A Unified View of FANOVA: A Comprehensive Bayesian Framework for Component Selection and Estimation

Yosra MARNISSI [1] **Maxime LEIBER** [1 2]

Abstract

This paper presents a comprehensive Bayesian framework for FANOVA models. We provide guidelines for tuning and practical implementation to improve scalability of learning and prediction. Our model is very flexible and can handle different levels of sparsity across and within decomposition orders, as well as among covariates. This flexibility enables the modeling of complex real-world data while enhancing interpretability. Additionally, it allows our model to unify diverse deterministic and Bayesian non-parametric approaches into a single equation, making comparisons and understanding easier. Notably, our model serves as the Bayesian counterpart of several deterministic methods allowing uncertainty quantification. This general framework unlocks potential for novel model developments that have been previously overlooked, such as the proposed Dirichlet mixing model that addresses limitations of existing models.

1. Introduction

In many applications, we aim to estimate the function that links a target output to input covariates. In this paper, we focus on Functional Analysis of Variance (FANOVA) models, which decompose this regression function into main and interactions terms (Wahba, 1990; Durrande et al., 2013; Ginsbourger et al., 2016; Chastaing & Le Gratiet, 2015) giving insights into the effects of individual covariates and combinations of covariates. To enhance interpretability, we often assume that the output depends on only a few key functional terms (Rudin et al., 2022). However, identifying the optimal set of these terms and characterizing their non-linear shapes poses a significant challenge. Two distinct approaches have emerged to address this problem.

A widely recognized method originating from the kernel community is the Component Selection and Smoothing Operator (COSSO) (Lin & Zhang, 2003; Storlie et al., 2011; Touzani & Busby, 2013; Ravikumar et al., 2009; Radchenko & James, 2010; Zhang & Lin, 2006; Wong et al., 2019). However, the existing literature on COSSO primarily focuses on estimation, largely overlooking uncertainty. This is primarily due to the absence of a Bayesian interpretation for this frequentist approach. In contrast, a separate line of research has proposed a Bayesian framework employing additive non-parametric models (Linkletter et al., 2006; Kaufman & Sain, 2010; Scheipl et al., 2012; Chen & Liao, 2020; Wei et al., 2020; Curtis et al., 2014; Lu et al., 2022; Duvenaud et al., 2011; Durrande et al., 2012; 2013). These approaches often assign a Gaussian process (GP) or a semi-parametric model to each functional component. Some variants incorporate sparsity models to limit the number of relevant terms. While both of these approaches aim to tackle the same problem, there has been a notable lack of connection between these two research communities.

Contributions: This paper proposes a new comprehensive Bayesian framework for FANOVA problems, offering several key contributions:

- We show that our model is *highly flexible*, accommodating different levels of sparsity to adapt to real-world data complexity. Thus, we believe that our contribution has the potential to inspire future model developments that may have remained overlooked until now. This paper introduces one potential model, the Dirichlet, that showed its efficiency in ultra-sparse settings.
- We demonstrate that our new model *unifies* deterministic and Bayesian state-of-the-art methods into a single equation, facilitating their comparison within a common framework. Notably, we demonstrate that our model serves as the Bayesian counterpart to COSSO-like approaches, enabling the quantification of uncertainty in traditionally deterministic methods.
- We address various practical implementation challenges and offer insightful implementation tricks and

[1]Safran Tech, Digital Sciences & Technologies Department, Châteaufort, France [2]INRIA, DI/ENS, PSL Research University, France. Correspondence to: Yosra Marnissi <marnissi.yosra@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

guidelines, covering topics such as prior choice, hyper-parameter tuning, scalable learning and prediction.

2. Comprehensive Bayesian framework for FANOVA model

This section begins by reviewing the FANOVA framework. We then present our proposed model, demonstrate its flexibility in handling multi-level sparsity and provide guidelines for its tuning and inference.

2.1. FANOVA model

Let $(\mathbf{x}^{(n)}, y^{(n)})$, for $n \in \{1, \dots, N\}$, be the observed data: $y^{(n)}$ is the scalar output and $\mathbf{x}^{(n)} = [x_1^{(n)}, \dots, x_D^{(n)}]^\top$ is a D-dimensional vector of potential covariates. The goal is to infer the regression function f such that $y^{(n)} = f\left(\mathbf{x}^{(n)}\right) + \varepsilon^{(n)}$ where $\varepsilon^{(n)}$ is a zero-mean noise. Within the FANOVA framework (Wahba, 1990), f writes as a sum of contributions of covariates and combinations of covariates:

$$f(\mathbf{x}) = b + \sum_{i=1}^{r} \sum_{c_j \in C_D^i} f_{c_j}(\mathbf{x}_{c_j}) \tag{1}$$

where $r \leq D$ is the decomposition order, b is a constant, C_D^i is the i-th combination of covariates without repetition i.e., subsets c_j with i distinct elements from $\{1, \dots, D\}$, and $\mathbf{x}_{c_j}$ denotes the subset of covariates in c_j. This results in $J = \sum_{i=1}^{r} \#C_D^i$ functional terms (excluding the constant component b) where $\#C_D^i$ denotes cardinal of C_D^i.

To improve model interpretability and explainability, it is common to truncate the sequence in (1) to include only low-order interactions. For instance, with $r = 1$, we only consider *main effects*, which are components depending on individual covariates while for $r = 2$, the model also includes *first-order interaction effects*, which are functions depending on two covariates and so on. However, as the covariates number and the model order increase, the number of terms still increase as $O(D^r)$ making the estimation problem challenging. Moreover, the model complexity impedes the interpretability and the understanding of results. Fortunately, the number of relevant terms is often much smaller than the total number of components.

2.2. Proposed Global-Local shrinkage hierarchical GP

GP are powerful non-parametric tools to define prior distributions over functions when the regression function is not limited to simple parametric forms. In this paper, we propose the following hierarchical FANOVA GP model providing flexible sparsity information through latent variables:

$$\begin{aligned} &\boldsymbol{\Lambda} \sim \mathcal{P}(\boldsymbol{\Lambda}) \\ &\boldsymbol{\Theta} \sim \mathcal{P}_{\mathbf{W}}(\boldsymbol{\Theta}) \\ &\forall i \in \{1, \dots, r\} \forall c_j \in C_D^i \quad f_{c_j}|_{\lambda_i, \theta_{c_j}} \sim \mathcal{GP}\left(0, \lambda_i \theta_{c_j} k_{c_j}\right) \end{aligned} \tag{2}$$

Functional component depending on covariates from subset c_j follows a GP of zero mean and kernel $\lambda_i \theta_{c_j} k_{c_j}$ where k_{c_j} is a given unweighted kernel function deriving the subspace of f_{c_j}. $\boldsymbol{\Lambda} = [\lambda_1, \dots, \lambda_r]^\top$ and $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1^\top, \dots, \boldsymbol{\Theta}_r^\top]^\top$ where $\boldsymbol{\Theta}_i$ is the vector containing parameters θ_{c_j} associated to components within the order i i.e., $\boldsymbol{\Theta}_i = [\theta_{c_j}, c_j \in C_D^i]^T$. $\mathcal{P}_{\mathbf{W}}(\boldsymbol{\Theta})$ and $\mathcal{P}(\boldsymbol{\Lambda})$ are then prior distributions that are affected to these vectors respectively and $\mathbf{W}$ are tuning hyper-parameters. In this paper, we will use "$\mathrm{p}(.)$" to denote the density function associated to the distribution $\mathcal{P}$.

Model (2) can be seen as an extension of finite-dimensional Bayesian *global-local shrinkage models* (Tang et al., 2018) to an infinite-dimensional setting. In fact, model (2) contains two kinds of latent variables. Each parameter λ_i acts as a global indicator of the strength of active components in a given order i, while variables θ_{c_j} control the sparsity of functional components within the same order i. If θ_{c_j} is too small, the component c_j will approach zero, ensuring sparsity within the set of components of the same order i if almost all θ_{c_j} are close to zero. In contrast, if λ_i is too small, the information carried by the components in order i may be considered as not relevant. For these reasons, parameters λ_i will be denoted as *global scales* while parameters θ_{c_j} will be denoted as *local scales*. Subsection 2.3 will particularly show how this local-global formulation yields a remarkable flexibility in managing multi-level sparsity.

2.3. Embracing flexibility: multi-level sparsity through Local-Global scales

A large family of models can be defined with appropriate choices of $\mathcal{P}_{\mathbf{W}}(\boldsymbol{\Theta})$ denoted as the *"mixing density"*.[1] The simplest case is when θ_{c_j} are assumed all a priori independent i.e., $\mathrm{p}_{\mathbf{W}}(\boldsymbol{\Theta}) = \prod_{i=1}^{r} \prod_{c_j \in C_D^i} \mathrm{p}_{w_{c_j}}(\theta_{c_j})$. Table 3 of Appendix A.1 presents examples of mixing densities including *discrete* (e.g., Spike and Slab) and *continuous* (e.g., Horseshoe (Carvalho et al., 2009)) models. While discrete models offer the optimal representation (many components are exactly zero), many experiences have highlighted several computational difficulties compared to continuous models (see (Malsiner-Walli & Wagner, 2018) for an example). The advantage of using a continuous model is that it allows for recent gradient-based approaches whether for optimization or for Bayesian sampling. To enhance the shrinkage behavior of continuous models and better replicate the behavior of discrete ones, several solutions can be extended from the literature. One can *adopt an adaptive strategy* inspired by

[1] With analogy to scale mixture of Gaussian in finite-dimensional models.

adaptive LASSO (Zou, 2006): penalize each component differently using prior information about relevant components from a rough solution or using techniques like mutual information between covariates. The relevance of the obtained model depends on these estimated weights. One can also *add a further layer in the hierarchical model* by assigning priors to these weights and estimating them with the other latent parameters (Wei et al., 2020) which increases the number of unknown variables. In this paper, we propose an alternative solution using a *non-separable mixing density*.

2.3.1. Component shrinkage with non-separable mixing

We specifically assume local parameters θ_{c_j} within the same family i to be dependent. Then, there are different ways to construct mixing priors on these dependant local scales. A natural way is to extend standard mixing densities to construct multivariate models. For instance, one can replace the univariate Exponential with:

$$\mathrm{p}_{\mathbf{W}}(\boldsymbol{\Theta}) \propto \exp\left(-\sum_{i=1}^{r}\sqrt{\boldsymbol{\Theta}_i^\top \mathbf{W}_i^{-1}\boldsymbol{\Theta}_i}\right) 1_{[0,+\infty[}(\boldsymbol{\Theta}) \quad (3)$$

where $\mathbf{W}_i$ is a positive definite matrix that can introduce correlations among specific values within $\boldsymbol{\Theta}_i$. Even when $\mathbf{W}i$ takes a diagonal form (the model is then closely linked to the $\ell_{1,2}$ loss), the parameters within $\boldsymbol{\Theta}_i$ remain dependent, albeit decorrelated. However, determining the appropriate diagonal values of $\mathbf{W}_i$ that describe the relative magnitudes of local scales remains an open problem.

In this paper, we propose a new alternative approach for promoting dependent local scales using the *Dirichlet prior*:

$$\mathrm{p}_{\mathbf{W}_i}(\boldsymbol{\Theta}_i) \propto \prod_{c_j \in C_D^i} \theta_{c_j}^{w_{c_j}} \quad (4)$$

with $\sum_{c_j \in C_D^i} \theta_{c_j} = 1$. Hyperparameters $\mathbf{W}_i$ indicate the relevance of each component. In the case of a symmetric Dirichlet (where all elements of $\mathbf{W}_i$ equal a given w_i), this hyperparameter acts as a concentration parameter. As w_i increases, the energy distribution becomes more evenly spread among components within order i. Conversely, with a smaller concentration parameter, the energy is more sparsely distributed, and the majority of components approach zero. The limiting case if when the energy inside the order i is concentrated onto a single component.

What makes the Dirichlet mixing particularly interesting is its explicit connection to the number of relevant components. In fact, when $w_i \leq 1$, it may be seen as an approximation for the expected percentage of relevant components inside the order i. In essence, knowing the expected percentage of relevant components allows for an effective setting of this mixing model. Then, this new model offers a high degree of flexibility for tuning, as it enables the construction of a structured prior over local scales while only adjusting a single hyperparameter. This is particularly interesting when we have a prior information regarding the number of relevant components or we are searching for a low-dimensional representation with a target number of relevant components at each order. To the best of our knowledge, the combination of GP and Dirichlet-like non-separable shrinkage proposed in our work is novel and results in an interesting structured prior on functional components. The resulting model approximates the discrete shrinkage behavior while being continuous, which is not common in other models.

2.3.2. Covariate selection with non-separable mixing

In many cases, particularly with numerous covariates, exploring a second level of sparsity through covariate selection significantly improves model interpretability. Restricting the active functional components in the final model to depend on only a small number of covariates can yield desirable sparsity properties. Notably, sparsity on covariates often translates to sparsity on selected components. Our model (2) can effectively handle this level of sparsity by using a non-separable mixing that considers dependencies between components sharing covariates. This can be achieved again by incorporating a correlation matrix for components sharing covariates in the same spirit of (3). In this paper, we propose instead to add a layer of covariate selection. The resulting mixing probability writes in an hierarchical manner:

$$\begin{aligned} &\boldsymbol{\eta} \sim \mathcal{P}(\boldsymbol{\eta}) \\ \forall i \in \{1,\dots,r\} \forall c_j \in C_D^i \quad &\theta_{c_j}|_{\eta_l, l\in c_j} \sim \mathcal{P}_{\eta_l, l\in c_j} \end{aligned} \quad (5)$$

The first line introduces the covariate selection variable $\boldsymbol{\eta} = [\eta_1,\dots,\eta_D]^T$ which can be assigned a prior distribution, such as a Dirichlet prior that can be tuned according to the desired number of relevant covariates. The second line deals with the mixing model, which is adjusted based on the values of η_l associated with the same covariates on which the interaction component depends. For instance, the prior distribution of $\theta_{\{1,2\}}$ is parameterized by the values of η_1 and η_2. Note that (5) results in a non-separable mixing. This means that when we integrate (5) with respect to $\boldsymbol{\eta}$, it leads to a dependency between components that share the same covariates, due to dependencies on the same variables η_l. However, adding this new layer increases the dimensionality of the problem with D variables. One approach is to set θ_{c_j} to be a deterministic function of θ_l for all covariates l in the subset c_j such as the product function. Then, (5) turns to:

$$\begin{aligned} &\boldsymbol{\eta} \sim \mathcal{P}(\boldsymbol{\eta}) \\ \forall i \in \{1,\dots,r\} \forall c_j \in C_D^i \quad &\theta_{c_j} = \prod_{l\in c_j} \eta_l \end{aligned} \quad (6)$$

This setup enforces a *hard shrinkage* effect on the interaction components when the corresponding main effects are not relevant. In other words, if the main effects do not significantly contribute to the model, the associated interaction components are strongly penalized, ultimately leading to their suppression. The primary advantage of this model is the reduced number of unknown parameters, as all local scales are expressed solely in terms of D variables η_l. However, a drawback arises when a covariate lacks relevance on its own but gains significance when associated with another covariate through an interaction component. In such cases, this approach may overlook important interactions or overestimate non-relevant main components.

Alternatively, while allowing η_l to directly control the values of main effects components (they are equal to main effect local scales), we propose to use individual η_l values to control interaction scales through their prior mixing. Equation (5) can be rewritten as follows:

$$\begin{aligned} & \boldsymbol{\eta} \sim \mathcal{P}(\boldsymbol{\eta}) \\ \forall c_j \in C_D^1 \quad & \theta_{c_j} = \eta_{c_j} \\ \forall i \in \{2, \dots, r\} \forall c_j \in C_D^i \quad & \theta_{c_j}|_{\eta_l, l \in c_j} \sim \mathcal{P}_{\eta_l, l \in c_j} \end{aligned} \tag{7}$$

This approach allows for a *soft shrinkage* effect on the interactions, addressing the limitation of the previous model while keeping the same number of parameters as model (2).

2.3.3. Different global scales

Global scales indicate the overall energy in the interaction level, promoting a third level of sparsity across orders. In model (2), we have set *a distinct global scale for each interaction order*, thereby increasing the number of unknown parameters. However, using different global scales can be useful in high-order decomposition settings (r is large), effectively shrinking all components in non-relevant interaction orders to zero. Moreover, even with a low decomposition order, employing a different global scale per interaction order may be necessary when the energy is not uniformly distributed across orders, as it will be shown in Section 4.

2.4. Hyperparameters r and k_{c_j}

2.4.1. Othogonal kernel function

The FANOVA framework (1) allows us to measure the contribution of each component to the overall model. For reliable interpretability, we need each component to have a unique effect, without redundancy. To address this issue, the concept of *kernel orthogonal additive models* has been introduced in previous literature. Examples of orthogonal kernels ensuring orthogonal subspaces are (Lu et al., 2022; Durrande et al., 2012) that add constraints on standard kernels to achieve identifiability. Alternatively, we can use kernels related to the Sobolev space including zero-mean functions with proper first derivatives (Storlie et al., 2011). Expressions of these kernels are provided in Appendix A.2.

Given that standard kernels such as RBF or Matern are not suitable in such framework, implementation tricks like (Pleiss et al., 2018; Liu et al., 2020; Wilson & Nickisch, 2015; Gardner et al., 2018) that use kernel interpolation to reduce complexity in learning and prediction, are no longer applicable. It becomes essential to explore alternative approaches that align with orthogonal kernels. In this context, we have observed that orthogonal kernels described above have a particular structure: they write as the sum of a stationary kernel and linear kernels. Therefore, it is still possible to extend these interpolation tricks to orthogonal kernels. Details about this extension are provided in Appendix B.1.

2.4.2. Decomposition order

In real-world problems, determining the true decomposition order can be challenging. If the order is overestimated (resulting in overfitting), the global scale can detect this issue by shrinking all components in higher orders to zero. This is feasible because we use different global parameters for different interaction levels. However, it is more common to underestimate the interaction order for interpretability reasons. In such cases, adding a *residual component* can help account for all terms above the chosen order (see Appendix A.3 for the expression). When this residual component has a significant energy, it suggests that the chosen decomposition order should be revisited and potentially increased to capture missing information. This adjustment can then allow the removal of the residual component, leading to better interpretability. Note that many papers, such as (Lu et al., 2022), have shown that a small decomposition order seems to be sufficient with orthogonal kernels to achieve competitive accuracy.

2.5. Learning and prediction

Let Ω denote the set of unknown variables. Three schemes can be used to estimate model parameters. One approach is the *expanded scheme*, which analyzes the model in terms of its functional components f_{c_j}, that is $\boldsymbol{\Omega} = \{\tau, b, f_{c_j}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}\}$ where $\mathbb{E}[\|\varepsilon^{(n)}\|^2] = \tau$. Another method is *partial marginalization*, which instead considers the latent function in (1) i.e., $f = b + \sum_{i=1}^{r} \sum_{c_j \in C_D^i} f_{c_j}$, then $\boldsymbol{\Omega} = \{\tau, b, f, \boldsymbol{\Theta}, \boldsymbol{\Lambda}\}$. A third technique is *full marginalization*, eliminating functional components so that $\boldsymbol{\Omega} = \{\tau, b, \boldsymbol{\Theta}, \boldsymbol{\Lambda}\}$. The resulting distributions are provided in Table 4 of Appendix A.4.

Each learning scheme has its own advantages and limitations, depending on the problem and available data. Marginalization reduces considerably the number of unknowns parameters $\boldsymbol{\Omega}$. For instance, full marginalization has $J + r + 2$ unknown variables, while the expanded and

partially marginalized models have $(N+1)J+r+2$ and $N+J+r+2$ respectively. However, marginalization can also break conjugacy properties, potentially complicating computations like the ones in Gibbs sampling. It also relies on the assumption of orthogonal functional subspaces, necessitating the careful selection of orthogonal kernels.

Once the learning scheme is chosen, one should select the inference method to estimate $\mathrm{p}(\boldsymbol{\Omega}|_{\mathbf{X},\mathbf{y}})$ and then deduce predictive densities. There are two primary approaches considered either to compute point estimates of parameters and components, or use sampling where a set of samples are generated to approximate the whole posterior distribution (Filippone et al., 2013; Filippone & Engler, 2015; Pinder et al., 2020).

In this paper, we leverage dimension reduction and consider the fully marginalized scheme. Let $\mathbf{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$ and $\mathbf{y} = (y^{(1)}, \dots, y^{(N)})$. We define the marginalized kernel as $k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}} = \sum_{i=1}^{r} \lambda_i \sum_{c_j \in C_D^i} \theta_{c_j} k_{c_j}$. Then, the predictive density of the the latent function writes

$$\mathrm{p}(f|_{\mathbf{X},\mathbf{y}}) = \int \mathrm{p}(f|_{\mathbf{X},\mathbf{y},\boldsymbol{\Theta}}) \mathrm{p}(\boldsymbol{\Omega}|_{\mathbf{X},\mathbf{y}}) d\boldsymbol{\Omega} \tag{8}$$

Very often, a *plug-in* approach is used to approximate this integral (Teckentrup, 2020): an estimate $\hat{\boldsymbol{\Omega}}$ from $\mathrm{p}(\boldsymbol{\Omega}|_{\mathbf{X},\mathbf{y}})$ is computed and the predictive distribution reduces to a GP: $f|_{\mathbf{X},\mathbf{y}} \sim \mathcal{GP}(\mu^*_{\hat{\boldsymbol{\Omega}}}, k^*_{\hat{\boldsymbol{\Omega}}})$ where

$$\mu^*_{\boldsymbol{\Omega}}(.) = b + k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}(., \mathbf{X}) \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\Theta},\boldsymbol{\Lambda}} (\mathbf{y} - b)$$

$$k^*_{\boldsymbol{\Omega}}(.,.) = k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}(.,.) - k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}(., \mathbf{X}) \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\Theta},\boldsymbol{\Lambda}} k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}(\mathbf{X}, .)$$

and $\boldsymbol{\Sigma}_{\boldsymbol{\Theta},\boldsymbol{\Lambda}} = \tau \mathbf{I}_N + k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}(\mathbf{X}, \mathbf{X})$. We can also show that the predictive distribution of a component f_{c_j} in a given order i is a GP i.e., $f_{c_j}|_{\mathbf{X},\mathbf{y}} \sim \mathcal{GP}(\mu^*_{\boldsymbol{\Omega}_{c_j}}, k^*_{\boldsymbol{\Omega}_{c_j}})$ where

$$\mu^*_{\boldsymbol{\Omega}_{c_j}}(.) = \gamma_{c_j} k_{c_j}(., \mathbf{X}) \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\Theta},\boldsymbol{\Lambda}} (\mathbf{y} - b)$$

$$k^*_{\boldsymbol{\Omega}_{c_j}}(.,.) = \gamma_{c_j} k_{c_j}(.,.) - \gamma^2_{c_j} k_{c_j}(., \mathbf{X}_{c_j}) \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\Theta},\boldsymbol{\Lambda}} k_{c_j}(\mathbf{X}_{c_j}, .)$$

where $\mathbf{X}_{c_j}$ denotes the subset of $\mathbf{X}$ in c_j and $\gamma_{c_j} = \lambda_i \theta_{c_j}$. In case of sampling, the marginal density $\mathrm{p}(\boldsymbol{\Omega}|_{\mathbf{X},\mathbf{y}})$ is approximated with S samples and the predictive distribution of the latent function (and components) writes as a mean and scale mixture of S GPs: $f|_{\mathbf{X},\mathbf{y}} \sim \frac{1}{S} \sum_{s=1}^{S} \mathcal{GP}(\mu^*_{\boldsymbol{\Omega}_s}, k^*_{\boldsymbol{\Omega}_s})$.

The resulting computational complexity in learning and prediction is discussed in Appendix B.2.

3. A unified framework

In this section, we will show how the flexibility offered by our model allows us to express almost state-of-the-art ANOVA approaches within the same equation (2).

3.1. Bayesian FANOVA approaches

Table 5 of Appendix B.3 shows how many Bayesian FANOVA models in the literature are special cases of our model (2). For example, additive variable selection such as (Scheipl et al., 2012; Antonelli & Dominici, 2018) fall under our model with $r = 1$. Rather than explicitly stating how each method is an instance of our model (as in Table 5), this section takes a more informative approach by categorizing these methods according to model tuning and learning.

3.1.1. Tuning-based categorization

Model tuning involves the selection of hyperparameters (namely decomposition order, kernel choice, mixing model), which are at the user's choice. Differences in tuning can break down to two points. The first one is the *Global-Local Scales Dilemma*. Most of Bayesian literature do not differentiate between local and global shrinkage. They typically employ local scales to weight each component differently while using a single global scale for all decomposition orders e.g., (Vo & Pati, 2017; Reich et al., 2009; Tang et al., 2023). We will show in our experiments that weighting all orders equally can be problematic especially when true relevant components are not distributed equally across orders. Other works have instead employed a different global scale per order but components within the same order are weighted equally e.g., (Durrande et al., 2012; Lu et al., 2022). However, our experiments will highlight the importance of local scales as they help the model to better adapt data complexity where some components might have a more significant influence and also ensure a more parsimonious solution in terms of relevant components. Only few works proposed joint covariate and component selection. For instance, several works promote their models as a form of variable selection by making slight adjustments to the kernel function. For example, selection may occur at the level of the lengthscale of the RBF kernel ensuring the automatic selection of covariates on which each component relies (Vo & Pati, 2017). In this regard, such an approach does not merely restrict the model to depend on a few set of covariates. Instead, it enables to infer a fixed number of components without specifying the order, as the model can directly infer the appropriate order. Then, this work can not be viewed as a covariate selection model but rather an automatic inference of the order. The only work we found that enables covariate selection is (Agrawal & Broderick, 2023) which is a particular instance of our generic model (5) using the hard shrinkage model (6). The second point of difference between Bayesian approaches is the *choice of global and local priors*. Most of works use discrete models and employ standard kernels for their GP models such as linear (Agrawal et al., 2019) and RBF (Tang et al., 2023). Only a few works discuss the importance of orthogonal kernels (Lu et al., 2022; Reich et al., 2009).

3.1.2. Learning-based categorization

Model learning focuses on the estimation procedure of the resulting model's unknown parameters (e.g., local/global scales). Differences in learning can also break down to two main points namely the *learning scheme* (expanded, partially or fully marginalized) and the *inference method* (plug-in, sampling etc.) as presented in Section 2.5. Most works rely on computing point estimations by alternate updates between parameters and latent function using a partially marginalized approach (Agrawal & Broderick, 2023; Tang et al., 2023). Discrete parameters are mainly updated using sampling steps. Some works employ the plug-in approach e.g., (Lu et al., 2022). In a fully Bayesian framework, the expanded scheme is typically used with Metropolis-Hasting or Gibbs (Reich et al., 2009; Kaufman & Sain, 2010).

3.2. Deterministic FANOVA approaches

There has also been much work on penalized methods for FANOVA models. Typically, f is estimated by minimizing a penalized function within a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$. Specifically, each functional component in (1) resides in a subspace $\mathcal{H}_{cj}$ of the orthogonal decomposition $\mathcal{H} = \bigoplus_{c_j} \mathcal{H}_{c_j}$. A prominent kernel solution in this context is the COSSO method (Lin & Zhang, 2003) which extends the standard kernel ridge solution known as Smoothing Spline ANOVA (SS-ANOVA) (Wahba, 1990; Gu & Gu, 2013) to component selection. In the same spirit of LASSO, COSSO modifies the traditional quadratic norm, replacing it with the (pseudo-) norm of the RKHS. The solution is the minimizer of the following loss function:

$$\frac{1}{N}\sum_{n=1}^{N}\left(y^{(n)} - f(\mathbf{x}^{(n)})\right)^2 + \alpha \sum_{i=1}^{r}\sum_{c_j \in C_D^i} \omega_{c_j} \|P_{c_j} f\|_{\mathcal{H}} \tag{9}$$

where P_{c_j} is the orthogonal projection of f on $\mathcal{H}_{c_j}$, $\omega_{c_j} \geq 0$ and $\alpha > 0$ are tuning hyperparameters. Note that when ω_{c_j} is tuned to penalize components differently, the solution is known as Adaptive COSSO (ACOSSO), in the same spirit as Adaptive LASSO. Some extensions of (9) have been also proposed for instance by replacing the ℓ_1 loss with another penalization such as $\ell_{1,2}$ (Radchenko & James, 2010). Methods for solving (9) can be found in (Touzani & Busby, 2013; Storlie et al., 2011).

While all Bayesian approaches in Section 3.1 are specific instances of our generic proposed model, no direct connection with deterministic approaches like COSSO has been discussed in these works. This lack of connection can be attributed to two primary reasons. First, despite the well-established link between GP and kernel ridge (Kanagawa et al., 2018), the use of sparse penalization obscures the explicit expression of COSSO-like solutions in terms of a GP and complicates their Bayesian interpretation. Second, there is another alternative in inference that was not explored in Bayesian non-parametric models which is the *marginalized posterior estimation*. This method involves integrating out kernel parameters, typically local scales, according to their prior models. This provides a different estimator from the one given by plug-in or alternate optimization. This unexplored alternative may also explain why a direct connection with deterministic approaches such as COSSO has not been established thus far, as we will show in Proposition 3.1.

Proposition 3.1. *The hierarchical GP model (2) under equal global scales and an exponential mixing for local scales, has as marginal Maximum A Posteriori (MAP) estimate, the COSSO solution (minimizer of (9)).*

Proof. We derive a tangent majorant for the negative marginal posterior likelihood of functional components and show that maximizing this tangent majorant is equivalent to COSSO solution. Details are provided in Appendix B.4. □

Extension to other COSSO models (Ravikumar et al., 2009; Radchenko & James, 2010) is straightforward (see Appendix B.4). This result is interesting as we have now a Bayesian representation of deterministic approaches, allowing to derive additional interesting quantities such as predictive uncertainty intervals.

4. Experiments

4.1. Set-up

This section studies the performance of the proposed model with different mixing priors in terms of **component selection** and **prediction accuracy** on simulated and real data. We particularly consider from Table 3 the proposed Bayesian COSSO (Exponential) and the Dirichlet models. Additional experiments using the Student't and the Horseshoe models are provided in Appendix C.3. Bayesian models were tested using two approaches: a plug-in with MAP estimation of parameters, and sampling using Hamiltonian Monte Carlo (HMC) (Hensman et al., 2015). To isolate the model's performance from hyper-parameters selection, we consider for our proposed models the same Sobolev kernel as in SS-ANOVA and COSSO. Comparisons with respect to the state-of-the-art are also provided (see Table 1)[2].

Metrics: The prediction accuracy is reported in terms of *root mean square error* (**RMSE**). To identify the most significant components contributing to the predicted model, we rely on variance analysis. The *number of active components* (**NAC**) corresponds to the minimal number of components required to achieve at least 99% of the total variance. For simulated data, since we know the true relevant components,

[2]Code is provided in https://github.com/ymarnissi/bayesfanova_paper.

Table 1. Benchmark methods. The term "Adaptive" refers to techniques whose mixing/weights were tuned based on SS-ANOVA.

Acronym	Method
SS	SS-ANOVA (Wahba, 1990)
COS.	COSSO (Lin & Zhang, 2003)
ACOS.	Adaptive COSSO (ACOSSO) (Storlie et al., 2011)
MARS	Mars (Friedman, 1991)
RBF-Ok	Additive models with orthogonal kernel (Lu et al., 2022)
BCOS.MAP	Bayesian COSSO with Plug-in (**ours**)
BCOS.HMC	Bayesian COSSO with sampling (**ours**)
Dir.MAP	Dirichlet model with Plug-in (**ours**)
Dir.HMC	Dirichlet model with sampling (**ours**)
ABCOS.MAP	Bayesian Adaptive COSSO with Plug-in (**ours**)
ABCOS.HMC	Bayesian Adaptive COSSO with sampling (**ours**)
ADir.MAP	Adaptive Dirichlet model with Plug-in (**ours**)
ADir.HMC	Adaptive Dirichlet model with sampling (**ours**)

the model selection performance is assessed using *True Positives* (correctly selected components), *False Positives* (incorrectly selected components), and *False Negatives* (missed components), which we use to compute the *F1 score*. For real data, only the NAC metric is used, with a smaller NAC indicating stronger model selection power. For Bayesian models, we also assess the accuracy of the estimated credible intervals by calculating the *coverage probability* (**Cover**). The latter represents the proportion of test points for which the true function lies within the estimated 95% credible interval. In order to get normalized metrics, we report in this section the **relative metric** with respect to the baseline method SS-ANOVA as follows: $m_{\text{rel}} = \frac{m - m_{\text{SS}}}{m_{\text{SS}}}$ where m is the metric (RMSE, F1, NAC) of the compared model and m_{SS} is the one obtained with SS-ANOVA.

Simulated data: We consider two examples with $r = 1$ and $r = 2$ and generate data using different realizations (Independent Uniform, Compound symmetry, Trimmed Autoregressive) with different correlation and noise levels (see Appendix C.1). Each experiment is repeated 10 times. The objective was to assess model performance with respect to mixing models, noise levels and covariate correlation. For simplification purposes and fair comparison, we restrict our experimental study on models that do not need initial estimates to tune their hyperparameters. Then, adaptive models are not included in these simulations. Detailed results for various simulated scenarios are presented in Tables 6 and 7 in the Appendices. Figure 1 offers a summarized view of these results, representing an aggregate of all conducted experiments using relative metrics. We also provide in Table 2 results where we investigate covariate selection and compare hard and soft shrinkage. Note that hard shrinkage is similar to the model in (Agrawal & Broderick, 2023).

Real data: We perform several regression tasks on UCI datasets using 5-fold cross-validation focusing on first-order interactions ($r = 2$) and compare to state-of-the-art including adaptive models. Information on prior and algorithm settings are provided in Appendix C.2. Figure 2 summarizes results on UCI data from Tables 8 and 9 in the Appendices.

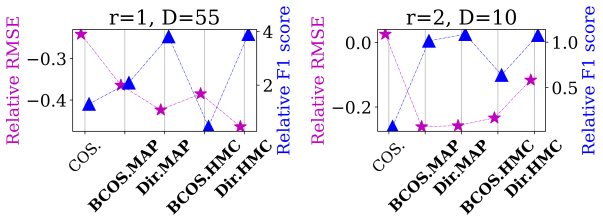


Figure 1. Relative RMSE and F1 score with respect to SS-ANOVA. Results are averaged over 210 runs including different noise and correlation levels. Individual results are provided in Tables 6 and 7 in the appendices. Lower relative RMSE indicate better predictions while higher relative F1 scores mean superior shrinkage.

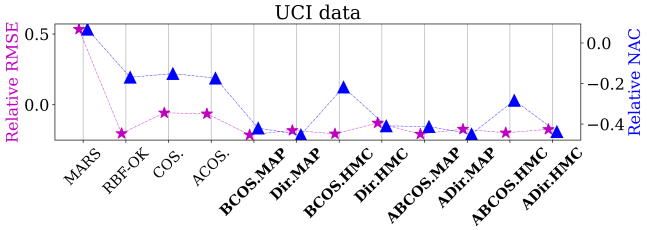


Figure 2. Relative RMSE and NAC with respect to SS-ANOVA. Results are averaged over all folds and datasets. Individual results are provided in Tables 8 and 9 in the appendices. Lower relative RMSE and NAC are indicative of better performance.

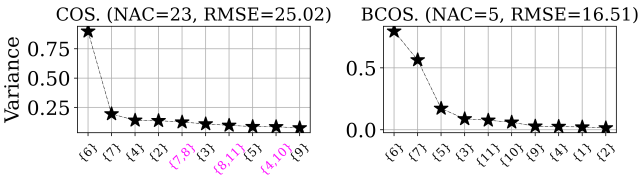


Figure 3. Most relevant components in the "Demand" dataset for COSSO and Bayesian COSSO. Using a unique global scale seems to introduce interaction components (in red) in COSSO not presented in Bayesian COSSO.

Table 2. Results with and without covariate selection with $r = 2$, $D = 30$ covariates and only 7 relevant components $f_{\{1\}}$, $f_{\{2\}}$, $f_{\{3\}}$, $f_{\{4\}}$, $f_{\{1,2\}}$, $f_{\{1,3\}}$, $f_{\{3,5\}}$ from a total of 465 components.

Metric	Without covariate selection	Hard shrinkage	Soft shrinkage
RMSE	1.66 ± 0.12	1.43 ± 0.42	1.46 ± 0.12
True Positives	5.60±0.30	5.00±0.50	5.40±0.40
False Positives	60.30±10.23	8.30±8.87	18.81±11.35
False Negatives	1.40±0.30	2.00±0.50	1.60±0.40

4.2. Results and discussions

In the following, we organize the key takeaways from our experiments into different sections.

Prediction accuracy and selection power are related: Figures 1 and 2 show that the proposed Bayesian models outperform state-of-the-art methods. In simulated data, the Dirichlet mixing exhibits superior performance by selecting fewer incorrect covariates. The reconstruction error being sensitive not only to correctly chosen covariates but also to noise introduced by false positives, sparse models like Bayesian Dirichlet, while occasionally missing one correct component, consistently outperform SS-ANOVA. Experiments on real data further validate the effectiveness of sparse models, achieving competitive performance with just a few components. For instance, in the "Housing" dataset, sparse models reduce by half the number of active components while achieving comparable RMSE to SS-ANOVA (see Table 8). Additionally, adapting mixing models using weights tuned from the SS-ANOVA estimator generally improves results. However, when the SS-ANOVA estimator is not accurate, this adaptation worsens estimates in several datasets. In this regard, using non-separable mixing models for improved shrinkage, such as the Dirichlet model, emerges as a more effective alternative.

The Dirichlet Model: intuitive and effective with strong prior knowledge or for dimensionality reduction: Additional experiments of Appendix C.3 support our findings from Figures 1 and 2. Beyond its discrete-like shrinkage behavior, the Dirichlet prior offers intuitive hyperparameter tuning, allowing direct control over the percentage of relevant components. In contrast, alternative priors like Exponential and Horseshoe require more complex tuning and lack the same straightforward interpretation of hyperparameters. However, the choice of the Dirichlet model should be carefully considered based on our objectives. It can be effective when a strong prior about the number of relevant components exists or when aiming for a low-dimensional representation with minimal components. Otherwise, it may not always optimize accuracy or prediction performance. In cases where the primary objective is accurate prediction or the sparsity level needs to be automatically inferred from data, alternative soft models like COSSO may be more suitable. It is important to mention that we also observed that algorithms with Dirichlet model requires slightly more iterations to converge compared to other independent models. This may be explained by the correlation between local scales within the Dirichlet model, which can slow down mixing and convergence especially in sampling.

Plug-in yields sparser solutions while Sampling provides more accurate uncertainty: Plug-in and fully Bayesian approaches yield comparable results. However, the plug-in approach consistently shows the best F1 score across simulated data. This can be attributed to its smallest number of false positives in most cases (see Tables 6 and 7). This observation aligns with established sparse reconstruction principles in finite dimensions, where the MAP estimate used in the plug-in approach tends to be more parsimonious compared to estimates based on sample means. But, the fully Bayesian approach appears to provide more accurate uncertainty intervals. In fact, the plug-in approach approximates $\mathrm{p}(\boldsymbol{\Omega}|_{\mathbf{y}})$ as a Dirac mass in a single point while the fully Bayesian approach incorporates the entire posterior distribution, providing a more reliable uncertainty. This observation is less evident in the UCI dataset (see Table 9), where both methods exhibit similar coverage results. This suggests that the marginal density of local and global parameters in the UCI data is well concentrated around the maximum, allowing it to be effectively represented by a Dirac distribution on the MAP estimators, as in plug-in.

Ultra-sparse models can accommodate correlated covariates: Despite a slight performance decrease for dependent covariates (see Tables 6 and 7), sparse models remain robust, effectively capturing relevant features. While SS-ANOVA tends to select more covariates in high correlation scenarios leading to increased false positives, COSSO, Bayesian COSSO, and Dirichlet models achieve good results with low false positives due to sparsity constraints. Although they may miss some relevant components with low energy, their ability to maintain low false positives is noteworthy.

Both global and local shrinkage are important: The observed differences between COSSO and Bayesian COSSO, particularly in the second simulated example ($r = 2$), stem from the Bayesian models' adaptive global shrinkage parameter. Contrary to SS-ANOVA and COSSO, which apply uniform penalties to both main and interaction effects, our Bayesian models dynamically adjust the global shrinkage parameter based on the observed energy in each order. This adaptation is advantageous when relevant components are unevenly distributed across orders. For instance, in the second example ($r = 2$), COSSO results align more closely with SS-ANOVA, possibly due to the inclusion of non-irrelevant interaction components in COSSO's estimate. In contrast, Bayesian models, with their adaptive global parameter, offer improved performance by effectively capturing relevant components and reducing the impact of irrelevant ones. For example, in the "Demand" UCI dataset (Figure 3), Bayesian COSSO primarily identifies main effects as relevant, while COSSO includes some fictitious interaction components primarily due to the global parameter. Enhancing the original COSSO by assigning different weights to the interaction order and repeating the experiments is possible. However, this introduces additional hyperparameters to tune through cross-validation. Similarly, when comparing RBF-OK (using a single local scale) to our proposed models, the importance of employing heterogeneous local scales that adapt the relevance of each component becomes evident. In fact, while both methods yield comparable RMSE in Figure 2, our models select fewer components.

Covariate selection is important, but caution is needed with hard shrinkage: Table 2 highlights the importance of covariate selection in providing a more parsimonious representation with minimal false positives. This emphasizes the importance of using structured models, such as the covariate selection model, in high-dimensional settings. However the hard shrinkage model often fails to identify component $f_{3,5}$ due to its dependence on the non-relevant main effect covariate x_5. Conversely, the soft shrinkage model shows the most balanced trade-off, achieving a low false positive rate while maintaining good accuracy. While the considered covariate selection model shows promising results in terms of parsimony, it also reveals the need for further improvement in capturing complex relationships with covariates such as a covariate that is only relevant when associated with another one through an interaction component.

5. Conclusion and future works

This paper offers a comprehensive framework for FANOVA models, with a large flexibility allowing to unify a broad range of models into a single equation, enabling for their comprehension and comparison within the same Bayesian framework. This has the particular advantage of quantifying uncertainty for traditionally deterministic models.

A limitation of our model is that the orthogonality of functional subspaces assumes implicitly independent covariates. Non-independent covariates can lead to non-identifiability. Our experiments showed that ultra-sparse models may address this issue. In fact, correlations between covariates can negatively affect support recovery: an inactive covariate with high correlation to an active one is more likely to be identified by a sparse model. However, handling non-independent features can be challenging in high-dimensional settings. Future work could extend our approach to non-independent input features and examine the effect of non-Gaussian noise and missing data using latent variable models. Another interesting research direction is scaling our method to large numbers of covariates and decomposition order in both estimation and prediction tasks.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agrawal, R. and Broderick, T. The SKIM-FA Kernel: High-Dimensional Variable Selection and Nonlinear Interaction Discovery in Linear Time. *Journal of Machine Learning Research*, 24(27):1–60, 2023.

Agrawal, R., Trippe, B., Huggins, J., and Broderick, T. The kernel interaction trick: Fast bayesian discovery of pairwise interactions in high dimensions. In *International Conference on Machine Learning*, pp. 141–150. PMLR, 2019.

Antonelli, J. and Dominici, F. A Bayesian semiparametric framework for causal inference in high-dimensional data. *arXiv preprint arXiv:1805.04899*, 2018.

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3):493–508, 2015.

Cao, J., Guinness, J., Genton, M. G., and Katzfuss, M. Scalable gaussian-process regression and variable selection using vecchia approximations. *The Journal of Machine Learning Research*, 23(1):15799–15828, 2022.

Carvalho, C. M., Polson, N. G., and Scott, J. G. Handling sparsity via the horseshoe. In *Artificial intelligence and statistics*, pp. 73–80. PMLR, 2009.

Chastaing, G. and Le Gratiet, L. ANOVA decomposition of conditional Gaussian processes for sensitivity analysis with dependent inputs. *Journal of Statistical Computation and Simulation*, 85(11):2164–2186, 2015.

Chen, C. and Liao, Q. ANOVA Gaussian process modeling for high-dimensional stochastic computational models. *Journal of Computational Physics*, 416:109519, 2020.

Curtis, S. M., Banerjee, S., and Ghosal, S. Fast Bayesian model assessment for nonparametric additive regression. *Computational Statistics & Data Analysis*, 71:347–358, 2014.

Durrande, N., Ginsbourger, D., and Roustant, O. Additive covariance kernels for high-dimensional Gaussian process modeling. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 21, pp. 481–499, 2012.

Durrande, N., Ginsbourger, D., Roustant, O., and Carraro, L. ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*, 115:57–67, 2013.

Duvenaud, D. K., Nickisch, H., and Rasmussen, C. Additive gaussian processes. *Advances in neural information processing systems*, 24, 2011.

Fang, Z., Kim, I., and Schaumont, P. Flexible variable selection for recovering sparsity in nonadditive nonparametric models. *Biometrics*, 72(4):1155–1163, 2016.

Filippone, M. and Engler, R. Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE). In *International Conference on Machine Learning*, pp. 1015–1024. PMLR, 2015.

Filippone, M., Zhong, M., and Girolami, M. A comparative evaluation of stochastic-based inference methods for Gaussian process models. *Machine Learning*, 93(1): 93–114, 2013.

Friedman, J. H. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. Gpytorch: Blackbox matrix-matrix Gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.

Ginsbourger, D., Roustant, O., Schuhmacher, D., Durrande, N., and Lenz, N. On anova decompositions of kernels and gaussian random field paths. In *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, pp. 315–330. Springer, 2016.

Girolami, M. and Calderhead, B. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Gu, C. and Gu, C. *Smoothing spline ANOVA models*, volume 297. Springer, 2013.

Hensman, J., Matthews, A. G., Filippone, M., and Ghahramani, Z. MCMC for variationally sparse gaussian processes. *Advances in Neural Information Processing Systems*, 28, 2015.

Hu, Z. and Dey, D. K. Generalized variable selection algorithms for gaussian process models by lasso-like penalty. *Journal of Computational and Graphical Statistics*, pp. 1–24, 2023.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Kaufman, C. G. and Sain, S. R. Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149, 2010.

Lin, Y. and Zhang, H. H. *Component selection and smoothing in smoothing spline analysis of variance models–COSSO*. 2003.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Ye, K. Q. Variable selection for Gaussian process models in computer experiments. *Technometrics*, 48(4): 478–490, 2006.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. When Gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems*, 31 (11):4405–4423, 2020.

Lu, X., Boukouvalas, A., and Hensman, J. Additive Gaussian Processes Revisited. In *International Conference on Machine Learning*, pp. 14358–14383. PMLR, 2022.

Malsiner-Walli, G. and Wagner, H. Comparing spike and slab priors for Bayesian variable selection. *arXiv preprint arXiv:1812.07259*, 2018.

Piironen, J. and Vehtari, A. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Artificial Intelligence and Statistics*, pp. 905–913, 2017.

Pinder, T., N., C., and Leslie, D. Stein variational gaussian processes. *arXiv preprint arXiv:2009.12141*, 2020.

Pleiss, G., Gardner, J., Weinberger, K., and Wilson, A. G. Constant-time predictive distributions for Gaussian processes. In *International Conference on Machine Learning*, pp. 4114–4123. PMLR, 2018.

Polson, N. G. and Scott, J. G. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

Radchenko, P. and James, G. M. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553, 2010.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.

Reich, B. J., Storlie, C. B., and Bondell, H. D. Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. *Technometrics*, 51(2):110–120, 2009.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

Scheipl, F., Fahrmeir, L., and Kneib, T. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532, 2012.

Storlie, C. B., Bondell, H. D., Reich, B. J., and Zhang, H.-H. Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica*, 21(2): 679, 2011.

Tang, T., Mak, S., and Dunson, D. Hierarchical shrinkage gaussian processes: applications to computer code emulation and dynamical system recovery. *arXiv preprint arXiv:2302.00755*, 2023.

Tang, X., Xu, X., Ghosh, M., and Ghosh, P. Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya A*, 80(2):215–246, 2018.

Teckentrup, A. L. Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1310–1337, 2020.

Timonen, J., Mannerström, H., Vehtari, A., and Lähdesmäki, H. LGPR: an interpretable non-parametric method for inferring covariate effects from longitudinal data. *Bioinformatics*, 37(13):1860–1867, 2021.

Touzani, S. and Busby, D. Smoothing spline analysis of variance approach for global sensitivity analysis of computer codes. *Reliability Engineering & System Safety*, 112:67–81, 2013.

Vo, G. and Pati, D. Sparse additive gaussian process with soft interactions. *Open Journal of Statistics*, 7(04):567, 2017.

Wahba, G. *Spline models for observational data*. Society for industrial and applied mathematics, 1990.

Wei, R. *Bayesian variable selection using continuous shrinkage priors for nonparametric models and non-Gaussian data*. PhD thesis, North Carolina State University, 2017.

Wei, R., Reich, B. J., Hoppin, J. A., and Ghosal, S. Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures. *Statistica Sinica*, 30(1):55–79, 2020.

Wilson, A. and Nickisch, H. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International conference on machine learning*, pp. 1775–1784. PMLR, 2015.

Wong, R. K., Li, Y., and Zhu, Z. Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, 114(525):406–418, 2019.

Ye, M., Ren, T., and Liu, Q. Stein self-repulsive dynamics: Benefits from past samples. *Advances in Neural Information Processing Systems*, 33:241–252, 2020.

Zhang, F., Chen, R.-B., Hung, Y., and Deng, X. Indicator-based bayesian variable selection for gaussian process models in computer experiments. *Computational Statistics & Data Analysis*, 185:107757, 2023.

Zhang, H. H. and Lin, Y. Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica*, pp. 1021–1041, 2006.

Zhu, H., Vannucci, M., and Cox, D. D. A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics*, 66(2):463–473, 2010.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

A. Bayesian FANOVA background

A.1. Mixing densities

Given N realizations of covariates $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})$, the conditional distribution of each component $f_{c_j}(\mathbf{X}) \in \mathbb{R}^N$, is a multivariate Gaussian distribution:

$$f_{c_j}(\mathbf{X}_{c_j})|_{\lambda_i, \theta_{c_j}, \mathbf{X}} \sim \mathcal{N}\left(0, \lambda_i \theta_{c_j} k_{c_j}(\mathbf{X}_{c_j}, \mathbf{X}_{c_j})\right) \tag{10}$$

where $k_{c_j}(\mathbf{X}_{c_j}, \mathbf{X}_{c_j}) \in \mathbb{R}^{N \times N}$ is the Gram matrix associated to the kernel k_{c_j} computed on covariates c_j of $\mathbf{X}$.

In the following, we will refer to **the marginal prior density** of each component as the prior density that has been integrated with respect to local shrinking parameters. This amounts to computing the integral of (10) with respect to $\mathcal{P}_{\mathbf{W}}(\mathbf{\Theta})$:

$$\mathrm{p}\left(f_{c_j}(\mathbf{X}_{c_j})|_{\lambda_i, \mathbf{X}}\right) \propto \int \exp\left(-\frac{f_{c_j}(\mathbf{X}_{c_j})^\top k_{c_j}(\mathbf{X}_{c_j}, \mathbf{X}_{c_j})^{-1} f_{c_j}(\mathbf{X}_{c_j})}{2\lambda_i \theta_{c_j}}\right) \theta_{c_j}^{-\frac{N}{2}} \mathrm{p}_{\mathbf{W}}(\mathbf{\Theta}) d\mathbf{\Theta} \tag{11}$$

Table 3 presents examples of mixing densities and the resulting models. Note that, for some (independent) mixing densities, the marginal prior (11) has a closed form. This is for example the case, when using a Bernoulli mixing, an Exponential mixing or an Inverse Gamma mixing models.

Table 3. Examples of mixing distributions.

Mixing density	**(Log) marginal prior**	**Proposed name**
$\mathcal{P}_{w_{c_j}}\left(\theta_{c_j}\right) = \text{Exp}\left(\frac{w_{c_j}^2}{2}\right)$	$\frac{w_{c_j}}{\sqrt{\lambda_i}}\sqrt{f_{c_j}\left(\mathbf{X}_{c_j}\right)^T k_{c_j}\left(\mathbf{X}_{c_j}, \mathbf{X}_{c_j}\right)^{-1} f_{c_j}\left(\mathbf{X}_{c_j}\right)}$	Exponential (Bayesian COSSO)
$\mathcal{P}_{w_{c_j}}\left(\theta_{c_j}\right) = \text{InvGamma}\left(\frac{\nu}{2}, \frac{\nu w_{c_j}^2}{2}\right)$	$\log\left(1 + \frac{w_{c_j}}{\nu\lambda_i} f_{c_j}\left(\mathbf{X}_{c_j}\right)^T k_{c_j}\left(\mathbf{X}_{c_j}, \mathbf{X}_{c_j}\right)^{-1} f_{c_j}\left(\mathbf{X}_{c_j}\right)^T\right)$	Student't with degree of freedom $\nu > 0$
$\mathcal{P}_{w_{c_j}}\left(\sqrt{\theta_{c_j}}\right) = \text{HalfCauchy}\left(w_{c_j}^2\right)$	Not explicit	Horshoe
$\mathcal{P}_{w_{c_j}}\left(\theta_{c_j}\right) = \beta\delta_0 + (1-\beta)\delta_{w_{c_j}}$	$f_{c_j}\left(\mathbf{X}_{c_j}\right) \sim \beta\delta_0 + (1-\beta)\mathcal{N}\left(0,\ \lambda_i w_{c_j} k_{c_j}\left(\mathbf{X}_{c_j}, \mathbf{X}_{c_j}\right)\right)$	Spike and Slab
$\mathcal{P}_{w_{c_j}}\left(\theta_{c_j}\right) = \beta\delta_0 + (1-\beta)\mathcal{U}(0, w_{c_j})$	Not explicit	Spike and Slab Uniform
$\mathcal{P}_{w_{c_j}}\left(\sqrt{\theta_{c_j}}\right) = \beta\delta_0 + (1-\beta)\text{HalfCauchy}\left(w_{c_j}^2\right)$	Not explicit	Spike and Slab Horsehoe
$\mathcal{P}_{\mathbf{W}}\left(\mathbf{\Theta}\right) = \text{Dirichlet}\left(\mathbf{W}^2\right)$ (see Equation (4))	Not explicit	Dirichlet

A.2. Expressions of Orthogonal kernel functions

A.2.1. Sobolev kernel

The $(K-1)^{th}$-order of Sobolev space S_{K-1} includes only functions that integrate to 0 and have $K-1$ proper derivatives:

$$S_{K-1} = \left\{g | g^{(1)}, \ldots, g^{(K-2)} \text{ are absolutely continuous, } \int_0^1 g(t)dt = 0 \text{ and } g^{(K-1)} \in L^2[0,1]\right\} \tag{12}$$

where $g^{(k)}$ is the k^{th} derivative of g. Most of papers working with this RKHS, set the corresponding reproducing kernel to

$$k_1(s,t) = \frac{B_{2K}\left(|s-t|\right)}{(-1)^{K+1}(2K)!} + \sum_{i=1}^{K} \frac{B_i(s)B_i(t)}{(i!)^2} \tag{13}$$

where B_i is the i^{th} Bernoulli polynomial. One can select (13) as a kernel function for the Gaussian processes modeling the main effect components while for first interaction components, one can select the kernel product

$$k_2\left((s_1,t_1),\ (s_2,t_2)\right) = k_1\left(s_1,t_1\right) k_1\left(s_2,t_2\right) \tag{14}$$

It is worth to note that in some papers, the employed Sobolev kernel is slightly different from the one in (13). In fact, one can use the same weighting coefficient in the non-stationary part of (13) i.e

$$k_1(s,t) = \frac{B_{2K}\left(|s-t|\right)}{(-1)^{K+1}(2K)!} + \frac{1}{c}\sum_{i=1}^{K} B_i(s)B_i(t) \tag{15}$$

for some positive constant c. The latter controls the linear for $K = 1$ (and the quadratic for $K = 2$) trends. Adjusting c to a high value gives vague yet priors to the linear/ quadratic trends. For the first order interaction effects with (15), the authors in (Reich et al., 2009) recommend adding a correction instead of using the kernel product as follows:

$$k_2\left((s_1,t_1),\,(s_2,t_2)\right) = k_1\left(s_1,t_1\right)k_1\left(s_2,t_2\right) + \left(\frac{1}{c}-1\right)k_{1,2}\left(s_1,t_1\right)k_{1,2}\left(s_2,t_2\right) \tag{16}$$

where $k_{1,2}(s,t) = \sum_{i=1}^{K} B_i(s)B_i(t)$. It is an important to mention that the Sobolev kernel is defined for covariates in the range of $[0,1]$. Therefore, when working with real datasets where the covariates may not naturally fall within this range, a transformation is required before applying the Sobolev kernel.

A.2.2. Orthogonal kernel from (Lu et al., 2022)

In order to ensure orthogonality, modifications are applied to conventional kernels, such as the squared exponential kernel, also known as the RBF kernel. The RBF kernel, denoted as $k_1(s,t) = \exp\left(-\frac{(s-t)^2}{2l^2}\right)$, is parameterized with a length scale hyperparameter l. To satisfy the orthogonality constraints, slight adjustments are made to these kernels. Specifically, constraints are imposed to enforce the integral of each function component with respect to the input measure to be zero (Durrande et al., 2012). When considering the RBF kernel with the orthogonality constraints, a closed-form expression can be obtained for the constrained kernel when the input density follows a Gaussian distribution or is approximated by a Gaussian mixture, uniform distribution, categorical distribution, or the empirical distribution. Assuming a Gaussian input measure, where each covariate $x^{(n)}$ is drawn from $\mathcal{N}(\mu,\delta)$, the constrained RBF kernel can be expressed as follows:

$$k_1(s,t) = \exp\left(-\frac{(s-t)^2}{2l^2}\right) - \frac{l\sqrt{l^2+2\delta^2}}{l^2+\delta^2}\exp\left(-\frac{(s-\mu)^2+(t-\mu)^2}{2(l^2+\delta^2)}\right) \tag{17}$$

To satisfy the assumption of a Gaussian input density, the authors in (Lu et al., 2022) propose the use of a normalizing flow technique. This approach involves transforming the continuous input features to approximate a Gaussian density through a sequence of bijective transformations. The parameters of these transformations are learned by minimizing the Kullback-Leibler divergence between a standard Gaussian distribution and the transformed input data. Once the parameters are determined, they are kept fixed, and the orthogonal kernel model is fitted on the transformed data with approximate Gaussian densities.

For interaction orders, the authors use product kernel functions similarly to (14).

A.3. Residual functional component

The residual component aims to account for non-considered interactions and prevent a loss in prediction accuracy. The corresponding kernel for this residual component can be expressed as follows when using a second order decomposition (main and first order interaction effects) (Reich et al., 2009):

$$k\left(\mathbf{s},\mathbf{t}\right) = \prod_{i=1}^{D}\left(1+k(s_i,\,t_i)\right) - 1 - \sum_{i=1}^{D} k_1\left(s_i,t_i\right) - \sum_{1\le l<m\le D} k_2\left((s_l,t_l)\,,\,(s_m,t_m)\right) \tag{18}$$

The expression for higher decomposition order is straightforward.

A.4. Learning schemes

Table 4. Different learning schemes. We define the marginalized kernel as $k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}} = \sum_{i=1}^{r} \lambda_i \sum_{c_j \in C_D^i} \theta_{c_j} k_{c_j}$.

	Expanded model	**Partially marginalized model**	**Marginalized model**
Dimension	$(N+1)J + r + 2$	$N + J + r + 2$	$J + r + 2$
Unknown Variables	$\boldsymbol{\Omega}_0 = \{\, \tau,\ b, f_{c_j}, \boldsymbol{\Theta}, \boldsymbol{\Lambda} \}$	$\boldsymbol{\Omega}_1 = \{\, \tau,\ b, f, \boldsymbol{\Theta}, \boldsymbol{\Lambda} \}$	$\boldsymbol{\Omega}_2 = \{\, \tau,\ b, \boldsymbol{\Theta}, \boldsymbol{\Lambda} \}$
$\mathrm{p}\left(\boldsymbol{\Omega} \middle\vert \mathbf{X}, \mathbf{y}\right) \propto$	$\exp\left(-\frac{1}{2\tau}\left\|\mathbf{y} - b - \sum_{j=1}^{J} \mathbf{f}_{c_j}(\mathbf{X})\right\|^2\right) \times$	$\tau^{-\frac{N}{2}} \exp\left(-\frac{1}{2\tau}\left\|\mathbf{y} - \mathbf{f}(\mathbf{X})\right\|^2\right) \times$	$\exp\left(-\frac{1}{2}(\mathbf{y} - b.1)^\top \left(\tau I_N + k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}(\mathbf{X},\mathbf{X})\right)^{-1}(\mathbf{y} - b.1)\right)$
	$\mathrm{p}(\tau)\mathrm{p}(b)\mathrm{p}_{\mathbf{W}}(\boldsymbol{\Theta})\,\mathrm{p}(\boldsymbol{\Lambda}) \times$	$\det\left(k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}(\mathbf{X},\mathbf{X})\right)^{-\frac{1}{2}} \times$	$\times \det\left(\tau I_N + k'_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}(\mathbf{X},\mathbf{X})\right)^{-\frac{1}{2}} \times$
	$\prod_{i=1}^{r}\prod_{c_j \in C_D^i} \theta_{c_j}^{-\frac{N}{2}} \lambda_{c_j}^{-\frac{N}{2}} \det\left(k_{c_j}\left(\mathbf{X}_{c_j}, \mathbf{X}_{c_j}\right)\right)^{-\frac{1}{2}} \times$	$\exp\left(-\frac{(\mathbf{f}(\mathbf{X})-b.1)^\top k'^{-1}_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}(\mathbf{X},\mathbf{X})(\mathbf{f}(\mathbf{X})-b.1)}{2}\right) \times$	$\mathrm{p}(\tau)\mathrm{p}(b)\mathrm{p}_{\mathbf{W}}(\boldsymbol{\Theta})\,\mathrm{p}(\boldsymbol{\Lambda})$
	$\tau^{-\frac{N}{2}} \exp\left(-\frac{\mathbf{f}_{c_j}(\mathbf{X})^\top k_{c_j}\left(\mathbf{X}_{c_j}, \mathbf{X}_{c_j}\right)^{-1} \mathbf{f}_{c_j}(\mathbf{X})}{2\theta_{c_j}\lambda_i}\right)$	$\mathrm{p}(\tau)\mathrm{p}(b)\mathrm{p}_{\mathbf{W}}(\boldsymbol{\Theta})\,\mathrm{p}(\boldsymbol{\Lambda})$	

B. Additional theoretical details

B.1. Highlighting the structured representation of orthogonal kernels and extension of interpolation methods

Let k_o be one the orthogonal kernels presented in Section A.2. One can see that it can be expressed as the sum of a stationary part, denoted by $k_s(t,u)$, and a linear part that involves a weighted sum of functions $v_i(t)v_i(u)$ for $i=1$ to L. The representation of k_o can be written as:

$$k_o(t,u) = k_s(t,u) + \sum_{i=1}^{L} \alpha_i v_i(t) v_i(u) \tag{19}$$

where α_i are some weights.

In the case of the orthogonal RBF kernel, the stationary part k_s corresponds to the standard RBF kernel, $L=1$, and the linear part involves $\alpha_1 = -\frac{l\sqrt{l^2+2\delta^2}}{l^2+\delta^2}$ and $v_1(.) = \exp\left(-\frac{(.-\mu)^2}{2(l^2+\delta^2)}\right)$. Similarly, for the Sobolev kernel, $k_s(t,u)$ and $v_i(t)$ are computed based on Bernoulli polynomials, on the distance $|t-u|$ and t, respectively.

Let's consider the Structured Kernel Interpolation (SKI) as an example (Pleiss et al., 2018; Liu et al., 2020; Wilson & Nickisch, 2015; Gardner et al., 2018). Given a SKI approximation for the stationary kernel

$$k_s(\mathbf{X},\mathbf{X}) \approx \boldsymbol{\omega}_{\mathbf{X}}^{\top} k_s(\mathbf{U},\mathbf{U}) \boldsymbol{\omega}_{\mathbf{X}}$$

where $\mathbf{U} \in \mathbb{R}^M$ are the interpolation points and $\boldsymbol{\omega}_{\mathbf{X}} \in \mathbb{R}^{M\times N}$ is the local interpolation weight matrix, we can maintain the same representation for the orthogonal kernel by setting:

$$k_{or}(\mathbf{X},\mathbf{X}) \approx V_{\mathbf{X}}^{\top} \mathbf{A}(\mathbf{U},\mathbf{U}) V_{\mathbf{X}}$$

where $\mathbf{A}(\mathbf{U},\mathbf{U}) \in \mathbb{R}^{(M+L)\times(M+L)}$ is the bloc diagonal matrix formed by two blocks: $k_s(\mathbf{U},\mathbf{U})$ and the diagonal matrix whose elements are α_i while $V_{\mathbf{X}} \in \mathbb{R}^{(L+M)\times N}$ is the matrix that results on the concatenation of $\boldsymbol{\omega}_{\mathbf{X}}$ and $v_i(\mathbf{X})$.

It is worth noting that since we have only computed interpolation approximation for stationary kernels (linear kernels are not approximated), the accuracy of the approximation depends solely on the interpolation accuracy of the stationary part of the orthogonal kernel.

B.2. Computational complexity

B.2.1. Matrix-vector product with interpolation of orthogonal kernels for main and interaction effects

It can be noted that, L in (19) is significantly smaller than M, with $L=1$ for the orthogonal RBF kernel and typically 1 or 2 for the Sobolev kernel. With this representation, the computational cost for a matrix-vector product remains similar to that of a stationary kernel, approximately $g(M) = O(N + M\log M)$, and the storage requirement is approximately $O(N+M)$.

Main effects can be expressed as the sum of individual effects for each input dimension i.e., $k_{main}(\mathbf{X},\mathbf{X}) = \lambda_1 \sum_{j=1}^{D} \theta_{\{j\}} k_{or}(\mathbf{X}_{\{j\}},\mathbf{X}_{\{j\}})$. Exploiting this additive structure, the cost of a matrix vector product for main effects is $O(Dg(M))$. Furthermore, the main effect kernel can still be represented in the same way as the individual effects by combining the individual orthogonal kernels into a single block-diagonal matrix. For interactions, leveraging the kernel's product structure and interpolated individual effect representation reduces matrix-vector multiplication complexity. Techniques like those in (Gardner et al., 2018) allow linear scaling with the number of terms in the product.

B.2.2. Computational complexity in learning

In a fully marginalized model, the number of unknown variables, while reduced, still scales as $O(D^r)$. However, by limiting to lower orders like $r=1$ or 2 for interpretability (as discussed in Section 2.4), we maintain a small number of unknowns. With a low decomposition order, the primary computational complexity during each iteration arises from manipulating $\boldsymbol{\Sigma}_{\boldsymbol{\Theta},\boldsymbol{\Lambda}}$, a weighted sum of Gram matrices. To mitigate this, we can employ black-box matrix-matrix multiplication, possibly incorporating structured representations of orthogonal kernels, to reduce time complexity. Tasks like inverse-matrix vector multiplication, trace, or logarithm of determinants, common in learning algorithms, can be efficiently approximated using a single optimization algorithm call, as described in (Gardner et al., 2018).

B.2.3. COMPUTATIONAL COMPLEXITY IN PREDICTION

Structured kernel representation can be employed to use (Pleiss et al., 2018). This is particularly useful for main effects or latent function in additive models ($r = 1$). This results in constant-time predictions regardless of the test dataset size after an initial pre-computation phase, which consumes $O(N + M \log M)$ time and requires $O(M)$ storage. For a sampling approach, this complexity scales linearly with the number of samples S.

B.3. Link to Bayesian additive models

Various Bayesian additive models proposed in the literature for selecting functional predictors can indeed be viewed as particular cases of the proposed model (2). Table 5 details this link for different models. It can be also noted that covariate selection models such as (Zou, 2006; Scheipl et al., 2012; Antonelli & Dominici, 2018) as well as semi-parametric regression with additive models such as the Multivariate Additive Regression Splines (Friedman, 1991) are also particular examples of our models.

Table 5. State of the art Bayesian approaches are particular instances of the proposed framework.

Method	**Tuning**				**Solution**	**Description**
	Global	*Local*	*Kernel*	*Order*		
(Vo & Pati, 2017)	N/A	ℓ^1 norm	$\begin{cases} \exp\left(-\sum_{l\in c_j}\tau_{jl}\lvert t_l-u_l\rvert^2\right) & \text{if all } \tau_{jl}>0 \\ 0 & \text{otherwise} \end{cases}$	Not specified	- Expanded scheme - MCMC for τ_{jl} - Optimization of conditional density of θ_{c_j} subject to τ_{jl}	- Parameters τ_{jl} are another layer for component shrinkage - The maximal order can be also inferred from τ_{jl} - No global scales are used to weight orders differently - Can be seen as extension of variable selection models (Zhang et al., 2023; Cao et al., 2022; Fang et al., 2016; Bobb et al., 2015; Hu & Dey, 2023) to additive GP models - Variable selection model using a different kernel (Linkletter et al., 2006) can be extended similarly to additive GP models
(Lu et al., 2022)	- Heterogeneous - Gamma prior	N/A	Orthogonal RBF	Specified	- Fully marginalized scheme - Plug-in approach	- No local scales are used to weight components differently - Can be seen as instances of GP additive models (Duvenaud et al., 2011) using orthogonal kernels - Can be seen as extension of additive covariance kernel (Durrande et al., 2012) to FANOVA framework
(Agrawal & Broderick, 2023)	- Heterogeneous - No prior	Bernoulli Uniform mixture	Linear kernel	Specified	- Partially marginalized scheme - Plug-in approach - Cross validation	- Function shrinkage is obtained by covariate selection using hard shrinkage (6) - Considers all interactions between selected covariates: does not assume sparsity of interactions between selected covariates - Can be seen as extension of additive kernel (Agrawal et al., 2019) (using Horseshoe mixing and Sampling) to higher interactions
(Reich et al., 2009)	N/A	Bernoulli Horseshoe mixture	Sobolev kernel	Specified	- Expanded scheme - Sampling	- Sampling is done with a Gibbs sampler updating all components sequentially which can be computationally expensive - Includes a residual component - Similar models (Scheipl et al., 2012; Curtis et al., 2014; Antonelli & Dominici, 2018) for r=1 and (Wei, 2017; Timonen et al., 2021) for r=2
(Tang et al., 2023)	N/A	Spike and Slab	Standard kernels	Specified	- Partially marginalized scheme - Sampling	- Sampling is done with a Gibbs sampler - Similar approaches (Zhu et al., 2010)
(Kaufman & Sain, 2010)	N/A	N/A	Standard kernels	Specified	- Expanded scheme - Sampling	- All (non-constant) components are weighted equally

B.4. Proof of Proposition 3.1

Let us assume that $\mathrm{p}(\lambda_i) = \delta_{\lambda\tau}$ (the same shrinkage global parameter for all the functional components) and that θ_{c_j} is unknown and $\mathrm{p}_{w_{c_j}}(\theta_{c_j})$ is the Exponential model from Table 3 (replacing $w_{c_j}^2$ with $\alpha_1 w_{c_j}^2/\tau$ for convenience purposes), then, rewrite the minus logarithm of the marginal posterior using the marginal prior of target components and after discarding all the deterministic variable priors and normalize with $N/2\tau$:

$$\mathcal{R}(f_{c_1}(\mathbf{X}_{c_1}),\ldots,f_{c_J}(\mathbf{X}_{c_J})) = \frac{1}{N}\|\mathbf{y}-b-\sum_{j=1}^{J} f_{c_j}(\mathbf{X}_{c_j})\|^2 + \alpha_N \sum_{j=1}^{J} w_{c_j}\sqrt{f_{c_j}(\mathbf{X}_{c_j})^\top k_{c_j}(\mathbf{X}_{c_j},\mathbf{X}_{c_j})^{-1} f_{c_j}(\mathbf{X}_{c_j})} + C_N \tag{20}$$

where $\alpha_N = 1/N\sqrt{\lambda\tau}$ and C_N is a constant independent of the target function. It can be noticed that the objective function (20) reduces to an adaptive group Lasso penalized problem.

In the following, we will demonstrate that the minimizer of (20) is the COSSO solution given in (Touzani & Busby, 2013; Storlie et al., 2011).

Lemma B.1. *Let us define*

$$\mathcal{M}(f_{c_1}(\mathbf{X}_{c_1}),\ldots,f_{c_J}(\mathbf{X}_{c_J}),\beta_{c_1},\ldots,\beta_{c_J}) = \frac{1}{N}\|\mathbf{y}-b-\sum_{j=1}^{J} f_{c_j}(\mathbf{X}_{c_j})\|^2 + \alpha_{0,N}\sum_{j=1}^{J}\beta_{c_j}^{-1} f_{c_j}^\top(\mathbf{X}_{c_j}) k_{c_j}(\mathbf{X}_{c_j},\mathbf{X}_{c_j})^{-1} f_{c_j}(\mathbf{X}_{c_j}) + \alpha_{1,N}\sum_{j=1}^{J} w_{c_j}^2\beta_{c_j} + \quad s.t\ \beta_{c_j}\geq 0 \quad \forall j \tag{21}$$

Then, there exist, $\alpha_{0,N}$ and $\alpha_{1,N}$ such that $\mathcal{M}$ is a tangent majorant of (20).

Proof. To simplify notation, let us denote by $\mathbf{u}_{c_j} = f_{c_j}(\mathbf{X}_{c_j})$. Since $\beta_{c_j}\geq 0$, then we can write β_{c_j} as a function of some vector $\mathbf{u}'_{c_j}$ (one possible value of $f_{c_j}(\mathbf{X}_{c_j})$) such that: $\beta_{c_j} = \sqrt{\mathbf{u}_{c_j}'^\top k_{c_j}(\mathbf{X}_{c_j},\mathbf{X}_{c_j})^{-1}\mathbf{u}'_{c_j}} \geq 0$. We can then write (21) equivalently as

$$\mathcal{M}(\mathbf{u}_{c_1}\ldots,\mathbf{u}_{c_J},\mathbf{u}'_{c_1}\ldots,\mathbf{u}'_{c_J}) = \frac{1}{N}\|\mathbf{y}-b-\sum_{j=1}^{J}\mathbf{u}_{c_j}\|^2 + \sum_{j=1}^{J}\mathcal{M}_{c_j}(\mathbf{u}_{c_j},\mathbf{u}'_{c_j}) \tag{22}$$

where

$$\mathcal{M}_{c_j}(\mathbf{u}_{c_j},\mathbf{u}'_{c_j}) = \alpha_{0,N}\frac{\mathbf{u}_{c_j}^\top k_{c_j}(\mathbf{X}_{c_j},\mathbf{X}_{c_j})\mathbf{u}_{c_j}}{\sqrt{\mathbf{u}_{c_j}'^\top k_{c_j}(\mathbf{X}_{c_j},\mathbf{X}_{c_j})^{-1}\mathbf{u}'_{c_j}}} + \alpha_{1,N} w_{c_j}^2\sqrt{\mathbf{u}_{c_j}'^\top k_{c_j}(\mathbf{X}_{c_j},\mathbf{X}_{c_j})^{-1}\mathbf{u}'_{c_j}} \tag{23}$$

We need to demonstrate that there exist some values of $\alpha_{0,N}$ and $\alpha_{1,N}$ such that $\mathcal{M}_{c_j}$ is a tangent majorant of $\mathcal{R}$ i.e

$$\mathcal{M}(\mathbf{u}_{c_1}\ldots,\mathbf{u}_{c_J},\mathbf{u}'_{c_1}\ldots,\mathbf{u}'_{c_J}) \geq \mathcal{R}(\mathbf{u}_{c_1}\ldots,\mathbf{u}_{c_J}) \quad \forall \mathbf{u}'_{c_J} \tag{24}$$

and that

$$\mathcal{M}(\mathbf{u}_{c_1}\ldots,\mathbf{u}_{c_J},\mathbf{u}'_{c_1}\ldots,\mathbf{u}'_{c_J}) = \mathcal{R}(\mathbf{u}_{c_1}\ldots,\mathbf{u}_{c_J}) \quad \text{for} \quad \mathbf{u}'_{c_j} = \mathbf{u}_{c_j} \quad \forall j \tag{25}$$

Let $v_{c_j} = \sqrt{\mathbf{u}_{c_j}^\top k_{c_j}(\mathbf{X}_{c_j},\mathbf{X}_{c_j})^{-1}\mathbf{u}_{c_j}} \geq 0$. We only need to show that given some values of $\alpha_{0,N}$ and $\alpha_{1,N}$ we have

$$\alpha_{0,N} v_{c_j}^2\beta_{c_j}^{-1} + \alpha_{1,N} w_{c_j}^2\beta_{c_j} \geq \alpha w_{c_j} v_{c_j} \quad \forall \beta_{c_j}\geq 0 \tag{26}$$

When $\beta_{c_j} = 0$, this implies $v_{c_j} = 0$ and then the inequality (26) holds with the assumption $0/0 = 0$.

For $\beta_{c_j} > 0$, we consider the second order polynomial $P = \alpha_{1,N} w_{c_j}^2 \beta_{c_j}^2 - \alpha w_{c_j} v_{c_j} \beta_{c_j} + \alpha_{0,N} v_{c_j}^2$. If the discriminant of polynomial P is negative, then the polynomial is always positive for all $\beta_{c_j} > 0$. It follows that if $\alpha_{1,N} = \dfrac{\alpha_N^2}{4\alpha_{0,N}}$ then (24) holds. □

Corollary B.2. *The minimizer of* (20) *with respect to* $f_{c_1}(\mathbf{X}_{c_1}), \ldots, f_{c_J}(\mathbf{X}_{c_J})$ *is equivalent to minimizing* (21) *with respect to* $f_{c_1}(\mathbf{X}_{c_1}), \ldots, f_{c_J}(\mathbf{X}_{c_J}), \beta_{c_1}, \ldots, \beta_{c_J}$ *where* $\alpha_{1,N} = \alpha_N^2 / 4\alpha_{0,N}$. *That is*

$$\underset{f_{c_1}(\mathbf{X}),\ldots,f_{c_J}(\mathbf{X}),\beta_{c_1},\ldots,\beta_{c_J}}{\textit{Argmin}} \mathcal{M}(f_{c_1}(\mathbf{X}), \ldots, f_{c_J}(\mathbf{X}), \beta_{c_1}, \ldots, \beta_{c_J}) = \underset{f_{c_1}(\mathbf{X}_{c_1}),\ldots,f_{c_J}(\mathbf{X}_{c_J})}{\textit{Argmin}} \mathcal{R}(f_{c_1}(\mathbf{X}_{c_1}), \ldots, f_{c_J}(\mathbf{X}_{c_J})) \tag{27}$$

Corollary B.3. *A solution for* (21) *(and equivalently for* (20)*) can be obtained through an alternating optimization process involving* f_{c_j} *and* β_{c_j}. *This process simplifies to solving a standard SS-ANOVA problem when* β_{c_j} *is fixed, and a nonnegative garrote optimization problem when* f_{c_j} *is fixed which is equivalent to the estimation strategy envolved in COSSO (Touzani & Busby, 2013; Storlie et al., 2011). Then, the MAP estimator of the hierarchical model* (2) *with an exponential mixing density reduces to the COSSO solution.*

Remark B.4. It is worth to note that additional parameters β_{c_j} play the role of selection parameters similarly to local shrinkage parameters θ_{c_j}. However, there is no reason that these two variables are equal. In fact, unlike θ_{c_j}, variables β_{c_j} have no Bayesian interpretation and are only added to the initial model for computational reasons.

Remark B.5. Very often, the work in (Reich et al., 2009), using a model similar to the Spike and Slab Horseshoe in Table 3, is considered in the literature as the Bayesian counterpart of COSSO. But this is actually a wrong connection since both methods are based on different models as we have shown in Proposition 3.1.

Extension to other COSSO models is straightforward. For instance, we show in the following proposition how (Ravikumar et al., 2009) is a particular instance of our proposed model.

Proposition B.6. *The hierarchical GP model* (2) *with* $r = 2$ *under equal global scales and mixing model* (28) *for local scales has as marginal MAP estimate, the solution in (Ravikumar et al., 2009).*

$$\begin{aligned}
\theta &\sim \textit{Exp}\left(\frac{w_2^2}{2}\right) \\
\alpha_{\{l\}} &\sim \textit{Exp}\left(\frac{w_1^2}{2}\right) \\
\theta_{\{l\}} &= \frac{1}{\frac{1}{\theta} + \frac{1}{\alpha_{\{l\}}}} \\
\theta_{\{l,l'\}} &= \theta
\end{aligned} \tag{28}$$

Proof. The marginal negative logarithm of the posterior density writes:

$$\frac{w_1}{\sqrt{\lambda}} \sum_{l=1}^{D} \sqrt{\psi_l} + \frac{w_2}{\sqrt{\lambda}} \sqrt{\sum_{l=1}^{D} \psi_l + \sum_{l=1}^{D-1} \sum_{l'=2}^{D} \psi_{l,l'}} \tag{29}$$

where

$$\psi_l = f_{\{l\}}\left(\mathbf{X}_{\{l\}}\right)^T k_{\{l\}}\left(\mathbf{X}_{\{l\}}, \mathbf{X}_{\{l\}}\right)^{-1} f_{\{l\}}\left(\mathbf{X}_{\{l\}}\right)$$

and

$$\psi_{l,l'} = f_{\{l,l'\}}\left(\mathbf{X}_{\{l,l'\}}\right)^T k_{\{l,l'\}}\left(\mathbf{X}_{\{l,l'\}}, \mathbf{X}_{\{l,l'\}}\right)^{-1} f_{\{l,l'\}}\left(\mathbf{X}_{\{l,l'\}}\right)$$

which can be seen as a ℓ_1 loss on main effects and $\ell_{1,2}$ on main and interactions effects which is similar to the penalization proposed in (Ravikumar et al., 2009). □

Remark B.7. In equation (20), we considered the marginal prior of the functional component, which is marginalized with respect to local scales. We did not consider the joint prior because our goal was to find the maximum a posteriori (MAP) estimate of the function. It is important to note the distinction between the marginal density $\mathrm{p}(f(\mathbf{X})|_{\mathbf{y}})$ and the joint

posterior density $\mathrm{p}(f(\mathbf{X}), \theta|_{\mathbf{y}})$. The marginal density $\mathrm{p}(f(\mathbf{X})|_{\mathbf{y}})$ represents the integrated probability over all possible values of the parameters θ, given the observed data $\mathbf{y}$. In other words, it accounts for the uncertainty in the local scales parameters. On the other hand, the joint posterior density $\mathrm{p}(f(\mathbf{X}), \theta|_{\mathbf{y}})$ represents the probability distribution over both the functional component $f(\mathbf{X})$ and the parameters θ, given the data $\mathbf{y}$. While we can derive the posterior mean estimate directly from the augmented model as $E_{f(\mathbf{X})|_{\mathbf{y}}}[f(\mathbf{X})] = E_{\theta|_{\mathbf{y}}}[E_{f(\mathbf{X})|_{\theta,\mathbf{y}}}[f(\mathbf{X})]]$, this is not necessarily true for the MAP estimate. The MAP estimate seeks to find the maximum of the marginal posterior density, which does not always correspond to the maximum of the joint posterior density:

$$\underset{f(\mathbf{X})}{\text{Argmax}}\ \mathrm{p}(f(\mathbf{X})|_{\mathbf{y}}) \neq \underset{f(\mathbf{X}),\theta}{\text{Argmax}}\ \mathrm{p}(f(\mathbf{X}), \theta|_{\mathbf{y}})$$

Therefore, if we minimize the negative logarithm of the joint posterior density with respect to the functional components and local scales, using techniques such as alternate minimization, it will result in a different estimator from the COSSO solution due to the distinct objectives they optimize.

C. Additional experiments and implementation details

C.1. Simulated examples set-up

C.1.1. Simulated examples in independent and dependent settings

We consider the following functions to generate our simulated data that take values on [0, 1]: $g_1(t) = t$, $g_2(t) = (2t-1)^2$, $g_3(t) = \dfrac{\sin(2\pi t)}{2-\sin(2\pi t)}$ and $g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos(2\pi t) + 0.5\sin^3(2\pi t)$. We use these functions to generate 2 target responses, one with only main effects and one with main and first order interaction effects:

$$f_1(x) = 5g_1(x_1) + 3g_2(x_2) + 4g_3(x_3) + 6g_4(x_4)$$

$$f_2(x) = f_1(x) = 5g_1(x_1) + 3g_2(x_2) + 4g_3(x_3) + 6g_4(x_4) + 4g_3(x_1x_2) + 6g_2\left(\frac{x_1+x_3}{2}\right) + 4g_1(x_3x_5)$$

We also follow (Lin & Zhang, 2003) and consider 3 realizations of training data:

- Uniform: draw covariates independently from a Uniform distribution on [0,1].
- Compound symmetry: draw $U, Z_1, \ldots, Z_D$ from a Uniform distribution on [0,1] then set the covariate as follows $x_i = \frac{Z_i + tU}{1+t}$ for some $t > 0$. Then, correlation $(x_i,\ x_j) = \frac{t^2}{1+t^2}$ for $i \neq j$.
- Trimmed Auto-Regressive (AR): draw $Z_1, \ldots, Z_D$ from a Gaussian distribution with zero mean and unit variance then set $x_1 = Z_1$, and $x_i = \rho x_{i-1} + \left(1-\rho^2\right)^{0.5} Z_i$ for $i > 1$, trim the covariates in $[-2.5,\ 2.5]$ and scale them to belong to $[0.1]$. Then, correlation $(x_i,\ x_j) \approx \rho$ for $i \neq j$.

We consider both models f_1 and f_2 with different scenarios to generate train data, namely the uniform, the component symmetry and the trimmed AR models. For f_1 , we set $D = 55$ and for f_2 we set $D = 10$. Then, for both examples, we have 55 functional components. With the first function, we have only 4 relevant components while for the second, we have 4 active main components and 3 active interaction components.

Benchmark methods :As benchmarking, we consider frequentist approaches SS-ANOVA (Wahba, 1990) and COSSO (Lin & Zhang, 2003). For Bayesian techniques, we focus on the proposed Dirichlet model. To provide a comprehensive benchmark, we also present results from the exponential model from Table 3, which serves as the Bayesian counterpart to COSSO. For simplification purposes and fair comparison, we restrict our experimental study on models that do not need initial estimates to tune their hyperparameters. Then, ACOSSO and Bayesian hierarchical models with heterogeneous hyperparameters w_i are not included in this section. We also consider the same kernel of all methods namely the Sobolev kernel with $K = 1$.

Details about prior settings: For the constant term b we adopt a Gaussian prior distribution with a mean of 0 and a variance of 10 while for the noise variance, we employ Half-Cauchy priors with a scale of 10. It is important to note that the priors for the mean and noise variance are intentionally chosen to be vague, aligning our Bayesian approach with the considered deterministic methods. This decision ensures that these parameters are solely informed by the data, without additional information from prior beliefs. As for the global scales, we acknowledge the significant impact of priors on the model's shrinkage behavior, a topic that has been extensively studied in the literature for finite dimension shrinkage problems (Polson & Scott, 2012; Piironen & Vehtari, 2017). In our experiments, we opt for vague distributions for these global parameters (a HalfCauchy prior with a large scale), indicating an absence of informative priors, similar to the reasons stated earlier. While this choice may not optimize performance, it allows for a direct comparison between deterministic and Bayesian methods, focusing solely on the effect of priors on local parameters (e.g., Dirichlet vs. Exponential). For the Dirichlet model, we set $w_i = 0.5$ for main effects and $w_i = 0.4$ for interaction effects. For Bayesian COSSO, we set $w_i = 1/D$ for main effects and $w_i = 1/D(D-1)$ for interaction effects. By employing these parameters, we aim to achieve consistent scales for the local parameters across different models. In fact, in main effects, which encompass D components, local scales generated from a Dirichlet prior have a sum equal to 1. Consequently, the prior mean for each main scale becomes $1/D$ for main effects and $1/D(D-1)$ for interactions effects.

Practical implementation: The selection policy adopted in these experiments relies on the variance analysis to identify the most significant components that contribute to the predicted model. Under the orthogonality assumption of the FANOVA

decomposition, the total variance of the target function is the sum of variances of all the summands. Therefore, the contribution of a component to the outcome can be measured in terms of the normalized variance: $\widetilde{V}_{c_j} = \frac{V_{c_j}}{\sum_{c_j} V_{c_j}}$ where V_{c_j} is the variance of component c_j. The higher this quantity is, the more informative is assumed to be the component. In order to select the components that we should include in the model, we search for the minimal number of components that achieves $t\%$ of the total energy where t is a threshold. We provide results for $t = 0.99$.

For estimation, we consider a sampling approach based on the Hamiltonian Monte Carlo (HMC) algorithm (Girolami & Calderhead, 2011). We specifically use the probabilistic programming framework Pyro [3]. For comparison, we also tested Stein Variational Gradient Descent (SVGD) (Pinder et al., 2020; Ye et al., 2020) but the results showed that while SVGD struggled to converge to the target parameter space, particularly with poor initialization, HMC performed well with the fully marginalized model. For the HMC algorithm, we use 5,000 iterations as burn-in. Prediction of the latent function is performed using 1,000 samples.

Results: Tables 6 and 7 provide the obtained results for each function with different scenarios (Independent, Dependant CS and Dependant AR) and different levels of noises. Each simulation study is replicated 10 times and then prediction metrics are averaged. For each experiment, we have highlighted the best method in bold based on the predictive results.

C.1.2. FANOVA MODELS WITH COVARIATE SELECTION

We consider the function f_2 with $D = 30$ covariates, where we have a total of 465 functional components but only 7 of them are relevant.

Details about prior settings: We compare three different models:

- Dirichlet prior on both main and interaction effects: we put a Dirichlet prior on both the main effects and the interaction effects with hyperparameter $w_i = 0.1$.
- Covariate selection model with the hard shrinkage model: we use a Dirichlet prior as the mixing density for the main effects. For the interactions, we adopt a hard shrinkage approach (6) which is similar to the model adopted in (Agrawal & Broderick, 2023).
- Covariate selection model with the soft shrinkage model: we use a Dirichlet prior as the mixing density for the main effects. For the interactions, we adopt a soft shrinkage approach (7).

Practical implementation: To evaluate the performance of the models, we conduct experiments using independent training data with $N = 500$ and assume a noise variance of $\tau = 1$. In order to make a fair comparison, we scale the local scales of the interaction components in the hard shrinkage model to have a sum equal to 1, which allows us to use the same global shrinkage parameter prior for both models. We use a Plug-in approach with a MAP estimator for the parameters. Each experiment is repeated 10 times to assess the consistency of the results.

C.2. UCI data set-up

We conducted regression experiments on UCI datasets using 5-fold cross-validation splits. We restrict our analysis to the first order interaction. It is worth to note that our assumption of restricting our models to low order interactions for the considered datasets can be motivated by the work in (Lu et al., 2022).

Benchmark methods: As benchmark models, we considered frequentist methods such as SS-ANOVA (Wahba, 1990), COSSO (Lin & Zhang, 2003), ACOSSO (Storlie et al., 2011), and the Multivariate Additive Regression Splines (Mars) method (Friedman, 1991). Additionally, we investigated Bayesian FANOVA using our Exponential and Dirichlet models. Notably, the Exponential model is the Bayesian counterpart of the COSSO model. For these Bayesian models, we explored an adaptive version where the weights/priors were fine-tuned based on SS-ANOVA, aligning with the spirit of ACOSSO. These methods are refered as Bayesian ACOSSO and Bayesian ADirichlet. To provide further context, we also included the Bayesian method proposed in (Lu et al., 2022) for comparison. The latter uses a Gaussian process with the orthogonal RBF kernel (RBF-OK) which is a generalization of the models (Kaufman & Sain, 2010; Duvenaud et al., 2011). However, we encountered challenges in obtaining satisfactory results with the SKIM-FA kernel (Agrawal & Broderick, 2023) in our

[3] https://docs.pyro.ai/

cross-validation UCI experiments. In certain cases, the SKIM-FA kernel selected no covariate, which ultimately resulted in negative R2 values. These unexpected outcomes indicate that the SKIM-FA kernel may not be suitable or well-suited for the specific UCI experiments conducted in our study, perhaps due to small number of covariates. The performance of the SKIM-FA kernel in these experiments fell short of our expectations and did not produce desirable results, then we did not include SKIM-FA results in table 8. We opted not to include the model (Vo & Pati, 2017) in our comparison because it does not provide the capability to predefine the target order, which is a specific requirement for our comparison. We did not also incorporate sampling approaches based on Gibbs sampling, as (Reich et al., 2009) and (Tang et al., 2018), due to their substantial computational demands. Instead, we included in our comparative analysis sampling methods that use Hamiltonian Monte Carlo (HMC) in conjunction with fully marginalized models, which are more computationally efficient.

Details about prior settings: Our proposed Bayesian methods use the Sobolev kernel, as presented in Appendix A.2. Consequently, the covariates are normalized to the range $[0, 1]$ before applying the model. To facilitate the choice of prior parameters for all datasets, we also normalize the output values y_i before training the model. Subsequently, the output is rescaled to its original scale prior before computing metrics. In the Dirichlet model, we assign a weight $w_i = 0.6$ to main effects and $w_i = 0.4$ to interaction effects. These hyperparameters values reflect our prior assumptions, indicating that approximately 60% of main effects and 40 % of interaction effects are likely to be relevant. For the constant function b, we adopt a Gaussian prior with a mean of 0 and a variance of 1. Regarding the noise variance, we employ HalfCauchy priors with a scale of 1. It is important to emphasize that we deliberately selected vague priors for the mean and noise variance for the same reasons explained in simulation part C.1. Similarly, We set a HalfCauchy prior with a large scale for global parameters to approximate a vague prior. To ensure comparability between the local parameters obtained from Dirichlet, we also employ the same strategy as for simulated examples in Appendix C.1 to set the hyperparameters of the Bayesian COSSO model. For the deterministic methods, we rely on cross-validation to choose the regularization hyperparameters from a grid as proposed in the original papers (Storlie et al., 2011).

Practical implementation: We provide the results obtained from our proposed models, using a HMC sampling approach with 2,000 iterations as burn-in. Prediction of the latent function is performed using 100 samples. We also provide the prediction results obtained with our models when using a plug-in approach, i.e., we compute MAP estimate of our parameters and we put these estimators into the Gaussian process model of the functional components.

In our study, we utilized the Python implementation of MARS from the library py-earth. For computing the functional components, we employed the implementation provided by the work (Agrawal & Broderick, 2023). The implementation can be found at the GitHub repository: `https://github.com/agrawalraj/skimfapaper`.

For the Orthogonal kernel RBF, we use the code provided by the authors in `https://github.com/amzn/orthogonal-additive-gaussian-processes`.

Results: Results are presented in Table 8 and 9. For each dataset, we have highlighted the best method in bold based on the predictive results. In cases where multiple methods yield the same results, we consider the method with the best shrinkage properties, as indicated by the smallest value of NAC, as the best method. We consistently observed superior performance of our Bayesian models compared to OK-RBF, with the exception of a single case. In the specific instance of the "Demand" dataset, OK-RBF outperformed all other methods. The noticeable discrepancy in RMSE between OK-RBF and the approaches based on the Sobolev kernel can be attributed to the normalization requirement imposed by the Sobolev kernel. In this case, the covariates were constrained to the interval [0,1]. This normalization process during training, may not effectively handle the out-of-range test covariates. This normalization with the small size of the dataset contribute to the observed variability (standard-deviation) and performance discrepancies in the RMSE results for these approaches.

C.3. Additional experiments on shrinking behaviour of mixing densities

The aim of this section is to compare the shrinkage behaviour of continuous mixing models from Table 3.

C.3.1. Set up

We consider the simple additive model f_1. The train data is generated according to the Uniform scenario. The energies (computed as the ℓ_2 norm) of these active components are then $E\left[5g_1\left(x_0\right)\right] = 9.69$, $E\left[3g_2\left(x_1\right)\right] = 1.80$, $E\left[4g_3\left(x_2\right)\right] = 3.49$ and $E\left[6g_4\left(x_3\right)\right] = 9.50$. The mean value is about 5.4. We consider a sample size of $N = 100$ and set the noise variance $\tau = 5.19$ which corresponds to a signal to noise ration of 8.24.

We consider 4 different hierarchical models by varying the mixing priors on local shrinkage parameters, namely, the

Bayesian COSSO (Exponential mixing) with $w_i = 1/D$, the Student't model (Inverse Gamma mixing) with $\nu = 1$ and $w_i = 1/D$ (which is a Cauchy model), the Horshoe model (Half Cauchy mixing) with $w_i = 1/D$ and the Dirichlet model with $w_i = 0.5$ in the first example and $w_i = 0.1$ in the second example. For the constant function, we assume a Gaussian prior $\mathcal{N}(0,\ 10)$, for the noise variance a Half-Cauchy prior $\mathcal{HC}(0, 10)$ and for the global shrinkage parameter, we set $\sqrt{\gamma} \sim \mathcal{HC}(200)$. For estimation, we consider a sampling approach based on HMC algorithm. We specifically use 5,000 iterations as burn-in and 100 samples for estimation. We consider the sampling approach to have an idea about the whole posterior marginal density of the scales parameters. To measure the shrinkage behaviour of the different models, we define the selection parameter as the product of the global and local scales normalized with the Gram matrix trace divided by N. With such a normalization, this selection parameter can be interpreted as a variance average.

C.3.2. RESULTS

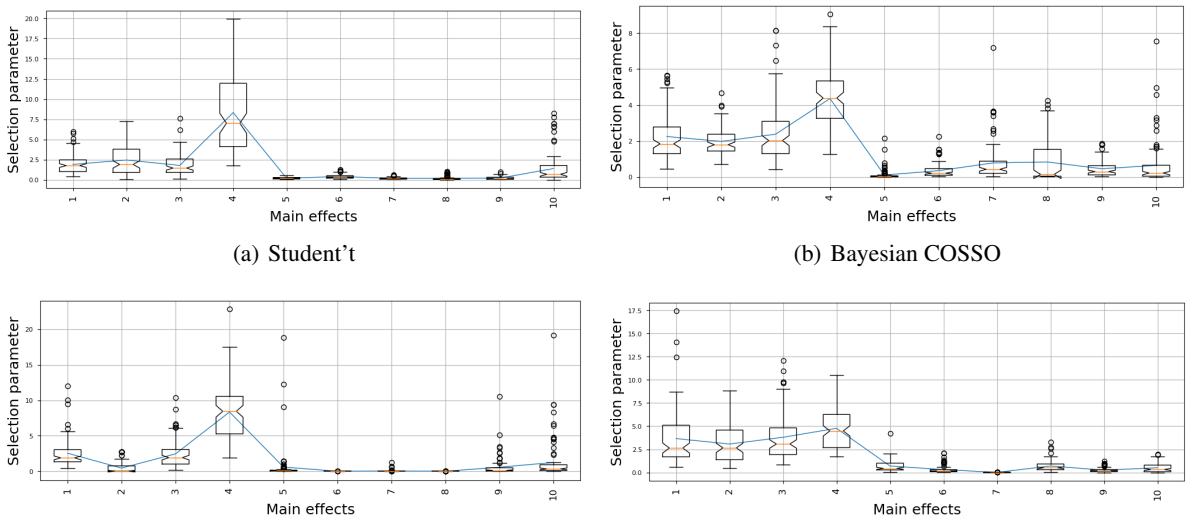
Example 1: In this first example, we set $D = 10$. The shrinkage parameters for these models are given in Figure 4. Overall, all the models were similar (almost exactly the same when discarding components with selection parameter under a certain threshold) in the number of active components that were selected. They have also well identified the correct relevant components. Bayesian COSSO tends to select less inactive components than the other methods. The difference between the later models in this regard was not large, but the Dirichlet model shrink slightly better irrelevant components to zero.

Example 2: We repeat the same experiments with $D = 55$. The model is then "utra-sparse" compared to the model of Example 1. The obtained selection parameters are given in Figure 5. Interestingly, all models except for Dirichlet fail to correctly distinguish the relevant components from the non-relevant ones in terms of interaction order. Figure 5 demonstrates that these models include many noise components in the estimated function. While Horseshoe and Student's priors exhibit better shrinkage behavior than Bayesian COSSO, they still miss some relevant components while incorrectly estimating a few non-relevant components with non-zero energy. In contrast, the Dirichlet model successfully recovers the correct significant components while effectively shrinking most of the non-informative ones. This observation is particularly notable in this scenario, as continuous models like Laplace (Bayesian COSSO) are not well-suited for ultra-sparse settings.

It is still however important to highlight that while the Dirichlet model demonstrates superior performance in terms of shrinkage compared to other methods, its effectiveness relies on the selection of the concentration parameter w_i. In our experiments, we opted to assign the same value to all components since we lacked prior information to favor one component over another. The value of w_i represents our prior belief regarding the number of relevant components. However, if this prior belief is inaccurate, it can lead to the exclusion of relevant components or result in noisy estimation with false irrelevant components. To address moderate shrinkage in the low-dimensional example 1, we set $w_i = 0.5$. This value strikes a balance between including relevant components and avoiding excessive noise in the estimation process. In example 2, where our aim was to obtain a low-dimensional representation of the data due to the large number of covariates, we set $w_i = 0.1$. This choice indicates our desire to include only 10% of the relevant covariates in the representation.

One slight limitation we observed with the Dirichlet model is its tendency to require slightly more iterations to converge in the HMC sampler compared to other independent models. This increased computational cost may be attributed to the correlation between local scales within the Dirichlet model, which can slow down the mixing and then the convergence process.

Finally, these results with all the different models are also sensitive to the setting of the global parameter. In the previous experiments, this problem was not explicitly considered, and a vague prior was used instead for this parameter.



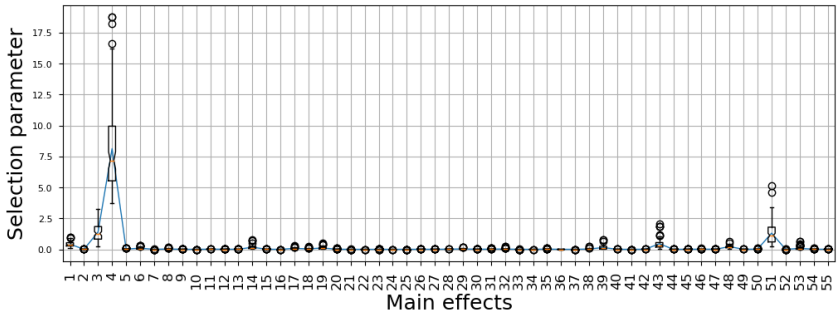
(a) Student't (b) Bayesian COSSO

(c) Horseshoe (d) Dirichlet

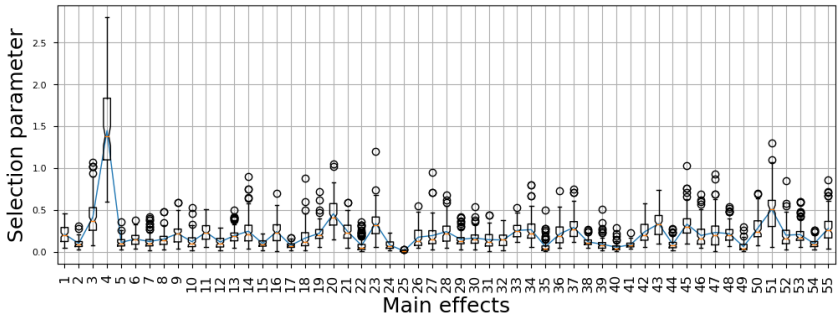
Figure 4. Shrinkage behaviour of the different models for $D = 10$.

Table 6. Regression metrics of FANOVA models on simulated data with $r = 1$ and $D = 55$. The number of relevant components is 4.

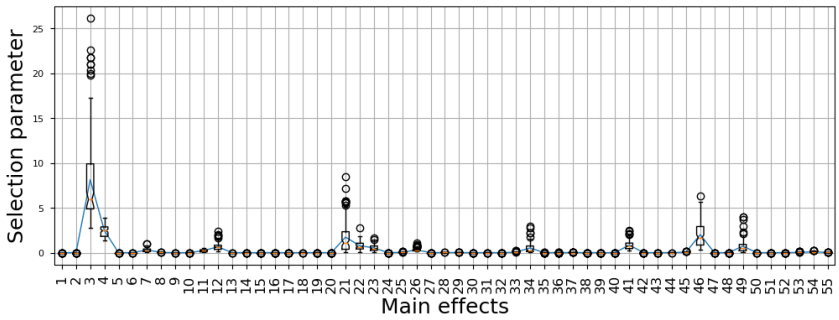
Scenario	Noise and correlation	Metric	SS	COS.	BCOS.MAP	BCOS.HMC	Dir.MAP	Dir.HMC
Inde.	$\tau = 1$	RMSE	3.01±0.34	1.42 ± 0.78	**1.05±0.14**	1.26±0.17	**1.03±0.18**	**0.95±0.26**
		TP	4.00±0.00	3.50±0.50	4.00±0.00	4.00±0.00	4.00±0.00	3.90±0.30
		FP	42.7±1.34	4.90±13.37	1.50±1.34	16.8±2.63	0.80±1.34	0.20±0.60
		FN	0.00±0.00	0.50±0.50	0.00±0.00	0.0±0.00	0.00±0.00	0.10±0.30
		Cover.	-	-	0.92±0.04	0.99±0.01	0.96±0.02	0.98±0.02
Inde.	$\tau = 2.19$	RMSE	3.27±0.21	2.20 ± 0.92	1.55±0.32	1.72±0.0.21	1.50±0.31	**1.16±0.30**
		TP	4.00±0.00	3.90±0.30	4.00±0.00	4.00±0.00	3.90±0.00	3.80±0.40
		FP	43.9±1.44	21.7±19.25	7.30±2.10	16.8±2.63	1.80±1.16	0.60±0.91
		FN	0.00±0.00	0.1±0.30	0.00±0.00	0.0±0.00	0.1±0.30	0.20±0.40
		Cover.	-	-	0.90±0.06	0.99±0.01	0.93±0.05	0.96±0.02
Inde.	$\tau = 5.19$	RMSE	3.43±0.18	2.87 ± 0.72	2.34±0.39	2.22±0.21	2.45±0.28	**1.79±0.78**
		TP	4.00±0.00	3.90±0.30	3.90±0.30	3.90±0.30	3.70±0.45	3.40±0.80
		FP	44.0±1.18	27.5±16.39	11.20±1.66	32.90±1.70	3.40±0.48	2.00±2.14
		FN	0.00±0.00	0.10±0.30	0.10±0.30	0.10±0.30	0.30±0.00	0.60±0.80
		Cover.	-	-	0.86±0.06	0.98±0.01	0.88±0.03	0.92±0.05
Dep. CS.	$\tau = 1$ $\rho = 0.5$	RMSE	3.18±0.34	1.79 ± 0.51	**1.54±0.19**	1.74±0.20	**1.42±0.40**	**1.52±0.77**
		TP	4.00±0.00	3.50±0.50	3.90±0.3	4.00±0.00	3.50±0.67	3.50±0.50
		FP	43.20±1.24	5.30±7.25	5.50±1.20	0.90±0.94	2.00±1.34	0.90±2.38
		FN	0.00±0.00	0.50±0.50	0.10±0.3	0.0±0.00	0.50±0.67	0.50±0.50
		Cover.	-	-	0.88±0.06	0.99±0.01	0.88±0.07	0.95±0.04
Dep. CS.	$\tau = 1$ $\rho = 0.8$	RMSE	3.56±0.33	2.48 ± 1.01	**1.56±0.18**	2.00±0.34	1.75±0.44	**1.51±0.34**
		TP	3.90±0.30	3.70±0.45	4.00±0.0	4.00±0.00	4.00±0.00	3.30±0.45
		FP	45.10±1.37	19.90±17.57	6.30±1.26	11.70±2.90	1.50±1.20	1.30±1.41
		FN	0.10±0.30	0.30±0.45	0.00±0.00	0.0±0.00	0.00±0.00	0.70±0.45
		Cover.	-	-	0.87±0.06	0.99±0.01	0.96±0.02	0.94±0.06
Dep. AR.	$\tau = 1$ $\rho = 0.5$	RMSE	3.23±0.27	2.49 ± 1.04	1.66±0.25	1.85±0.22	**1.42±0.40**	**1.41±0.39**
		TP	3.80±0.40	3.40±0.66	3.90±0.30	4.00±0.00	3.80±0.40	3.50±0.70
		FP	44.00±1.20	16.70±18.30	5.50±1.20	2.60±3.20	0.90±1.94	0.00±0.00
		FN	0.20±0.40	0.60±0.66	0.01±0.30	0.00±0.00	0.20±0.40	0.50±0.70
		Cover.	-	-	0.80±0.09	0.98±0.03	0.10±0.30	0.94±0.05
Dep. AR.	$\tau = 1$ $\rho = 0.8$	RMSE	2.94±0.27	1.87 ± 0.38	1.68±0.19	1.70±0.16	**1.43±0.22**	4.03±7.09
		TP	4.00±0.00	2.90±0.70	4.00±0.00	4.00±0.00	4.00±0.00	3.30±0.64
		FP	44.00±1.51	6.10±13.21	0.00±0.00	19.20±3.48	0.80±0.87	0.70±1.00
		FN	0.00±0.00	1.10±0.70	0.01±0.30	0.00±0.00	0.00±0.00	0.70±0.64
		Cover.	-	-	0.83±0.09	0.98±0.03	0.94±0.05	0.87±0.09



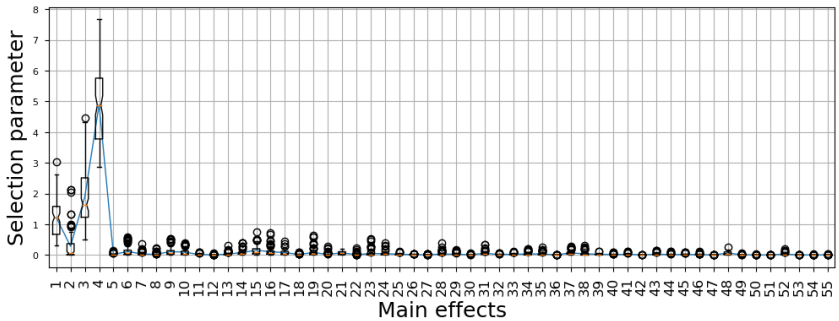
(a) Student't



(b) Bayesian COSSO



(c) Horseshoe



(d) Dirichlet

Figure 5. shrinkage behaviour of the different models for $D = 55$.

Table 7. Regression metrics of ANOVA models on simulated data with $r = 2$ and $D = 10$. The number of relevant components is 7.

Scenario	Noise and correlation	Metric	SS	COS.	BCOS.MAP	BCOS.HMC	Dir.MAP	Dir.HMC
Inde.	$\tau = 1$	RMSE	2.32±0.22	1.96 ± 0.48	**1.10±0.13**	1.42±0.21	1.37±0.18	1.69±0.29
		TP	6.50±0.50	6.10±0.83	5.70±0.78	6.00±0.63	5.20±0.60	4.50±0.92
		FP	26.70±1.26	17.90±13.38	1.90±10.94	4.80±1.24	2.0±1.18	0.70±0.90
		FN	0.50±0.50	0.90±0.83	0.00±0.00	1.30±0.678	1.80±0.60	2.50±0.92
		Cover.	-	-	0.97±0.02	0.98±0.01	0.97±0.03	0.94±0.05
Inde.	$\tau = 2.19$	RMSE	2.43±0.31	2.28 ± 0.44	**1.66±0.34**	**1.63±0.25**	1.87±0.39	1.89±1.08
		TP	6.70±0.45	6.30±1.00	5.60±0.66	6.10±0.70	5.30±0.45	6.10±1.13
		FP	29.00±2.00	23.50±12.36	5.00±1.78	11.30±4.24	3.20±1.53	4.30±1.00
		FN	0.30±0.45	0.70±1.00	1.40±0.66	0.0±0.00	0.90±0.70	0.90±1.13
		Cover.	-	-	0.94±0.04	0.99±0.01	0.94±0.05	0.94±0.04
Inde.	$\tau = 5.19$	RMSE	2.71±0.22	2.75 ± 0.23	**2.16±0.31**	**2.11±0.21**	2.32±0.36	2.83±1.56
		TP	6.60±0.48	6.70±0.45	5.80±0.60	6.20±0.60	5.30±0.78	4.50±1.11
		FP	31.30±1.73	32.70±4.36	7.30±2.41	18.40±3.66	3.80±1.99	1.90±1.57
		FN	0.40±0.48	0.30±0.45	1.20±0.60	0.80±0.60	1.70±0.78	2.50±1.11
		Cover.	-	-	0.95±0.03	0.98±0.02	0.95±0.04	0.96±0.01
Dep. CS.	$\tau = 1$ $\rho = 0.5$	RMSE	2.97±0.25	3.05 ± 0.53	**2.09±0.22**	**2.06±0.21**	**2.11±0.28**	2.20±0.84
		TP	6.50±0.80	5.20±0.87	3.80±0.6	4.60±0.66	3.60±0.48	4.40±1.01
		FP	22.20±2.27	17.40±9.83	1.90±1.44	4.10±2.02	7.70±1.37	0.90±0.83
		FN	0.50±0.80	1.80±0.87	3.20±0.6	0.0±0.00	2.40±0.66	2.60±1.01
		Cover.	-	-	0.89±0.06	0.96±0.03	0.86±0.06	0.91±0.07
Dep. CS.	$\tau = 1$ $\rho = 0.8$	RMSE	4.64±1.10	5.59 ± 1.63	**2.43±0.04**	**2.19±0.26**	**2.21±0.24**	4.49±4.80
		TP	6.10±0.30	5.50±0.80	3.80±0.4	3.80±0.40	3.30±0.45	3.50±0.67
		FP	28.40±5.64	20.30±2.14	1.00±0.44	2.70±0.78	0.90±0.70	1.10±0.83
		FN	0.90±0.30	1.50±0.80	3.20±0.40	3.20±0.40	3.70±0.45	3.50±0.67
		Cover.	-	-	0.85±0.06	0.95±0.03	0.79±0.08	0.83±0.14
Dep. AR.	$\tau = 1$ $\rho = 0.5$	RMSE	2.77±0.32	2.32 ± 0.36	**1.72±0.40**	2.11±0.30	**1.83±0.29**	1.85±1.57
		TP	6.50±0.50	6.20±0.87	5.50±1.02	5.50±1.25	5.60±0.66	5.30±1.10
		FP	29.30±2.05	14.70±10.35	1.80±0.97	6.00±2.08	2.40±1.68	3.80±0.74
		FN	0.50±0.50	0.80±0.87	1.50±1.02	1.50±1.25	1.40±0.66	1.70±1.10
		Cover.	-	-	0.89±0.07	0.93±0.05	0.90±0.04	0.90±0.06
Dep. AR.	$\tau = 1$ $\rho = 0.8$	RMSE	2.49±0.24	2.35 ± 0.21	**1.34±0.24**	1.71±0.20	1.62±0.24	1.89±0.31
		TP	6.10±0.53	5.70±0.90	5.60±0.48	5.20±0.60	5.20±0.97	4.60±0.91
		FP	30.60±2.05	20.50±12.88	2.00±1.18	4.70±1.41	2.50±0.80	0.80±0.97
		FN	0.90±0.53	1.30±0.90	1.40±0.48	1.80±0.60	1.80±0.97	2.40±0.91
		Cover.	-	-	0.95±0.03	0.98±0.1	0.92±0.05	0.93±0.05

Table 8. Regression metrics of FANOVA models on various UCI datasets. Proposed models are run using the Sobolev kernel.

Data set (N, D, J)	Metric	SS	MARS	COS.	ACOS.	BCOS. HMC	ABCOS. HMC	Dir. HMC	ADir. HMC	RBF-OK
	RMSE	7.95±0.45	8.90±2.44	7.49±0.42	7.65±0.30	**4.09±0.32**	4.32±0.47	4.33±0.38	4.35±0.25	4.23±0.36
concrete	MAE	6.23±0.33	6.12±0.38	5.88±0.33	6.02±0.23	2.75±0.18	2.88±0.19	3.01±0.36	2.99±0.18	2.80±0.18
(1000,	R2	0.77±0.03	0.68±0.20	0.79±0.02	0.79±0.02	0.94±0.01	0.93±0.01	0.93±0.01	0.93±0.01	0.93±0.01
8, 36)	NAC	21.8±0.4	18.6±0.80	10.6±2.7	15.0±1.7	22.3±1.1	19.5±2.1	17.6±1.4	14.6±1.0	23.6±3.38
	Cover.	-	-	-	-	0.95±0.03	0.98±0.01	0.93±0.02	0.92±0.05	0.38±0.02
	RMSE	30.30±7.35	14.51±16.75	25.03±9.23	23.80±6.27	11.95±4.68	19.17±11.30	34.55±16.17	17.95±5.74	**3.11±3.64**
demand	MAE	16.76±4.39	7.34±6.23	13.23±3.98	12.20±2.95	6.44±1.86	9.41±4.51	18.76±6.38	9.48±1.79	1.25±1.48
(60,	R2	0.86±0.04	0.79±0.38	0.91±0.04	0.92±0.02	0.97±0.02	0.94±0.06	0.81±0.11	0.95±0.03	0.99±0.00
12, 78)	NAC	42.0±4.9	6.8±2.72	18.0±6.2	14.8±4.5	4.4±1.5	6.0±1.8	8.0±3.6	6.40±2.9	2.00±0.0
	Cover.	-	-	-	-	0.98±0.03	0.95±0.04	0.97±0.04	0.95±0.04	0.10±0.03
	RMSE	0.47±0.05	1.83±0.57	0.50±0.05	0.47±0.05	**0.45±0.05**	**0.45±0.05**	**0.45±0.05**	**0.45±0.04**	**0.47±0.03**
energy	MAE	0.34±0.03	1.36±0.35	0.37±0.02	0.34±0.03	0.33±0.02	0.33±0.02	0.33±0.03	0.33±0.01	0.34±0.02
(768,	R2	0.99±0.00	0.96±0.02	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
8, 36)	NAC	11.0±0.6	12.0±2.19	9.4±1.4	9.2±0.8	6.4±1.3	6.4±1.0	6.4±1.3	6.4±1.0	6.00±0.00
	Cover.	-	-	-	-	0.93±0.02	0.94±0.04	0.93±0.02	0.95±0.04	0.98±0.01
	RMSE	2.98±0.34	3.67±0.45	3.06±0.43	2.99±0.42	3.06±0.42	**3.05±0.46**	**3.07±0.56**	3.12±0.52	3.30±0.51
housing	MAE	2.05±0.14	2.44±0.25	2.07±0.20	2.06±0.19	2.05±0.15	2.06±0.18	2.08±0.21	2.13±0.22	2.13±0.25
(506,	R2	0.89±0.03	0.83±0.03	0.88±0.04	0.89±0.04	0.88±0.04	0.88±0.04	0.88±0.05	0.87±0.04	0.86±0.03
13, 91)	NAC	71.2±0.8	24.40±2.93	59.2±2.3	46.2±6.1	58.4±6.7	36.2±2.7	29.0±7.48	30.2±2.4	35.8±2.63
	Cover.	-	-	-	-	0.93±0.01	0.92±0.03	0.95±0.03	0.93±0.03	0.25±0.15
	RMSE	1.86±0.83	4.16±0.51	1.73±0.60	1.75±0.71	**0.51±0.05**	**0.50±0.06**	0.56±0.07	0.59±0.12	0.60±0.26
yacht	MAE	1.06±0.42	3.45±0.26	1.06±0.32	1.01±0.26	0.27±0.02	0.26±0.02	0.32±0.06	0.34±0.09	0.33±0.10
(306,	R2	0.98±0.01	0.91±0.01	0.98±0.01	0.98±0.01	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
6, 21)	NAC	1.2±0.40	2.60±0.80	1.0±0.0	1.00±0.0	1.2±0.40	1.2±0.4	1.2±0.4	1.2±0.4	1.6±1.35
	Cover.	-	-	-	-	0.97±0.02	0.97±0.02	0.96±0.01	0.96±0.01	0.89±0.06
	RMSE	2.66± 0.30	2.85± 0.19	2.63±0.28	2.63±0.28	2.64±0.31	2.67±0.30	2.78±0.26	**2.69±0.30**	2.81±0.35
autompg	MAE	1.90 ± 0.17	2.06 ± 0.14	1.90±0.16	1.90±0.16	1.89±0.17	1.90±0.15	2.00±0.14	1.91±0.14	2.00±0.22
(392,	R2	0.88±0.02	0.86±0.02	0.88±0.02	0.88±0.02	0.88±0.02	0.87±0.02	0.87±0.01	0.87±0.02	0.86±0.02
7, 28)	NAC	17.8±0.7	8.80±1.32	16.8±0.4	16.6±0.5	14.0±1.54	12.4±1.95	7.8±0.40	9.0±2.09	14.20±2.48
	Cover.	-	-	-	-	0.95±0.02	0.94±0.02	0.93±0.01	0.94±0.02	0.31±0.05
	RMSE	4.18±0.28	3.39±0.14	3.22±0.17	3.20±0.19	3.24±0.22	3.29±0.23	3.39±0.33	**3.17±0.23**	3.76±0.80
pumadyn	MAE	3.36±0.21	2.62±0.12	2.52±0.09	2.50±0.11	2.55±0.19	2.56±0.18	2.69±0.32	2.45±0.18	3.05±0.75
(1000,	R2	0.43±0.05	0.61±0.05	0.65±0.05	0.66±0.05	0.65±0.05	0.63±0.06	0.61±0.09	0.66±0.05	0.51±0.18
8, 36)	NAC	25.0±1.3	13.4±2.57	10.6±2.73	9.8±1.9	19.2±2.25	16.6±2.33	8.6±2.93	5.4±1.35	3.80±0.74
	Cover.	-	-	-	-	0.94±0.02	0.94±0.03	0.94±0.03	0.94±0.01	0.40±0.48
	RMSE	0.60±0.11	0.77±0.18	0.57±0.11	0.58±0.12	**0.51±0.07**	0.55±0.11	0.54±0.10	0.60±0.09	0.55±0.06
servo	MAE	0.38±0.05	0.57±0.09	0.36±0.05	0.36±0.05	0.33±0.05	0.33±0.05	0.32±0.04	0.37±0.06	0.35±0.03
(166,	R2	0.85±0.04	0.74±0.06	0.86±0.04	0.85±0.05	0.88±0.02	0.86±0.04	0.87±0.04	0.84±0.03	0.86±0.01
4, 10)	NAC	9.00±0.00	7.0±0.63	8.0±0.63	7.4±1.0	7.2±0.43	6.6±0.48	6.2±1.16	4.4±0.48	7.0±0.0
	Cover.	-	-	-	-	0.96±0.01	0.96±0.01	0.96±0.01	0.95±0.01	0.95±0.01
	RMSE	0.52±0.01	0.60±0.09	0.53±0.01	0.53±0.01	0.53±0.01	0.53±0.01	**0.53±0.02**	0.53±0.02	0.54±0.06
stock	MAE	0.39±0.01	0.43±0.03	0.40±0.01	0.40±0.01	0.40±0.02	0.40±0.02	0.39±0.00	0.40±0.01	0.41±0.04
(536,	R2	0.74±0.04	0.65±0.12	0.73±0.04	0.73±0.04	0.73±0.05	0.73±0.03	0.74±0.03	0.74±0.04	0.72±0.03
8, 36)	NAC	19.2±0.8	15.2±1.6	16.4±0.8	16.4±0.8	10.8±1.60	11.6±1.85	7.2±0.97	9.4±0.80	10.4±4.58
	Cover.	-	-	-	-	0.93±0.01	0.93±0.01	0.93±0.01	0.93±0.02	0.91±0.16

Table 9. Regression metrics of our FANOVA models on various UCI datasets. Comparison between plug-in and sampling approaches.

Data set (N, D, J)	Metric	BCOS.MAP	ABCOS.MAP	Dir.MAP	ADir.MAP	BCOS.HMC	ABCOS.HMC	Dir.HMC	ADir.HMC
	RMSE	4.15±0.29	4.16±0.32	**4.13±0.26**	4.34±0.34	4.09±0.32	4.32±0.47	4.33±0.38	4.35±0.25
concrete	MAE	2.83±0.20	2.82±0.20	2.82±0.20	2.98±0.22	2.75±0.18	2.88±0.19	3.01±0.36	2.99±0.18
(1000,	R2	0.93±0.00	0.93±0.00	0.93±0.00	0.93±0.00	0.94±0.01	0.93±0.01	0.93±0.01	0.93±0.01
8, 36)	NAC	18.0±1.41	18.6±1.01	13.2±1.46	13.8±1.16	22.3±1.1	19.5±2.1	17.6±1.4	14.6±1.0
	Cover.	0.94±0.01	0.94±0.01	0.94 ±0.01	0.93 ±0.00	0.95±0.03	0.98±0.01	0.93±0.02	0.92±0.05
	RMSE	15.76±16.11	15.97±16.01	19.91±14.83	22.48±12.62	**11.95±4.68**	19.17±11.30	34.55±16.17	17.95±5.74
demand	MAE	6.42±7.24	6.50±7.24	10.07±7.06	11.66±3.53	6.44±1.86	9.41±4.51	18.76±6.38	9.48±1.79
(60,	R2	0.93±0.10	0.93±0.10	0.91±0.10	0.90±0.08	0.97±0.02	0.94±0.06	0.81±0.11	0.95±0.03
12, 78)	NAC	2.80.0±0.40	2.80±0.40	3.20±0.74	3.40±0.80	4.4±1.5	6.0±1.8	8.0±3.6	6.40±2.9
	Cover.	0.96±0.04	0.95±0.04	0.95±0.04	0.90±0.08	0.98±0.03	0.95±0.04	0.97±0.04	0.95±0.04
	RMSE	**0.43±0.04**	0.43±0.04	0.44±0.03	0.44±0.03	0.45±0.05	0.45±0.05	0.45±0.05	0.45±0.04
energy	MAE	0.32±0.01	0.32±0.07	0.33±0.01	0.32±0.01	0.33±0.02	0.33±0.02	0.33±0.03	0.33±0.01
(768,	R2	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
8, 36)	NAC	4.0±0.0	4.20±0.40	4.80±0.40	7.0±0.0	6.40±1.3	6.40±1.0	6.40±1.3	6.4±1.0
	Cover.	0.96±0.02	0.95±0.02	0.95±0.01	0.95±0.01	0.93±0.02	0.94±0.04	0.93±0.02	0.95±0.04
	RMSE	3.15±0.56	3.22±0.51	3.05±0.51	3.17±0.43	3.06±0.42	3.05±0.46	**3.07±0.56**	3.12±0.52
housing	MAE	2.07±0.19	2.10±0.15	2.04±0.19	2.12±0.18	2.05±0.15	2.06±0.18	2.08±0.21	2.13±0.22
(506,	R2	0.87±0.05	0.86±0.05	0.88±0.04	0.87±0.04	0.88±0.04	0.88±0.04	0.88±0.05	0.87±0.04
13, 91)	NAC	21.20±0.74	21.0±1.54	59.2±2.3	24.6±1.2	58.4±6.7	36.2±2.7	29.0±7.48	30.2±2.4
	Cover.	0.92±0.02	0.91±0.01	0.93±0.01	0.94±0.01	0.93±0.01	0.92±0.03	0.95±0.03	0.93±0.03
	RMSE	0.58±0.08	0.58±0.08	0.70±0.19	0.57±0.71	**0.51±0.05**	**0.50±0.06**	0.56±0.07	0.59±0.12
yacht	MAE	0.30±0.08	0.30±0.04	0.35±0.26	0.32±0.01	0.27±0.02	0.26±0.02	0.32±0.06	0.34±0.09
(306,	R2	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
6, 21)	NAC	1.2±0.40	1.2±0.4	1.2±0.4	1.2±0.40	1.2±0.40	1.2±0.4	1.2±0.4	1.2±0.4
	Cover.	0.95±0.02	0.95±0.02	0.93±0.02	0.95±0.01	0.97±0.02	0.97±0.02	0.96±0.01	0.96±0.01
	RMSE	2.70± 0.28	2.72±0.29	2.71±0.34	**2.66±0.29**	2.64±0.31	2.67±0.30	2.78±0.26	2.69±0.31
autompg	MAE	1.94 ± 0.15	1.96±0.16	1.96±0.16	1.90±0.15	1.89±0.17	1.90±0.15	2.00±0.14	1.91±0.14
(392,	R2	0.87±0.02	0.87±0.02	0.87±0.02	0.88±0.02	0.88±0.02	0.87±0.02	0.87±0.01	0.87±0.02
7, 28)	NAC	10.40±1.49	9.6±1.35	7.80±1.45	8.0±0.63	14.0±1.54	12.4±1.95	7.8±0.40	9.0±2.09
	Cover.	0.94±1.49	0.93±0.02	0.93±0.02	0.94±0.02	0.95±0.02	0.94±0.02	0.93±0.01	0.94±0.02
	RMSE	3.19±0.20	3.20±0.21	3.25±0.20	**3.17±0.23**	3.24±0.22	3.29±0.23	3.39±0.33	**3.17±0.23**
pumadyn	MAE	2.46±0.0.15	2.48±0.16	2.51±0.16	2.45±0.17	2.55±0.19	2.56±0.18	2.69±0.32	2.45±0.18
(1000,	R2	0.66±0.05	0.65±0.05	0.64±0.05	0.66±0.05	0.65±0.05	0.63±0.06	0.61±0.09	0.66±0.05
8, 36)	NAC	5.60±1.85	5.80±1.72	4.80±0.97	5.40±1.49	19.2±2.25	16.6±2.33	8.6±2.93	5.4±1.35
	Cover.	0.93±0.01	0.93±0.01	0.93±0.02	0.94±0.01	0.94±0.02	0.94±0.03	0.94±0.03	0.94±0.01
	RMSE	0.56±0.10	0.56±0.10	0.61±0.11	0.61±0.11	**0.51±0.07**	0.55±0.11	0.54±0.10	0.60±0.09
servo	MAE	0.34±0.06	0.34±0.06	0.37±0.06	0.38±0.08	0.33±0.05	0.33±0.05	0.32±0.04	0.37±0.06
(166,	R2	0.86±0.04	0.86±0.04	0.84±0.04	0.83±0.04	0.88±0.02	0.86±0.04	0.87±0.04	0.84±0.03
4, 10)	NAC	6.6±0.48	6.60±0.48	5.60±0.48	4.60±0.48	7.2±0.43	6.6±0.48	6.2±1.16	4.4±0.48
	Cover.	0.95±0.02	0.95±0.02	0.95±0.02	0.94±0.03	0.96±0.01	0.96±0.01	0.96±0.01	0.95±0.01
	RMSE	0.53±0.01	0.53±0.01	0.52±0.01	0.52±0.02	0.53±0.01	0.53±0.01	**0.53±0.02**	0.53±0.02
stock	MAE	0.39±0.01	0.40±0.01	0.39±0.01	0.40±0.02	0.40±0.02	0.40±0.02	0.39±0.00	0.40±0.01
(536,	R2	0.73±0.04	0.73±0.04	0.74±0.04	0.74±0.05	0.73±0.05	0.73±0.03	0.74±0.03	0.74±0.04
8, 36)	NAC	10.4±0.8	10.0±1.42	10.0±0.63	8.80±0.4	10.8±1.60	11.6±1.85	7.2±0.97	9.4±0.80
	Cover.	0.93±0.01	0.91±0.02	0.93±0.02	0.93±0.02	0.93±0.01	0.93±0.01	0.93±0.01	0.93±0.02