Design of Experiment for Bayesian transferred model

Loïc Iapteff, Julien Jacques, Benoit Celse, Victor Lameiras Franco da Costa

Design of Experiment for Bayesian transferred model

Loïc Iapteff [1], Julien Jacques [2], Benoit Celse [3] & Victor Costa [4]

[1] *loic.iapteff @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*
[2] *julien.jacques @univ-lyon2.fr, 5 Avenue Pierre Mendès France, 69500 Bron*
[3] *benoit.celse @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*
[4] *victor.costa @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*

Résumé. Le groupe IFP commercialise des catalyseurs et doit s'engager sur leur performance. Il est donc nécessaire de disposer de modèles prédictifs fiables pour chaque nouvelle génération de catalyseurs. Ces modèles sont construits à partir de données expérimentales très couteuses. Afin d'optimiser les coûts, notre ambition est de réduire le nombre d'expérimentations nécessaires pour estimer un modèle associé à un nouveau type de catalyseur. De précédents travaux ont montré qu'une approche de transfert Bayésien permettait d'améliorer la qualité des modèles lorsque le nombre d'observations est réduit. Dans cet article, les plans d'expériences sont étudiés afin de déterminer comment sélectionner ce nombre réduit d'observations permettant d'obtenir les meilleurs modèles par transfert Bayésien. Cet article montre que l'algorithme Kennard and Stone peut, sous certaines conditions, offrir de meilleurs résultats que des plans optimaux.

Mots-clés. Plan d'expérience, Transfer learning, inférence bayésienne

Abstract. IFP group develops catalysts and has to guarantee their performances. It is therefore crucial to have good predictive models for all new catalysts. These models are built upon very expensive experimental data. In order to minimize costs, we aim at reducing the number of new data points to measure to fit a model on the new catalyst. Previous work has shown that a Bayesian transfer approach can improve the quality of models when the number of observations is reduced. In this paper, experimental designs are studied in order to determine how to select this reduced number of observations to obtain the best models by Bayesian transfer. This article shows that the Kennard and Stone algorithm can, under certain conditions, offer better results than optimal designs.

Keywords. Design of Experiment, Transfer learning, Bayesian inference

1 The challenge

IFP Group develops and sells catalysts to chemical and biochemical producers. Catalysts are solids that make the reaction feasible, faster, and/or at a lower temperature and pressure. Performance must be guaranteed and it is therefore crucial to have good predictive

models for all new catalysts. These models are built on very expensive experimental data, therefore the aim is to use a small number of data and select them carefully to build a good model.

For the modeling of hydrocracking process, previous work has shown that the Bayesian transfer approach significantly reduces the number of points needed to obtain robust models (Iapteff et al. 2020). The method consists in estimating a model for a new catalyst using Bayesian models, whose priors depend on the old catalyst generation, for which important amounts of data are available. Celse, J.-J. D. Costa, and V. Costa 2016 shows the effectiveness of D-optimal designs for modeling the hydrocracking process using the kinetic model.

This paper focuses on hydrocracking modeling using Bayesian linear model (1):

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\beta X} + \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}), \\ \boldsymbol{\beta} &\sim \pi(\cdot), \end{aligned} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of model parameters, assumed to be a random variable of prior density $\pi(\cdot)$, $\boldsymbol{X}$ is the matrix of observations and $\boldsymbol{y}$ is the variable of interest. For the experiments, an industrial dataset from hydrocracking facilities is available. The prior $\pi(\cdot)$ is built with an old dataset that uses another catalyst.

The goal of this work is to show that when the sample size is not too small, basic geometric experimental design offer better results than optimal designs. This paper is composed of two parts: a first section presents the Design of Experiment (DoE) and a second section shows the results on our industrial dataset, using some of the most commonly employed DoE approaches. This section also explains why the results are not necessarily the best when using optimal designs.

2 Design of Experiment

The purpose of the Design of Experiments (DoE) is to organise the experiments carried out for scientific research or industrial studies in the best possible way. When experiments are used to study the relation between a quantity of interest and other features, DoE aims to maximise the information on this quantity from a reduced amount of data. The benefits of DoE are multiple, for instance it allows to control bias, to reduce random variation, to increase the precision of parameter estimates, to improve predictions of future observations, or to make a choice between competing models. In this section, we focus on most common DoE approaches for linear models.

2.1 Model independent approaches

The Kennard and Stone algorithm (Kennard and Stone 1969) is an iterative algorithm for selecting the best observations regardless of the type of model. This algorithm is a sequential construction for a minimax space-filling design and is based on a set of possible observations in order to extract the optimal points. It is initialized with a training dataset composed of the two most distant points of the complete dataset and a candidate set composed of the remaining points. Then, at each iteration, one observation $\boldsymbol{x}_i$ is moved from the candidate set to the training set such as $\boldsymbol{x}_i = \underset{\boldsymbol{x}_i \in \text{Candidate}}{argmax} \left(\underset{\boldsymbol{x}_{i'} \in \text{Training}}{min} dist(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) \right)$ until the training dataset reaches a predefined size. Notice that in the original Kennard and Stone algorithm, the euclidean distance is considered. With such an approach, we try to obtain an uniform distribution over the features space for the training set.

Other DoE model-independent approaches are well studied, such as modified Kennard and Stone algorithm, Latin hypercube sampling or full factorial designs but are not developped in this paper. The model independent DoE approaches are also often used to initialise algorithms for obtaining optimal designs, allowing faster convergence.

2.2 Optimal DoE for Linear Model

In this section, DoE approaches for the linear model (1) is considered. For this model, the most commonly used criteria are the $\mathcal{D}/\mathcal{A}/\mathcal{E}/\mathcal{G}$-optimal designs (Fedorov 1972, De Aguiar et al. 1995, Wong 1994). When the objective is to improve the prediction's quality of the model, a popular criterion is the $\mathcal{G}$-optimal criterion, which aims to minimize the maximum prediction uncertainty: the design must minimize $\underset{\boldsymbol{x}_i}{max}(\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1})\boldsymbol{x}_i)$. The $\mathcal{D}$, $\mathcal{A}$ and $\mathcal{E}$ optimal criteria use the dispersion matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ which is proportional to the covariance matrix of parameters. The idea is that minimizing the uncertainty in the parameters helps to minimize the uncertainty in prediction.

The $\mathcal{D}$-optimal design is the design that minimizes $det((\boldsymbol{X}^T\boldsymbol{X})^{-1})$. It is the most widely used criterion due to its efficiency and ease of implementation. The Fedorov's algorithm (Fedorov 1972) gives an exact $\mathcal{D}$-optimal design with an easy implementation. This algorithm ensures convergence to a local minimum, so different initializations can be considered. An usual initialization is to take a design obtained via the Kennard and Stone algorithm.

The $\mathcal{A}$-optimal design and $\mathcal{E}$-optimal design minimize the trace of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and the largest eigenvalue of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, respectively.

2.3 Bayesian Experimental Design

Since the objective of DoE is often to correctly estimate the model parameters, it can be interesting to provide a prior information on the value of these parameters. This approach is called Bayesian Experimental Design (Chaloner and Verdinelli 1995) and also has the

advantage that a single observation can be sufficient to build the model, the information matrix being always non-singular with a proper prior. Bayesian Experimental Design is commonly based on an utility function $U(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y})$, defining the contribution of the choice of design $\boldsymbol{d}$, yielding the $\boldsymbol{y}$-values for the dependent variable and with the model's parameters $\boldsymbol{\theta}$. The optimal design $\boldsymbol{d}_{opt}$ is thus obtained by maximizing the maximum expected utility:

$$\boldsymbol{d}_{opt} = \underset{d}{argmax} \int_{\mathcal{Y}} \int_{\Theta} U(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{y}, \tag{2}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})$ is the likelihood of the model and $p(\boldsymbol{\theta})$ the prior information, which is assumed not to be affected by the choice of the design.

In the Bayesian linear regression case, with a Gaussian prior for parameters $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and σ fixed, the posterior distribution of the parameters is known:

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}((\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{y} + \sigma^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}), \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}).$$

In our case of study, the parameters $\boldsymbol{\beta}$ are considered uncertain and σ is assumed to be known and identical to the parameter of the source model since we seek to model similar phenomena with identical tolerance.

Equivalently to the classical DoE framework, Bayesian alphabetical criteria are developed and use the matrix $(\boldsymbol{X}^T\boldsymbol{X}+\sigma^2\boldsymbol{\Sigma}^{-1})^{-1}$, proportional to the covariance of the Bayesian linear model parameters, instead of the matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Then, Bayes $\mathcal{D}/\mathcal{A}/\mathcal{E}/\mathcal{G}$-optimal designs minimize respectively $det((\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1})$, the trace of $(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}$, the largest eigenvalue of $(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}$ and $\underset{\boldsymbol{x}_i}{max}(\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{x}_i)$.

The Bayesian $\mathcal{D}$-optimal design can be derived from the mutual information for the Gaussian linear model while the Bayesian $\mathcal{A}$-optimal design can be obtained with the quadratic loss. Concerning the $\mathcal{E}$-optimal design and $\mathcal{G}$-optimal design, no utility function has yet been found whose maximisation leads to the construction of such a model.

Notice that Bayesian DoE is not only used if one wishes to build a Bayesian predictive model but can also be used in a non-Bayesian framework.

3 Use case: hydrocraking dataset

The dataset used for this experiment is obtained from two refineries using a new generation catalyst and is composed of 1004 observations described with 12 features. Let $\boldsymbol{X}$ be the matrix of observations and $\boldsymbol{y}$ the vector of the actual value of Diesel density to be predicted. In this experiment, different DoE are tested for a varying size of sample n. The aim is to build a design that allows to obtain the most efficient models with the fewest possible points for learning. The objective is to combined the DoE approach with the Bayesian transfer method (Iapteff et al. 2020), which has proven its efficiency in

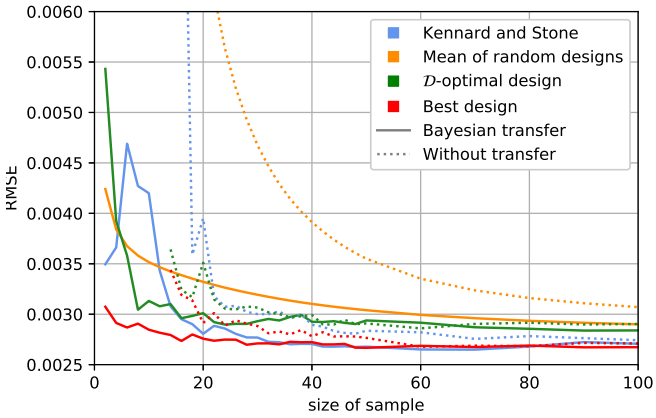

Figure 1: Comparison of the different designs for the Linear model. The red lines give the best possible design obtained by taking for each sample size n the best RMSE score among the thousand random samples.

improving the quality of the model. The model fitted with the Bayesian transfer approach is compared to a classical "non-transfer" model, for which $\boldsymbol{\beta}$ is assumed to be no longer random and estimated by maximum likelihood. The "Bayesian transfer model" is the model learned with prior on $\boldsymbol{\beta}$: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_{prior}, \boldsymbol{\Sigma}_{prior})$. As a reminder, for the Bayesian approach, the prior is built from a dataset from refineries using an old generation catalyst. To compare different methods, the target dataset is split: 600 observations are used as candidate for the design and the 404 remaining observations are used for testing the model. The RMSE score is thus always evaluated only on these 404 points whatever the size n of the sample.

For both models with and without transfer, three different designs are considered:

- Mean of random designs: samples are randomly selected from the candidates dataset, a thousand times, and scores are then averaged.
- Kennard and Stone,
- $\mathcal{D}$-optimal design: the Fedorov's algorithm is used in order to build the $\mathcal{D}$-optimal design for each sample size. For the model without transfer, classical $\mathcal{D}$-optimal design is considered while Bayesian $\mathcal{D}$-optimal design is considered for the model fit with Bayesian transfer.

Note that other alphabetical criteria have been tested and offer less good results than the $\mathcal{D}$-optimal criterion for our prediction task and are therefore not presented here. The results are plotted in Figure 1.

For all designs, the Bayesian transfer approach improves the results compared to a non-transfer approach. Nevertheless, the non-transfer approach with a $\mathcal{D}$-optimal design

provides very good results fairly quickly . The best results with a reduced number of points are obtained with a $\mathcal{D}$-optimal design coupled with the Bayesian transfer approach. As the number of points increases, above 15, the Bayesian transfer approach coupled with the design obtained via Kennard and Stone is better.

In the present case study, the good performance of the Kennard and Stone method in comparison with the $\mathcal{D}$-optimal approach can be explained by the fact that data came from two refineries. Indeed, the dataset is then not exactly homogeneous, and the discrete $\mathcal{D}$-optimal design may be far from the theoretical optimum. Additional experiments on simulated dataset confirmed this behavior.

Another important point is that the design obtained with Kennard and Stone algorithm allows to reach the best possible score when number of observation is sufficient, around 25 observations for our application with 12 features.

4 Conclusion

This study shows that in some practical cases, building optimal designs is not the best solution. It is the case when the features space distribution is not homogeneous and the chosen model does not perfectly describes the phenomenon. Then, the Kennard and Stone algorithm allows to keep the important information when the number of observation is sufficient, and become more effective than optimal design. However, when the number of observations is low, $\mathcal{D}$-optimality remains the most efficient.

References

Celse, Benoit, Jean-Jérôme Da Costa, and Victor Costa (2016). "Experimental Design in Nonlinear Case Applied to Hydrocracking Model: How Many Points Do We Need and Which Ones?" In: *International Journal of Chemical Kinetics* 48.11, pp. 660–670.

Chaloner, Kathryn and Isabella Verdinelli (1995). "Bayesian experimental design: A review". In: *Statistical Science*, pp. 273–304.

De Aguiar, P. F. et al. (1995). "D-optimal designs". In: *Chemometrics and intelligent laboratory systems* 30, pp. 199–210.

Fedorov, Valerii Vadimovich (1972). *Theory of Optimal Experiments Designs*. Academic Press.

Iapteff, Loıc et al. (2020). "Reducing the number of experiments required for modelling the hydrocracking process with kriging through Bayesian transfer learning". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Kennard, R.W. and L.A. Stone (1969). "Computer aided design of experiments". In: *Technometrics*, pp. 137–148.

Wong, W. K. (1994). "Comparing robust properties of A, D, E and G-optimal designs". In: *Computational statistics & data analysis* 18, pp. 441–448.