Bayesian Nonparametric Predictive Modeling for Personalized Treatment Selection

Matteo Pedone, Raffaele Argiento, and Francesco C. Stingo

Abstract We develop a Bayesian nonparametric predictive model to establish personalized therapeutic strategies for oncology patients. We leverage characteristics of both the patient and disease to support decision making in the selection of the optimal treatment. The core component of the model is a product partition model with covariates (PPMx) that induces clusters of observations that are more homogeneous with respect to predictive biomarkers. We conduct a simulation study to evaluate different modeling choices regarding PPMx in the framework of personalized treatment selection.

Keywords Product partition models · Nonparametric Bayes · Model-based clustering · Personalized medicine

1 Introduction

Our approach is motivated by an open problem in cancer genomics and personalized medicine. Personalized medicine's mission is to tailor treatment to individual patient characteristics leveraging various sources of heterogeneity. The distinctive mark of statistical inference under the personalized medicine paradigm is to disregard heterogeneity as nuisance to inference, but rather to take advantage of it to improve therapeutic strategies [2]. Cancer is a complex process and, to understand underlying biological phenomena, heterogeneity in both patients and disease must be accounted.

M. Pedone (✉) · F. C. Stingo
Universitá degli Studi di Firenze, viale Morgagni 65, 50134 Firenze, Italy
e-mail: matteo.pedone@unifi.it

F. C. Stingo
e-mail: francescoclaudio.stingo@unifi.it

R. Argiento
Universitá degli Studi di Bergamo, via Salvecchio 19, 24129 Bergamo, Italy
e-mail: raffaele.argiento@unibg.it

R. Argiento et al. (eds.), *New Frontiers in Bayesian Statistics*, Springer Proceedings
in Mathematics & Statistics 405, https://doi.org/10.1007/978-3-031-16427-9_10

We develop a method for personalized treatment selection that leverages prognostic and predictive biomarkers.

Prognostic biomarkers impact the likelihood of achieving a therapeutic response regardless of the selected treatment. By contrast, predictive biomarkers determine which patients are likely or unlikely to benefit from a particular class of treatment regimes. Since cancer is an inherently heterogeneous disease, each tumor is unique and hence, for predictive covariates, patients should not be regarded as exchangeable [3]. Given genomic signatures and a set of prognostic markers, building on [4] we leverage prognostic determinants to measure how likely a patient is to reach a given clinical response. Predictive biomarkers are exploited to drive patients clustering within each treatment. This is done to typify the extent of benefit offered by a specific therapeutic strategy on groups of patients characterized by close profiles in predictive determinants. We are assuming to know which biomarkers are prognostic and which are predictive. Although this assumption seems restrictive, it remains crucial. Biomarkers, in order to lead to optimal treatment selection, need to be validated on completely independent data set not used during development. That is, rather than develop prognostic/predictive biomarkers, our goal is personalized treatment selection employing validated biomarkers.

The Bayesian framework naturally handles model-based clustering assuming as random parameter of the model the partition of the sample subjects. In particular, we adopt the product partition model with covariates (PPMx) [5] to induce clusters of observations that are more homogeneous with respect to predictive covariates, building partitions that are only partially exchangeable. The class of PPMx models is a powerful Bayesian nonparametric tool to incorporate covariates' information into the prior for the random partition. Indeed, under this class of models, patients with similar covariates are a priori more likely to be clustered together. This feature enables us to quantify the effectiveness of each competing therapeutic strategy for patients with similar genetic profiles.

Finally, the posterior predictive distribution of this model arises as a natural way to assess the extent to which a new untreated patient is likely to attain a level of clinical response for competing treatments. We elicit response utility weights and evaluate utility expectation for each therapy [3]. The treatment with the largest mean predictive utility is considered the optimal treatment.

The goal of this paper is to provide guidance regarding the specification of the prior distribution for the random partition in the framework of optimal treatment selection. In fact, as the number of predictive biomarkers grows, the influence of PPMx models on clustering tends to overwhelm the information from the response, negatively affecting inference and out-of-sample prediction. In order to calibrate the influence that covariates have on partition probabilities we follow [8]'s strategy to temper covariate impact on clustering. The evaluation of different calibrations is empirically done through simulations based on gene expression data from a leukemia study [1].

The remainder of the article is organized as follows. In Sect. 2 we state the proposed model, focusing on the aspects addressed in the simulation study. We also give some details on the computational strategy. In Sect. 3 we describe the predictive utility approach adopted for treatment selection. We report and discuss the results of the simulation study in Sect. 4 and Sect. 5 concludes the paper.

2 The Model

Let $a = 1, \dots, T$ index candidate therapies to whom $n = \sum_{a=1}^{T} n^a$ patients are assigned to, where n^a denotes the number of patients treated with therapy a. A common choice to characterize varying levels of treatment response is to evaluate it in terms of the extent of residual disease after a given clinically relevant post-therapy follow-up duration. Let y_i^a be the random variable of the $i-$th patient's response to treatment a among K possible levels of increasing treatment benefit, where $y_i^a = k$ for $i = 1, \dots, n^a$ and $k = 1, \dots, K$. In addition, let $\boldsymbol{\pi}_i^a = (\pi_{i1}^a, \dots, \pi_{iK}^a)$ denote the vector such that π_{ik}^a is the probability of observing outcome k for the $i-$th patient under treatment a. The treatment response is an ordinal-valued random variable and y_i^a follows a multinomial distribution $y_i^a \mid \boldsymbol{\pi}_i^a \overset{\text{ind}}{\sim} \text{Multinomial}(1, \boldsymbol{\pi}_i^a)$. For each treatment, we consider a training dataset of n^a patients, $(y_i^a, \boldsymbol{z}_i^a, \boldsymbol{x}_i^a)$ where $i = 1, \dots, n^a$ and $\boldsymbol{z}_i^a$ and $\boldsymbol{x}_i^a$ are a $P-$dimensional and $Q-$dimensional vector of prognostic and predictive features, respectively.

As mentioned in Sect. 1, to relax exchangeability among observations, we adopt a model for random partition depending on predictive markers. We denote with $\rho^a = \{S_1^a, \dots, S_{C^a}^a\}$ the treatment-specific partition of the indices $\{1, \dots, n^a\}$, where C^a is the number of clusters among patients treated with therapy a and $n_j^a = \mid S_j^a \mid$ is the cardinality of cluster j, for $j = 1, \dots, C^a$. Finally, cluster-specific quantities are denoted with the super script "$\star$". For example, when considering the $j-$th cluster for treatment a, the response vector is $\boldsymbol{y}_j^{a\star} = \{y_i^a : i \in S_j^a\}$ while $\boldsymbol{x}_j^{a\star} = \{\boldsymbol{x}_i^a : i \in S_j^a\}$ is the partitioned covariate matrix. Using a conjugate prior for $\boldsymbol{\pi}_i^a$, we assume the following hierarchical model for $a = 1, \dots, T$:

$$y_i^a | \boldsymbol{\pi}_i^a \overset{\text{ind}}{\sim} \text{Multinomial}(1, \boldsymbol{\pi}_i^a)$$

$$\boldsymbol{\pi}_1^a, \dots, \boldsymbol{\pi}_{n_a}^a \mid \boldsymbol{\eta}_1^{a\star}, \dots, \boldsymbol{\eta}_{C_a}^{a\star}, \rho^a, \boldsymbol{\beta} \sim \prod_{j=1}^{C^a} \prod_{i \in S^a} \text{Dirichlet}(\boldsymbol{\pi}_i^a; \boldsymbol{\gamma}_i^a(\boldsymbol{\eta}_j^{a\star}, \boldsymbol{\beta}, \boldsymbol{z}_i^a)),$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$ is a $P \times K$ matrix of regression parameter shared across levels of response and individuals. The K-dimensional vectors $\boldsymbol{\eta}_1^{a\star}, \dots, \boldsymbol{\eta}_{C_a}^{a\star}$ are cluster-specific parameters, that is, $\boldsymbol{\eta}_j^{a\star}$ is a parameter shared by all the individual in cluster S_j^a. Finally, $\boldsymbol{\gamma}_i^a(\boldsymbol{\eta}_j^{a\star}, \boldsymbol{\beta}, \boldsymbol{z}_i^a) = (\gamma_{i1}^a(\eta_{j1}^{a\star}, \boldsymbol{\beta}_1, \boldsymbol{z}_i^a), \dots, \gamma_{iK}^a(\eta_{jK}^{a\star}, \boldsymbol{\beta}_K, \boldsymbol{z}_i^a))$, is a vector of log-linear functions on the prognostic marker and cluster-specific parameters defined as follows:

$$\log(\gamma_{ik}^a(\eta_{jk}^{a\star}, \boldsymbol{\beta}_k, \boldsymbol{z}_i^a)) = \eta_{jk}^{a\star} + \beta_{1k} z_{i1}^a + \dots + \beta_{Pk} z_{iP}^a.$$

2.1 Priors

The choice of a covariate-dependent prior on the random partition enables predictive biomarkers to drive the clustering. Priors for $\{\rho^a\}$ and $\{\boldsymbol{\eta}_j^{a\star}\}$, are defined independent across treatments. In fact, we want to allow the response probabilities to change from treatment to treatment even for subject with similar genetic profile. This independence assumption prevents the model from inducing a partition that implies the same response probability for genetically similar subjects that have received different treatments. The joint law of $(\rho^a, \boldsymbol{\eta}_j^{a\star})$ is assigned hierarchically as:

$$P(\rho^a = \{S_1^a, \ldots, S_{C^a}^a\} \mid \boldsymbol{x}^a) \propto \prod_{j=1}^{C^a} c(|S_j^a|)g(\boldsymbol{x}_j^{a\star}), \tag{1}$$

$$\boldsymbol{\eta}_1^{a\star}, \ldots, \boldsymbol{\eta}_{C^a}^{a\star} \mid C^a \overset{\text{iid}}{\sim} p_0.$$

In Equation (1) the prior on the random partition is given via *cohesion* function c and *similarity* function g.

The cohesion function acts on clusters, depending only on the cluster size. Following [5], we choose a commonly adopted cohesion function, that is $c(S_j^a) = \alpha\Gamma(|S_j^a|)$, $\alpha > 0$, corresponding to the marginal partition model available from a Dirichlet process.

The *similarity* g is a non-negative function that measures how homogeneous patients in the same cluster are, with respect to predictive markers. It plays a crucial role since it increases the probability that patients with close genetic profiles are co-clustered. In Sect. 2.2 we list and describe two *similarity* function g along with strategies designed to temper the covariates' influence on clustering.

Following [10], for p_0 we adopted a conjugate Normal-Inverse Wishart. The posterior distribution for $\boldsymbol{\eta}^{a\star}$ results in C^a independent multivariate normal densities.

The priors for the parameters $\{\boldsymbol{\beta}_k\}$ are assumed to be independent and, to enhance predictive performance, we specified horseshoe priors: $\beta_{pk} \sim N(0, \sigma_{pk}^2)$, for $p = 1, \ldots, P$, where $\sigma_{pk}^2 = \lambda_{pk}^2 \cdot \tau_k^2$, with $\lambda_{pk}, \tau_k \sim \text{HalfCauchy}(0, 1)$.

2.2 Similarity Function

Predictive biomarkers drive the clustering process trough the *similarity* function, that measure the homogeneity of the $x_i \in \boldsymbol{x}_j^\star$. In theory any non-negative function that produces larger values for more close covariates is suitable. In order to evaluate the influence of this choice on the response to treatment prediction we present the two *similarity functions* that are compared in the simulation study. As mentioned before, in order to counteract the strong effect that a large number of covariates may have on partition probabilities, we adopt a strategy to temper their effects. In particular, we briefly discuss the coarsening of the *similarity* function.

The original *similarity* function proposed by [5] is to choose $g(\boldsymbol{x}_j^{a\star})$ as the marginal probability of an auxiliary probability model. It takes the form

$$g(\boldsymbol{x}_j^{a\star}) = \int \prod_{i \in S_j^a} q(x_i^a \mid \boldsymbol{\xi}_j^{a\star}) q(\boldsymbol{\xi}_j^{a\star}) \mathrm{d}\boldsymbol{\xi}_j^{a\star}. \tag{2}$$

Note that $\{x_i^a\}$ are not considered random: this structure is convenient because the correlation induced by the cluster-specific parameters $\{\boldsymbol{\xi}_j^{a\star}\}$ leads to large values of $g(\boldsymbol{x}_j^{a\star})$ for close $\{x_i^a\}$.

For continuous covariates [5] suggests as default choice for $g(\boldsymbol{x}_j^{a\star})$ the marginal distribution of $x_j^{a\star}$ under a normal sampling model. A conjugate pair for $q(\cdot \mid \boldsymbol{\xi}^{a\star})$ and $q(\boldsymbol{\xi}^{a\star})$ greatly facilitates the evaluation of $g(\boldsymbol{x}_j^{a\star})$: $q(\cdot \mid \boldsymbol{\xi}_j^{a\star}) = N(\cdot | m_j^{a\star}, \upsilon_j^{a\star})$ and $q(\boldsymbol{\xi}_j^{a\star}) = q(m_j^{a\star}, \upsilon_j^{a\star}) = NIG(m_j^{a\star}, \upsilon_j^{a\star} | m_0, k_0, \upsilon_0, n_0)$, that are the Normal and Normal-Inverse-Gamma density functions, respectively. A simplified version of this conjugate model forces covariate clusters to have the same variance: $\upsilon_j^{a\star} = \upsilon^{a\star}$ and results in $q(\boldsymbol{\xi}_j^{a\star}) = N(m_j^{a\star} \mid m_0, s_0^2)$. We will refer to this latter formulation as the “Auxiliary NN” and the first one as the “Auxiliary NNIG”. Note that we focus here on continuous covariates. A major advantage offered by similarities of the form of (2) is that they easily account also for categorical, ordinal and count covariates [6].

[9] propose a variation of (2), defining $g(\boldsymbol{x}_j^{\star})$ as the posterior predictive distribution of $\boldsymbol{x}_j^{\star}$ in cluster S_j:

$$g(\boldsymbol{x}_j^{a\star}) = \int \prod_{i \in S_j^a} q(x_i^a \mid \boldsymbol{\xi}_j^{a\star}) q(\boldsymbol{\xi}_j^{a\star} | \boldsymbol{x}_j^{a\star}) \mathrm{d}\boldsymbol{\xi}_j^{a\star}, \tag{3}$$

with $q(\boldsymbol{\xi}_j^{\star} | \boldsymbol{x}_j^{\star}) \propto \prod_{i \in S_j} q(x_i | \boldsymbol{\xi}_j^{\star}) q(\boldsymbol{\xi}_j^{\star})$. Since the covariates are used twice, this function is called “Double-dipper". The rationale for this formulation, that has the same form as (2), is to give more weight to the local covariate structure. This is pursued by weighting $\boldsymbol{x}_j^{\star}$s “likelihood” with the “posterior distribution” of $\boldsymbol{\xi}_j^{\star}$ instead of its “prior”.

As for the *auxiliary similarity*, when x is continuous we can have the “Double-dipper NN” or “Double-dipper NNIG”. Finally note that, for multivariate $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iQ})$, as in our case of study, we use $g(\boldsymbol{x}_j^{\star}) = \prod_q^Q g(\boldsymbol{x}_{jq}^{a\star})$.

As an alternative to variable selection or to reducing the dimensionality of the covariate space through the use of sufficient statistics, [8] proposes to calibrate the influence of covariates on clustering. In particular we consider the *coarsened similarity function*:

$$\tilde{g}(\boldsymbol{x}_j^{a\star}) = g(\boldsymbol{x}_j^{a\star})^{1/Q}. \tag{4}$$

In order to shrink the degree of *coarsening* we want to induce on the partition probabilities, we also consider a small variation of (4) which will be referred to as *shrunk coarsened similarity*: $\tilde{g}(\boldsymbol{x}_j^{a\star}) = g(\boldsymbol{x}_j^{a\star})^{1/\sqrt{Q}}$.

2.3 Posterior Computation

A MCMC procedure is used to fit the PPMx model. The core part of the algorithm is the updating of the cluster labels. The computation associated with fitting Equation (1) is based on [7]'s Algorithm 8, where applying a Gibbs sampling to a state augmented by the addition of auxiliary parameters greatly facilitates the update of the partition. Conditional on the updated cluster labels, all the remaining parameters are easily updated with Gibbs sampler or Metropolis-Hastings steps.

3 Treatment Selection

In order to select the optimal treatment for a new, untreated patient $\tilde{i}$, we are interested in the predictive probability of $y_{\tilde{i}}$. Given the observed responses for the n^a patients previously treated with therapy a, that is $\boldsymbol{y}^a$, the predictive probability of response level k under treatment a is

$$p(y_{\tilde{i}} = k \mid \boldsymbol{y}^a, \boldsymbol{z}^a, \boldsymbol{x}^a, \boldsymbol{z}_{\tilde{i}}, \boldsymbol{x}_{\tilde{i}}),$$

where $\boldsymbol{z}_{\tilde{i}}$ and $\boldsymbol{x}_{\tilde{i}}$ denote the P and Q dimensional vectors containing prognostic and predictive markers for the new patient. To facilitate treatment selection for multinomial ordinal outcomes, we adopt utility weights. In clinical oncology response categories are ordinal and consider changes in tumor size and/or distant migration after the treatment. We establish utility weights that turn a multinomial setting into a one-dimensional selection criterion considering the relative importance of each level of the ordinal response. Let $\boldsymbol{\omega}$ be a $K-$dimensional vector denoting the utility assigned to tumor response levels. To make $\boldsymbol{\omega}$ reflect clinical importance of each level (non respondent, partially respondent and respondent), we set $\boldsymbol{\omega} = (0, 40, 100)^\top$, following [4]. We can then compute the mean predictive utility for patient $\tilde{i}$ as:

$$\varphi^a(\tilde{i}) = \sum_{k=1}^{K} \omega_k p(y_{\tilde{i}} = k \mid \boldsymbol{y}^a, \boldsymbol{z}^a, \boldsymbol{x}^a, \boldsymbol{z}_{\tilde{i}}, \boldsymbol{x}_{\tilde{i}}).$$

The $\tilde{i}-$th patient will be assigned to the therapy ensuring the largest predictive utility, that can be considered to be optimal among the competing treatments.

4 Illustrative Example

To empirically assess the performance of the coarsened similarity function presented in Sect. 2.2, we conduct a simulation study. To compare model fit and treatment selection we generate synthetic data adopting the processes designed by [4] (see Scenario 2), with the only difference that we use 10 predictive markers (instead of 90), while we consider the same two prognostic covariates.

This procedure yields $n = 152$ patients that are assigned to $T = 2$ competing treatment. We consider 3 levels for the ordinal-valued response variable. We standardize all predictive biomarkers.

The hyperparameters for Auxiliary NN and Double-dipper NN similarities are $(m_0 = 0, s_0^2 = 1)$. For hyperparameters needed when the NNIG model is employed in the similarities, on the ground of the results obtained by [8] in their extensive simulation study and sensitivity analysis, we set $(m_0 = 0, k_0 = 1, v_0 = 10, n_0 = 2)$.

For each similarity function we run the PPMx for 150, 000 iterations, descarding the first 50, 000 due to burn-in and keeping each 10−th draw from the posterior distribution. To compare the goodness-of-fit we report the log pseudomarginal likelihood (LPLM). To evaluate the predictive performances we adopt the same metrics as in [4]:

(i) MOT, that is the number of misassigned patients;
(ii) $\%\Delta$MTU, it measures the relative gain in treatment utility with respect to the other treatment; note that it is defined only for the case of two alternative treatments. It ranges from -1 to 1 ($\%\Delta\text{MTU} = 1$ only in the case of optimal treatment assignment rule);
(iii) NPC that is the number of correctly predicted outcomes.

Prediction is based on a leave-one-out cross-validated strategy. The numerical results reported in Table 1 are averaged over 100 data sets generated for each case. Standard deviations are given in brackets. The best performance for each metric is reported in bold.

The Double-dipper similarity outperforms the Auxiliary similarity function. Double-dipper best performances are probably due to the larger weight given to the covariates in the model-based clustering process.

Focusing on the lower pane of Table 1, we notice that the Double-dipper function delivers better results when the NNIG model is assumed. In fact, NNIG offers a greater flexibility than NN, as it does not force clusters to share the same variance.

Restricting our focus to the Double-dipper NNIG similarity function, Table 1 offers a last comparison between Coarsening and Shrunk Coarsening. The former achieves better performances in terms of goodness-of-fit, while the latter is to be preferred according to those metrics evaluating prediction. Shrunk Coarsened similarity outperforms Coarsened similarity assigning fewer patient to the non optimal treatment (15.18 vs 24.13) and reaching a larger relative gain in treatment utility (82% versus 64%). Coarsening, on the other hand, yields slightly better performances in terms of number of correctly predicted outcome and LPML.

Given the focus on treatment selection rather than inference on model parameters, Shrunk Coarsened Double-dipper NNIG is the similarity function best suited for our model.

Table 1 Simulation study on *similarity functions*

Similarity	MOT	%ΔMTU	NPC	LPML
Coarsened Auxiliary NN	34.33	0.47	80.63	−129.98
	(4.71)	(0.05)	(6.06)	(4.33)
Coarsened Auxiliary NNIG	28.50	0.58	80.41	−129.18
	(5.79)	(0.08)	(5.92)	(4.47)
Shrunk Coarsened Auxiliary NN	55.70	0.30	74.28	−155.51
	(33.72)	(0.41)	(7.00)	(3.81)
Shrunk Coarsened Auxiliary NNIG	70.82	0.10	67.00	−156.75
	(7.64)	(0.10)	(6.75)	(3.95)
Coarsened Double-dipper NN	31.93	0.50	79.91	−124.08
	(4.71)	(0.05)	(5.95)	(4.17)
Coarsened Double-dipper NNIG	24.13	0.64	**81.70**	**−121.26**
	(6.66)	(0.09)	(5.87)	(3.65)
Shrunk Coarsened Double-dipper NN	19.83	0.73	77.40	−141.58
	(9.03)	(0.10)	(6.50)	(4.42)
Shrunk Coarsened Double-dipper NNIG	**15.98**	**0.82**	77.08	−146.17
	(8.06)	(0.09)	(6.05)	(4.44)

5 Conclusion

Employing PPMx to cluster together patients with close genetic profiles and then evaluate the effectiveness of competing treatments on groups of similar patients shows promise. In this paper we focus on the choice of the similarity function, that is pivotal in PPMx models, in the framework of optimal treatment selection. We find the Double-dipper similarity to perform particularly well when a shrunk coarsening is employed and the NNIG model is adopted.

Several extension are currently under investigation, with a sharp focus on similarity functions that could enable us to include a larger number of predictive markers.

References

1. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**, 531–537 (1999)
2. Kosorok, M.R., Laber, E.B.: Precision medicine. Annu. Rev. Stat. Appl. **6**, 263–286 (2019)
3. Ma, J., Stingo, F.C., Hobbs, B.P.: Bayesian predictive modeling for genomic based personalized treatment selection. Biometrics **72**, 575–583 (2016)
4. Ma, J., Stingo, F.C., Hobbs, B.P.: Bayesian personalized treatment selection strategies that integrate predictive with prognostic determinants. Biometrical J. **61**, 902–917 (2019)
5. Müller, P., Quintana, F., Rosner, G.L.: A product partition model with regression on covariates. J. Comput. Graph. Stat. **20**, 260–278 (2011)

6. Müller, P., Quintana, F., A., Jara, A., Hanson, T.: Bayesian Nonparametric Data Analysis. Springer, Heidelberg (2015)
7. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**, 249–265 (2000)
8. Page, G.L., Quintana, F.: Calibrating covariate informed product partition models. Stat. Comput. **28**, 1009–1031 (2018)
9. Quintana, F., Müller, P., Papoila, A.L.: Cluster-specific variable selection for product partition models. Scand. J. Stat. **42**, 1065–1077 (2015)
10. West, M., Müller, P., Escobar, M.D.: Hierarchical priors and mixture models, with applications in regression and density estimation. In: Aspects of Uncertainty: A Tribute to D.V. Lindley, pp. 363–386 (1994)