The Annals of Applied Statistics
2021, Vol. 15, No. 1, 1–21
https://doi.org/10.1214/20-AOAS1372

DISCOVERING INTERACTIONS USING COVARIATE INFORMED RANDOM PARTITION MODELS

BY GARRITT L. PAGE[1], FERNANDO A. QUINTANA[2] AND GARY L. ROSNER[3]

[1]*Department of Statistics, Brigham Young University, page@stat.byu.edu*
[2]*Departamento de Estadística, Pontificia Universidad Católica de Chile, quintana@mat.uc.cl*
[3]*Department of Biostatistics, Johns Hopkins University, grosner1@jhmi.edu*

Combination chemotherapy treatment regimens created for patients diagnosed with childhood acute lymphoblastic leukemia have had great success in improving cure rates. Unfortunately, patients prescribed these types of treatment regimens have displayed susceptibility to the onset of osteonecrosis. Some have suggested that this is due to pharmacokinetic interaction between two agents in the treatment regimen (asparaginase and dexamethasone) and other physiological variables. Determining which physiological variables to consider when searching for interactions in scenarios like these, minus a priori guidance, has proved to be a challenging problem, particularly if interactions influence the response distribution in ways beyond shifts in expectation or dispersion only. In this paper we propose an exploratory technique that is able to discover associations between covariates and responses in a general way. The procedure connects covariates to responses flexibly through dependent random partition distributions and then employs machine learning techniques to highlight potential associations found in each cluster. We provide a simulation study to show utility and apply the method to data produced from a study dedicated to learning which physiological predictors influence severity of osteonecrosis multiplicatively.

1. Introduction. In studies that collect covariate measurements on subjects/experimental units in addition to a response, it is of principal interest to determine which covariates influence the response and in what way. This seemingly benign statistical problem has been seriously considered for many years with methods ranging from exploratory techniques to more model-based procedures. Perhaps the reason why research dedicated to this problem persists is that, in practice, building a statistical model is as much an art as a science. This is a consequence of important or "significant" associations depending completely on the factors included in the model. Therefore, knowing which effects to include in a model is crucial, but this information is rarely known a priori. Since this information is rarely known, an approach that is commonly used in practice proposes fitting a saturated model (a model containing all possible covariates or predictors) and then employing some type of multiplicity test correction, model selection criterion, shrinkage method or stochastic search to locate "significant" factors (see, e.g., Chung and Dunson (2009), George and McCulloch (1997), Ishwaran and Rao (2005), Mitra, Müller and Ji (2017), Scott and Berger (2010), Smith and Kohn (1996), Tibshirani (1996)). These methods have been shown to work well in many instances (Barrera-Gómez et al. (2017)). However, when multiplicative effects and/or nonlinear associations are present and are of interest, a fairly common scenario (see, e.g., Hu, Joshi and Johnson (2009), Lim and Hastie (2015)), the process of identifying associations becomes much more problematic as the saturated model can quickly become unwieldy. This happens to be the case in the study under consideration, where we examine which factors affect the

Received July 2019; revised March 2020.
Key words and phrases. Multiplicative associations, dependent random partition models, nonparametric Bayes, exploratory data analysis.

severity of osteonecrosis in children treated for acute lymphoblastic leukemia (ALL). Interest lies in learning how time-varying physiological and other baseline covariates, such as gender, triglycerides and body mass index (BMI) influence disease severity in an additive or multiplicative fashion. Employing the saturated model approach here with all possible two-way and/or three-way interactions becomes computationally expensive and inferentially difficult, so knowing which covariates to consider is important.

Adding to the complexity of the scenario just described, methods used to identify associations (either additive or multiplicative) are geared toward discovering covariates that only influence the mean of the response distribution. It is completely plausible that interactions also influence the variance or even the shape of the response distribution, and most methods are not equipped to detect these types of associations. Regression methods that allow covariates to influence the entire response (or error) distribution have been developed and are commonly known as density regression (see Dunson, Pillai and Park (2007), Fan, Yao and Tong (1996)). There has been work in density regression that simultaneously carries out variable selection (Shen and Ghosal (2016), Tokdar, Zhu and Ghosh (2010)), but they require that all effects of interest be included in a model. Though it may be possible to extend work done in density regression to incorporate multiplicative effects, a priori information would be necessary to guide which of these effects to consider. In light of this, an exploratory procedure that is able to discover possible associations in a general way, prior to model fitting, would be appealing.

There is a growing literature dedicated to methods developed with the main purpose of explicitly identifying interactions without restricting themselves to variable selection. Reich et al. (2012) develop a statistical emulator and devise a procedure that allows them to learn how the inputs of a stochastic computer model influence the output (which is a distribution). Bien, Taylor and Tibshirani (2013) use the Lasso, and Lim and Hastie (2015) utilize the group Lasso to detect interactions from hierarchically coherent models (i.e., two-way interactions are present only if both main effects are present). Kapelner and Bleich (2016) employ Bayesian additive regression trees (BART) to identify interactions while Du and Linero (2018) develop methods based on Bayesian decision tree ensembles that incorporate an additive component. XGBoost (Chen and Guestrin (2016)) is a scalable machine learning system for tree boosting that has also been employed to discover interactions. In more recent work, Ferrari and Dunson (2019) use a factor model to induce models that include interactions and Agrawal et al. (2019) search for interactions in high dimensions.

The methods just cited are developed for specific scenarios while what we propose is more general and can be employed with any data model, making the procedure essentially "model free." This is important for our application; as the response is ordinal, rendering many of the works just cited unavailable. Thus, the exploratory procedure we propose discovers associations where anything is fair game, in the sense that associations could be additive and/or multiplicative and could influence any aspect of the response density (i.e., mean, spread, shape, etc.) regardless of the data model that will eventually be employed. Our goal is ambitious and, admittedly, being able to discover all possible interactions in the general way we are proposing is, presumably, not possible. Thus, we do not claim that the exploratory approach detailed in the sequel discovers all "significant" interactions; rather it provides guidance to practitioners by highlighting possible interactions with very little overhead.

We finish our literature review noting that there is a growing literature dedicated to so-called subgroup analysis, that is, the study of heterogenous treatment effects among subgroups of a study population. Subgroups are typically identified or defined based on specific values in the covariate space. Since subpopulation treatment effects are of principal interest, focus is placed on studying the interaction between a covariate and a treatment (see, e.g.,

Berger, Wang and Shen (2014), Henderson et al. (2020), Liu et al. (2017), Schnell et al. (2016), Simon (2002), Su et al. (2018), Varadhan and Wang (2014)). Our approach is more general in nature but can be used as an exploratory tool to help discover which covariates interact with the treatment and provide guidance by spotting subpopulations of potential interest.

As a point of terminology, in what follows we use the term "interaction" to denote something more general than what is referred to in a linear model setting. Here, an interaction exists if the response distribution is in some way influenced by specific combinations of covariate values. This general conception includes the special case where all the influence is carried by only one of the covariates. Thus, we recommend carefully investigating the "interactions" detected by our method to gain insight into how they (and the covariates they contain) influence the response distribution. This highlights the exploratory nature of our proposal.

The remainder of the article is organized as follows. In Section 2 we detail the data collected from the study that motivated this work. Section 3 provides an overview of our approach with some background on dependent random partition distributions and association rules. Section 4 provides more details of our approach, and Section 5 describes two simulation studies. In Section 6 we apply the procedure to the osteonecrosis data, and Section 7 contains a brief discussion. All publicly available source code used to implement the methods we develop are available in the Supplementary Material (Page, Quintana and Rosner (2021b)).

2. Osteonecrosis study. The study we consider was designed to learn more about factors affecting risk for osteonecrosis in children suffering from acute lymphoblastic leukemia (ALL). With combination chemotherapeutic regimens, five-year survival is around 85% overall for childhood ALL, with some subgroups' rates well above 90% (American Cancer Society (2018)). These regimens include the drug asparaginase and the steroid dexamethasone. Some have suggested that there is a pharmacokinetic interaction between these two agents, leading to greater interpatient variability and severe adverse events (Kawedia et al. (2011)). The principal aim of the motivating study was to learn about relationships between physiological characteristics and treatment characteristics and how these relationships influence susceptibility to osteonecrosis as a result of therapy. The analysis considers a number of physiological covariates measured on each patient, including low-density lipoprotein (LDL), high-density lipoprotein (HDL) levels, body mass index (BMI) and others. In addition to these covariates, the data include plasma levels of dexamethasone, cortisol and asparaginase at various times during each patient's course of treatment (baseline, week 7, week 8) to explore how the pharmacokinetics (PK) of these substances influenced the risk and severity of osteonecrosis and whether other factors interact with the drugs' PK. Finally, demographic variables include age at diagnosis, gender and race. In total, 23 predictors were considered with numerical summaries provided in Tables 1 and 2. Out of the 400 patients in the study, we have complete data vectors for 234, and we focus on these.

TABLE 1
Number of patients in each gender by race category

Gender	Asian	Black	Hispanic	Other	White
Female	1	21	8	7	96
Male	2	19	8	3	69

TABLE 2
Numerical summary of the 19 *continuous covariates measured on each subject in the osteonecrosis study*

Covariate	Min.	Q1	Q2	Mean	Q3	Max.
AgeAtDiagnosis	1.02	3.20	5.23	6.77	9.68	18.84
DexClWk07	1.99	9.03	14.69	21.83	25.03	458.07
DexClWk08	1.62	8.09	11.67	14.07	16.62	61.71
CortisolBaseline	0.25	4.07	5.67	7.46	8.22	59.41
CortisolWk07	0.11	4.93	7.12	9.00	10.34	55.16
CortisolWk08	0.00	0.32	0.58	1.11	1.01	30.27
HDLBaseline	11.30	39.05	48.60	56.28	59.25	271.00
HDLWk07	10.70	41.92	52.50	59.10	67.20	198.00
HDLWk08	1.90	29.45	44.00	50.31	61.50	238.00
LDLBaseline	13.00	62.00	82.00	83.95	104.75	195.00
LDLWk07	15.00	57.25	80.00	82.04	100.38	218.00
LDLWk08	5.00	50.25	68.00	73.44	91.50	179.00
TriglyceridesBaseline	20.00	58.62	87.50	104.60	124.75	472.00
TriglyceridesWk07	20.00	55.02	74.50	114.79	132.50	885.00
TriglyceridesWk08	15.90	75.25	122.50	220.75	286.00	1029.50
AlbuminBaseline	3.30	4.00	4.20	4.17	4.40	5.10
AlbuminWk07	2.10	3.20	3.90	3.75	4.30	4.80
AlbuminWk08	2.20	3.10	3.60	3.56	4.00	4.80
BMIBasline	12.19	15.68	16.77	18.05	19.23	38.48
AsparaginaseAntibodyAUC	0.42	1.07	3.54	14.20	24.83	93.85

Table 1 displays the gender-by-race distribution with each gender being equally represented. A third binary covariate (called LowRisk) which indicates the prognosis of the patient's leukemia (low or high risk of relapse) is also included. This variable should be highly influential because exact treatment regimens are based on patients' risk factors relating to prognosis of their leukemia. Of the 234 subjects, 109 were in the low-risk group while 125 were in the high risk. Table 2 provides a numerical summary of continuous covariates. Notice that a few of them are highly right skewed. Empirical correlations between time-varying covariates suggest that temporal dependence is also present. This dependence can be accommodated in our approach and is formally considered in the model developed by Barcella et al. (2018). We briefly comment that week 12 measurements were also collected, but very irregularly. This resulted in many incomplete covariate vectors, and for this reason we only consider measurements taken up to week 8.

The response measured in this study reflects the severity of osteonecrosis ranging on an ordinal scale from 0 for none, up to 4 for high grade. Table 3 contains the number of patients by grade of osteonecrosis. We expect, but do not force, the covariates' influences on osteonecrosis grade to be multiplicative, but it is not clear which covariates to pair together when exploring multiplicative effects. This is something we hope to discover.

TABLE 3
Grade of osteonecrosis and the number of patients diagnosed for each one

	Osteonecrosis Grade				
	0	1	2	3	4
Number of Subjects	73	115	27	17	2

3. Background and preliminaries. The exploratory procedure we develop consists of three stages. In this section we introduce each one and then provide notation and background information necessary to make ideas concrete.

The first step consists of connecting covariates to the response by way of a dependent random partition model. These types of models create partitions of the data taking into account homogeneity among covariates, and, in particular, some of these models, such as the PPMx by Müller, Quintana and Rosner (2011), aim specifically at constructing priors for which two individuals with similar covariate values are more likely to cocluster. The resulting partition is, a posteriori, weighted by the corresponding cluster sizes, the likelihood component and the similarity of covariates that belong to a cluster. This feature has proven to be quite useful in capturing many types of nonlinearities present in the data, including combined changes in shape, location and scale of the predictive distribution as the predictors change. Since these types of models are closely related to discrete random probability models of the type used in Bayesian nonparametrics, which are well known for their flexibility (see, e.g., Müller et al. (2015)), in theory, any statistical model that produces (explicitly or implicitly) a covariate dependent partition may be employed. This includes the dependent Dirichlet process mixture (DDP) model of MacEachern (2000) and its variants (e.g., ANOVA DDP of De Iorio et al. (2004)). As explained below, we adopt here the PPMx model.

In the second step the estimated partition from step 1 is used to identify potentially interesting interactions. One possible way of employing the partition to pinpoint potential interactions is to identify covariates that seem to influence cluster formation. This would imply that an association exists among the influential covariates and the response (a necessity for an interaction to exist). Identifying influential covariates can be done by determining which covariates have the same (or similar) values for many individuals in a cluster. An unsupervised learning technique developed to carry out this type of search establishes so-called association rules (Han, Kamber and Pei (2012), Chapter 6). There is precedence to using association rules as a tool to identify interactions (see Heba et al. (2014)), but we use them to rank potential interactions based on their "importance." Ranking interactions based on their "importance" is, implicitly, something all procedures described in Section 1 do. Many of the procedures, however, provide very little guidance or criteria on how to determine which interactions are simply noise. This motivates the inclusion of a third step to our exploratory approach.

In the third step of our procedure, we verify that a detected interaction indeed affects the response distribution. This is done by comparing posterior predictive densities that are based on specific values from covariates that were identified by an association rule. If the posterior predictive densities do not change as a function of the covariates found in the association rule, then their association with the response is weak at best. (As a side note, Gabry et al. (2019) advocate using a posterior predictive distribution as an exploratory or confirmatory tool.) We now provide more pertinent background information for dependent random partition models and association rules.

3.1. *Dependent random partition models.* We begin by introducing some notation. Let $i = 1, \ldots, m$ index the m experimental units in a designed experiment or m subjects in an observational study. Further, let $\rho_m = \{S_1, \ldots, S_{k_m}\}$ denote a partitioning (or clustering) of the m units into k_m subsets such that $i \in S_j$ implies that unit i belongs to cluster j. A common alternative notation that specifies a partitioning of the m units into k_m clusters is to introduce m cluster labels $s_1, \ldots, s_m$ such that $s_i = j$ implies $i \in S_j$. We will use Y_i to denote the ith subject's response variable with $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ denoting an m dimensional response vector and $\boldsymbol{Y}_j^\star = \{Y_i : i \in S_j\}$ the jth cluster's response vector. Similarly, let $\boldsymbol{X} = (X_1, \ldots, X_m)$ denote a covariate vector and $\boldsymbol{X}_j^\star = \{X_i : i \in S_j\}$ a partitioned covariate vector. When p covariates are measured on each individual, both continuous and qualitative,

then $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$ will denote the ith individual's p-dimensional covariate vector and, with a slight abuse of notation, set $\boldsymbol{X}^\star_j = (\boldsymbol{X}^\star_{j1}, \ldots, \boldsymbol{X}^\star_{jp})$ where $\boldsymbol{X}^\star_{jh} = \{X_{ih} : i \in S_j\}$ for $h = 1, \ldots, p$. Thus, depending on context, $\boldsymbol{X}^\star_j$ could possibly be a super vector of stacked patients' covariate vectors.

We now introduce notation associated with the model that will be employed to connect $\boldsymbol{X}$ to $\boldsymbol{Y}$ through ρ_m. A dependent random partition prior distribution is assigned to ρ_m. This prior distribution parametrized by $\boldsymbol{\eta}$, will be denoted using $\mathrm{RPM}_X(\boldsymbol{\eta})$. Once a prior for ρ_m is specified, we will make use of $f(\boldsymbol{Y} \mid \rho) = \prod_{j=1}^{k_n} f_j(\boldsymbol{Y}^\star_j)$ as a data model where $f_j(\boldsymbol{Y}^\star_j) = \int \prod_{i \in S_j} f_j(Y_i \mid \boldsymbol{\theta}^\star_j)\, dG_0(\boldsymbol{\theta}^\star_j)$, $f_j(\cdot \mid \boldsymbol{\theta}^\star_j)$ denotes the likelihood for $\boldsymbol{Y}^\star_j$ and G_0 a prior for cluster specific parameters $\boldsymbol{\theta}^\star_j$. Note that unit-specific parameters can be connected to their cluster-specific counterpart via $\boldsymbol{\theta}_i = \boldsymbol{\theta}^\star_{s_i}$. Alternatively, the data model can be written hierarchically using the cluster labels $s_1, \ldots, s_n$ in the following way:

$$
\begin{aligned}
Y_i \mid \boldsymbol{\theta}^\star, s_i &\overset{ind}{\sim} f_{s_i}\big(\boldsymbol{\theta}^\star_{s_i}\big) \quad \text{for } i = 1, \ldots, m, \\
\boldsymbol{\theta}^\star_\ell &\overset{iid}{\sim} G_0 \quad \text{for } \ell = 1, \ldots, k_m, \\
\rho_m \mid \boldsymbol{X} &\sim \mathrm{RPM}_X(\boldsymbol{\eta}).
\end{aligned} \tag{3.1}
$$

Notice that covariates do not appear in the data model so that neither prespecified associations nor their forms are required. As a result, $\boldsymbol{Y}$ is only connected to $\boldsymbol{X}$ through the posterior distribution of ρ_m (denoted by $\pi(\rho_m \mid \boldsymbol{Y}, \boldsymbol{X})$) which will facilitate our interaction search. There are a number of computational techniques that have been developed to fit model (3.1). Most are some variant of MCMC that depends on the exact specification of $\mathrm{RPM}_X(\boldsymbol{\eta})$. We opt to employ algorithm 8 of Neal (2000) when fitting model (3.1), since it is a general algorithm that can be used in a variety of settings.

As already discussed, there are many possibilities for $\mathrm{RPM}_X(\boldsymbol{\eta})$. In what follows we employ the covariate dependent random partition model (PPMx) described in Müller, Quintana and Rosner (2011) (and further explored in Page and Quintana (2018)) because of its flexibility in being able to easily incorporate all types of covariates. The exact form of $\mathrm{RPM}_X(\boldsymbol{\eta})$ when adopting the PPMx is

$$
\mathrm{RPM}_X(\boldsymbol{\eta}) \propto \prod_{j=1}^{k_m} c(S_j) g\big(\boldsymbol{X}^\star_j \mid \boldsymbol{\eta}\big),
$$

where $c(S)$ is a set function that measures the chance that elements in S co-cluster a priori, and $g(\boldsymbol{X}^\star \mid \boldsymbol{\eta}) = \int \prod_{i \in S_j} q(\boldsymbol{X}_i \mid \boldsymbol{\xi}^\star) q(\boldsymbol{\xi}^\star \mid \boldsymbol{\eta})\, d\boldsymbol{\xi}^\star$ is a similarity function that produces higher values for $\boldsymbol{X}^\star$'s that contain covariate values that are more similar. Here, $q(\cdot \mid \cdot)$ denotes an auxiliary probability model (likelihood and prior) that has no effect in the data model and is used only to introduce the desired effect in the similarity function. When multiple covariates of different types are available, Müller, Quintana and Rosner (2011) suggest using the following multiplicative form:

$$
g\big(\boldsymbol{X}^\star_j \mid \boldsymbol{\eta}\big) = \prod_{\ell=1}^{p} g\big(\boldsymbol{X}^\star_{j\ell} \mid \boldsymbol{\eta}\big). \tag{3.2}
$$

3.2. *Association rules.* Association rules are used to discover patterns or relations among a large collection of variables. They are typically denoted using $\{A\} \Rightarrow \{B\}$ where A and B define a subset of the covariate space that does not share any variables. Connecting the idea to the osteonecrosis study, a possible association rule is $\{A = \mathrm{AgeAtDiagnosis} \in [3, 5)\} \Rightarrow \{B = \mathrm{DexClWk07} \in [9, 12)\}$ which would indicate that if *AgeAtDiagnosis* is between three and five, then *DexClWk07* is between nine and 12.

There are two criteria to evaluate the importance of $\{A\} \Rightarrow \{B\}$. The first, called *support*, is the proportion of patients whose covariate vector contains both A and B. The second criterion, called *confidence*, is the fraction of patients that display B among those that display A. Our criteria to determine which covariate pairs to consider when looking for interactions will be those with large support and confidence. For a more detailed overview of association rules, we refer interested readers to Hastie, Tibshirani and Friedman (2009), Chapter 14, or Han, Kamber and Pei (2012), Chapter 6.

The typical notation used for association rules highlights the fact that they are, by nature, directional. However, our only purpose in using association rules is to identify pairs of covariates that are associated with each other and possibly influence the distribution of Y in a nonadditive way. Because of this, we will use $X_1 \Leftrightarrow X_2$ to denote that, at least, one association rule contained the pair X_1, X_2 in some fashion as the direction of the association is immaterial for our purposes.

4. Interaction discovery procedure. In this section we provide more details of our three-step interaction discovery procedure, after which we provide two simulation studies. In what follows we will refer to the three-step procedure as the Random partition, Association rule, Interaction Discovery procedure (or the RAID procedure).

A natural way of employing $\pi(\rho_m \mid \boldsymbol{Y}, \boldsymbol{X})$ in the second stage of the exploratory procedure is to produce a point estimate of ρ_m using methods found in Dahl (2020) (or some other alternative) and then apply association rules to each of the resulting $\boldsymbol{X}_j^\star$s. However, as noted by Wade and Ghahramani (2018), there might be substantial variability associated with the partition point estimate. Therefore, we instead apply association rules to the resulting $\boldsymbol{X}_j^\star$s for each (or a subset) of the MCMC draws collected. This approach provides a means of propagating uncertainty associated with ρ_m through the exploratory procedure. Since the interactions identified will be based on specific clusters, they are local in the same way that treatment effects are local in subset analysis. That is, the interaction may not remain consistent across the entire population. (We explore this in the simulation study of Section 5.2.) We briefly note that applying association rules to each of the $\boldsymbol{X}_j^\star$'s is crucial to finding interactions, as they are connected to $\boldsymbol{Y}$ through $\pi(\rho_m \mid \boldsymbol{Y}, \boldsymbol{X})$. Applying association rules directly to $\boldsymbol{X}$ would make interaction detection impossible as there is no connection to $\boldsymbol{Y}$.

As stated previously, the third step consists of using posterior predictive densities to confirm interactions. To make this step concrete, consider an example with three binary covariates. Let $p_{ijk}(Y_0 \mid \boldsymbol{Y}, \boldsymbol{X})$ denote the posterior predictive density for $X_1 = i$, $X_2 = j$ and $X_3 = k$ with $i, j, k \in \{0, 1\}$, and assume that, within a cluster, the association rule with highest total support and confidence is $X_1 \Leftrightarrow X_2$. To verify that there does indeed exist what we call an interaction between (X_1, X_2), we test the following hypothesis (after fixing $X_3 = k$ to its empirical median):

$$H_0 : p_{00k}(Y_0 \mid \boldsymbol{Y}) = p_{01k}(Y_0 \mid \boldsymbol{Y}) = p_{10k}(Y_0 \mid \boldsymbol{Y}) = p_{11k}(Y_0 \mid \boldsymbol{Y}). \tag{4.1}$$

If the hypothesis is rejected, then we conclude that X_1 and X_2 interact. That is, the predictive distribution of Y is, in some way, influenced by X_1 and/or X_2, and the specific way in which this occurs can be later explored separately. A similar approach would be employed if association rules highlighted any of the other possible two-way interactions (i.e., $X_1 \Leftrightarrow X_3$, $X_2 \Leftrightarrow X_3$). We also consider the possibility of a three-way interaction in Section 6.

There are a number of procedures that might be selected to test the hypothesis found in (4.1). The Pólya tree procedure, outlined in Chen and Hanson (2014), is quite flexible and powerful, and thus we opt to use it in what follows. However, any other procedure that is able to formally test the hypothesis in (4.1) is completely valid. Just as with Chen and Hanson's (2014) method, most methods employed to test (4.1) will produce a "p-value." We use the

"p-values" as a validation tool, and thus there is no guarantee that Type 1 error rates are preserved. That said, the simulation in the Section 5.2 suggests that error rates are not far from the size of the test.

5. Simulation studies. In this section we detail two simulation studies carried out to explore the RAID procedure's ability to detect interactions. The first is based on a toy example that is included to provide insight into each of the three steps of our approach. The second is a more realistic setting that is based on the covariate structure found in the osteonecrosis data set. In both simulation studies a continuous response variable is used, even though the response in the osteonecrosis study is ordinal. This was done so that we could compare our approach to other procedures found in the literature. As far as we are aware, no other procedure is able to accommodate an ordinal response. This highlights the "model-free" property of our approach.

In order to connect $\boldsymbol{Y}$ to $\boldsymbol{X}$ through $\pi(\rho_m \mid \boldsymbol{Y}, \boldsymbol{X})$, we fit the following particular case of model (3.1) to each generated synthetic data set in both simulation studies:

$$
\begin{aligned}
Y_i \mid \boldsymbol{\mu}^\star, \boldsymbol{\sigma}^\star, s_i &\overset{ind}{\sim} N\big(\mu^\star_{s_i}, \sigma^\star_{s_i}\big) \quad \text{for } i = 1, \ldots, m, \\
\big(\mu^\star_\ell, \sigma^\star_\ell\big) &\overset{iid}{\sim} N\big(\mu_0, \sigma_0^2\big) \times \mathrm{Unif}(0, A) \quad \text{for } \ell = 1, \ldots, k_m, \\
\rho_m \mid \boldsymbol{X} &\sim \mathrm{RPM}_X(\boldsymbol{\eta}),
\end{aligned} \tag{5.1}
$$

with $\mu_0 \sim N(0, 10^2)$, $\sigma_0 \sim \mathrm{Unif}(0, 10)$ and $A = 1$. For the $\mathrm{RPM}_X(\boldsymbol{\eta})$ we used the PPMx with cohesion $c(S) = M \times (|S| - 1)!$ connecting the partition model to that induced by a Dirichlet process (DP) mixture and the "rich get richer" property (i.e., a model that favors a small number of large clusters). M has connections with the DP dispersion parameter, and we used $M = 1$. For the similarity we employed the auxiliary similarity function (see Müller, Quintana and Rosner (2011), Page and Quintana (2018)). As a result, for qualitative variables, $q(\cdot \mid \boldsymbol{\xi}^\star)$ and $q(\boldsymbol{\xi}^\star \mid \boldsymbol{\eta})$ correspond to multinomial and Dirichlet density functions, respectively, and for continuous variables they take on a Gaussian and Gaussian-Inverse-Gamma density functions. Thus, for qualitative covariates $\boldsymbol{\eta}$ is a vector of Dirichlet shape parameters that we set to 0.1, and for continuous covariates, $\boldsymbol{\eta}$ corresponds to the $m_0 =$ center, $k_0 =$ scale, $\nu_0 =$ shape and $\kappa_0 =$ rate parameters of Gaussian-Inverse-Gamma distribution. After standardizing continuous covariates, we employed $m_0 = 0$, $k_0 = 0.5$, $\nu_0 = 1$, $\kappa_0 = 2$. In this model the prior parameters that seem to have the most influence over $\pi(\rho_m \mid \boldsymbol{Y}, \boldsymbol{X})$ are A and k_0. These parameters determine how much weight is placed on $\boldsymbol{Y}$ or $\boldsymbol{X}$ when forming clusters. We explore sensitivity to the specification of these prior parameters in Section 6.1.

5.1. *Simulation study*: *Toy example*. In this simulation, data sets consisting of three binary covariates are generated (i.e., $X_i \in \{0, 1\}$, for $i = 1, 2, 3$). The response distribution $f(Y \mid X_1, X_2, X_3)$ is made to explicitly depend on $\boldsymbol{X} = (X_1, X_2, X_3)$ in the three ways that are listed in Table 4. Each row of the table represents a possible covariate combination, and the columns correspond to a particular response distribution. Note first from the table that only X_1 and X_2 influence the distribution of Y, and they do so multiplicatively (i.e., they interact in the linear model sense). As a control, we consider the case where the distribution of Y does not depend on $\boldsymbol{X}$, in any way, which is denoted by $f_0(Y \mid \boldsymbol{X})$. From Table 4 note that $f_0(Y \mid \boldsymbol{X})$ corresponds to a standard normal regardless of the values of $\boldsymbol{X}$. For $f_1(Y \mid \boldsymbol{X})$, the variance and shape remain the same regardless of the value of $\boldsymbol{X}$, but the mean changes based on the values of X_1 and X_2. For $f_2(Y \mid \boldsymbol{X})$, the mean and shape remain the same, but the variance changes, depending on the values of X_1 and X_2. Lastly, for $f_3(Y \mid \boldsymbol{X})$ the mean and variance do not change, but the values X_1 and X_2 influence the shape. The scenario

TABLE 4

List of distributions used in the simulation study. Here, $N(a, b)$ denotes a normal distribution with mean a and standard deviation b, $SN(a, b, c)$ denotes a skew-normal distribution with location ($a = 10$), scale ($b = 1$) and skew ($c = 20$) parameters. For the two-component mixture we set $p_1 = p_2 = 0.5$, $s_1^2 = s_2^2 = 1/16$ and $m_1 = -m_2 = \sqrt{15}/4$

X_1	X_2	X_3	$f_0(Y \mid \boldsymbol{X})$	$f_1(Y \mid \boldsymbol{X})$	$f_2(Y \mid \boldsymbol{X})$	$f_3(Y \mid \boldsymbol{X})$
1	1	1	$N(0, 1)$	$N(0, 1)$	$N(0, 1)$	$N(0, 1)$
1	1	0	$N(0, 1)$	$N(0, 1)$	$N(0, 1)$	$N(0, 1)$
1	0	1	$N(0, 1)$	$N(2, 1)$	$N(0, 3)$	$SN(a, b, c)$
1	0	0	$N(0, 1)$	$N(2, 1)$	$N(0, 3)$	$SN(a, b, c)$
0	1	1	$N(0, 1)$	$N(0, 1)$	$N(0, 1)$	$N(0, 1)$
0	1	0	$N(0, 1)$	$N(0, 1)$	$N(0, 1)$	$N(0, 1)$
0	0	1	$N(0, 1)$	$N(4, 1)$	$N(0, 6)$	$\sum_{j=1}^{2} p_j N(m_j, s_j^2)$
0	0	0	$N(0, 1)$	$N(4, 1)$	$N(0, 6)$	$\sum_{j=1}^{2} p_j N(m_j, s_j^2)$

under $f_1(Y \mid \boldsymbol{X})$ follows the traditional definition of an interaction and thus can be detected using any number of procedures. The other two scenarios would not be detected, even if an interaction between X_1 and X_2 is explicitly included in a linear model.

For each scenario we generated 1000 data sets each with 500 observations and fit model (5.1). At each MCMC iterate of ρ_m, cluster-specific association rules were gathered using the `apriori` function found in the `arules` package (Hahsler et al. (2015)) of the statistical software `R` (R Core Team (2017)). This function is based on the a priori algorithm of Agrawal et al. (1996). We considered association rules from clusters that contained at least 10 subjects/experimental units. After association rules were gathered using a lower bound support of 0.25 and confidence of 0.5, we identified the covariate tandem that had the highest total support and confidence. If this resulted in a tie, then both covariate tandems were considered. Then, for each association rule the Pólya tree-based testing procedure of Chen and Hanson (2014) was used to test the hypothesis in (4.1). The X variable absent from the association rule was fixed to its empirical median value. The p-values from this testing procedure are derived from permutation tests and values in Table 5 are based on 500 permutations.

Lastly, since the number of posterior predictive draws collected to test (4.1) can be selected somewhat arbitrarily, the adage "statistical vs. practical significance" is important here, as it would be undesirable to reject (4.1) for differences among predictive densities that are inconsequential. Thus, when carrying out the test of Chen and Hanson (2014), we considered $N \in \{50, 100, 250\}$ posterior predictive draws.

In Section 1.2 of the online Supplementary Material (Page, Quintana and Rosner (2021a)), we provide more details of the simulation results for each of the 24 possible association rules, but here we focus on results for testing (4.1) which are found in Table 5. Each column of the

TABLE 5

p-values averaged over 1000 synthetic datasets from testing hypothesis (4.1) using the Pólya tree method of Chen and Hanson (2014)

	$N = 50$				$N = 100$				$N = 250$			
arules	f_0	f_1	f_2	f_3	f_0	f_1	f_2	f_3	f_0	f_1	f_2	f_3
$X_1 \Leftrightarrow X_2$	0.53	0.00	0.00	0.02	0.48	0.00	0.00	0.00	0.44	0.00	0.00	0.00
$X_1 \Leftrightarrow X_3$	0.47	0.54	0.50	0.48	0.51	0.38	0.50	0.45	0.44	0.37	0.43	0.41
$X_2 \Leftrightarrow X_3$	0.49	0.00	0.00	0.36	0.50	0.00	0.00	0.25	0.43	0.00	0.00	0.14

table contains the p-values produced by the Chen and Hanson (2014) Pólya tree method averaged across the 1000 data sets. The Monte Carlo errors, associated with entries found in Table 5, are all small with the largest being 0.01. From Table 5 it seems that, for this data-generating mechanism, it is sufficient to only consider 50 posterior predictive draws when testing the hypothesis for all three interaction types.

From Table 5 it also seems that, under f_0, the average p-value is large in all cases, indicating that the posterior predictive densities in (4.1) are essentially the same regardless of the number of posterior predictive draws and association rule. This was expected. Similarly, $X_1 \Leftrightarrow X_3$ under f_1, f_2 and f_3 and $X_2 \Leftrightarrow X_3$ under f_3 produced large p-values which was also expected. Since there is an interaction between X_1 and X_2, the small p-values associated with $X_1 \Leftrightarrow X_2$ under f_1, f_2, and f_3 were expected. However, notice that, under f_1 and f_2, the association rule that includes X_2 and X_3 also produced small p-values for (4.1). This was unexpected. Since X_2 and X_3 do not interact, fixing X_1 makes it so that (4.1) amounts to testing for the "marginal effect" of X_2. Thus, small p-values associated with hypothesis (4.1) can indicate an interaction or a marginal effect. To correctly determine which—further testing can be carried out or the posterior predictive densities being compared in (4.1)—should be carefully examined. Here, we carried out the later by providing the posterior predictive densities in Figure 1. Estimated densities in the figure correspond to the first synthetic data set generated based on f_1. The first row of the figure corresponds to densities when X_3 is fixed, the second row when X_2 is fixed and the third when X_1 is fixed. Notice that the densities in the first row are noticeably different and clearly depend on the levels of X_1 and X_2 (i.e., there is an interaction between X_1 and X_2). Close inspection of rows 2 and 3 of the figure shows that differences seen in the posterior predictive densities are not due to X_3 at all, but entirely to X_1 for the second row and X_2 for the third. Thus, graphs like Figure 1 are able to help determine if interactions correspond to "main effects" or "interaction effects," as they are typically defined in a linear model setting. Ultimately, the take-home message of Table 5 and Figure 1 is that the exploratory procedure is able to identify that X_1 and X_2 interact regardless of the influence that they have on the distribution of Y.

Next, we compared the performance of our method to five alternative approaches that detect interactions. To this end, to each generated data set we applied the following procedures:

1. Linear model with all main effects and two-way interactions;
2. Spike & slab model of Ishwaran and Rao (2005) with all main effects and two-way interactions;
3. The hierarchical group-Lasso (hgLASSO) procedure in Lim and Hastie (2015);
4. The BART procedure detailed in Kapelner and Bleich (2016);
5. Procedure based on XGBoost (Karbowiak and Biecek (2019)).

Each of the listed methods were fit in `R` using the following: `lm` function for linear model, the `spikeslab` package (Ishwaran, Rao and Kogalur (2013)) for spike & slab, the `glinternet` package (Lim and Hastie (2019)) for hgLASSO, the `bartMachine` package (Kapelner and Bleich (2016)) for BART and the `xgboost` package (Chen et al. (2019)) coupled with the `EIX` package (Karbowiak and Biecek (2019)) for XGBoost. To apply the procedures just listed, some type of user input was required to confirm that an interaction was detected. For the sake of conciseness, we provide specific details of how procedures were implemented in Section 1.1 of the online Supplementary Material (Page, Quintana and Rosner (2021a)).

To study the impact that the RPM prior might have on results, we also used the RAID procedure coupled with the induced random partition distribution from the ANOVA DDP. The particular ANOVA DDP we employed is that for which the atoms were made to depend on group membership (as defined by the three categorical covariates), but weights did not. Lastly, we checked the sensitivity of results to posterior convergence by identifying interactions based on an MCMC chain that was run for only 500 iterations

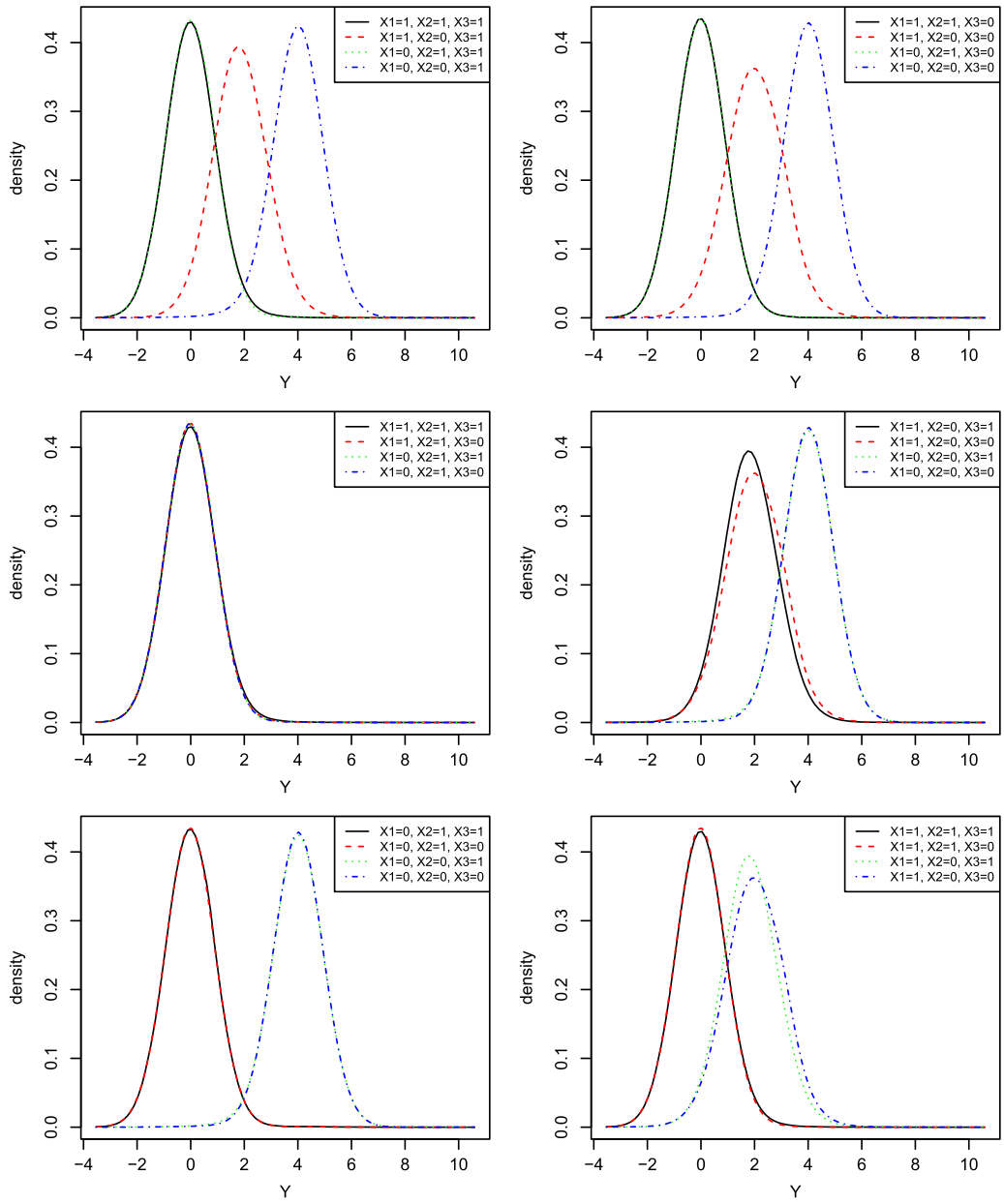

FIG. 1. *Posterior predictive densities for the six combinations of X_1, X_2 and X_3 based on a synthetic data set used in the simulation study.*

The values listed in Table 6 are the proportion of data sets for which the interaction between X_1 and X_2 was correctly identified. The Monte Carlo standard errors associated with proportions in Table 6 were all very small with 0.09 (associated with "RAID with ANOVA DDP, $N = 100$") being the largest. Notice that, when no interaction was present, all procedures performed similarly, except for XGBoost and RAID with ANOVA DDP. These two procedures seemed to be more liberal in their interaction detection. When the interaction was based on the expectation of the response distribution (column f_1), all procedures detected the interaction except for XGBoost. However, if the interaction was based on the variance or shape of the response distribution, the RAID procedure with PPMx was the only one that

TABLE 6
Percent of data sets in the simulation study for which the interaction between X_1 and X_2 was correctly identified

Procedure	f_0	f_1	f_2	f_3
Linear Model	0.04	1.00	0.05	0.06
Spike & Slab	0.08	1.00	0.05	0.09
hgLASSO	0.09	1.00	0.12	0.10
BART	0.00	0.99	0.00	0.00
XGBoost	0.27	0.54	0.37	0.40
RAID with ANOVA DDP, $N = 50$	0.19	1.00	1.00	0.23
RAID with ANOVA DDP, $N = 100$	0.44	1.00	1.00	0.47
RAID with ANOVA DDP, $N = 250$	0.61	1.00	1.00	0.87
RAID with PPMx, $N = 50$	0.06	1.00	1.00	0.96
RAID with PPMx, $N = 100$	0.04	1.00	1.00	1.00
RAID with PPMx, $N = 250$	0.08	1.00	1.00	1.00
Results with only 500 MCMC iterates				
RAID with PPMx, $N = 50$	0.10	1.00	1.00	0.23
RAID with PPMx, $N = 100$	0.28	1.00	1.00	0.73
RAID with PPMx, $N = 250$	0.13	1.00	1.00	1.00

consistently detected it. Lastly, the RAID procedure still performed surprisingly well, even when the number of collected MCMC iterates was small.

5.2. *Simulation study*: *Osteonecrosis study covariate structure*. We now describe a simulation study where the number of covariates in each synthetic data set is similar to that found in the osteonecrosis study. More specifically, each synthetic data set contained 21 covariates, 19 of which were continuous and two binary. The binary covariate values were generated using a Bernoulli distribution with probability of success equal to 0.5. The continuous covariates were generated using a continuous uniform distribution with support $(-1, 1)$. We created interactions that impact the mean, spread and shape of the response distribution in two ways. The first employed the setup established in Table 4, except that instead of one noise covariate that was categorical, there are 19 continuous noise covariates.

The second method of generating data sets was based on interactions that depend on continuous covariates. This was done by using the following data generating mechanisms to generate response values:

- $Y \sim N(5X_1X_2, \sigma)$ (interaction affecting mean);
- $Y \sim N(0, \exp\{5X_1X_2\}\sigma)$ (interaction affecting spread);
- $Y \sim \pi(X_1, X_2)N(0, \sigma) + (1 - \pi(X_1, X_2))[0.5N(-0.5, \sigma) + 0.5N(0.5, \sigma)]$ where $\pi(X_1, X_2) = 1/(1 + \exp\{200X_1X_2\})$ (interaction affecting shape).

For this setting X_1 and X_2 were continuous covariates generated using a uniform distribution with support $(-1, 1)$. Similar to what was done before, 19 "noise" covariates were also included, 17 of which were continuous (coming from a uniform distribution with support $(-1, 1)$) and two binary (coming from a Bernoulli distribution with probability of success equal to 0.5).

We explored how the ratio of signal to noise affected the ability to detect interactions by considering $\sigma \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. Finally, we also considered the case when only 60% and 25% of observations exhibited the interaction. This was done by using a $N(0, 1)$ distribution to generate response values for 40% (or 75%) of the units.

To each synthetic data set, the RAID PPMx procedure was fit along with the six competitors introduced in the previous section. For each procedure we recorded if the interaction was detected (True Positive) and also enumerated the number of falsely detected interactions (False Positives). We briefly note that, since association rules require discrete variables, before employing association rules for each cluster, the continuous covariates were dichotomized such that an equal number of observations belonged to each interval.

Here, we only describe results for the case when the interaction was present in 100% of the population and include results when the interaction was present in only 60% or 25% of the population in the online Supplementary Material (see Figures S.1–S.4). Trends were similar in all scenarios, except that it became more challenging to detect interactions as the percent of population that exhibits interaction decreased, with the RAID method being the least impacted when the interaction only applies to a subset of the population.

In Figure 2, notice that, except for XGBoost, all methods did reasonably well in recovering the interaction that affects the mean between two categorical covariates, with the RAID and LM methods decaying more when the signal-to-noise ratio decreased. When interactions based on continuous covariates affected mean, the RAID method did the worst (only detecting the interaction in about 50% of the synthetic datasets). However, when interactions affected spread or shape, the RAID procedure is the only method that was able to detect the interaction with any type of regularity. The BART method won when interaction between continuous covariates affected the spread, and the XGBoost method tended to perform the worst.

Figure 3 displays the false positives. It seems that the spike & slab and RAID procedures tended to produce the most amount of false positives, while the LASSO and BART method did reasonably well in avoiding this. Upon further investigation, many of the false positives detected by the RAID procedure included one of X_1 or X_2 (covariates that participated in the

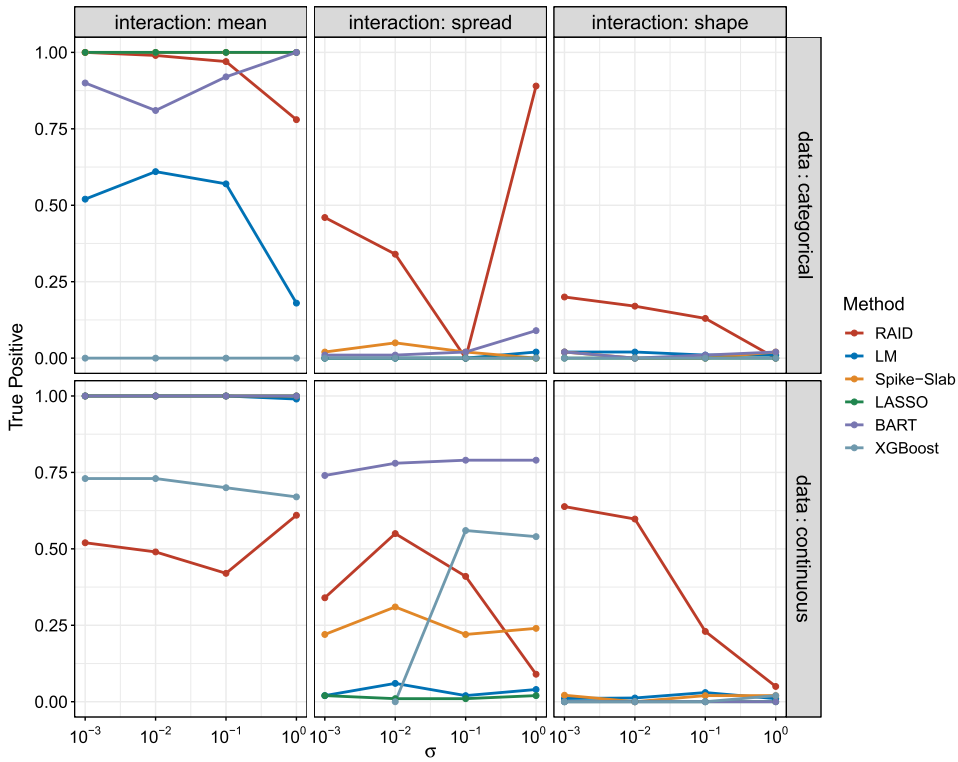

FIG. 2. *Results from the second simulation study*: *Percentage of synthetic datasets for which the interaction was detected. In each figure the top row corresponds to an interaction that is based on categorical covariates and the bottom row to continuous covariates.*

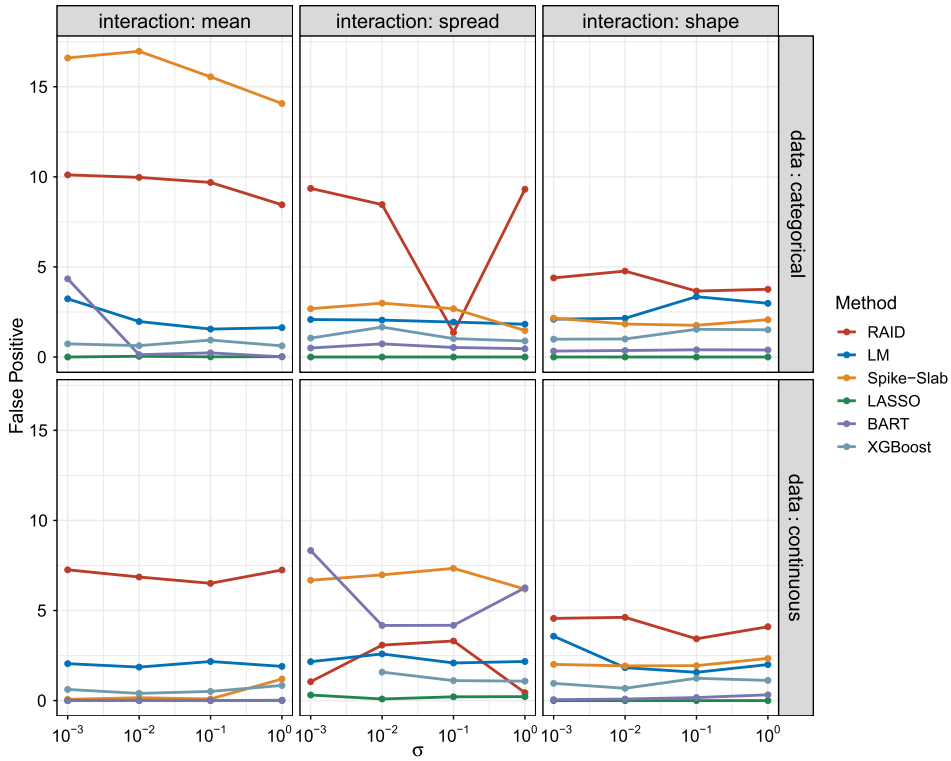

FIG. 3. *Results from the second simulation study*: *Average number of falsely detected interactions. In each figure the top row corresponds to an interaction that is based on categorical covariates and the bottom row to continuous covariates.*

interaction). In fact, if interactions that included either X_1 or X_2 were not classified as false positives, then the number of false positives detected by the RAID procedure, averaged over all data generating scenarios, was only 1.01. Thus, as discussed in the previous section, the RAID procedure detects any combination of covariates that affect the response distribution in some way.

6. Interactions in the osteonecrosis study. We now turn our attention to the osteonecrosis study. Since the response is an ordinal variable, it is natural to consider latent variable ordinal data models, as those described in Bao and Hanson (2015) and in Kottas, Müller and Quintana (2005). These types of models are clearly more complex than model (3.1), but they permit demonstrating the flexibility of the exploratory procedure. In fact, any data model can be employed so long as the response is connected to the covariates through a dependent random partition model (or analogous predictor dependent clustering procedures).

For the sake of completeness, we detail the ordinal model in its entirety. Let $Y_i \in \{0, 1, 2, 3, 4\}$ be the ordinal response measured on the ith patient $i = 1, \ldots, 234$. Further, let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{i22})$ denote subject i's 22-dimensional covariate vector. As is done in Bao and Hanson (2015), real-valued latent scores $Z_1, \ldots, Z_n$ are introduced such that, for $-\infty = \gamma_0 < \gamma_1 < \cdots \gamma_4 < \gamma_5 = \infty$,

$$Y_i = \ell \quad \Leftrightarrow \quad \gamma_\ell < Z_i \leq \gamma_{\ell+1} \quad \text{for } \ell = 0, \ldots, 4.$$

The appeal of employing the methods described in Bao and Hanson (2015) and Kottas, Müller and Quintana (2005) is that the values selected for γ are immaterial, so long as the model is flexible enough to assign sufficient probability mass to each $(\gamma_\ell, \gamma_{\ell+1}]$ interval. Since assigning a PPMx prior to ρ_m results in modeling the latent ordinal scores with a mixture (which

is essentially the same data model arising from marginalizing the random measure in the approach by Bao and Hanson (2015)), we conclude that the model we specify affords sufficient flexibility and, therefore, the γ can be selected arbitrarily. In light of this, we set $\gamma_1 = 0$, $\gamma_2 = 1/3$, $\gamma_3 = 2/3$, $\gamma_4 = 3/3$. The remainder of the model is expressed hierarchically after introducing latent cluster labels $s_1, \ldots, s_n$

$$
\begin{aligned}
Z_i \mid \boldsymbol{\mu}^\star, \boldsymbol{\sigma}^{2\star}, s_i &\sim N\big(\mu^\star_{s_i}, \sigma^{2\star}_{s_i}\big), \\
\big(\mu^\star_j, \sigma^\star_j\big) &\sim N\big(\mu_0, \sigma_0^2\big) \times \mathrm{Unif}(0, A), \\
\Pr(\rho \mid \boldsymbol{X}) &\propto \prod_{j=1}^{k} c(S_j) g\big(\boldsymbol{X}^\star_j\big),
\end{aligned}
\tag{6.1}
$$

where $c(S_j)$ and $g(\boldsymbol{X}^\star_j)$ are set to functions detailed in Section 3.1. Further, we assume $\mu_0 \sim N(0, 10^2)$, $\sigma_0 \sim \mathrm{Unif}(0, 10)$ and set $A = 0.1$.

This model was fit to the osteonecrosis data by collecting 1000 MCMC iterates after discarding the first 10,000 as burn-in and thinning by 15. Afterward, we identified association rules for each MCMC iterate of ρ_m. Covariate values for those variables not identified in the association rule are set to the overall median. Testing (4.1) was done by using Chen and Hanson (2014)'s method based on 500 permutations and 100 posterior predictive draws. This testing procedure was replicated five times with a unique set of 100 random posterior predictive draws, and the average p-value is reported. (We stress again that p-values are used only as a validation summary, and no notion of Type I error is implied.) As in the simulation study, each continuous covariate was dichotomized such that an equal number of observations belonged to each interval when employing association rules. Also, we explored the impact that trichotomizing continuous covariates might have on the RAID exploratory procedure. Results are provided in Section 2.1 of the online Supplementary Material (Page, Quintana and Rosner (2021a)).

Since the study was conducted specifically to explore if dexamethasone clearance (DexCl) and/or asparaginase (AsparaginaseAntibodyAUC) interact with other physiological measurements, we restrict attention to those association rules that contain at least one of these two covariates. Results for interactions that appeared in at least 50% of the MCMC iterates are provided in Table 7. The column "*Pr*" corresponds to the percent of MCMC iterates that identified the particular association rule, "Supp." is the support, "Conf." the confidence, $|S|$ corresponds to average cluster size from which association rule was identified and "p-value" indicates the result of testing (4.1). Details associated with the "Prior#" column are provided in Section 6.1. Notice that DexClWk07 interacts with LowRisk and AlbuminWk07 while

TABLE 7

*Association rules from the osteonecrosis data with continuous covariates being dichotomized. Here, only association rules that contain either DexClwk*07, *DexClwk*08, *or AsparaginaseAntibodyAUC are considered*

Association Rule	*Pr*	Supp.	Conf.	$\lvert S\rvert$	p-value	Prior#
DexClWk07 ⇔ LowRisk	1.00	0.62	0.91	43.44	0.00	16
DexClWk07 ⇔ AlbuminWk07	0.99	0.58	0.85	39.65	0.01	15
AsparaginaseAntibodyAUC ⇔ AlbuminWk07	0.93	0.84	0.95	39.35	0.00	14
AsparaginaseAntibodyAUC ⇔ HDLWk07	0.81	0.72	1.00	10.40	0.01	7
AsparaginaseAntibodyAUC ⇔ LowRisk	0.81	0.87	0.98	40.55	0.18	17
AsparaginaseAntibodyAUC ⇔ LDLWk08	0.55	0.53	0.88	15.71	0.00	6
DexClWk08 ⇔ HDLWk08	0.53	0.51	0.88	16.36	0.00	8
DexClWk08 ⇔ AlbuminWk08	0.50	0.83	0.98	10.66	0.85	3

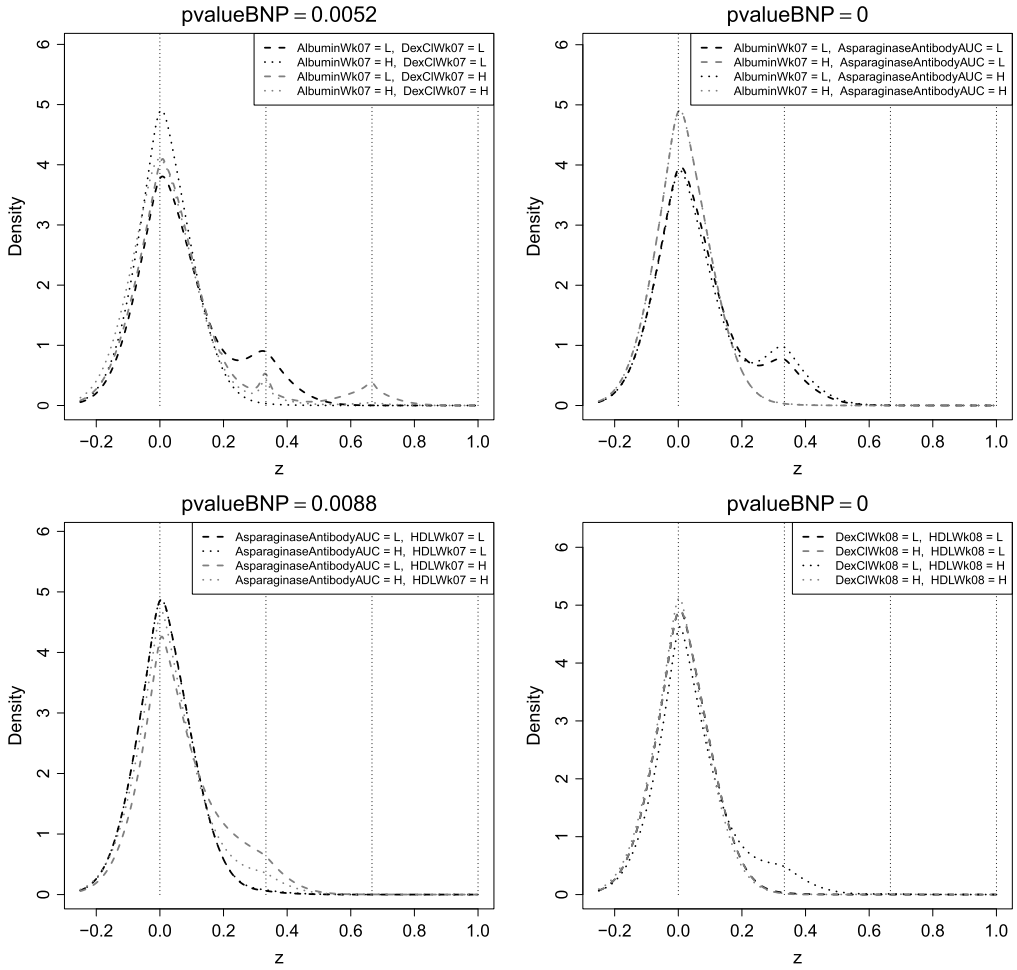

FIG. 4. *Posterior predictive densities for the latent variable z and used in hypothesis test* (4.1) *for the first four association rules in Table* 7 *that do not include LowRisk. The vertical lines correspond to* $\gamma_1, \ldots, \gamma_4$.

DexClWk08 interacts with HDLWk08. On the other hand, Asparaginase seems to interact with AlbuminWk07, HDLWk07 and LDLWk08. Figure 4 contains the posterior predictive densities corresponding to the first four association rules listed in Table 7 that do not contain LowRisk. Interestingly, note that the interaction between AlbuminWk07 and AsparaginaseAntibodyAUC influences the shape of the predictive densities, verifying that there is indeed an interaction between them and that the interaction seems to impact the shape of the response distribution (as evidenced by changes in shape of the predictive distributions).

Being able to interpret results at the latent level is not always straightforward. Thus, it would be appealing to determine how interactions influence the risk of osteonecrosis. Table 8 contains predictive probabilities associated with each osteonecrosis grade for specific levels of covariates found in interactions displayed in Figure 4 that contain DexCl. It appears that low level of DexCl results in higher risk of grade 2 or higher osteonecrosis, but the exact risk depends on the levels of the other covariates (AlbuminWk07 and HDLWk08). The combination of low DexClWk07 and high AlbuminWk07 results in the highest risk of osteonecrosis. This leads one to hypothesize that a three-way interaction between Albumin, HDL and DexCl may be present. We explored this by testing a version of (4.1) that includes posterior predictive densities for all possible combinations of the three covariates (eight in total). Doing this resulted in rejecting the null (with a p-value of 0) that all densities are equal, and, hence, we

TABLE 8
Posterior probabilities for each severity grade of osteonecrosis corresponding to two association rules found in Table 7

Covariates	Osteonecrosis Grade				
	0	1	2	3	4
AlbuminWk07: L, DexClWk07: L	0.124	0.818	0.058	0.000	0.000
AlbuminWk07: L, DexClWk07: H	0.193	0.805	0.002	0.000	0.000
AlbuminWk07: H, DexClWk07: L	0.282	0.643	0.061	0.014	0.000
AlbuminWk07: H, DexClWk07: H	0.414	0.569	0.017	0.000	0.000
DexClWk08: L, HDLWk08: L	0.200	0.799	0.001	0.000	0.000
DexClWk08: L, HDLWk08: H	0.195	0.805	0.000	0.000	0.000
DexClWk08: H, HDLWk08: L	0.287	0.682	0.030	0.001	0.000
DexClWk08: H, HDLWk08: H	0.330	0.667	0.003	0.000	0.000

conclude that a three-way interaction exists. More details are provided in Section 2.2 of the online Supplementary Material (Page, Quintana and Rosner (2021a)).

Section 2.3 of the online Supplementary Material also provides details on how the RAID procedure can accommodate the temporal structure that exists among the covariates through the PPMx prior. This is done by considering a multivariate similarity. As can bee seen in Section 2.3 of the online Supplementary Material, including temporal dependence in the partition prior affects partition configuration and, as a result, interactions that are identified by the RAID procedure.

6.1. *Sensitivity to prior specification.* Even though we developed the RAID procedure to be as automatized as possible (after standardizing covariates and possibly the response), "tuning" parameters that may affect the list of possible interactions exist. Those that are most influential are prior parameter specifications that have direct impact on $\pi(\rho_m \mid \boldsymbol{Y}, \boldsymbol{X})$. In Section 5 we mentioned that, for model (5.1), these are A and k_0, as they regulate the influence that $\boldsymbol{Y}$ and $\boldsymbol{X}$ have on cluster formation. Note that as $A \to 0$, clusters become more heterogeneous with respect to $\boldsymbol{Y}$ while as $k_0 \to 0$ clusters become more heterogeneous with respect to $\boldsymbol{X}$. To explore sensitivity to these three prior specifications, we ran the RAID procedure for the osteonecrosis data based on $A \in \{0.1, 1, 10\}$ and $k_0 \in \{0.1, 1, 10\}$. Since the cohesion function also has an effect on $\pi(\rho_m \mid \boldsymbol{Y}, \boldsymbol{X})$, we ran the RAID procedure using $c(S_j) = 1$ (a uniform type cohesion function) in addition to $c(S_j) = M \times (|S| - 1)!$. This resulted in 18 prior configurations and, as a result, 18 runs of the RAID procedure. On average, it took 15 minutes to apply the RAID procedure to each of the 18 prior configurations. In Section 2.4 of the online Supplementary Material (Page, Quintana and Rosner (2021a)), we provide Tables S.5-S.22 that list the potential interactions for each of the separate 18 runs. Here, we briefly summarize the results. Values in the column labeled "Prior#" in Table 7 are the number of prior configurations that detected the corresponding interaction. Counting up the total number of unique interactions detected across the 18 runs of the RAID procedure resulted in 23 unique interactions. Each of the interactions listed in Table 7 were among the unique 23. There was a general agreement in the interactions detected among the 18 runs with four of the interactions in Table 7 appearing in the majority of the prior configurations. Overall, it seems that results are fairly robust to prior specifications.

Lastly, as a type of follow-up analysis, using the `polr` function found in the `MASS` R-package (Venables and Ripley (2002)), we fit an ordered logistic regression model that included all eight two-way interactions listed in Table 7 and their corresponding main effects

along with the three-way interaction detailed in Section 2.2 of the online Supplementary Material. (We remark that fitting a saturated ordered logistic regression model with all two-way interactions for these data is not possible.) This model did provide some evidence to suggest that an interaction between DexClWk08 and HDLWk08 along with `AsparaginaseAntibodyAUC` and HDLWk07 was present in addition to the three-way interaction but failed to identify any of the other interactions. This suggests that the added flexibility of our exploratory approach may be useful to detect interactions that would not be possible under more restricted alternatives.

7. Conclusions. We have detailed an exploratory procedure that, in a general way, is able to identify interactions. The definition of interaction that we espouse is more general than that typically associated with linear models. Here, we conceptualize an interaction as influencing any aspect of the response distribution rather than focusing solely on the first moment. Further, the procedure is able to identify interactions without a priori information regarding which (multiplicative) effects to include in a data model. This was seen when applying the procedure to the osteonecrosis data set. In the application, relationships that were known to exist (e.g., Low DexCl and higher risk of osteonecrosis) were identified without considering them explicitly in the procedure. In addition, a three-way interaction, also not explicitly included in the data model, was highlighted (see Section 2.2 Supplementary Material). This interaction was not previously known by the investigators, and its scientific relevance has yet to be determined.

We also demonstrated that the procedure can be easily employed regardless of the type of response being measured. Indeed, the procedure is essentially "data model free," as it can be employed regardless of the data model so long as some type of random partition model is employed. Note, however, that our exploratory procedure only identifies potential interactions. That is, it does not estimate their effect. If this is desired, then any number of statistical procedures that include the specific covariates identified by our procedure can be employed (i.e., some form of regression).

One limitation of the RAID procedure (which is more an MCMC limitation) is that computation associated with the PPMx does not tend to scale well as p and/or n grows. Future research will be dedicated to implementing methods in the RAID procedure that permits it to scale well.

Finally, we provide some general recommendations when employing the RAID procedure. For the first step we recommend using the PPMx as the RPM. To automate the prior specification of model (5.1), we suggest standardizing the response and covariates and setting $A = 0.5$ and $k_0 = 0.5$. This essentially puts equal weight on $\boldsymbol{Y}$ and $\boldsymbol{X}$ when forming clusters. Decisions associated with step two determine how big a "net" one wants to cast when searching for interactions with the understanding that a larger "net" could result in more false positives. Decreasing support and confidence will cast a larger "net" as will decreasing the number of MCMC iterates in which an association must appear. Informal explorations led us to set support to 0.25, confidence of 0.5 and that a particular association rule should appear in 50% of MCMC iterates before being formally considered. For step three we suggest using 100 posterior predictive draws when comparing predictive densities and a cut-off of 0.01 for "p-values" if the Chen and Hanson (2014) method is used. This is somewhat arbitrary, but seemed to identify interactions well in the simulations. Lastly, we recommend using posterior predictive densities to visualize the interactions. Doing so will highlight how the interaction is affecting the response distribution, be it a shift in expectation as in Figure 1 or a change in shape as in Figure 4. If it is determined that the interaction affects the expectation of the response distribution, then a follow-up analysis to quantify interaction effects can be carried out using generalized linear models. If the interaction affects the response distribution in

some other way, then formal additional analysis is not available and quantifying the interaction effects will require other (perhaps ad-hoc) approaches similar to our approach with the osteonecrosis data set.

Acknowledgments. We thank all the reviewers for comments and suggestions that helped us to greatly improve this manuscript. We thank Drs. Yuhui Chen and Timothy Hanson for kindly sharing code that carries out the Pólya Tree permutation test. Garritt L. Page was also affiliated with Basque Center of Applied Mathematics and was partially supported by the Basque Government through the BERC 2018-2021 program, by the Spanish Ministry of Science, Innovation and Universities through BCAM Severo Ochoa accreditation SEV-2017-0718. Fernando A. Quintana is also affiliated to ANID—Millennium Science Initiative Program—Millennium Nucleus Center for the Discovery of Structurs in Complex Data. This work was supported by grant FONDECYT 1180034 and by ANID—Millennium Science Initiative Program—NCN17_059. Gary L. Rosner was partially funded by grants GM092666 and P30CA006973.

SUPPLEMENTARY MATERIAL

Supplement to "Discovering interactions using covariate informed random partition models" (DOI: 10.1214/20-AOAS1372SUPPA; .pdf). Supplement contains additional details and results associated with simulations studies and the osteonecrosis application.

Computer codes for 'Discovering interactions using covariate informed random partition models" (DOI: 10.1214/20-AOAS1372SUPPB; .zip). R-scripts for methods described in this paper and the osteonecrosis dataset.

REFERENCES

AGRAWAL, R., MANNILA, H., SRIKANT, R., TOIVONEN, H. and VERKAMO, A. I. (1996). Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds.) 307–328. American Association for Artificial Intelligence, Menlo Park, CA, USA.

AGRAWAL, R., TRIPPE, B., HUGGINS, J. and BRODERICK, T. (2019). The kernel interaction trick: Fast Bayesian discovery of pairwise interactions in high dimensions. In *Proceedings of the* 36*th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.). *Proceedings of Machine Learning Research* **97** 141–150. PMLR, Long Beach, CA.

AMERICAN CANCER SOCIETY (2018). Survival Rates for Childhood Leukemias. https://www.cancer.org/cancer/leukemia-in-children/detection-diagnosis-staging/survival-rates.html [Accessed: 18 March 2018].

BAO, J. and HANSON, T. E. (2015). Bayesian nonparametric multivariate ordinal regression. *Canad. J. Statist.* **43** 337–357. MR3388321 https://doi.org/10.1002/cjs.11253

BARCELLA, W., DE IORIO, M., FAVARO, S. and ROSNER, G. L. (2018). Dependent generalized Dirichlet process priors for the analysis of acute lymphoblastic leukemia. *Biostatistics* **19** 342–358. MR3984987 https://doi.org/10.1093/biostatistics/kxx042

BARRERA-GÓMEZ, J., AGIER, L., PORTENGEN, L., CHADEAU-HYAM, M., GIORGIS-ALLEMAND, L., SIROUX, V., ROBINSON, O., VLAANDEREN, J., GONZÁLEZ, J. R. et al. (2017). A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environ. Health* **16** 74.

BERGER, J. O., WANG, X. and SHEN, L. (2014). A Bayesian approach to subgroup identification. *J. Biopharm. Statist.* **24** 110–129. MR3196130 https://doi.org/10.1080/10543406.2013.856026

BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805 https://doi.org/10.1214/13-AOS1096

CHEN, T. and GUESTRIN, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD* '16 785–794. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2939672.2939785

CHEN, Y. and HANSON, T. E. (2014). Bayesian nonparametric k-sample tests for censored and uncensored data. *Comput. Statist. Data Anal.* **71** 335–346. MR3131974 https://doi.org/10.1016/j.csda.2012.11.003

CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V., TANG, Y., CHO, H., CHEN, K., MITCHELL, R., CANO, I. et al. (2019). xgboost: Extreme gradient boosting. R package version 0.90.0.2.

CHUNG, Y. and DUNSON, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Amer. Statist. Assoc.* **104** 1646–1660. MR2750582 https://doi.org/10.1198/jasa.2009.tm08302

DAHL, D. B. (2020). salso: Sequentially-allocated latent structure optimization. R package version 0.1.11.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. and MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99** 205–215. MR2054299 https://doi.org/10.1198/016214504000000205

DU, J. and LINERO, A. R. (2018). Interaction detection with Bayesian decision tree ensembles. arXiv:1809.08524.

DUNSON, D. B., PILLAI, N. and PARK, J.-H. (2007). Bayesian density regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 163–183. MR2325270 https://doi.org/10.1111/j.1467-9868.2007.00582.x

FAN, J., YAO, Q. and TONG, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83** 189–206. MR1399164 https://doi.org/10.1093/biomet/83.1.189

FERRARI, F. and DUNSON, D. B. (2019). Bayesian factor analysis for inference on interactions. arXiv:1904.11603v1.

GABRY, J., SIMPSON, D., VEHTARI, A., BETANCOURT, M. and GELMAN, A. (2019). Visualization in Bayesian workflow. *J. Roy. Statist. Soc. Ser. A* **182** 389–402. MR3902665 https://doi.org/10.1111/rssa.12378

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.

HAHSLER, M., BUCHTA, C., GRUEN, B. and HORNIK, K. (2015). arules: Mining association rules and frequent itemsets. R package version 1.3-1.

HAN, J., KAMBER, M. and PEI, J. (2012). *Data Mining Concepts and Techniques*, 1st ed. Elsevier.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HEBA, I., AMANY, A., AHMED, S. E. and AMR, S. (2014). Novel data-mining methodologies for detecting drug-drug interactions: A review of pharmacovigilance literature. In *Advances in Environmental Sciences*, *Development and Chemistry* (W. L. Staff, ed.) 301–314. Wseas LLC.

HENDERSON, N. C., LOUIS, T. A., ROSNER, G. L. and VARADHAN, R. (2020). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics* **21** 50–68. MR4043845 https://doi.org/10.1093/biostatistics/kxy028

HU, J., JOSHI, A. and JOHNSON, V. E. (2009). Log-linear models for gene association. *J. Amer. Statist. Assoc.* **104** 597–607. MR2751441 https://doi.org/10.1198/jasa.2009.0025

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158 https://doi.org/10.1214/009053604000001147

ISHWARAN, H., RAO, J. S. and KOGALUR, U. B. (2013). spikeslab: Prediction and variable selection using spike and slab regression. R package version 1.1.5.

KAPELNER, A. and BLEICH, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *J. Stat. Softw.* **70** 4.

KARBOWIAK, E. and BIECEK, P. (2019). EIX: Explain interactions in 'XGBoost'. R package version 1.0.

KAWEDIA, J. D., KASTE, S. C., PEI, D., PANETTA, J. C., CAI, X., CHENG, C., NEALE, G., HOWARD, S. C., EVANS, W. E. et al. (2011). Pharmacokinetic, pharmacodynamic, and pharmacogenetic determinants of osteonecrosis in children with acute lymphoblastic leukemia. *Blood* **117** 2340–2347.

KOTTAS, A., MÜLLER, P. and QUINTANA, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comput. Graph. Statist.* **14** 610–625. MR2170204 https://doi.org/10.1198/106186005X63185

LIM, M. and HASTIE, T. (2015). Learning interactions via hierarchical group-lasso regularization. *J. Comput. Graph. Statist.* **24** 627–654. MR3397226 https://doi.org/10.1080/10618600.2014.938812

LIM, M. and HASTIE, T. (2019). glinternet: Learning interactions via hierarchical group-lasso regularization. R package version 1.0.10.

LIU, J., SIVAGANESAN, S., LAUD, P. W. and MÜLLER, P. (2017). A Bayesian subgroup analysis using collections of ANOVA models. *Biom. J.* **59** 746–766. MR3672694 https://doi.org/10.1002/bimj.201600064

MACEACHERN, S. N. (2000). Dependent Dirichlet processes. Technical Report, Ohio State Univ., Dept. Statistics.

MITRA, R., MÜLLER, P. and JI, Y. (2017). Bayesian multiplicity control for multiple graphs. *Canad. J. Statist.* **45** 44–61. MR3623463 https://doi.org/10.1002/cjs.11305

MÜLLER, P., QUINTANA, F. and ROSNER, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20** 260–278. Supplementary material available online. MR2816548 https://doi.org/10.1198/jcgs.2011.09066

MÜLLER, P., QUINTANA, F. A., JARA, A. and HANSON, T. (2015). *Bayesian Nonparametric Data Analysis*. *Springer Series in Statistics*. Springer, Cham. MR3309338 https://doi.org/10.1007/978-3-319-18968-0

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

PAGE, G. L. and QUINTANA, F. A. (2018). Calibrating covariate informed product partition models. *Stat. Comput.* **28** 1009–1031. MR3835631 https://doi.org/10.1007/s11222-017-9777-z

PAGE, G. L., QUINTANA, F. A. and ROSNER, G. L. (2021a). Supplement to "Discovering interactions using covariate informed random partition models." https://doi.org/10.1214/20-AOAS1372SUPPA

PAGE, G. L., QUINTANA, F. A. and ROSNER, G. L. (2021b). Source code for "Discovering interactions using covariate informed random partition models." https://doi.org/10.1214/20-AOAS1372SUPPB

R CORE TEAM (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

REICH, B. J., KALENDRA, E., STORLIE, C. B., BONDELL, H. D. and FUENTES, M. (2012). Variable selection for high dimensional Bayesian density estimation: Application to human exposure simulation. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 47–66. MR2877584 https://doi.org/10.1111/j.1467-9876.2011.00772.x

SCHNELL, P. M., TANG, Q., OFFEN, W. W. and CARLIN, B. P. (2016). A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics* **72** 1026–1036. MR3591587 https://doi.org/10.1111/biom.12522

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

SHEN, W. and GHOSAL, S. (2016). Adaptive Bayesian density regression for high-dimensional data. *Bernoulli* **22** 396–420. MR3449788 https://doi.org/10.3150/14-BEJ663

SIMON, R. (2002). Bayesian subset analysis: Application to studyingtreatment-by-gender interactions. *Stat. Med.* **21** 2909–2916.

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

SU, X., PEÑA, A. T., LIU, L. and LEVINE, R. A. (2018). Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Stat. Med.* **37** 2547–2560. MR3824516 https://doi.org/10.1002/sim.7660

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TOKDAR, S. T., ZHU, Y. M. and GHOSH, J. K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.* **5** 319–344. MR2719655 https://doi.org/10.1214/10-BA605

VARADHAN, R. and WANG, S.-J. (2014). Standardization for subgroup analysis in randomized controlled trials. *J. Biopharm. Statist.* **24** 154–167. MR3196132 https://doi.org/10.1080/10543406.2013.856023

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, Fourth ed. Springer, New York. ISBN 0-387-95457-0. MR1337030 https://doi.org/10.1007/978-1-4899-2819-1

WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.* **13** 559–626. With discussion and a reply by the authors. MR3807860 https://doi.org/10.1214/17-BA1073