Reducing the number of experimental points to fit kinetic models : A Bayesian approach

Loïc Iapteff, Julien Jacques, Benoit Celse, Victor Costa

Reducing the number of experimental points to fit kinetic models: A Bayesian approach

Loïc Iapteff,[†,‡] Julien Jacques,[‡] Benoit Celse,[†] and Victor Costa[*,†]

†*IFPEN, Rond-point de l'echangeur de Solaize 69360 Solaize, France*

‡*Laboratoire ERIC, Univ Lyon 2, 5 Av. Pierre Mendes France 69500 Bron, France*

E-mail: victor.costa@ifpen.fr

Abstract

Hydrocracking is a crucial refinery process that transforms heavy molecules (i.e. vacuum gas oil (VGO)) into lighter and highly-valued products such as naphtha, kerosene and diesel. It is a two-step process. The hydrotreatment (HDT) reactor uses a more robust catalyst, which essentially serves to remove heteroatoms from the VGO feed in order to satisfy product quality constraints and avoid poisoning of the more delicate zeolite-based HCK catalysts. The second, hydrocracking (HCK) reactor uses a commercial zeolite catalyst with a carefully selected balance of acid and metallic sites. For hydrotreatment simulation, the kinetic model is decomposed in several ODE (Ordinary Differential Equation). Catalyst vendors develop more and more catalysts. For each new catalyst (new generation), the kinetic parameters must be refitted. This task is costly and time consuming.

In this paper, in order to reduce the required number of experimental points, a Bayesian transfer approach is proposed to fit the parameters of catalyst (n+1), using the past knowledge of catalyst (n) to add more information. A method for the choice of the prior is proposed and can be used for any type of parametric model. This approach is applied and shows an improvement in the prediction performance and robustness

compared to a classical fitting method. In our case, only 10 pilot plant points on catalyst (n+1) are requested to refit a HDN kinetic model.

1 Introduction

In a classical hydrocracker, a mixture of hydrocarbon feed and hydrogen is heated and injected into a reactor vessel containing a hydrotreating catalyst. This catalyst accelerates the reactions that remove sulfur and nitrogen from the hydrocarbon and open up and saturate aromatics rings. The entire output from this reactor is then injected into a second reactor containing a hydrocracking catalyst, which helps the reactions that crack apart the hydrocarbons while saturating them with hydrogen. The resulting mix of converted and unconverted hydrocarbon is then separated. Unconverted hydrocarbon can then be recycled to the hydrocracking step for further conversion, sent to a second hydrocracking vessel, or sent to another conversion unit as feed (e.g., an FCC). Diesel range material can also be drawn off at the separation steps to maximize diesel production, or it can be processed further (through recycling or second-step hydrocracking) to maximize naphtha production. Some hydrocrackers are single-stage units with just one reactor that is usually filled with hydrocracking catalyst, but the rest of the process is the same. The HCK is generally run at high temperatures (up to 420°C) and at high hydrogen pressures (>90 bar) on a bi-functional catalyst.

The hydrocracker is particularly valuable in a refinery that is trying to maximize diesel production and reduce residual fuel oil. The hydrocracker yields a high volume of kerosene and diesel of good quality (high cetane and low sulfur). However, its volume yield of naphtha is low and of low quality (low RON).

Hydrotreating before the hydrocracking reactor ensures better operation of the hydrocracking catalyst because the catalyst is sensitive to these heteroatoms. For example, neutralisation of acid sites by basic nitrogen compounds or poisoning of metal sites by sulphur or

metals. Without pretreatment, hydrocracking catalysts would deactivate quickly and would have to be replaced more often.

The most widely used method for modeling hydrotreating problems are kinetic-rate models. These models provide a mathematical representation of the chemical reactions taking place. They are commonly based on differential equations. Experimental data are used to fit the model parameters. The advantage of this method is that it offers a good understanding of the process. Many research has been carried out on kinetic model construction for hydrotreating, and reviews[1,2] on kinetic modeling for these processes are proposed. The kinetic modeling can be divided in two different approaches: the detailed kinetic modeling and the lumping approach. Due to the many compounds that occur in the feedstock and the very complex reactions in the process, a detailed kinetic model is very difficult to develop. The lumping approach is the most used. It groups the oil mixture into different components according to their chemical properties. Becker et al.[3,4] used continuous lumping for hydrotreating and hydrocracking modeling. The authors reported that the model gives good conversion and yield structure predictions. Sánchez et al.[5] proposed a kinetic model for hydrocracking of heavy oils using the lumping approach. The products are defined as 5 different lumps which are unconverted residue (> 538 °C), vacuum gas oil (343-538 °C), distillates (204-343 °C), naphtha (Initial Boiling Point-204 °C) and gases. The model is reported to provide a good prediction with an average absolute error of less than 5 percents. Sadighi et al.[6] developed a 6-lump kinetic model including a catalyst decay for a commercial VGO hydrocracker. The 6 lumps are light naphtha (40-90 °C), heavy naphtha (90-150 °C), kerosene (150-260 °C), diesel (260-380 °C), Vaccum Gas Oil and unconverted oil (> 380 °C).

Catalyst vendors develop more and more catalysts. For each new catalyst (new generation), the kinetic parameters must be refitted.

For each catalyst generation, the following workflow is used to fit the developed kinetic models[7]:

1. Definition of an experimental design

2. Execution of the pilot plant tests

3. Validation of the pilot plant tests

4. Kinetic parameters estimation

Overall, this is very time consuming and costly. The definition of an experimental design is usually a function of the allocated experimental time and the project priorities. As for the tests themselves, they can last for 4 to 6 weeks each, producing 4 to 8 experimental points. A full campaign for a catalyst generation can involve 5 to 10 tests during 6 to 12 months. Once the test is finished and the analytical results are available, a fine analysis (trend evolution, outlier detection, comparison with previous tests, ...) is required to process and validate the results. This may take up to 3 months. At the end of the line, the raw validated data undergoes further treatment by the kinetic model engineer. A model is then defined and the parameters are fitted. This stage normally requires 3 months of work. As a rule of thumb, we consider that a new generation of catalysts arrives every couple of years.

The purpose of this article is to reduce the requested time for a catalyst of generation (n+1) by using the information of catalyst of generation (n).

The aim of our work is then to improve the quality of our hydrodenitrogenation (HDN) model using Transfer Learning. To define Transfer Learning, let's start by defining a domain as $\mathcal{D} = (\mathcal{X}, P(\mathcal{X}))$ with $\mathcal{X}$ a feature space and $P(\mathcal{X})$ its probability distribution, and an associated task $\mathcal{T} = (Y, f)$ with f the function used to predict $y \in Y$ given $\boldsymbol{x} \in \mathcal{X}$. The target data is the data from the phenomenon to be modeled, while the source dataset refers to a linked dataset, used to improve the modeling of the target. We name $\mathcal{D}_s$ and $\mathcal{T}_s$ the domain and task of source data and $\mathcal{D}_t$ and $\mathcal{T}_t$ the domain and task of the target data. Transfer Learning is intended to improve the learning of the target predictive function f_t using the knowledge in $\mathcal{D}_s$ and $\mathcal{T}_s$, where $\mathcal{D}_s \neq \mathcal{D}_t$, or $\mathcal{T}_s \neq \mathcal{T}_t$. Reviews[8–10] on Transfer Learning can be read for those interested in a complete analysis of the state of the art in this domain.

The work presented in this paper is an extension of the method proposed in a previous work[11] for the modeling of the density of the hydrocracking diesel cut using a kriging model (gaussian process). The aim of the current paper is to use the information from the previous generation catalyst (n) to reduce the required time to fit the current catalyst generation (n+1) model. This method relies on Bayesian statistics,[12] which is based on the Bayes rule (Equation (1)):

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \quad = \quad \frac{\pi(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}, \tag{1}$$

where $f(y|\boldsymbol{\theta})$ is a parametric model with parameters $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$ is the prior distribution and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior distribution. Bayesian inference assumes external knowledge about the parameters, without having seen the data. This external information is contained in the prior distribution $\pi(\boldsymbol{\theta})$ and have to be chosen by the user. In this work, we propose to solve the transfer learning problem by choosing a prior depending on the source model parameters.

Section 2 shows the material and the developed methods to tackle the problem.

Section 3 shows the obtained results.

Section 4 concludes the work.

2 Material and Methods

2.1 Data Presentation

In this paper, we focus on the hydrodenitrogenation (HDN) modeling. For our application, two datasets are available. The first data set corresponds to the previous catalyst generation (n). The second data set corresponds to the current catalyst generation (n+1). The aim is to fit the denitrogenation model of catalyst(n+1).

The dataset from the pilot plants using a catalyst (n) is called the source dataset. A first sorting is done in order to eliminate observations with values outside predefined intervals.

In addition, an outlier detection is carried out using the Local Outlier Factor method.[13] The source dataset is then composed of 61 observations and will be used as a basis to obtain the prior knowledge.

The dataset from the pilot plants using a new catalyst (n+1) is called the target dataset. It is composed of 126 observations. The same methodology is applied to remove outliers.

The variable of interest is the nitrogen content (N) after the hydrotreating stage (Nslip). In order to model its value, an ODE-based kinetic model with a structure showing to be efficient for the modeling of hydrodenitrogenation[14] is used and the influential features are presented in Table 1. This model is very simple but enough to show the methodology.

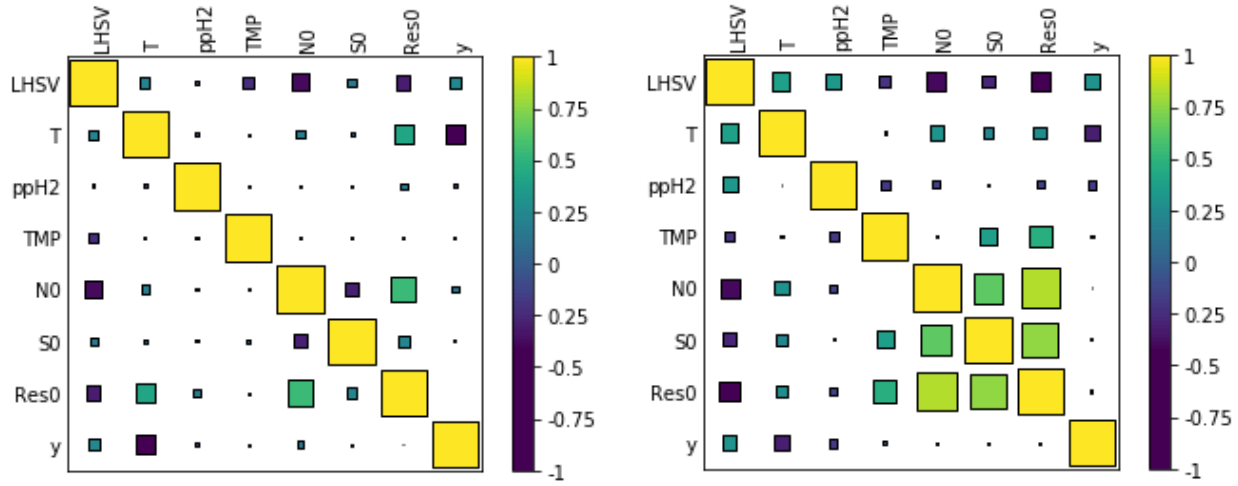
These features are all quantitative variables, including the output to be predicted. The features space is the same for both datasets, and the variability is similar for the source and the target dataset.

Table 1: Features description Range.

Feature	**Description**	**Range**
$LHSV$	Liquid Hourly Space Velocity: the ratio of liquid volume flow per hour to reactor volume (in h^{-1}). It is the inverse of the residence time.	[0.5;4 h-1]
T	Temperature of the hydrotreating reactor (in °C).	[350;410 °C]
ppH_2	Hydrogen partial pressure (in bar).	[80;160 bar]
TMP	Weighted average of the simulated distillation (in °C): $TMP = \frac{1}{7}$(FEED_SimDis05 + 2 × FEED_SimDis50 + 4 × FEED_SimDis95)	[400;650 °C]
N_0 (FEED_NIT)	Nitrogen content in feedstock (in ppm).	[800;4000 ppm]
S_0 (FEED_SULF)	Sulfur content in feedstock (in mass percent).	[0.5;3 %m/m]
Res_0	Resines content in feedstock (in mass percent).	[4;15 %m/m]
N (to be predicted)	Nitrogen content after hydrotreating (in ppm). The variable of interest that we want to model using other features.	[3;400ppm]

The experimentation to obtain the observations were carried out using 32 different feed-

stocks for the source dataset and 20 for the target dataset. The features correlations are quite similar for both datasets (Figure 1). The higher correlation between the feedstock features for the target dataset than for the source dataset shows that the target feedstocks, in addition to being fewer in number, are more similar to each other. Other features have low correlations, with an absolute value of less than 0.4, and are comparable for both datasets.



(a) source dataset

(b) target dataset

Figure 1: Correlation matrix for hydrotreating source and target datasets. The correlation coefficient use is the Pearson's product-moment coefficient. The size and color of the squares both represent the value of the correlation coefficient. The larger the square the higher the absolute value of the coefficient.

2.2 ODE-based kinetic modeling

An ODE-based kinetic model is considered and the model structure is obtained by solving the following differential equation:

$$\begin{aligned} \frac{dy}{dt} &= -k_0 \frac{exp(-\frac{E_a}{R_g}(\frac{1}{T}-\frac{1}{Tref}))(\frac{ppH_2}{ppH_{2,ref}})^m y^n}{(1+A_0 Res_0)(1+\frac{C_0 N_0}{1+S_0})} \times \\ & (1-u\cdot exp(-\frac{b}{R_g}(\frac{1}{T}-\frac{1}{Tref}))(\frac{ppH_2}{ppH_{2,ref}})^a(\frac{WADT}{WADT_{ref}})^v y^r), \end{aligned} \tag{2}$$

where k_0 and u are the kinetic pre-exponential rate constants, E_a and b are the Activation Energies (cal) and R_g is the universal gas constant (cal/mol/cal). In our case, y stands for nitrogen (ppm), t for the residence time (h), T for the temperature (K), ppH_2 for the partial pressure of Hydrogen, Res_0 for the Feed Resin (% m/m), N_0 for the Feed Nitrogen (ppm), S_0 for the Feed Sulfur (% m/m) and $WADT$ for the weighted average distillation temperature of the feedstock (K).

This rate equation can be decomposed into two terms. The first one corresponds to the removal of nitrogen-containing compounds whereas the second term takes into account the reverse reactions that might occur. Indeed, in some conditions, the aromatics hydrogenation thermodynamic equilibrium is reached. This leads to an inversion of the HDN reactions which involve a hydrogenation step.

In this model, the direct kinetic term is composed of a pre-exponential rate constant k_0, an Arrhenius-type term ($E_a/R_g/T$), a power law associated to hydrogen (m is the partial reaction order) and to nitrogen (n is the partial reaction order) and two inhibitors, modeled as Langmuir-type adsorption curves: resins (Res_0) and nitrogen over sulfur content in the feedstock $\frac{N_0}{1+S_0}$. A_0 and C_0 are the inhibition adsorption constants to be fitted.

For the second term, the inverse kinetic term is composed of a pre-exponential rate constant u, an Arrhenius-type term ($b/R_g/T$), a power law associated to hydrogen (a is the partial reaction order) and to nitrogen (r is the partial reaction order). An empirical term, related to the feedstock composition ($WADT$) was also added to the return term using a power law: v.

The value of the reference terms are displayed in Table 2.

For a given value of the parameters, the resolution of the differential equation is done numerically using a method of backward differentiation formula. Then, for each observation $(\boldsymbol{x}_i)_i$, a numeric estimation of the nitrogen content evolution over time is obtained. The residence time of the feed in the reactor being known ($LHSV^{-1}$), we obtain the prediction of the nitrogen at the reactor outlet, which we note $f_\theta(\boldsymbol{x}_i)$ for an observation i.

Table 2: References values of the kinetic model.

Symbol	Value
R_g	1.987215583 cal/mol/K
T_{ref}	649.15 K
$ppH_{2,ref}$	32.5 bar
$WADT_{ref}$	643.15 K

The objective is to optimize parameters $\boldsymbol{\theta} = (k_0, E_a, m, n, a, b, A_0, C_0, u, r, v)$ such that the cost function $\sum_{i=1}^{K} \frac{(f_\theta(\boldsymbol{x}_i)-y_i)^2}{y_i}$, with K the sample size, is minimal. Moreover, boundaries are fixed for the parameters in order to keep a physical sense (Table 3).

Table 3: Boundaries of kinetic model parameters.

Param	k_0	E_a	m	n	a	b	A_0	C_0	u	r	v
Min	0	$1 \cdot 10^4$	0,3	0,3	-10	$-4 \cdot 10^4$	0	-5	0	-10	-10
Max	10^3	$8 \cdot 10^4$	10	10	0	0	10	5	3	10	10

Minimize the cost function in $\boldsymbol{\theta}$ is equivalent to maximize the likelihood of the statistical model (3):

$$\begin{aligned} y_i &= f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + \epsilon_i, \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_i^2), \\ \sigma_i^2 &= \sigma \cdot y_i. \end{aligned} \tag{3}$$

Indeed, by noting $\mathbf{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_{K_t}^2)$, we obtain:

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

Where $\boldsymbol{\theta}_{ML}$ is the maximum likelihood estimator of (3).

2.3 Model fitting

To perform the Bayesian inference to fit the target model, we need to choose the prior distribution. The first step is therefore to fit a performing model on the source dataset to extract the knowledge we have from it. The model (3) is considered. In order to have an estimation of the parameters distribution, the optimisation of the parameters is done using a MCMC algorithm. The MCMC algorithm used is a Metropolis Hastings within Gibbs algorithm[15]. A flat prior is considered (constant over the parameter space), so no knowledge on the prior distribution of the parameters is brought, and thus the likelihood is optimized. All available source observations are used to fit and evaluate the model. The MCMC algorithm is used (10000 iterations). After removing a burn-in period of 2000 observations, the estimation of $\hat{\boldsymbol{\theta}}_s$ and $Var(\hat{\boldsymbol{\theta}}_s)$ are obtained respectively as the mean and the covariance over the remaining 8000 iterations.

The structure for the target kinetic model is the same as for the source model (equation (3)). The fitting is done by using the Bayesian transfer approach, by transferring the parameters knowledge using the prior. To build the prior distribution, it is assumed that the

target parameters $\boldsymbol{\theta}_t$ follow a multivariate Gaussian distribution:

$$\pi(\boldsymbol{\theta}_t) \quad \sim \quad \mathcal{N}(\hat{\boldsymbol{\theta}}_s, gVar(\hat{\boldsymbol{\theta}}_s)), \tag{4}$$

where $\hat{\boldsymbol{\theta}}_s$ are the estimated source parameters.

Due to the structure of the model, the variances of the parameters are really distinct from each other, which leads to a badly scaled covariance matrix. In order to avoid numerical issues, as with the inversion of the covariance matrix, parameters are normalized. The ODE is thus rewritten, modifying each θ_i by $\sqrt{Var(\hat{\theta}_{i,s})}\theta_i$, with $\boldsymbol{\theta} = (\theta_i)_{i=0,...,10} = (k_0, E_a, m, n, a, b, A_0, C_0, u, t, v)$ and $Var(\hat{\theta}_{i,s})$ the i^{th} diagonal element of $Var(\hat{\boldsymbol{\theta}}_s)$. It should be noted that this modification has no impact on the model and only affects the values taken by the parameters. In the following, this structure for the model is considered and by doing so, the diagonal of $Var(\hat{\boldsymbol{\theta}}_s)$ is composed of ones.

The main challenge with the Bayesian transfer approach is to choose a suitable g value to adapt the prior's impact. It should be noticed that $\hat{\boldsymbol{\theta}}_t \underset{g\to 0}{\to} \hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_t \underset{g\to+\infty}{\to} \hat{\boldsymbol{\theta}}_{t,ML}$, where $\hat{\boldsymbol{\theta}}_{t,ML}$ is the maximum likelihood estimator on the target sample only, without transfer. The heuristic method to chose the g value proposed for Bayesian kriging transfer[11] is not adapted for this case. Indeed, the model is more complex and we cannot know the range in which parameters will lead to a high value of the likelihood. Two effective methods to choose it are then proposed.

Selection of the value of g by cross-validation The first method is to use leave-one-out cross-validation to determine which value of g offers the best score. For a given sample of size n_t and for each g value tested, cross-validation is performed on the training set (the given sample). n_t model are thus fitted using $n_{sample} - 1$ observations and the score is evaluated on the remaining observations ($\frac{(\hat{y}_i - y_i)^2}{y_i}$ where i is the observation not used to fit the model). The mean of the n_t test scores (scores on remaining observations) is calculated. Finally, the value of g with the lowest averaged score is kept. The drawback of this approach is that

several models have to be fitted, thus if the model is complex and the sample size is quite large, it becomes time consuming.

Selection of the value of g by "bound on training score" The second approach to chose the value of g, which we call "bound on training score", is based on the score obtained for the different g values on the training dataset. This method relies on the score we wish to obtain on the whole dataset. With g near to 0, the model is close to the source model and the score is far from expectation. With an high value of g, the model moves closer to the target model without transfer and the score monotonically reaches the score without transfer as the value of g increases. The aim is to keep a maximum of prior information without getting a model leading to bad prediction on the training set. Then, we look for the smallest g value among the tested value $\mathcal{G}$ for which the training score is lower than a fixed value for the expectation score:

$$g_{selected} = \begin{cases} \underset{g \in \mathcal{G}}{min}(g) \\ \sum_{i \in Train} \frac{(\hat{y}_i^g - y_i)^2}{y_i} \leq S_{exp} \end{cases}$$

where S_{exp} is the chosen expectation score and $\hat{y}_i^g$ is the prediction obtained using $gVar(\hat{\boldsymbol{\theta}}_s)$ as prior variance. In that way, it is ensured that the prior information does not degrade the model quality, and maximum knowledge from the prior is kept. This method implies the knowledge of the performance expectation, that is the score expected on the target dataset with a good model. If we assumed that the target model will lead to similar performances as the source model, the score obtained with the source model on the source dataset can be a good estimation. For these reason, it was used for this particular application.

Once the prior is chosen, the target Bayesian model can be fitted. In order to compare results with the current method, we also fit a model "without transfer". To optimize the parameters of the Bayesian transferred model, we first use a MCMC algorithm starting from the prior mean, which is the source parameters estimation. However, the posterior

distribution is non Gaussian, with multiple modality, and a good estimation of the posterior is not attained with such an approach because the Markov Chain remains blocked in modalities close to the initialization.

Therefore, another approach is considered to optimize the parameters values. A gradient descent is performed, using the L-BFGS-B algorithm[16], in order to find parameters $\hat{\boldsymbol{\theta}}_t$ that maximize the posterior (Equation (5)). By doing so, we do not obtain an estimate of the posterior distribution, but an estimate of the posterior maximum, sufficient to fit the model parameters.

$$\pi(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y}) \quad \propto \quad p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{5}$$

where $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})$, $\boldsymbol{\Sigma} = diag(\sigma_1^2,\ldots,\sigma_{K_t}^2)$ with $\sigma_i^2 = \sigma_t \cdot max(5, y_i)$ and σ_t supposed to be known and equal to $\hat{\sigma}_s$.

Maximizing the posterior corresponds to solve the Equation (6):

$$\begin{aligned}\hat{\boldsymbol{\theta}}_t &= \arg\max_{\boldsymbol{\theta}} \exp(-\frac{1}{2}(\boldsymbol{y}-f_{\boldsymbol{\theta}}(\boldsymbol{X}))^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-f_{\boldsymbol{\theta}}(\boldsymbol{X})))exp(-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)^T gVar(\hat{\boldsymbol{\theta}}_s)^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)) \\ &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2}(\boldsymbol{y}-f_{\boldsymbol{\theta}}(\boldsymbol{X}))^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-f_{\boldsymbol{\theta}}(\boldsymbol{X})) + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)^T gVar(\hat{\boldsymbol{\theta}}_s)^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_s)\end{aligned} \tag{6}$$

To ensure that the global minimum (or at least a good local minimum) is found, several initializations are considered. 100 distinct initializations are randomly sampled using the distribution $\mathcal{N}(\hat{\boldsymbol{\theta}}_s, 10 \times gVar(\hat{\boldsymbol{\theta}}_s))$.

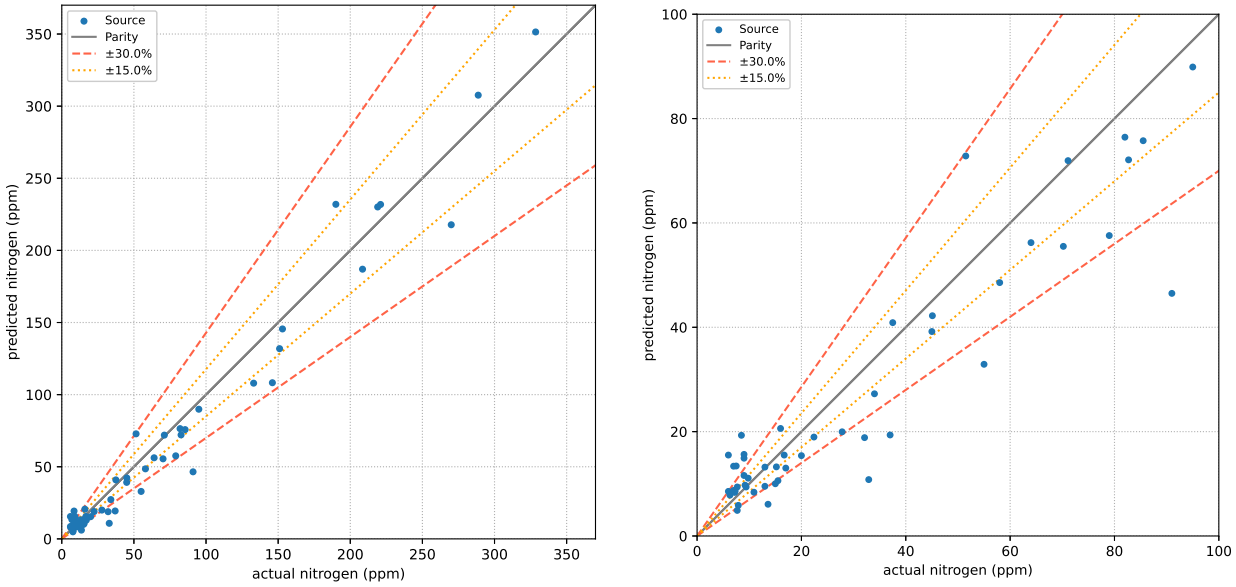
For the sake of calculation time, the same initializations are used for the "without transfer" approach. As a consequence, we do not cover the space of possible values for the parameters. Thus we indirectly constrain the "without transfer" model parameters to stay close to the estimated source parameters values. It is possible that a combination of parameters far from the initialization leads to an higher likelihood value. This option was not explored. This would mean a better "without transfer" model on the training set, but it does not

imply a better model for the test set, particularly with a small training set. Actually, it provides beneficial information here. With few observations and initialization far from source parameters estimation, we can build a "without transfer" model that perform better on the training set but is worse on the test set compared to the "without transfer" model we use. The comparison between the Bayesian transferred and the classical fitted model therefore remains relevant.

3 Results and Discussion

3.1 Source modelling

Firstly, the model fitted on the source dataset is studied. The parity plots of the results are presented in Figure 2. The fitted model offers satisfying results.



(a) $N \in [0, 400]$ (b) $N \in [0, 100]$

Figure 2: Parity plot of the fitted source model with 61 observations. The dashed lines represent the intervals for which the prediction error is less than 15% and 30% of the actual value. On the right, a zoom for nitrogen values below 100 ppm is drawn.

An additional criterion used to evaluate the quality of the model is the Delta T error (Figure 3). It represents the difference in temperature of the hydrotreating reactor T needed to

obtain the experimental value of the nitrogen content (all the other inputs remain constant):

$$f_{\hat{\boldsymbol{\theta}}_s}(\boldsymbol{x}_i + (DeltaT_i, 0, 0, 0, 0, 0, 0)) = y_i, \tag{7}$$

with the first element of $\boldsymbol{x}_i$ being the temperature of the reactor. For example, a Delta T of 5°C means that with a value of reactor temperature 5°C higher than the actual value, the nitrogen content estimation will be the actual value. It is an interesting measure because the reactor temperature furnace is the only operating condition that can easily be changed by the refiner. A similar conclusion to the analysis of parity graphs is reached: satisfying results are obtained.

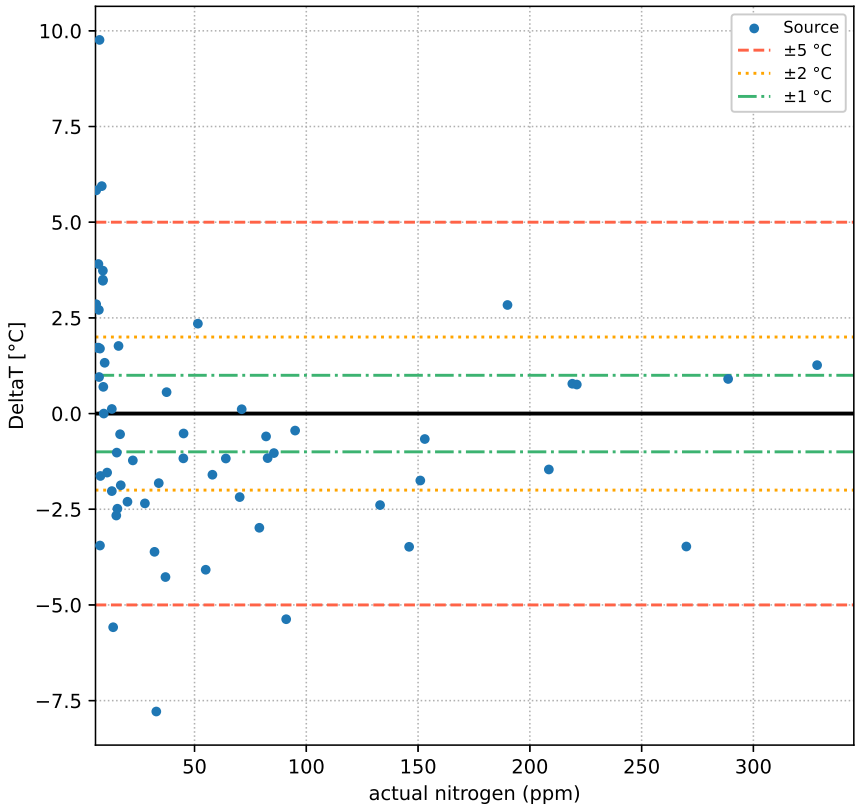


Figure 3: Delta T for the fitted source model with the 61 observations. The dashed lines represent the intervals for which the delta T is less than 5 and 2 and 1°C.

The quantified summary of the performances obtained with the source model is as follows:

- $\sum_{i=1}^{61} \frac{(f_\theta(\boldsymbol{x}_i) - y_i)^2}{y_i} = 2.34,$

- 90% of the observations have an absolute Delta T lower than 5°C,
- 54% of the observations have an absolute Delta T lower than 2°C.

The model obtained is considered of satisfying quality and it will be considered as the model to transfer to the target dataset.

Since it is not possible to plot the corresponding 12-dimensional Gaussian prior, Figure 4 shows the marginal histograms and the marginal distribution of each dimension of $\hat{\boldsymbol{\theta}}_s$. The correlation between source parameters over the 8000 iterations is also plotted (Figure 5). It can be noticed that for some parameters the Gaussian distribution does not represent well the stationary distribution of the Markov Chain (parameters a, u and t for example) and another prior might be more appropriate for these parameters. Furthermore, due to the bounds, for some parameters there is a spike on the bar graph for the boundary values (parameters m, a, b and v) and then the Gaussian prior does not fit perfectly once again. However, the choice to maintain this Gaussian prior is made to conserve the covariance information between all parameters.

Before fitting a new target model, we are interested in the results of the source model applied to the target data. The results obtained are, as expected, not good and the Nitrogen prediction is higher than the Nitrogen experiments because the catalyst (n+1) is more active than the catalyst (n) (Figure 6).

Similar performance is obtained with a Delta T (defined in equation 7) higher than 10 degrees Celsius (Figure 7).

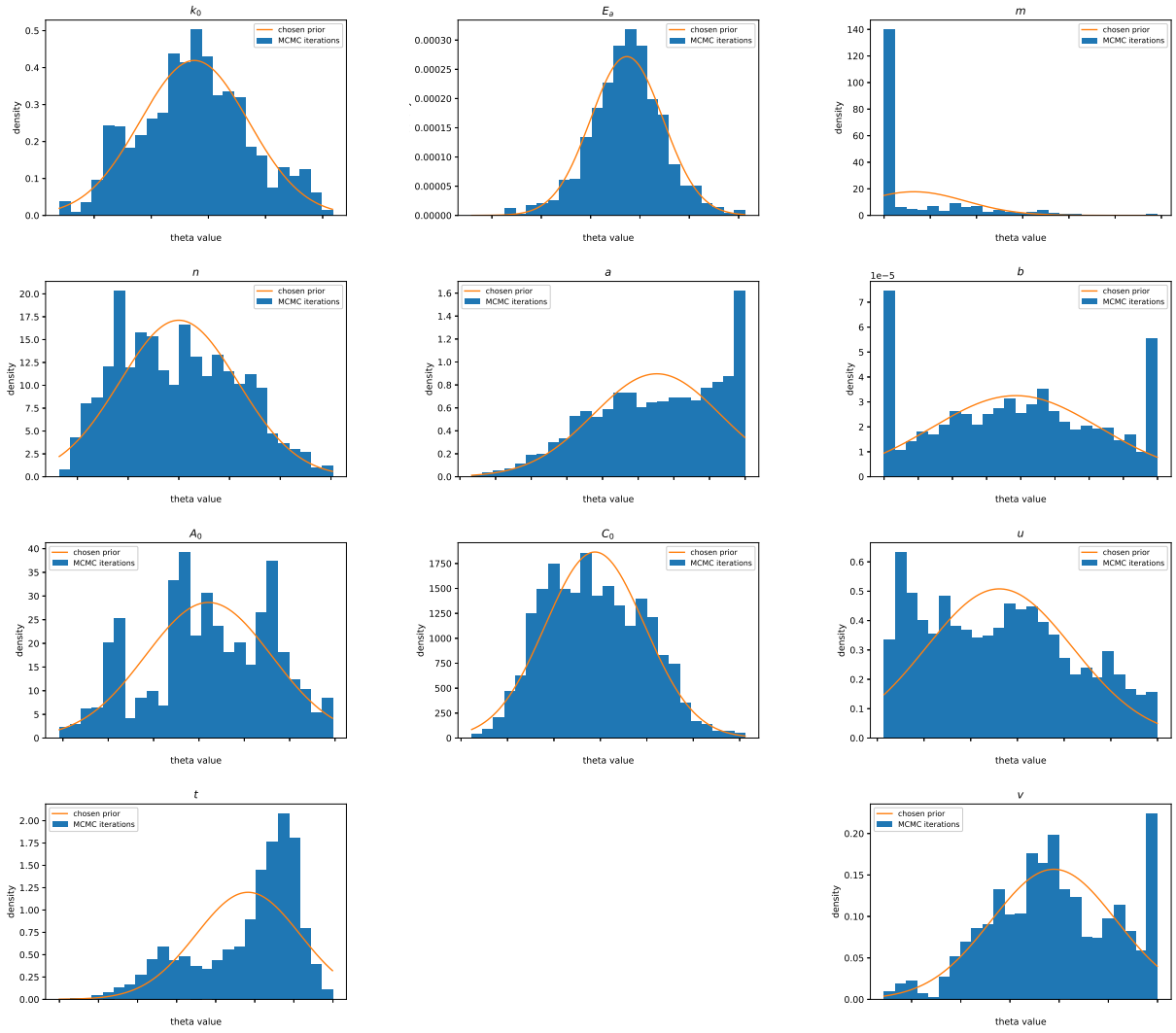


Figure 4: Univariate chosen prior and MCMC sample for the different parameters. Values are hidden for confidentiality reasons.

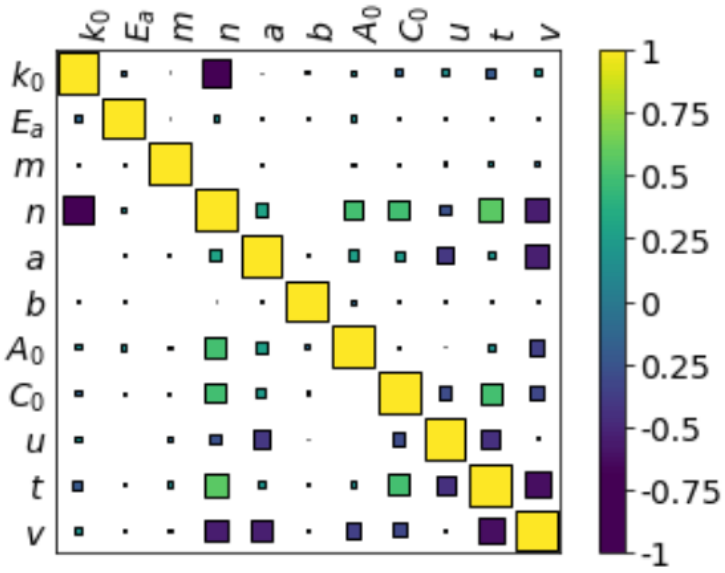


Figure 5: Correlation between source parameters over the 8000 iterations of the MCMC algorithm, burn in period being removed.

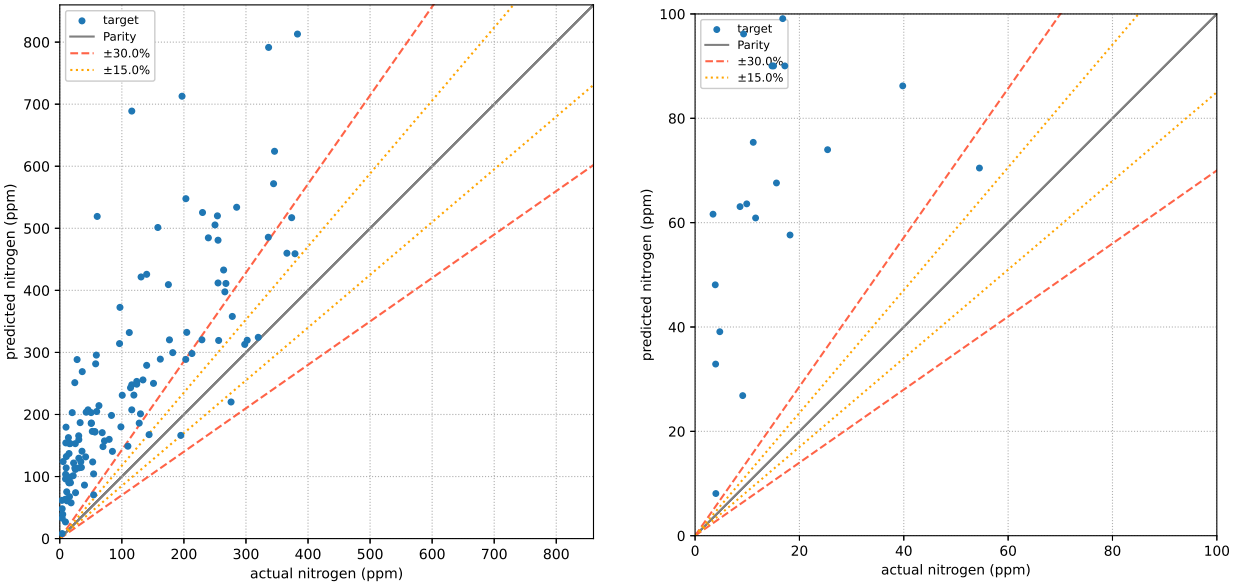


Figure 6: Parity plot of source model applied to target dataset. The dashed lines represent the intervals for which the prediction error is less than 15% and 30% of the actual value. On the right, a zoom for nitrogen values below 100 ppm is drawn.

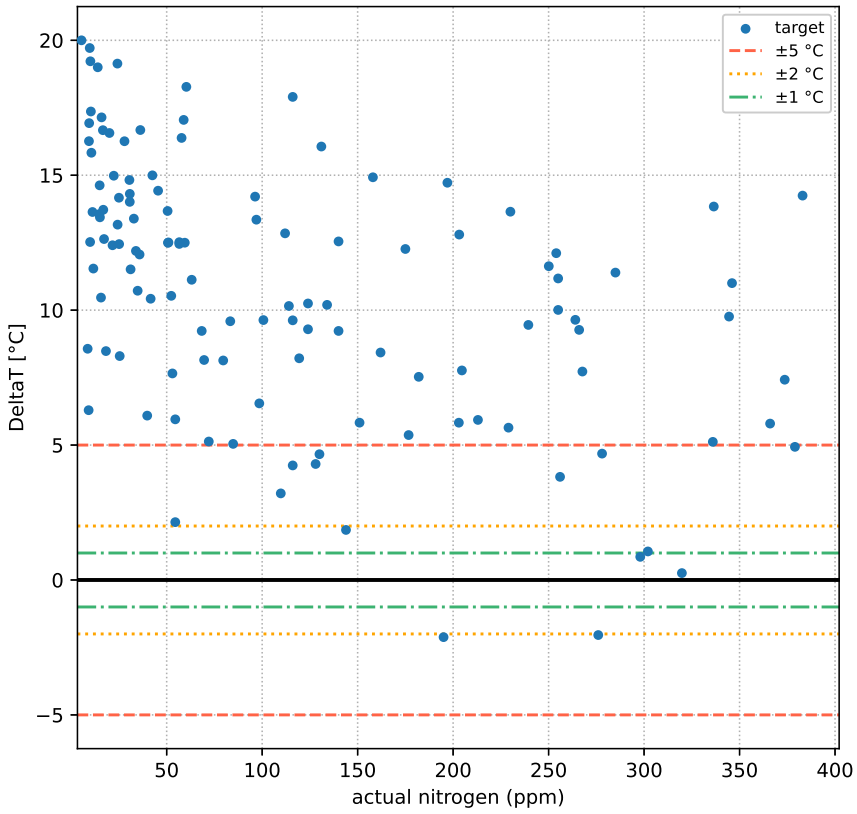


Figure 7: Delta Temperature for the source model applied to the target dataset. The dashed lines represent the intervals for which the delta T is less than 1 and 2 and 5°C.

3.2 Target modelling

Different target samples of different sizes are considered, randomly sampled, in order to evaluate the Bayesian transfer, to compare it to the "without transfer" approach, and to study the impact of the scalar g. For each sample size $n_t \in \{5, 10, 15, 20\}$, 10 samples are drawn. A "without transfer" model and Bayesian transfer models with different value for g ($g \in \{1, 10, 100, 1000, 10000\}$) are fitted for each sample.

Figure 8 illustrate the impact of the value of g. It shows, for a sample size $n_t = 15$, the averaged score obtained over the 10 samples, with different g value for the Bayesian transfer, on the training and on the test set (continuous line). The dashed lines are the score with a non transfer model fitted on the training data set (blue) and tested on the test set (orange). Without transfer, the training score is very low (around 1.6) but the test score is high (around 3.4) because no a priori information is added to the fit.

Due to the Bayesian definition, with an high g value the prior is neglected and the score obtained with the Bayesian transfer and that obtained without transfer are similar. This is the case on both the training set (the size 15 target sample in the example of Figure 8) and the test set (the entire target dataset) (see for example $g = 10^4$). For the training set, when the value of g decreases, the training score monotonously increases to tend to the score using source model when g is near to zero. Concerning the test set, the evolution of the score as a function of g is not monotonous. A well chosen g value increases model quality and conversely a badly chosen g value decreases it. It is thus crucial to define a method to chose a good g value which is a trade off between the train and the test score. In our case, a g value around 10 is optimal.

To deepen the example of this sample, the parity plots for the "without transfer model" and the Bayesian models with a bad value of g and a well chosen g value are plotted (Figure 9).

As 15 observations are sufficient to fit a satisfying model without transfer for this application, similar results are obtained with the Bayesian transfer method as shown in Figure

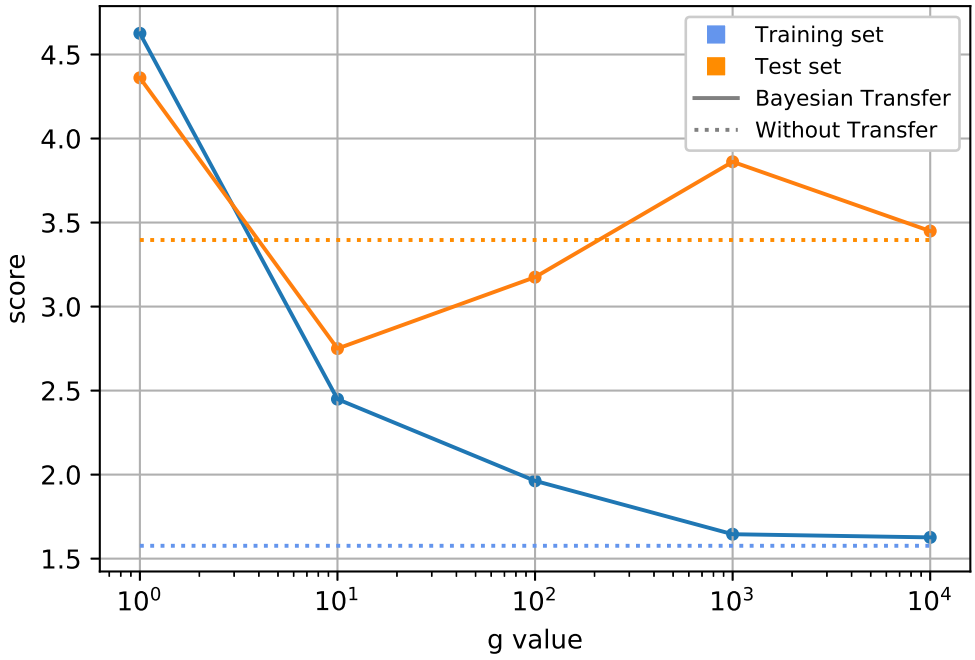


Figure 8: Example of g value impact for a random sample of size 15 (average). Blue continuous line: Training set score with transfer, Orange continuous line: Test set score with transfer, Blue dashed line: Training set score without transfer, Orange dashed line: Test set score without transfer

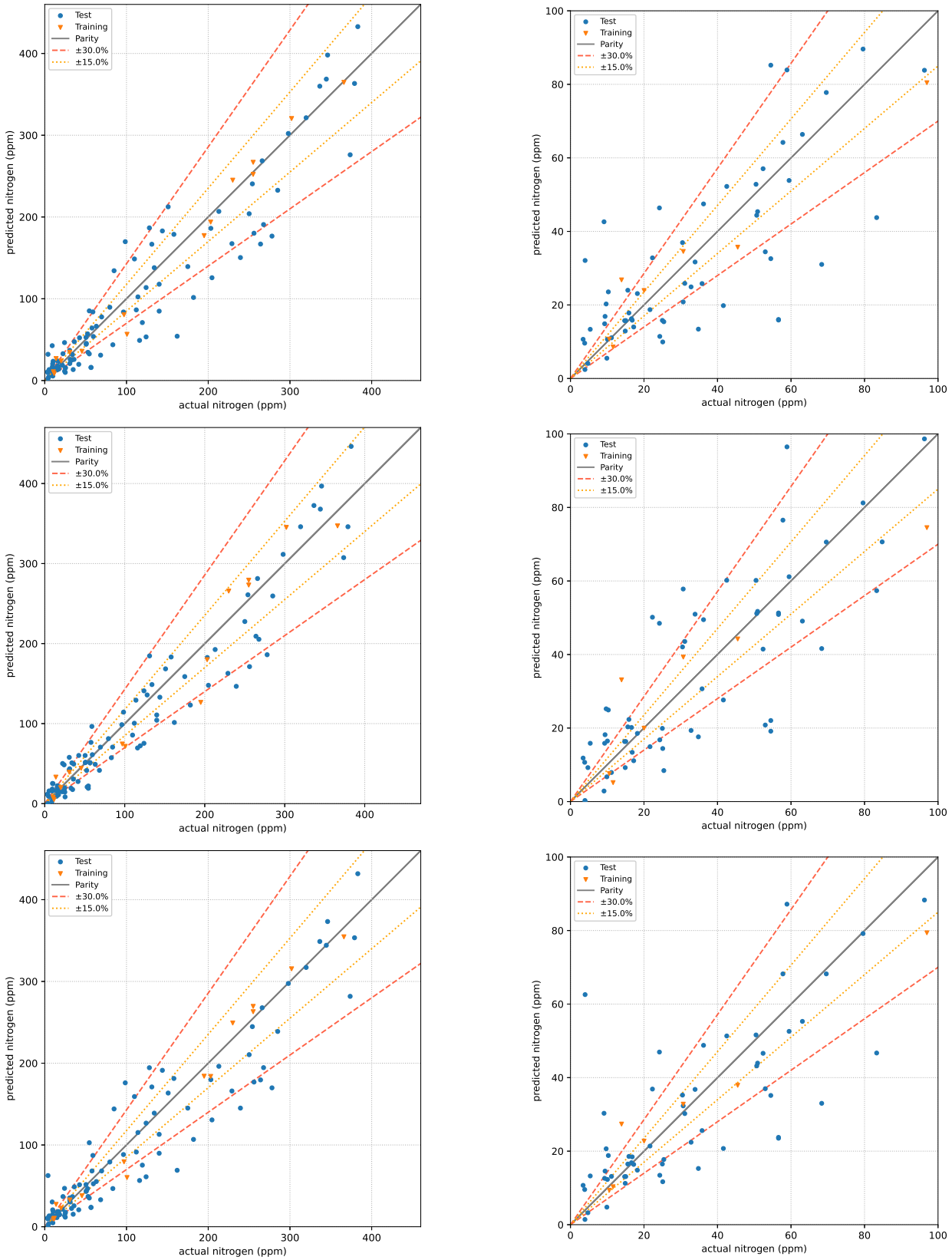


Figure 9: Parity plot for a random sample of size 15. On the right graphs, a zoom is performed for the range [0ppm;100ppm]. Top: Without transfer, Center: Bayesian transfer with g=10 (good choice), Bottom: Bayesian transfer with g=1000 (bad choice).

9. Nevertheless, a slight improvement is obtained with Bayesian transfer with a good choice for the g value (center graphs in Figure 9). If less observations are available from the target dataset, the classical approach can lead to really bad models. For example, the size 10 random sample presented in Figure 10 leads to poor results with the classical approach, but by using the Bayesian transfer model with a well chosen g value, the results are satisfying and thus a great improvement is provided.

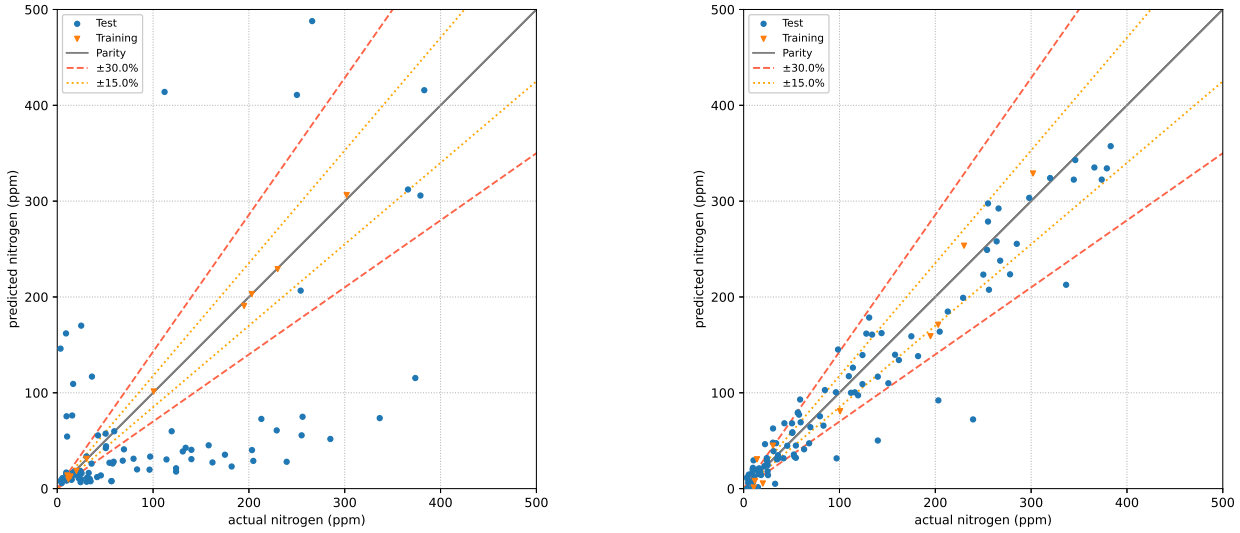


Figure 10: Parity plot for a random sample of size 10. Left: Without transfer, Right: Bayesian transfer with g=10.

We see that the Bayesian transfer, with a correct choice of the g value, leads to an improvement of the model's prediction performance. Thus, two methods are proposed to define the optimal g value to chose.

Selection of the g value by cross-validation The first method is to use leave-one-out cross-validation to determine which value of g offers the best score.

In order to compare the Bayesian transfer models with the classical fitted models, this method is tested on the different sample sizes and the averaged scores and the minimum-maximum score interval (the minimum and the maximum score obtained over the 10 random samples) are studied for each sample size (Figure 11). The method leads to an improvement

of the score with the Bayesian transfer, especially for small designs, and thus to a good choice for the g value. Moreover, the Bayesian transfer model is less affected by the design quality, as it leads to a smallest maximum score even if the sample size is low. This method is then more robust to a bad design choice than classical method. It shows that using only 10 pilot plant points is enough to refit the HDN model.

However, the main drawback of this method is that several models have to be fitted, thus if the model is complex and the sample size is quite large, it may become time consuming.

Selection of the g value by "bound on training score" The second approach to chose the value of g, which we call "bound on training score", is based on the score obtained for the different g values on the training dataset.

This method is tested on the different samples (Figure 12). As for the first method, it leads to a good choice for the g value and thus to an improvement of the score with the Bayesian transfer and a smallest maximum score. The scores are slightly less good than those with the cross validation method (Figure 13), but has the advantage of being less time consuming. Again only 10 pilot points are requested to fit the kinetic model.

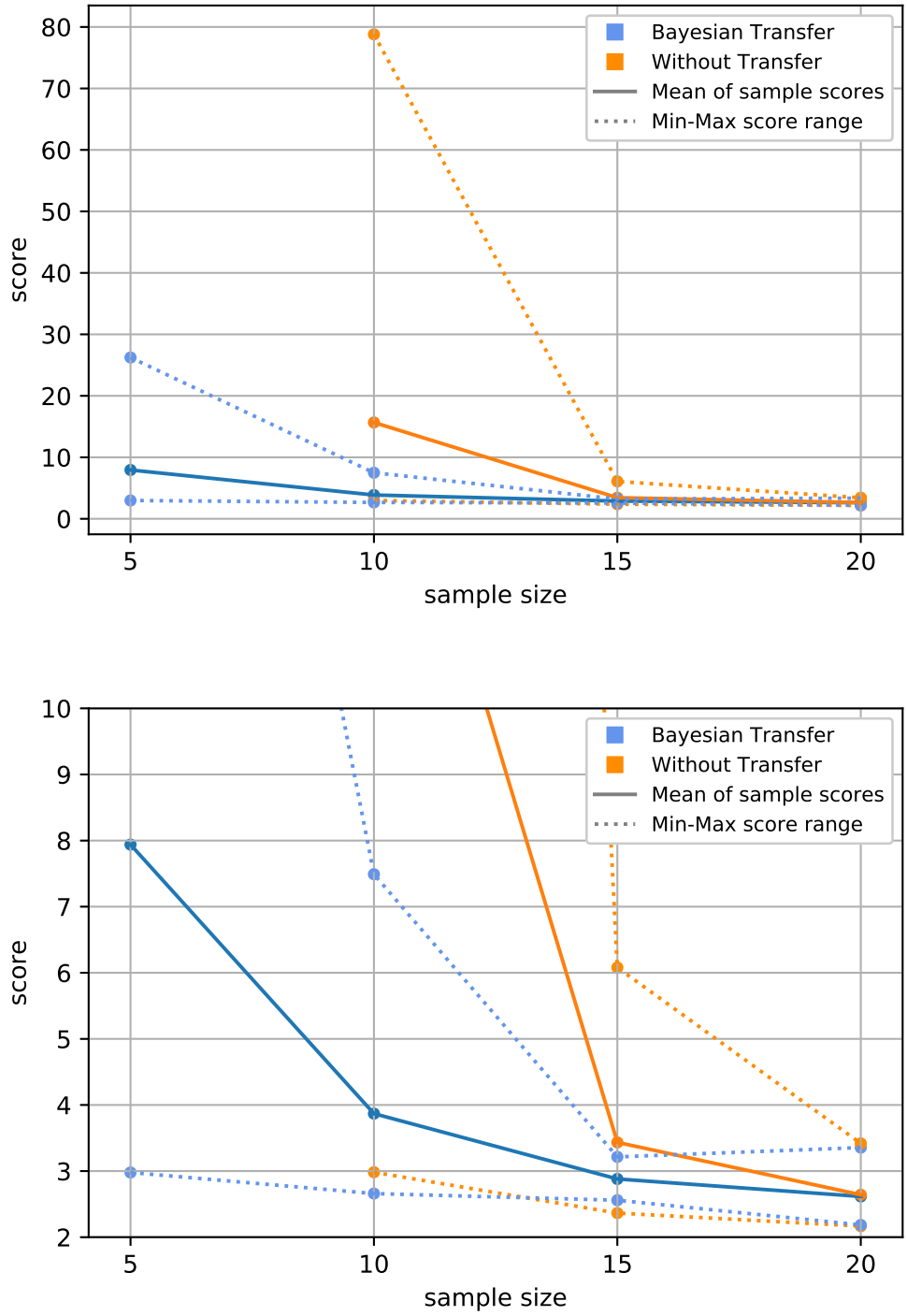


Figure 11: Comparison of classical fitted model and Bayesian model, g chosen with the cross validation method, for different size of the training sample. The score $\sum_{i=1}^{K} \frac{(\hat{y}_i - y_i)^2}{max(5, y_i)}$ evolution is plot according to training sample size. The mean and the minimum-maximum score range over the 10 samples are plotted. On the bottom, a zoom is applied on the score. Blue continuous line: average score with Bayesian Transfer, Orange continuous line: average score without transfer, Blue dashed line: Min/Max score with transfer, Orange dashed line: Min/Max score without transfer

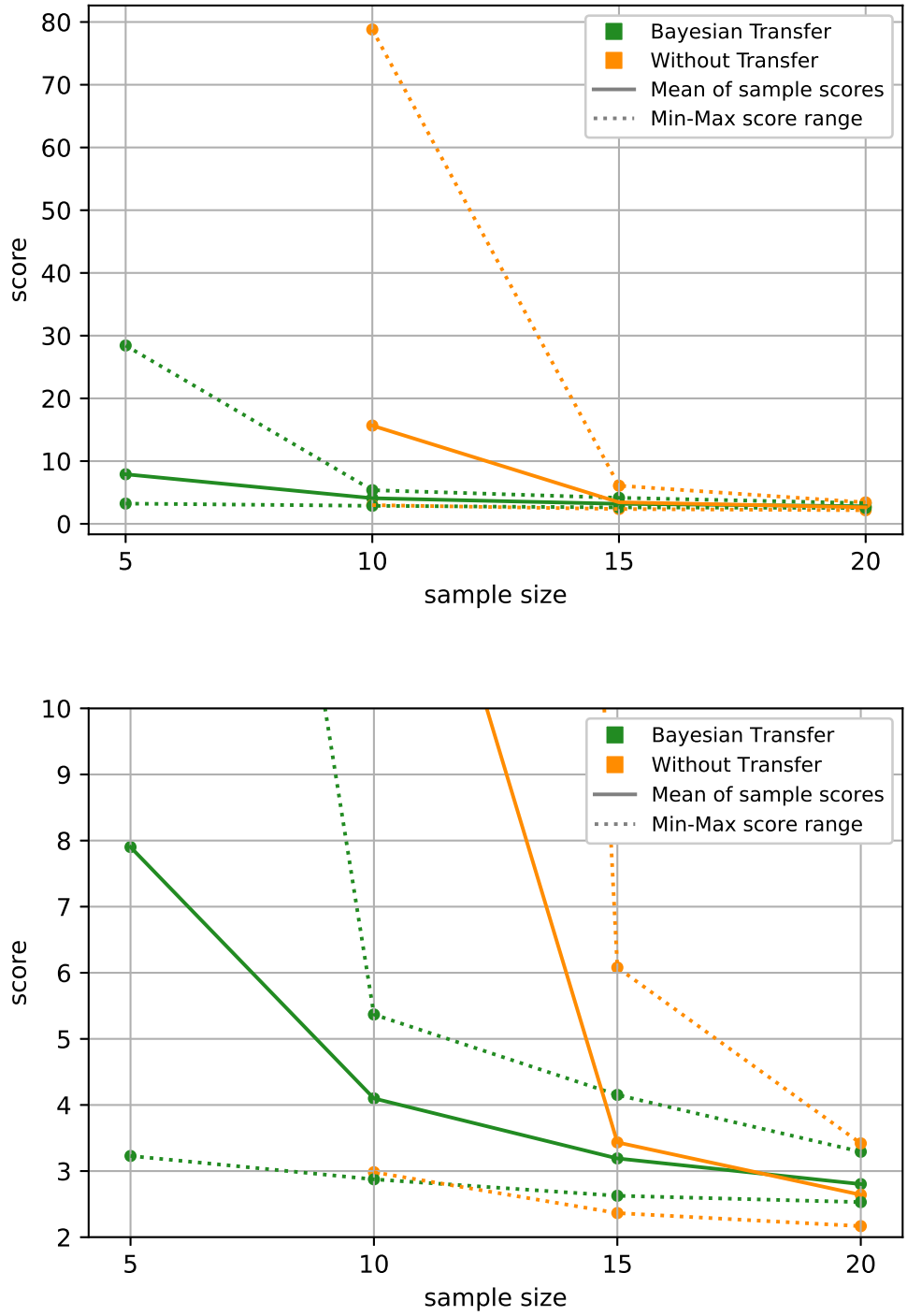


Figure 12: Comparison of classical fitted model and Bayesian model, g chosen with the bound on training score method, for different size of the training sample. The score $\sum_{i=1}^{K} \frac{(\hat{y}_i - y_i)^2}{max(5, y_i)}$ evolution is plot according to training sample size. The mean and the minimum-maximum score range over the 10 samples are plotted. On the bottom, a zoom is applied on the score. Green continuous line: average score with Bayesian Transfer, Orange continuous line: average score without transfer, Green dashed line: Min/Max score with transfer, Orange dashed line: Min/Max score without transfer.

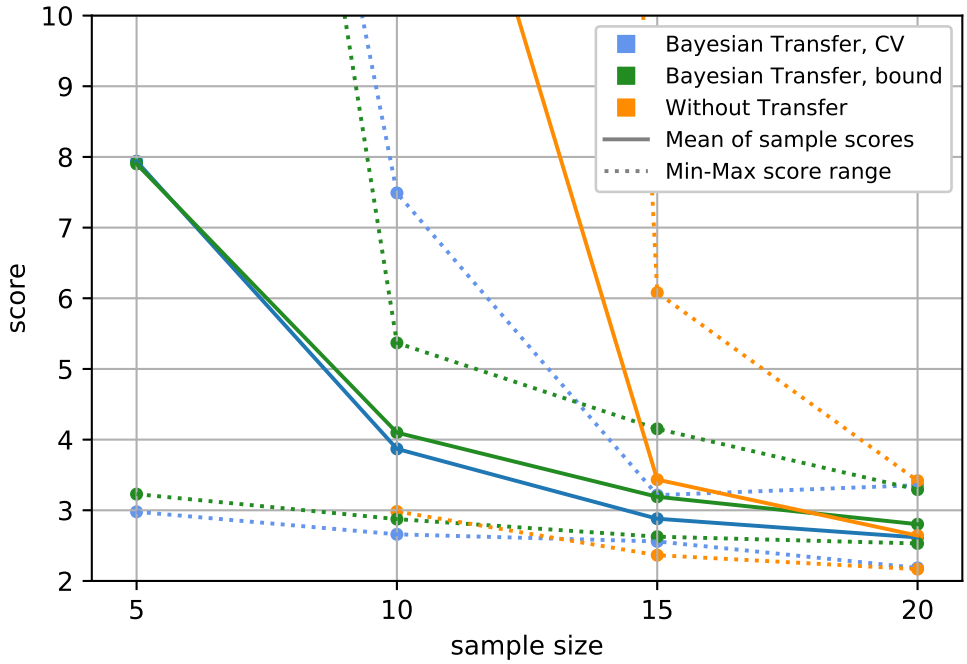


Figure 13: Comparison of the score evolution for the different methods. Mean, minimum and maximum scores of the 10 samples are plot. Blue continuous line: average score with Bayesian Transfer (Cross Validation), Green continuous line: average score with Bayesian Transfer (Bound), Orange continuous line : average score without transfer, Blue dashed line : Min/Max score with transfer (Cross Validation), Green dashed line : Min/Max score with transfer (Bound), Orange dashed line : Min/Max score without transfer

4 Conclusion

In this paper, a methodology to fit a kinetic model for the hydrotreating process for the new catalyst generation (n+1) catalyst using few observations is proposed. To add some information and then to decrease the number of requested pilot plant points, data of previous catalyst generation (n) are used. A Bayesian transfer method, which leads to more robust models, is proposed. It is less impacted by the design quality when compared to the classical model. Also, the predictions performance is improved, especially for small designs where it could not be possible to fit a satisfying model without such an approach. In our case, only 10 pilot points are enough to refit a HDN kinetic model. To obtain a good predictive model, a good choice of the prior variance and thus of the g value is crucial. Two methods for choosing it efficiently are proposed and tested. The cross-validation method is recommended for the best performance, but it may be time consuming depending on the size of the training set. The "bound on training score" method is a good alternative. This work is performed on a particular application but the transfer method developed can be used for any kinetic or any parametric model. It thus can be used in many other areas, enabling the valorization of past knowledge for many applications.

References

(1) Ancheyta, J.; Sánchez, S.; Rodríguez, M. A. Kinetic modeling of hydrocracking of heavy oil fractions: A review. *Catalysis Today* **2005**, 76–92.

(2) de Oliveira, L. P.; Hudebine, D.; Guillaume, D.; Verstraete, J. J. A review of kinetic modeling methodologies for complex processes. *Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles* **2016**, *71*, 45.

(3) Becker, P. J.; Celse, B.; Guillaume, D.; Costa, V.; Bertier, L.; Guillon, E.; Pirngru-

ber, G. Hydrotreatment modeling for a variety of VGO feedstocks: A continuous lumping approach. *Fuel* **2015**, *139*, 133–143.

(4) Becker, P. J.; Celse, B.; Guillaume, D.; Costa, V.; Bertier, L.; Guillon, E.; Pirngruber, G. A continuous lumping model for hydrocracking on a zeolite catalysts: model development and parameter identification. *Fuel* **2016**, 73–82.

(5) Sánchez, S.; Rodríguez, M. A.; Ancheyta, J. Kinetic model for moderate hydrocracking of heavy oils. *Industrial & engineering chemistry research* **2005**, *44*, 9409–9413.

(6) Sadighi, S.; Ahmad, A.; Mohaddecy, S. R. S. 6-Lump kinetic model for a commercial vacuum gas oil hydrocracker. *International Journal of Chemical Reactor Engineering* **2010**, *8*.

(7) Davy, L.; Castro, J.; Wessels, R.; Rey-Bayle, M.; Merdrignac, I.; Celse, B. Global Methodology for Catalyst Screening and Optimization Process. Application to Mild Hydrocracking. *Industrial & Engineering Chemistry Research* **2020**, *59*, 21133–21143.

(8) Pan, S.; Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **2010**, 1345–1359.

(9) Tsung, F.; Zhang, K.; Cheng, L.; Song, Z. Statistical transfer learning: A review and some extensions to statistical process control. *Quality Engineering* **2018**, 115–128.

(10) Yang, Q.; Zhang, Y.; Dai, W.; Pan, S. J. *Transfer learning*; Cambridge University Press: Cambridge, 2020.

(11) Iapteff, L.; Jacques, J.; Rolland, M.; Celse, B. Reducing the number of experiments required for modelling the hydrocracking process with kriging through Bayesian transfer learning. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **2021**, *70*, 1344–1364.

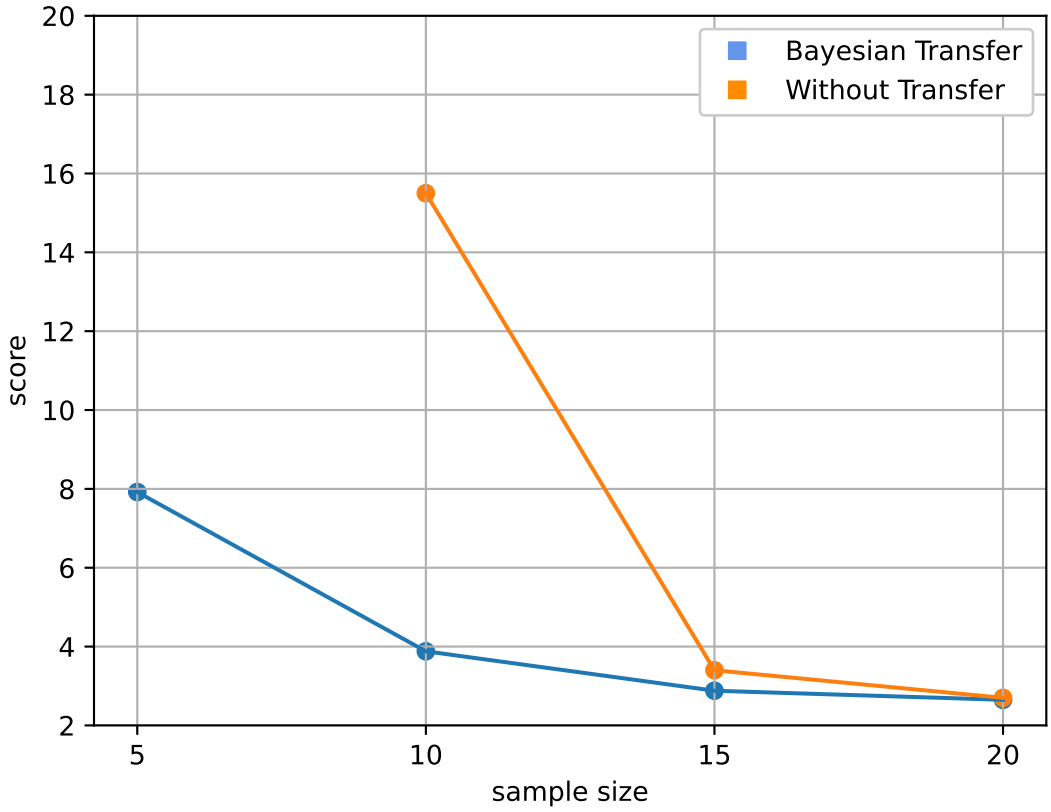
(12) Robert, C. P., et al. *The Bayesian choice: from decision-theoretic foundations to computational implementation*; Springer: New York, 2007; Vol. 2.

(13) Breunig, M.; Kriegel, H.; Ng, R.; Sander, J. LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000; pp 93–104.

(14) Cao, N. Y. P.; Celse, B.; Guillaume, D.; Guibard, I.; Thybaut, J. W. Accelerating Kinetic Parameter Identification by Extracting Information from Transient Data: A Hydroprocessing Study Case. *Catalysts* **2020**, 361.

(15) Tierney, L. Markov chains for exploring posterior distributions. *the Annals of Statistics* **1994**, 1701–1728.

(16) Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing* **1995**, *16*, 1190–1208.


Bayesian Transfer
Without Transfer
score
sample size
20
18
16
14
12
10
8
6
4
2
5
10
15
20