Automatic relevance determination of categorical variables for mixed variables Gaussian process regression.

Théo Rabut, Thomas Galeandro-Diamant, Hamamache Kheddouci

Automatic relevance determination of categorical variables for mixed variables Gaussian process regression

Theo Rabut
Université de Lyon, Université Lyon 1,
LIRIS UMR CNRS 5205
Lyon, France

Thomas Galeandro-Diamant
DeepMatter
Glasgow, United Kingdom

Hamamache Kheddouci
Université de Lyon, Université Lyon 1,
LIRIS UMR CNRS 5205
Lyon, France

Abstract

Categorical variables combined with continuous variables are gaining interest in multiple applications of Gaussian processes. Different covariance functions can be used to handle categorical variables, but we found in a previous study that one of the most versatile was initially proposed in an optimization method called COCABO. Choosing an adequate prior when using a Gaussian process regression can drastically improve the regression performances. We propose in this paper a modification to the COCABO covariance function to allow the Gaussian processes to automatically find the relevance of each categorical variable. This mechanism, is inspired by Automatic Relevance Determination (ARD) for the continuous variables, can effectively be used with covariance functions that use Hamming distance to treat categorical variables. While we use the Categorical-ARD (CATARD) mechanism with the CoCaBO covariance function, it can be generalized to every covariance function that does not relax categorical variables to calculate the covariance. We used both synthetic benchmarks and real world data in order to establish the performances of the CATARD covariance function.

CCS Concepts

• **Applied computing** → **Chemistry**; • **Computing methodologies** → **Gaussian processes**; **Uncertainty quantification**.

Keywords

Categorical variables, Gaussian processes, Regression

ACM Reference Format:
Theo Rabut, Thomas Galeandro-Diamant, and Hamamache Kheddouci. 2025. Automatic relevance determination of categorical variables for mixed variables Gaussian process regression. In *Proceedings of ACM SAC Conference (SAC'25).* ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/xx.xxx/xxx_x

1 Introduction

Gaussian processes (GPs) have become widely recognized in scientific literature, particularly in the realms of machine learning, statistics, and probabilistic modeling. Gaussian processes offer a versatile framework for capturing intricate, non-linear patterns in data. The main feature that makes Gaussian processes popular is their ability to predict a predictive distribution for each input which can be interpreted as a prediction value (mean of the distribution) and an associated uncertainty (variance).

The work of Christopher K. I. Williams and Carl E. Rasmussen in Gaussian Processes for Machine Learning (GPML) [2] played a tremendous part in the democratization of Gaussian process in the machine learning community.

Gaussian processes find applications across diverse scientific fields such as astronomy, physics or biology with researchers continually exploring new methodologies and advancements [1, 3, 11]. Furthermore, GPs have shown remarkable performances when applied to noisy data or time related problems.

Gaussian process regression is a kernel based regression method. The covariance function used to build the kernel constitutes the prior that the user has on the problem. Hence, the design of covariance functions and their hyper-parameters can strongly influence the model quality.

For a new input $\mathbf{X}^*$, a fitted GP predicts, after a series of joint and marginalization operations, a distribution over functions on the new input. Then, the mean and the variance of this distribution are respectively considered as the prediction and the associated uncertainty. The prediction for a new input $\mathbf{X}^*$ is given as follow :

$$\mu(\mathbf{X}^*) = K(\mathbf{X}^*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1}\mathbf{y} \tag{1}$$

Whereas the uncertainty is given by :

$$\sigma(\mathbf{X}^*) = K(\mathbf{X}^*, \mathbf{X}^*) - K(\mathbf{X}^*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1}K(\mathbf{X}, \mathbf{X}^*) \tag{2}$$

where:

- $K(\mathbf{X}^*, \mathbf{X})$ is the covariance vector between the prediction point $\mathbf{X}^*$ and the observation points $\mathbf{X}$.
- $K(\mathbf{X}, \mathbf{X})$ is the covariance matrix between the observation points $\mathbf{X}$.
- σ_n^2 is the noise variance (if the observations are noisy).
- I is the identity matrix of dimension $n \times n$.
- $\mathbf{y}$ is the vector of observations.

Over the past decade, the significance of categorical variables in data analysis and statistical modeling has grown exponentially [8, 12]. The increased availability of diverse and complex datasets has underlined the importance of effectively handling categorical information.

Facing continuous and categorical features in a regression problem can be quite challenging, the standard approach when facing categorical variables being to numerically encode these variables. These encoding techniques and specially the one-hot encoding technique has been used along with continuous kernels to model datasets with mixed variable inputs using Gaussian processes [5, 19].

SAC'25, March 31 –April 4, 2025, Sicily, Italy

ACM ISBN 979-8-4007-0629-5/25/03
https://doi.org/xx.xxx/xxx_x

Saves *et al.* propose the use of a specific distance metric (e.g. Gower distance), encapsulating both categorical distance and continuous distance in a more standard covariance function equation [15]. Others rather use, as a main covariance function for the Gaussian processes, the combination of covariance functions that are designed for several data types [4, 18].

We propose a modification of the covariance function proposed in an optimization method called COCABO [14] in order to improve the performances on different regression problems that include continuous and categorical variables. With this modification, we intend to automatically capture the relevance of each categorical variable by using an hyper-parameter for every categorical dimension. Our interest is mainly drawn from Bayesian optimization, and more specifically, from chemical formulation and chemical reaction optimization where the impact of categorical variables is significant and heterogeneous.

2 The COCABO covariance function

The covariance function proposed in COCABO [14] is composed by two sub-kernels, one for the continuous variables and one for the categorical variables. This covariance is defined as:

$$K(\mathbf{z},\mathbf{z}') = \sigma_g((1-\lambda)\times(K_{cont}(\mathbf{x},\mathbf{x}')+K_{cat}(\mathbf{h},\mathbf{h}'))+\lambda\times K_{cont}(\mathbf{x},\mathbf{x}')\times K_{cat}(\mathbf{h},\mathbf{h}')) \tag{3}$$

w.r.t.:

$$K_{cont}(\mathbf{x},\mathbf{x}') = \text{Matérn}_{5/2}(\mathbf{x},\mathbf{x}') \tag{4}$$

$$K_{cat}(\mathbf{h},\mathbf{h}') = \sigma\sum_{i=0}^{D}\alpha(h_i,h_i') \tag{5}$$

$$\alpha(h_i,h_i') = \begin{cases} 1 & \text{if } h_i = h_i' \\ 0 & \text{otherwise} \end{cases}$$

and :

- $\mathbf{z} = (\mathbf{x},\mathbf{h})$
- $\mathbf{x}$ the continuous variables
- $\mathbf{h}$ the categorical variables
- σ_g the main amplitude hyper-parameter
- λ the product-sum balance hyper-parameter
- D is the number of categorical variables

This covariance function combines the two sub-kernels by summing them and/or multiplying them. The balance of the product and the sum is ruled by an hyper-parameter λ. This hyper-parameter, along with the others of the covariance, is fitted to the data following the maximization of the log marginal likelihood.

Such a balance allows the model to capture different relations between variables. On the one hand, the sum will capture an offset of one variable type over the other. On the other hand, the product can capture more complex relations such as one variable type that amplify the other. This balance between the operators enables a desirable versatility when facing heterogeneous datasets (or evolving ones) [13].

While the COCABO covariance function is capable of handling a wide variety of regression problems, it does not make any difference of categorical dimensions since the K_{cat} (Eq. 5) treats every categorical variable the same way. Many problems possess categorical variables that impact the output variables with different magnitudes.

3 CATARD covariance function

One way to capture differences between categorical variables is to modify the covariance function. Thus, we propose a covariance function called CATARD which refers to the Automatic Relevance Determination mechanism applied to CATegorical variables. The CATARD covariance function is an extension of the COCABO covariance function given at the equation 3. The only difference lies within the categorical sub-covariance function.

Usually, in a continuous kernel, the distance between data points is weighted by the length-scale hyper-parameter in order to tune the relevance of the distance in the resulting covariance values.

The ARD mechanism enables the ability to treat each dimension differently by assigning an independent length-scale to each of these dimensions. It is mainly used for problems where the scale of each continuous dimension differ from one to another [10].

This ARD mechanism can address the main problem of the COCABO covariance function presented in the previous section but instead of re-scaling continuous variables independently, we propose to use an hyper-parameter on each categorical dimension. The equation of the categorical specific sub-covariance function is then given by:

$$K_{cat}(\mathbf{h},\mathbf{h}') = \sum_{i=0}^{D}\sigma_i\alpha(h_i,h_i') \tag{6}$$

where σ_i is the hyper-parameter that is specific to each categorical variable. The hyper-parameter values are tuned following a maximization of the log marginal likelihood.

Maximizing the Log Marginal Likelihood

Gaussian processes are fitted to the data by finding the optimal hyper-parameters that maximize the log marginal likelihood (LML). A standard method (and the one we use) to deal with this maximization is the Limited-memory BFGS (L-BFGS) [9].

The gradients of the log marginal likelihood are easy to calculate so its maximization can be done with gradient descent based algorithms. We are looking for the partial derivatives of the LML with respect to each hyper-parameter we are tuning. Following Rasmussen's and Williams's GPML [2], we have:

$$\frac{\partial}{\partial\theta_j}\log p(\mathbf{y}|X,\theta) = \frac{1}{2}tr((\alpha\alpha^T - K^{-1})\frac{\partial K}{\partial\theta_j})$$

We then need to calculate the gradient of the kernel K with respect to each hyper-parameter θ_j.

$$\frac{\partial K}{\partial\sigma_g} = \frac{K}{\sigma_g}$$

$$\frac{\partial K}{\partial\lambda} = \sigma_g\times(K_{cont}K_{cat} - K_{cont} - K_{cat})$$

$$\frac{\partial K}{\partial\sigma_{\text{Matérn}_{5/2}}} = \sigma_g\times((1-\lambda)\frac{2K_{cont}}{\sigma_{\text{Matérn}_{5/2}}} + \lambda - \frac{2K_{cont}}{\sigma_{\text{Matérn}_{5/2}}}\times K_{cat})$$

$$\frac{\partial K}{\partial\ell} = \sigma_g\times\frac{5r^2\sigma^2_{\text{Matérn}_{5/2}}}{3\ell^3}\exp(\frac{-\sqrt{5}r}{\ell})[1+r\frac{\sqrt{5}}{\ell}]\lambda(\frac{1}{\lambda}-1+K_{cat})$$

Name	Cont.var.	Cat.var.	Number of categories
Styblinski-Tang	4	4	[2, 3, 5, 10]
Ackley	4	4	[2, 3, 5, 10]
Rosenbrock	4	4	[2, 3, 5, 10]

Table 1: Synthetic problems

Formulation name	Outputs	Cont.var.	Cat.var.	Nb. of categories	Size
Lubricant	1	7	3	[2, 3, 6]	38
Hair dye	2	9	3	[5, 5, 3]	28
Rubber	3	10	3	[4, 3, 3]	59
Polycarbonate resin	3	11	5	[3, 20, 4, 3, 3]	87

Table 2: Real world problems

$$\frac{\partial K}{\partial \sigma_i} = \sigma_g \alpha(h_i, h_i') \times (1 - \lambda + \lambda \times K_{cont}(\mathbf{x}, \mathbf{x}'))$$

These partial derivatives are quick to compute. Hence, the optimization of the hyper-parameter can be straightforward using gradient based methods. We used a multi-started gradient descent to maximize the LML. The number of starting points depends on the number of hyper-parameters, which for the proposed model, is related to number of categorical variables. In our experiments, we used 10 starting points because we have a limited (3 to 5) number of categorical variables. We observed that increasing the number of starting points do not significantly increase the performance of the model.

4 Experiments

4.1 Synthetic data

First, in order to evaluate the effectiveness of the CATARD approach on diverse problems, we propose to use well-known synthetic functions to generate datasets. These functions were taken from the optimization community and often serve as benchmarks for optimization methods [16]. For this study to be the most generic possible, we selected functions with diverse shapes: Rosenbrock function (valley), Ackley function (peaked) and Styblinski-Tang (bowl). Since these synthetic functions contain only continuous inputs, we discretized a subset of the continuous continuous dimensions to have both continuous and categorical inputs.

Then, we sampled data from these functions and we used the resulting datasets to train the Gaussian processes. In order to obtain a representative dataset we sampled the data using a method for each variable type and concatenate the two sets of variables.

For the continuous variables, we used a Latin Hypercube Sampling (LHS) technique that ensure a low discrepancy between data points [7]. For the categorical variables, we drew a subset of all possible combinations with a method called Generalized Subset Design (GSD) [17] because GSD ensures a well balanced categorical dataset (no categories are over or under represented). We used the quotient of the Euclidian division of the number of categorical combinations by the size of the subset we need as the reduction factor (method parameter).

4.2 Real world data

Secondly, we harvested public chemical formulation datasets (described in Table 2). These datasets are particularly suited for this study because they have multiple continuous and categorical variables and the importance of the different categorical variables is heterogeneous.

The lubricant formulation problem consists of predicting the coefficient of dynamic friction with 10 continuous variables and 3 categorical ones.

The hair dye formulation problem consists of 2 outputs (the coloration intensity and the rinsing sensation) with 9 continuous variables and 3 categorical ones.

The rubber formulation problem consists of 3 outputs (the Mooney viscosity, the energy dissipation factor $\tan(\delta)$ and the crack resistance growth). 10 continuous variables are used along with 3 categorical variables.

The polycarbonate resin formulation problem consists of predicting 3 output variables (the elasticity modulus, the impact resistance and the heat resistance) from 11 continuous variables and 5 categorical ones. We address this problem as 3 different regression problems (one for each output).

4.3 Methods

We propose to compare our model against Gaussian processes with other covariance functions.

First, the method denoted as "OHE" refers to a one-hot encoding of categorical variables with a $\text{Matérn}_{5/2}$ covariance function. While this can be considered as the standard approach to handle categorical variables it has many drawbacks. They are mainly due to the number of input dimensions that grows with the number of categories.

The COCABO method refers to the covariance function proposed in the optimization method COCABO [14] and described in section 2. This covariance function is based upon a combination of a sum and a product of two covariance functions designed and applied on continuous and categorical variables separately.

The "GOWER" method refers to Gaussian Processes with a Gower distance based $\text{Matérn}_{5/2}$ covariance function. The Gower distance [6] enables covariance calculation between two points without the supplementary dimensions that an encoding imposes. Many recent works uses the Gower distance to quantify the correlation between mixed variables.

4.4 Uncertainty based Metrics

While the standard machine learning metrics, such as the mean squared error (MSE) or the coefficient of determination (R^2), are designed to resume the capacity of the model to generalize over a dataset, they do not include the notion of uncertainty. Hence, in order to capture the uncertainty in the metrics we used, we chose to use both of the log marginal likelihood (LML) and the mean standardized log loss [2] (MSLL) as they provides different insights on the quality of the fit.

4.4.1 Log Marginal Likelihood. The log marginal likelihood correspond to the probability of the observed data integrated over the model parameters. This metric is composed by three terms that are respectively built upon the difference between predictions and the

ground truth (referred as error term), the uncertainty (complexity term), and a constant depending on the number of input data points:

$$\log p(\mathbf{y}|X,\theta) = -\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \quad (7)$$

Where $\mathbf{K}$ is the covariance matrix, and $\mathbf{y}$ is the observations. The first term is the only one that involves the observations (the error term). The complexity term penalizes data dispersion given by the covariance. This metric is usually used without a cross validation. Thus, without noise, the LML is only composed by the complexity and the constant terms.

4.4.2 Mean Standardized Log Loss. The Mean Standardized Log Loss refers to the mean of the log probability of the predictions. The MSLL does not use the determinant of the kernel but we can identify two main components in the equation 8. The first term is based on the uncertainty and it reflects the complexity of the model. The second term is the prediction error squared is divided by the variance. This second term makes the MSLL more sensitive to overfitting than the LML.

$$-\log p(\hat{y}|\mathbf{X},\mathbf{y},\hat{x}) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{(y-\hat{y})^2}{2\sigma^2} \quad (8)$$

Where σ is the uncertainty at test points and $(y-\hat{y})^2$ is the squared prediction error. The Mean Standardized Log Loss is built following a cross validation. In this work we used a leave-one out cross validation on smaller datasets (size ≤ 50), we used a 5-fold cross validation otherwise.

5 Results

We compare the CATARD covariance based GP to three other GPs with different covariance functions: a Matérn$_{5/2}$ covariance function with a one-hot encoding of the categorical variables, the COCABO covariance function, and a Matérn$_{5/2}$ covariance function that uses the Gower distance instead of an Euclidian distance. In order to measure the performance of each model, we use the LML and the MSLL as regression metrics because they include the prediction uncertainty in their calculations. We use the same approach to construct and fit the Gaussian Processes on the synthetic regression problems and all four of the chemical formulation datasets.

5.1 Synthetic problems

The next tables show the metrics for each selected methods on the three synthetic datasets we built (respectively Ackley 3, Rosenbrock 4 and Styblinski-Tang 5). Each table is splitted following the size of the dataset we generated (50, 100 and 150 points).

	50		100		150	
	LML	MSLL	LML	MSLL	LML	MSLL
OHE	-70.87	1.59 ± 1.95	-133.82	1.35 ± 0.26	-190.31	1.35 ± 0.48
GOWER	-70.47	**1.54** ± 1.98	-133.65	1.38 ± 0.24	-188.37	1.38 ± 0.48
COCABO	-69.2	1.54 ± 2.29	**-112.93**	**1.06** ± 0.27	-162.982	**1.14** ± 0.55
CATARD	**-67.24**	1.61 ± 2.35	-122.87	1.18 ± 0.28	**-161.24**	1.21 ± 0.65

Table 3: Ackley Log marginal likelihood and Mean Standardized Log Loss for the presented models on the Ackley problem with 3 different sample sizes: 50, 100, 150.

	50		100		150	
	LML	MSLL	LML	MSLL	LML	MSLL
OHE	-69.64	1.46 ± 1.21	-136.65	1.40 ±0.18	-202.53	1.34 ± 0.28
GOWER	- 69.60	1.43 ± 1.06	-136.55	1.42 ± 0.20	-198.62	1.37 ± 0.29
COCABO	-60.93	1.30 ± 1.52	-98.87	0.75 ± 0.08	-136.91	**0.78** ± 0.21
CATARD	**-55.481**	**0.939** ± 1.02	**-95.02**	**0.74** ± 0.08	**-134.40**	0.81 ± 0.23

Table 4: Rosenbrock Log marginal likelihood and Mean Standardized Log Loss for the presented models on the Rosenbrock problem with 3 different sample sizes: 50, 100, 150.

	50		100		150	
	LML	MSLL	LML	MSLL	LML	MSLL
OHE	-69.53	1.41 ± 1.20	-134.97	1.39 ± 0.21	-196.47	1.33 ± 0.25
GOWER	-69.52	1.43 ± 1.17	-134.87	1.41 ± 0.17	-199.62	1.50 ± 0.59
COCABO	-67.65	1.49 ± 1.22	-120.37	**1.11** ± 0.11	-176.33	**1.11** ± 0.20
CATARD	**-62.52**	**1.29** ± 1.19	**-119.43**	1.23 ± 0.11	**-175.23**	1.17 ± 0.24

Table 5: Styblinski-Tang Log Marginal Likelihood (LML) and Mean Standardized Log Loss (MSLL) for the presented models on the Styblinski-tang problem with 3 different sample sizes: 50, 100, 150.

The CATARD covariance based GP does not provide satisfactory MSLL on the Ackley benchmark datasets as shown in the Table 3. Moreover the two nearly identical LML on both of the datasets with 50 and 150 samples and the higher LML on the dataset with 100 samples also favor the COCABO covariance function over the CATARD one. We argue that the ability of the CATARD-based Gaussian process to determine the influence of each categorical variable separately on the output is not relevant in the Ackley scenario as the discretization of the complex structure of the Ackley function leads to highly similar categorical variables. Therefore, these results expose the limitations of the CATARD covariance function.

As shown in Table 4, the CATARD covariance function outperforms the other approaches on the Rosenbrock regression problem. However, as the size of the dataset increases, the COCABO covariance function find itself to be a viable alternative to the CATARD approach.

On the Styblinski-Tang regression problem (Table 5), the CATARD shows the higher LML with the smallest MSLL by a large margin on the smallest dataset. Once more, as the size of the dataset increases, the performance gap between the CATARD and COCABO covariance functions diminishes, ultimately favoring the COCABO covariance function.

Overall, the CATARD covariance function tends to have a higher log marginal likelihood than the other methods. It indicates that the covariance function needs less kernel amplitude to explain the observations. This is mainly due to the focus on relevant categorical variables that the ARD mechanism provides. The metrics we used in our benchmarking strategy provide insights into the quality of the fit of the GPs, making them relevant metrics for choosing a covariance function over a dataset. Consequently, they are also relevant metrics for determining whether or not to incorporate the ARD mechanism into the covariance function.

5.2 Real world problems

The next tables provide the results of the 4 covariance functions on 4 datasets that comes from chemical formulation experiments. The size of these datasets range from 28 to 87. Their distribution is sparse and may contain noise. We did not include a noise specific hyper-parameter on any of the Gaussian processes but it can be the subject of further experiments.

Models	Dynamic friction coef.	
	LML	MSLL
OHE	-31.71	**0.21** ± 0.75
GOWER	-44.56	0.68 ± 1.57
COCABO	-30.68	0.66 ± 3.20
CATARD	**-26.58**	0.64 ± 3.64

Table 6: Lubricant formulation Log Marginal Likelihood and Mean Standardized Log Loss of the presented models on the only dependent variable of the lubricant formulation

	Coloration intensity		Rinsing sensation	
	LML	MSLL	LML	MSLL
OHE	-34.15	1.98 ± 3.09	-22.82	1.15 ± 2.22
GOWER	-26.41	1.29 ± 3.12	-15.41	0.77 ± 1.84
COCABO	-21.638	1.18 ± 2.71	-12.31	**0.15** ± 1.52
CATARD	**-20.78**	**1.10** ± 2.33	**-11.57**	0.18 ± 1.64

Table 7: Hair dye formulation Log Marginal Likelihood and Mean Standardized Log Loss of the presented models on the two dependent variables of the hair dye formulation.

	Mooney viscosity		Energy dissipation		Crack resistance growth	
	LML	MSLL	LML	MSLL	LML	MSLL
OHE	-83.26	1.52 ± 0.71	-83.57	1.50 ± 0.68	-71.79	1.19 ± 1.28
GOWER	-77.07	0.95 ± 0.99	-81.79	0.902 ± 0.99	-66.79	0.75 ± 1.49
COCABO	-64.30	**0.89** ± 1.50	-67.69	0.81 ± 1.10	- 64.29	0.76 ± 1.66
CATARD	**-62.43**	0.97 ± 1.59	**-66.18**	**0.68** ± 1.22	**-58.11**	**0.73** ± 1.68

Table 8: Rubber formulation Log Marginal Likelihood and Mean Standardized Log Loss of the presented models on the two dependent variables of the rubber formulation.

	Elasticity		Impact resist.		Heat resist.	
	LML	MSLL	LML	MSLL	LML	MSLL
OHE	-68.40	1.44 ±0.65	-94.46	1.23 ± 0.26	-98.00	1.52 ± 0.86
GOWER	-70.01	1.34 ± 0.63	-81.83	2.38 ± 0.61	-84.94	1.35 ± 0.86
COCABO	-24.67	**1.10** ± 0.65	-76.38	1.21 ± 2.51	-64.24	**1.04** ± 0.68
CATARD	**-18.59**	1.12 ± 0.65	**-65.792**	**1.20** ± 0.82	**-60.66**	1.37 ± 0.57

Table 9: Polycarbonate resin formulation Log Marginal Likelihood and Mean Standardized Log Loss of the presented models on the three dependent variables of the polycarbonate resin formulation

The results on real world data (Tables 6, 7, 8, 9) show an improvement for all LML.

On the lubricant formulation dataset (Table 6), the Matérn$_{5/2}$ kernel with one-hot encoding yields the lowest MSLL by a large margin. We emphasize that the error term of the MSLL is divided by the squared variance of the prediction, which makes the MSLL highly sensitive to outliers. Therefore, the one-hot encoding approach is capable of effectively modeling the outliers present in the lubricant dataset.

The regression performances on the hair dye formulation (Table 7) show that the CATARD covariance function produces a GP with a higher LML than the other approaches. While these metrics indicate a superior modeling capacity of the CATARD-based GP, the MSLL (and the associated standard deviation) on the rinsing sensation dependent variable shows the COCABO-based GP as the most accurate.

On the rubber formulation dataset (Table 8), the CATARD-based GP outperforms the other GPs on every LML and almost every MSLL. The only instance where the COCABO covariance function achieves a lower MSLL is for the regression on Mooney viscosity, indicating a more accurate model.

On the polycarbonate formulation dataset (Table 9), the CATARD-based GP offers the higher LML for the three dependent variables. For both Elasticity and Heat Resistance, using the COCABO covariance function results in a smaller MSLL. However, for the Impact Resistance regression, the MSLL of the GP built upon the COCABO covariance function fluctuates more, with a standard deviation of 2.51, which makes the CATARD covariance approach preferable.

The results on the real world datasets suggest that selecting the CATARD function for Gaussian process regression can be effective, particularly when there is a need to differentiate the relevance of categorical variables and when the dataset contains few outliers.

6 Conclusion

In this paper, we have shown that a covariance function capable of capturing differences of each categorical variables in terms of impact on the output leads to models with better metrics. The synthetic benchmarking strategy we used is inspired from the optimization literature with a discretization of a subset of the input variables and a DoE approach to sample these highly practical functions.

While our results shows an improvement in the modeling of uncertainty, prediction error based metrics aren't drastically improved. Further experiments could involves new metrics, comparison with more and newer models such as Bayesian hierarchical models or

an extension of the benchmarking strategy to consolidate the performances of the CATARD covariance function.

As well as the continuous ARD mechanism, the ARD mechanism applied to categorical variables can be applied for many covariance functions. We believe that this works paves the way to multiple kernel regression methods with mixed variables and in particular Gaussian Processes applications. Our motivation stems from Bayesian optimization, where the selection of the covariance function significantly enhances the performance of the optimizers [13]. Consequently, the application of the CATARD covariance function within a Bayesian optimization framework to optimize chemical will be the focus of future research.

Acknowledgments

This work was supported by the R&D Booster SMAPI project 2020 of the Auvergne-Rhône-Alpes Region. It is also partially funded by the French National Agency for Research (ANR) on the reference GRADIENT ANR-22-CE23-0009.

References

[1] Suzanne Aigrain and Daniel Foreman-Mackey. 2023. Gaussian process regression for astronomical time series. *Annual Review of Astronomy and Astrophysics* 61 (2023), 329–371.

[2] Christopher K. I. Williams Carl Edward Rasmussen. 2006. *Gaussian processes for machine learning*. MIT press.

[3] Wei Dang, Shengjun Liao, Bo Yang, Zhengtong Yin, Mingzhe Liu, Lirong Yin, and Wenfeng Zheng. 2023. An encoder-decoder fusion battery life prediction method based on Gaussian process regression and improvement. *Journal of Energy Storage* 59 (2023), 106469.

[4] David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. 2013. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*. PMLR, 1166–1174.

[5] Eduardo C Garrido-Merchán and Daniel Hernández-Lobato. 2020. Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. *Neurocomputing* 380 (2020), 20–35.

[6] John C Gower. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 3-4 (1966), 325–338.

[7] Jon C Helton and Freddie Joe Davis. 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety* 81, 1 (2003), 23–69.

[8] Bernhard O Josephus, Ardianto H Nawir, Evelyn Wijaya, Jurike V Moniaga, and Margaretha Ohyver. 2021. Predict mortality in patients infected with COVID-19 virus based on observed characteristics of the patient using logistic regression. *Procedia computer science* 179 (2021), 871–877.

[9] Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming* 45, 1 (1989), 503–528.

[10] Kailong Liu, Yi Li, Xiaosong Hu, Mattin Lucu, and Widanalage Dhammika Widanage. 2019. Gaussian process regression with automatic relevance determination kernel for calendar aging prediction of lithium-ion batteries. *IEEE Transactions on Industrial Informatics* 16, 6 (2019), 3767–3777.

[11] Ian C McDowell, Dinesh Manandhar, Christopher M Vockley, Amy K Schmid, Timothy E Reddy, and Barbara E Engelhardt. 2018. Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLoS computational biology* 14, 1 (2018), e1005896.

[12] Daniel Powers and Yu Xie. 2008. *Statistical methods for categorical data analysis*. Emerald Group Publishing.

[13] Theo Rabut, Hamamache Kheddouci, and Thomas Galeandro-Diamant. 2022. Categorical-Continuous Bayesian Optimization Applied to Chemical Reactions. In *International Conference on Optimization and Learning*. Springer, 226–239.

[14] Binxin Ru, Ahsan Alvi, Vu Nguyen, Michael A Osborne, and Stephen Roberts. 2020. Bayesian optimisation over multiple continuous and categorical inputs. In *International Conference on Machine Learning*. PMLR, 8276–8285.

[15] Paul Saves, Youssef Diouane, Nathalie Bartoli, Thierry Lefebvre, and Joseph Morlier. 2023. A mixed-categorical correlation kernel for Gaussian process. *Neurocomputing* (2023), 126472.

[16] S. Surjanovic and D. Bingham. [n. d.]. Virtual Library of Simulation Experiments: Test Functions and Datasets. Retrieved February 16, 2024, from http://www.sfu.ca/~ssurjano.

[17] Izabella Surowiec, Ludvig Vikstrom, Gustaf Hector, Erik Johansson, Conny Vikstrom, and Johan Trygg. 2017. Generalized subset designs in analytical chemistry. *Analytical chemistry* 89, 12 (2017), 6491–6497.

[18] Laura P Swiler, Patricia D Hough, Peter Qian, Xu Xu, Curtis Storlie, and Herbert Lee. 2014. Surrogate models for mixed discrete-continuous variables. *Constraint Programming and Decision Making* (2014), 181–202.

[19] Yichi Zhang, Siyu Tao, Wei Chen, and Daniel W Apley. 2020. A latent variable approach to Gaussian process modeling with qualitative and quantitative factors. *Technometrics* 62, 3 (2020), 291–302.