically reaches the score without transfer as the value of g increases. The aim is to maximise the prior impact without getting a bad model on the training set.
For a given sample:

- Take the lowest g-value so that the training score is lower than the expectation score
- Need to know the performance expectation: source model score on source dataset is used

The chosen g-value is not the same for the different designs

Results with cross validation

- g chosen using cross validation
- Score improve with Bayesian transfer, particularly with small designs
- Smallest min-max interval with Bayesian transfer: less impacted by design quality

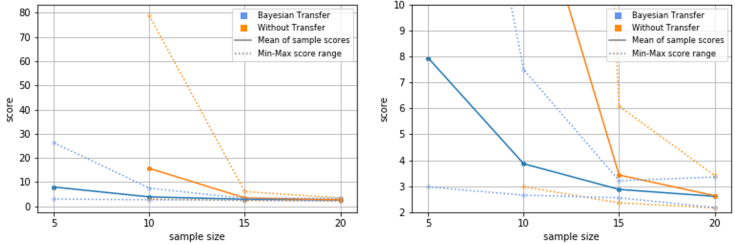


Figure: Score evolution according to sample size. The mean and the minimum-maximum score range over the 10 samples are plotted. On the right, a zoom is applied.

Results with bound on training score

- Similar results for the second method for selecting the g-value
- Both methods provide a good choice of g
- Which one to chose:
  - Cross validation: time consuming
  - Bound on training score: hyperparameter to fix

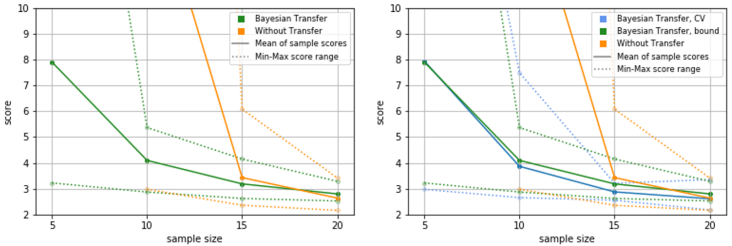


Figure: Score evolution for different sample size. On the right, the different approaches are compared.

Conclusion

- Conclusion:
  - Bayesian transfer leads to more robust model, less impacted by the design quality
  - The prediction performance is improved, especially for small designs
  - A good choice of g is crucial for good performance: the cross-validation method is recommended
- Perspective:
  - Test more g-value to refine its chosen value
  - Couple Bayesian transfer with Design of Experiment for ODE based kinetic model
  - Apply Bayesian transfer on other parametric model and application

Any questions?

Thanks for your attention!

Loïc IAPTEFF, PhD student at IFP Energie Nouvelles, France
PhD guidance:
Julien JACQUES, Université Lyon 2
Benoît CELSE, IFPEN
Victor COSTA, IFPEN
Mail: loic.iapteff@ifpen.fr