Supplementary material to Bayesian personalized treatment selection strategies that integrate predictive with prognostic determinants

Junsheng Ma [1], Francesco C. Stingo,[2,*], Brian P. Hobbs[1]

[1] Department of Biostatistics, The University of Texas MD Anderson Cancer Center

Unit 1411 1400 Pressler Street, Houston, TX 77030, U.S.A.

[2] Department of Statistica, Informatica, Applicazioni "G.Parenti", University of Florence,

Florence, 50134, Italy

* Corresponding Author, email: fra.stingo@gmail.com

S.1 Theoretical considerations

This section extends section 2.3 to further explore theoretical aspects of the proposed Bayesian predictive selection rules to elucidate the manner in which treatment selection is impacted by prognostic features. Following the same notation of the main manuscript, ϕ_j refers to the mean utility for subject i when only predictive features are considered, whereas ϕ'_j refers to the mean utility for subject i when both predictive and prognostic features are considered. Considering the case wherein clinical response can be adequately characterized by a binary variable, we have Lemma 1.

Lemma S.1.1. *For binary outcomes incorporating prognostic features fails to alter the resultant decision for treatment selection.*

Proof. We can show that $\phi_1 - \phi_2 = w_1(p_{11} - p_{21})$ and

$$
\begin{aligned}
\phi_1' - \phi_2' &= w_1 d_1 \left(\frac{p_{11}}{C_1} - \frac{p_{21}}{C2}\right) = w_1 d_1 \left(\frac{p_{11}}{d_0 p_{10} + d_1 p_{11}} - \frac{p_{21}}{d_0 p_{20} + d_1 p_{21}}\right) \\
&= \frac{w_1 d_1}{C_1 C_2}\{p_{11}(d_0 - d_0 p_{21} + d_1 p_{21}) - p_{21}(d_0 - d_0 p_{11} + d_1 p_{11})\} = \frac{w_1 d_1 d_0}{C_1 C_2}(p_{11} - p_{21}).
\end{aligned}
$$

The sign of $\phi_1' - \phi_2' = \frac{w_1 d_1 d_0}{C_1 C_2}(p_{11} - p_{21})$ depends on $(p_{11} - p_{21})$, and thus consistent with the sign of $\phi_1 - \phi_2 = w_1(p_{11} - p_{21})$. Thus, an optimal treatment resulting from $\phi_1 - \phi_2$ in the absence of prognostic covariates would be consistent with the optimal treatment incorporating prognostic covariates $\phi_1' - \phi_2'$. □

By way of contrast, for ordinal outcomes (i.e. $K > 2$) integrating prognostic features with the Bayesian predictive approach may alter the sign of $\phi_1' - \phi_2'$, and thereby may yield differential optimal treatments when compared to models that utilize predictive features alone. To elucidate, we consider responses comprised of trinary ordinal-valued outcomes. Based on predictive features only, the difference in mean treatment utility is $\phi_1 - \phi_2 = w_1(p_{11} - p_{21}) + w_2(p_{12} - p_{22})$. Integrating both prognostic and predicative features yields the following difference in expected mean utility

$$
\begin{aligned}
\phi_1' - \phi_2' &= w_1 d_1 \left(\frac{p_{11}}{d_0 p_{10} + d_1 p_{11} + d_2 p_{12}} - \frac{p_{21}}{d_0 p_{20} + d_1 p_{21} + d_2 p_{22}}\right) + \\
&\quad w_2 d_2 \left(\frac{p_{12}}{d_0 p_{10} + d_1 p_{11} + d_2 p_{12}} - \frac{p_{22}}{d_0 p_{20} + d_1 p_{21} + d_2 p_{22}}\right) \\
&= \frac{w_1 d_1}{C_1 C_2}\{d_0(p_{11} - p_{21}) + (d_2 - d_0)(p_{11} p_{22} - p_{21} p_{12})\} + \\
&\quad \frac{w_2 d_2}{C_1 C_2}\{d_0(p_{12} - p_{22}) + (d_1 - d_0)(p_{12} p_{21} - p_{11} p_{22})\}
\end{aligned}
$$

With some algebra, we can show that

$$
\phi_1' - \phi_2' = \frac{d_0}{C_1 C_2}\{w_1 d_1(p_{11} - p_{21}) + w_2 d_2(p_{12} - p_{22})\} + \frac{p_{11} p_{22} - p_{21} p_{12}}{C_1 C_2}\{d_1 d_2(w_1 - w_2) + d_0(w_2 d_2 - w_1 d_1)\},
$$

where $C_1 = d_0 p_{10} + d_1 p_{11} + d_2 p_{12}$ and $C_2 = d_0 p_{20} + d_1 p_{21} + d_2 p_{22}$. Thereby sign of quantity $\phi_1' - \phi_2'$ depends on d_k, w_k, p_{jk} for $k = 0, 1, 2$ and $j = 1, 2$. There exist a few special cases where the treatment selection decisions are invariant to integration of the prognostic features. These are

described in Lemmas 2 and 3.

Lemma S.1.2. *(i) Assume that $p_{11} \geq p_{21}$ and $p_{12} \geq p_{22}$, then $\phi_1 - \phi_2 \geq 0$ and $\phi_1' - \phi_2' \geq 0$ if $\frac{d_1}{d_0} \to 0$ and $\frac{d_1}{d_2} \to 0$.*

(ii) Assume that $p_{11} \leq p_{21}$ and $p_{12} \leq p_{22}$, then $\phi_1 - \phi_2 \leq 0$ and $\phi_1' - \phi_2' \leq 0$ if $\frac{d_1}{d_0} \to 0$ and $\frac{d_1}{d_2} \to 0$.

Proof. For (i) it is clear that $\phi_1 - \phi_2 \geq 0$. We can show that

$$
\lim_{\frac{d_1}{d_0} \to 0, \frac{d_1}{d_2} \to 0} \frac{\phi_1' - \phi_2'}{d_0 d_2} = \frac{w_2}{C_1 C_2}(p_{12} - p_{22} + p_{11}p_{22} - p_{21}p_{12})
$$

$$
\geq \frac{w_2}{C_1 C_2}(p_{12} - p_{22} + p_{21}p_{22} - p_{21}p_{12}) = \frac{w_2}{C_1 C_2}(p_{12} - p_{22})(1 - p_{21}) \geq 0
$$

that obviously implies $\phi_1' - \phi_2' > 0$. Point (ii) can be analogously proven. Note that $\frac{d_1}{d_0} \to 0$ and $\frac{d_1}{d_2} \to 0$ imply that the probability of observing response level $y_{i^\star} = 1$ tends to 0. $\square$

Lemma S.1.3. *(i) Assume that $p_{11} \geq p_{21}$ and $p_{12} \geq p_{22}$, then $\phi_1 - \phi_2 \geq 0$ and $\phi_1' - \phi_2' \geq 0$ if $\frac{d_2}{d_0} \to 0$ and $\frac{d_2}{d_1} \to 0$.*

(ii) Assume that $p_{11} \leq p_{21}$ and $p_{12} \leq p_{22}$, then $\phi_1 - \phi_2 \leq 0$ and $\phi_1' - \phi_2' \leq 0$ if $\frac{d_2}{d_0} \to 0$ and $\frac{d_2}{d_1} \to 0$.

Proof. For (i) it is clear that $\phi_1 - \phi_2 \geq 0$. We can show that

$$
\lim_{\frac{d_2}{d_0} \to 0, \frac{d_2}{d_1} \to 0} \frac{\phi_1' - \phi_2'}{d_0 d_1} = \frac{w_1}{C_1 C_2}(p_{11} - p_{21} - p_{11}p_{22} + p_{21}p_{12})
$$

$$
\leq \frac{w_1}{C_1 C_2}(p_{11} - p_{21} - p_{11}p_{12} + p_{21}p_{12}) = \frac{w_1}{C_1 C_2}(p_{11} - p_{21})(1 - p_{12}) \geq 0
$$

that obviously implies $\phi_1' - \phi_2' > 0$. The proof for (ii) follows the same approach. Again, $\frac{d_2}{d_0} \to 0$ and $\frac{d_2}{d_1} \to 0$ indicate that the probability of observing the outcome of $y_{i^\star} = 2$ tends to 0. $\square$

When the probability of observing response level $y_{i^\star} = 0$ tends to 0, however, we fail to obtain similar results. Specifically, we can show that $\lim_{\frac{d_0}{d_1} \to 0, \frac{d_0}{d_2} \to 0} \frac{\phi_1' - \phi_2'}{d_1 d_2} = \frac{p_{11}p_{22} - p_{21}p_{12}}{C_1 C_2}\{(w_1 - w_2)\}$. However when $p_{11} \geq p_{21}$ and $p_{12} \geq p_{22}$, this quantity (and consequently $\phi_1' - \phi_2'$) can be either positive or negative; on the contrary, $\phi_1 - \phi_2 \geq 0$, potentially leading to differences in the selection of the treatment.

3

S.2   Simulating 152 patients and generating treatment responses

Genomic features, such as expression levels of molecular quantities arising from sequencing data, tend to exhibit correlation structures that are poorly characterized by smooth parametric models. Our simulation study was devised to emulate the dependence structure observed in a well-known dataset of leukemia (http://www.pnas.org/content/101/12/4164.full?tab=ds) containing gene expression levels for a total of 5,000 genes across 38 patients, 11 diagnosed with acute myelogenous leukemia (AML) and 27 were diagnosed with acute lymphoblastic leukemia (Gentleman *et al.*, 1999). To obtain a comparable sample size to that used in our case study presented in Section 4 (where $n = 158$) we expanded the dataset to yield a total of 152 simulated patients, $38 \times 4 = 152$, each with 92 features. In addition, to reflect the current clinical cancer context, wherein a potentially large-set of molecular features are considered as actionable therapeutic targets, while a small number of established clinically-derived factors are utilized for prognostication, we selected the first 90 features as predictive, and reserved the remaining 2 as prognostic.

To simulate 152 patients, we clustered the top 1,000 varied genes (the maximum minus the minimum level of observed gene expression), and selected 92 clusters containing at least four genes. Assuming the first 4 highly correlated genes were exchangeable, we stacked them as one feature and hence converted each patient into four. This process closely follows the schema proposed by (Ma *et al.*, 2016); more details can be found in the following paragraphs as well as in the supplementary materials of (Ma *et al.*, 2016).

Generating treatment responses.   We simulated the ordinal outcome variables using two separate continuation-ratio logistic functions. The first,

$$r_{kj}(\boldsymbol{x}_i, A_i = j) = ln\{\frac{P(y_i = k|\boldsymbol{x}_i, A_i = j)}{p(y_i < k|\boldsymbol{x}_i, A_i = j)}\} = \eta_{jk} + \boldsymbol{\beta}_{1jk}\psi(d_i) \tag{1}$$

to characterize the effects of the predictive features. The second,

$$r_{kj}(\boldsymbol{z}_i, A_i = j) = ln\{\frac{P(y_i = k|\boldsymbol{z}_i, A_i = j)}{p(y_i < k|\boldsymbol{z}_i, A_i = j)}\} = \delta_{jk} + \boldsymbol{\beta}_{2k}\boldsymbol{z}_i \tag{2}$$

to define response-level probabilities for the prognostic features, where x_i and z_i represent predictive and prognostic features, respectively. The information carried by the many predictive features x_i can be summarized via dimension reduction techniques, such as principal component analysis (PCA) in our implementation, and stored in a new variables d_i. With $\psi(d_i)$ we denote a one-dimensional function of a summary of the predictive covariates, such as the first two principal components in our implementation. By definition, values of the coefficients for the prognostic features, β_{2k}, were identical for all treatments.

The regression coefficients β_{2k} and β_{1jk} were set to values that could produce realistic response rates. Probabilities for each level of ordinal response variable, referred to hereafter as ordinal response probabilities or ORP, were generated for each patient in proportion to the pointwise product of (1) and (2). For example, given probabilities (0.19,0.59,0.32) and (0.5,0.3,0.2), respectively, for a particular patient, the true ORP for generating the ordinal-valued response variable would be obtained as $\frac{0.19 \times 0.5}{0.19 \times 0.5 + 0.59 \times 0.3 + 0.32 \times 0.2}$.

Figure S.1 depicts differences in the underlying true mean treatment utilities for the 152 cases used in our simulation study, each determined by 90 predictive and 2 prognostic features under two general scenarios The simulated cases are intentionally heterogenous such that the extent of gain obtained from the optimal treatment (in relation to the non-optimal treatment) varies by patient. For example, as characterized by DMTU, gains obtained from correctly assigning the optimal treatments for patients 113 and 117 were 54 and 3, respectively.

In all simulation studies we did not generate the simulated data from the proposed model. Within our model framework, we can not simulate prognostic features from mixed normal distributions since the conditional probability of belonging to a given response group depends on both prognostic and predictive features. For simulation purposes, the strategy of generating prognostic features with fixed response probabilities fails when predictive features are incorporated (Franzén, 2008). We took an alternative strategy and used the observed prognostic features from one of our case studies to generate the response variables. We power transformed these features, selecting the coefficients in (2) such that the resulting generated data have a clustering pattern that resembles the pattern of data generated from normal-mixture models (Franzén, 2008). Figure S.2 presents

the scatter and conditional density plots for the transformed features based on one simulated data set under scenario 2. We can see that the generated data are well clustered in terms of responses although the power transformed features are not normally distributed. Note that while we generated only two scenarios, in Section 3.1 of the manuscript we repeated our analysis using 1) the prognostic covariates with power transformations 2) the prognostic covariates in the original scale 3) random sampling from independent normal distributions with mean and variance estimated from the original data. The scatter and density plots for the original and random features under scenario 2 are presented in Figures S.3 and S.4 respectively. We can see that the conditional distribution of the original features are clearly apart from the normality assumption. This simulation strategy worked very well for our simulation studies, although the features were not generated from mixture normal distributions. As described in the main manuscript, the proposed approach worked reasonably well across all scenarios.

S.3 Additional simulation results

Our modeling framework can simultaneously integrate high-dimensional omics-type predictive and a reasonable large number of prognostic variables. Our approach performed well with 90 predictive features in this study, as well as in our previous study with 200 predictive features in both simulation and real data analysis, (Ma *et al.*, 2016). On the other hand, to investigate the scale of our approach in handling prognostic features, we conducted additional simulation study in scenarios with relatively large number of prognostic features. Specifically, we extended simulation scenario 2 to scenario 2.1 with the 2 features and 8 noise random variables, and scenario 2.2 with the 2 features and 6 random noise variables. These noise variables were generated from independent normal distributions and were not used to generate the outcome variables. Results are presented in the following, Table S1.

In general, results from scenario 2.1 and 2.2 are comparable to those from scenario 2, except that the numbers of patients for which the model correctly predicted their simulated outcomes are relatively small. To our experience, this is common for data analysis using high dimensional data with some noise variables. To improve model accuracy in predicting the outcome, data are

often pre-processed for analysis. In our data analysis, we described a straightforward data pre-processing method to select predictive and prognostic variables, while more advanced approaches can be found in (Yu *et al.*, 2003) and its references.
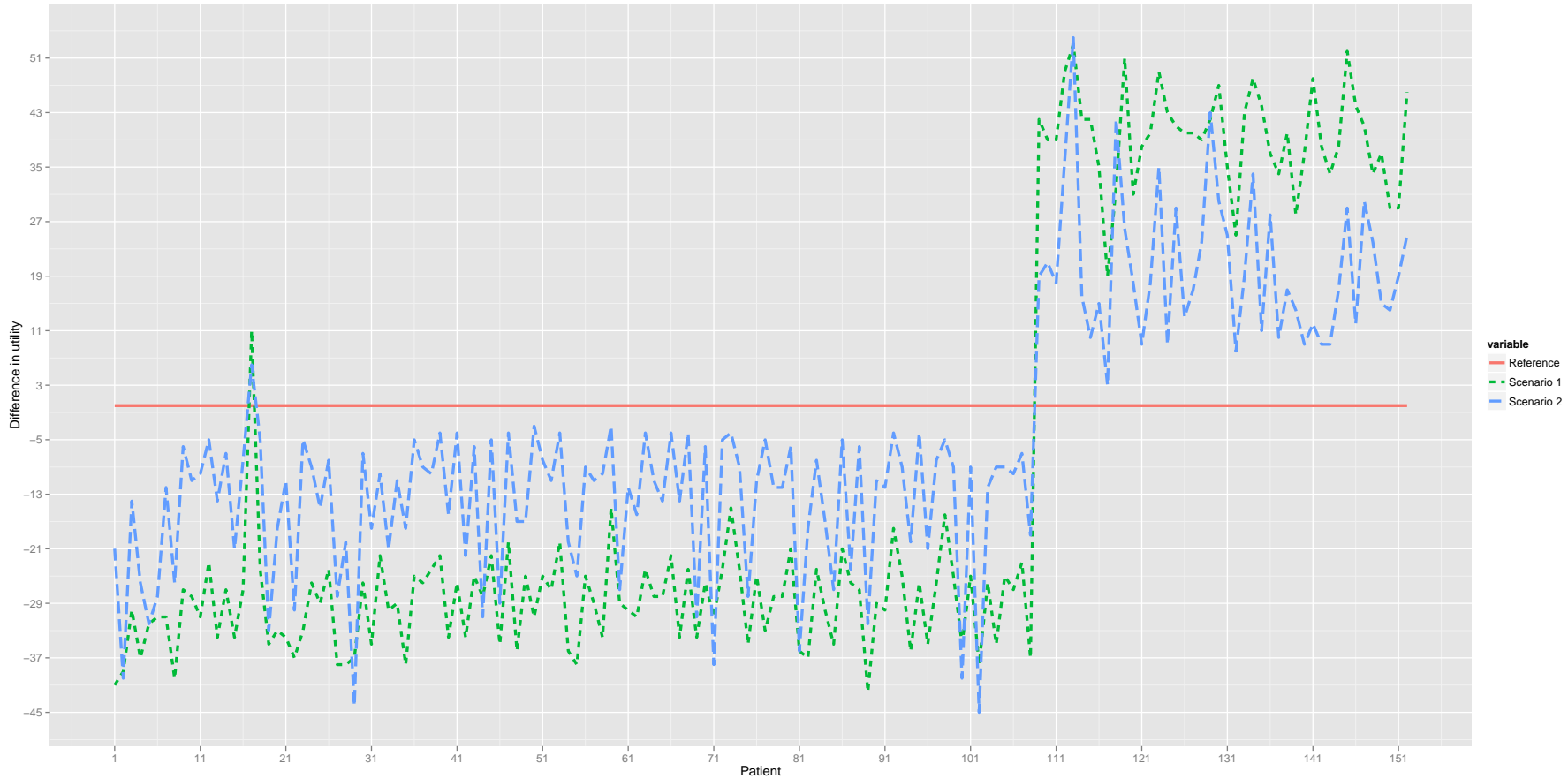
Table S1: Simulation results based on scenario 2. Scenario 2.1: 10 prognostic variables including 2 features and 8 random normal noise variables; scenario 2.2: 8 prognostic variables including 2 features and 6 random normal noise variables. The table provides means values and standard deviations (SD) obtained for the summary measures of $\%\Delta_g = \%\Delta\text{MTU}_g$, MOT, and CPO. Results are based on 100 duplicated data sets.

Method	MOT(SD)	$\%\Delta_g$(SD)	CPO(SD)
Scenario 2.1			
HC-BPP	10.0 (7.0)	0.853 (0.11)	68.9 (7.0)
KM-BPP	22.0 (9.3)	0.679 (0.15)	67.8 (6.9)
PAM-BPP	25.0 (9.3)	0.651 (0.14)	65.4 (7.0)
LASSO	47.0 (8.6)	0.462 (0.12)	99.1 (9.5)
Ridge	28.0 (5.9)	0.573 (0.10)	75.5 (6.6)
LassoINT	44.0 (10.2)	0.439 (0.15)	104.7 (9.8)
RidgeINT	34.0 (8.6)	0.482 (0.16)	75.9 (6.5)
Scenario 2.2			
HC-BPP	10.0 (8.0)	0.859 (0.12)	71.5 (7.5)
KM-BPP	22.0 (9.2)	0.686 (0.15)	70.4 (7.3)
PAM-BPP	24.0 (9.8)	0.662 (0.15)	67.7 (7.5)
LASSO	46.0 (9.0)	0.471 (0.12)	99.3 (9.7)
Ridge	28.0 (6.0)	0.574 (0.10)	75.8 (6.7)
LassoINT	44.0 (10.4)	0.438 (0.15)	104.8 (9.3)
RidgeINT	34.0 (8.8)	0.476 (0.16)	76.2 (6.7)

References

Franzén, J. (2008). Assessment of variations in control of asthma over time. Bayesian Cluster Analysis: Some Extensions to Non-standard Situations (Doctoral dissertation, Statistiska institutionen)

Golub, T. R. and Slonim, D. K. and Tamayo, P. and Huard, C. and Gaasenbeek, M. and Mesirov, J. P. and Coller, H. and Loh, M. L. and Downing, J. R. and Caligiuri, M. A. and others (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Ma, J. and Stingo, F. C. and Hobbs, B. P. (2016). Bayesian predictive modeling for genomic based personalized treatment selection. *Biometrics* **72**, 575–583.

Yu, L. and Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03) 2003; 856-863.

Figure S.1: Differences in the true mean treatment utilities for simulated patients. Negative (positive) values indicate enhanced effectiveness for treatment 1 (treatment 2).

Figure S.2: Scatter (top) and density (bottom) plots for the transformed features based on one simulated data set under scenario 2.
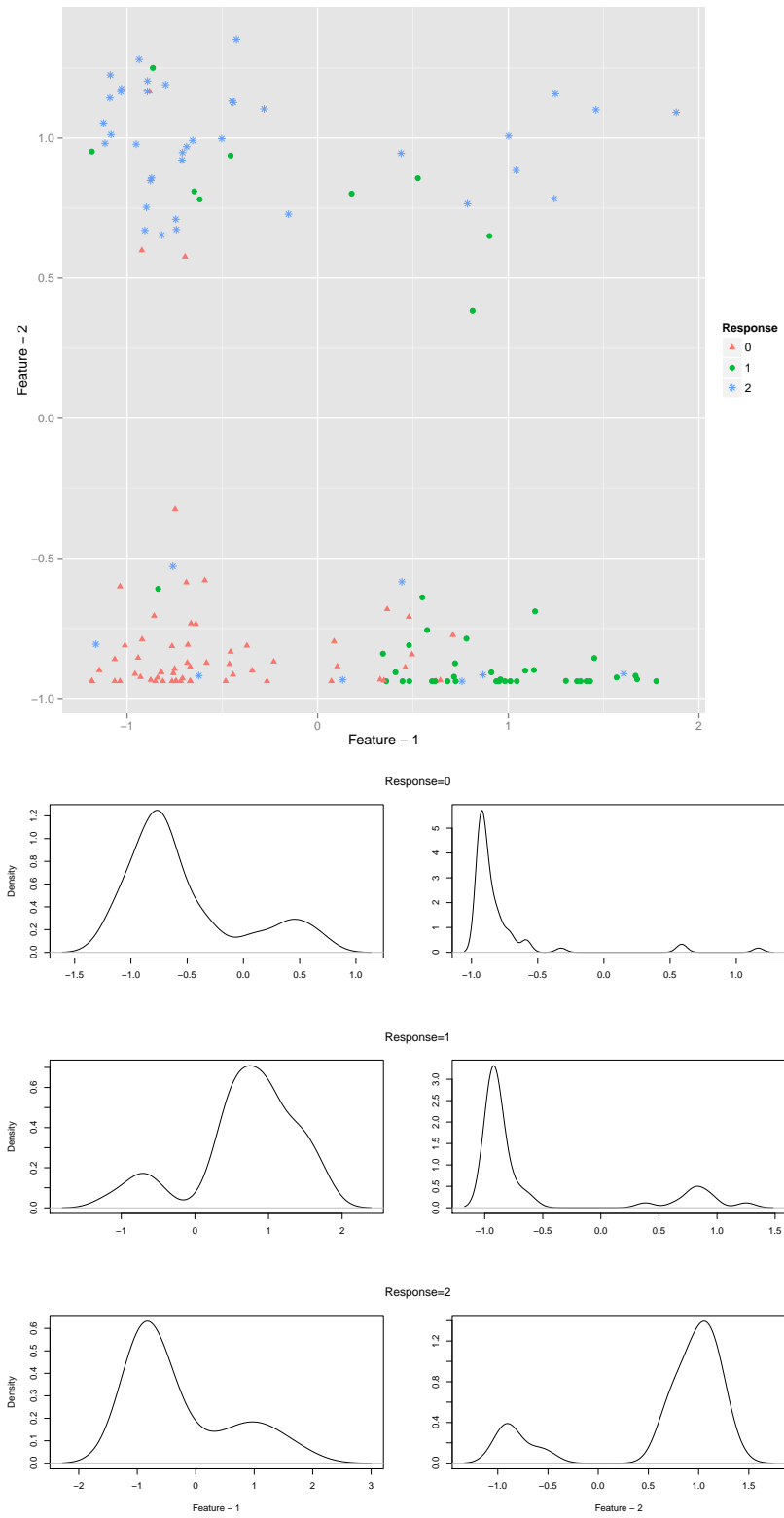
Figure S.3: Scatter (top) and density (bottom) plots for the orginal features based on one simulated data set under scenario 2.
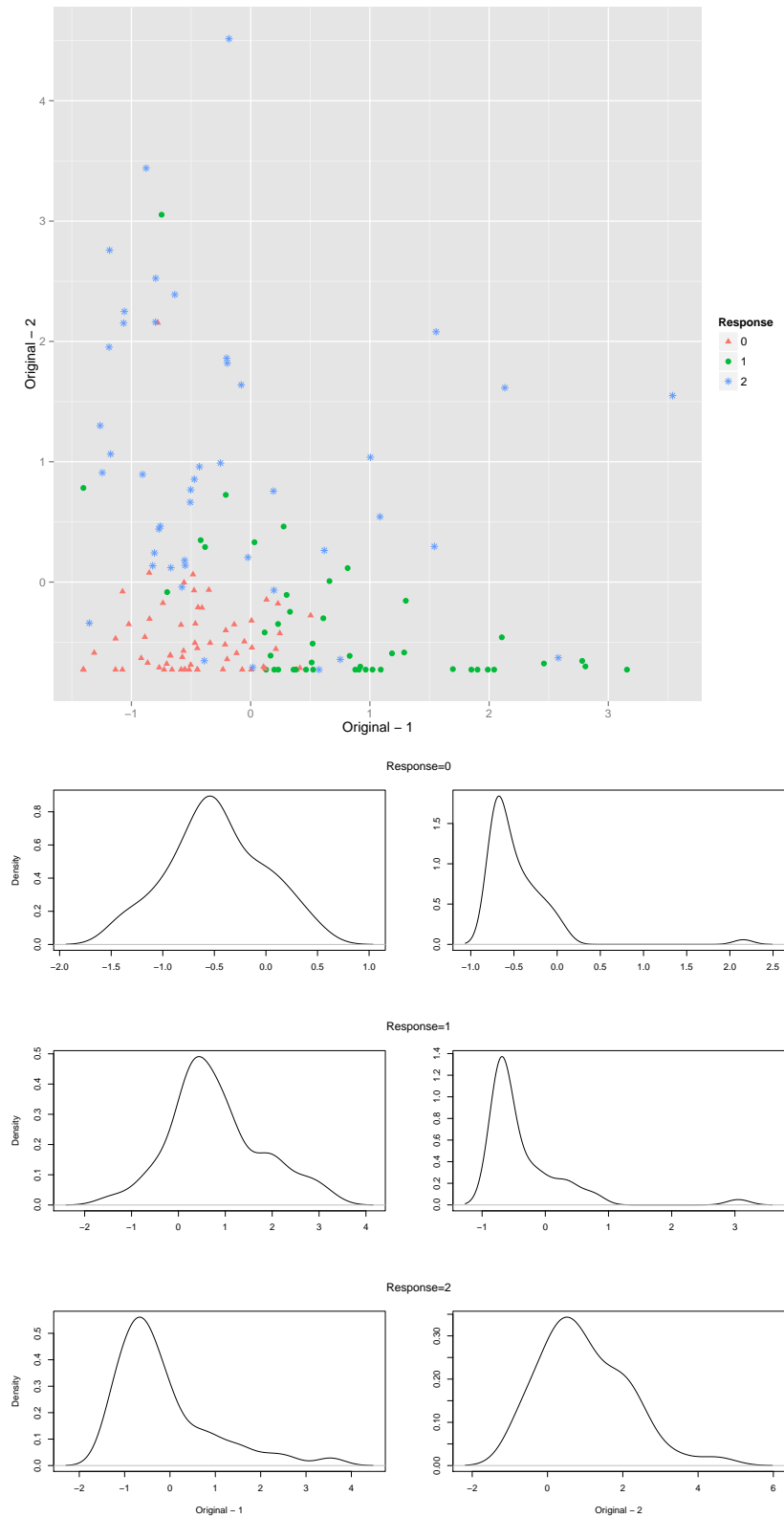
Figure S.4: Scatter (top) and density (bottom) plots for the random features based on one simulated data set under scenario 2.