4 Simulation Study

To assess the efficacy of the proposed methodology, we applied GRRF to simulated genetic data and compared its results with two well-known variable selection methods: varSelRF and LASSO logistic regression. The simulations were implemented in R using the "RRF", "varSelRF", and "glmnet" packages for the GRRF, varSelRF, and LASSO methods, respectively. In line with (Deng), we fixed $\texttt{mtry} = \sqrt{\text{number of variables in pathway}}$ and $\texttt{ntree} = 1000$ for the GRRF and varSelRF methods; we also held $\lambda_0 = 1$ so that γ was the only parameter for GRRF. Beyond this, the default settings for each of the methods were used.

4.1 Pathway Rankings

We simulated a genetic dataset with ten pathways, two of which were significant, and then measured the performance of GRRF, varSelRF, and LASSO in distinguishing significant from non-significant pathways.

To accomplish this, we first generate 140 iid random variables sampled from $\mathcal{U}(0,1)$. Pathways 1–2 are each assigned ten variables: $P_1, \ldots, P_{10}$ and $Q_1, \ldots, Q_{10}$, respectively. The remaining 120 variables are distributed evenly amongst Pathways 3–10. Within Pathways 1–2, the first five variables are designated significant. Following (Deng), we use the significant variables within each pathway to create five redundant variables: $P_{i+10} \leftarrow P_i$ and $Q_{i+10} \leftarrow Q_i$ for $i = 1, \ldots, 5$.

A response Y is then generated via one of the following response models:

RM1: An extension of the model found in (Friedman). Note $\epsilon \sim \mathcal{N}(0,1)$.

$$Y = 10\sin(\pi P_1 P_2) + 20(P_3 - 0.5)^2 + 10P_4 + 5P_5 \\ + 15\sin(\pi Q_1 Q_2) + 25(Q_3 - 0.5)^2 + 20Q_4 + 10Q_5 + \epsilon$$

RM2: A variant of RM1 that includes interaction across pathways.

$$Y = 10\sin(\pi P_1 Q_1) + 20(P_2 Q_2 - 0.5)^2 + 10(P_3 + Q_3) + 5(P_4 + Q_5) + 10(P_5 + Q_4) + \epsilon$$

RM3: A simple linear model.

$$Y = \sum_{i=1}^{5} 10(P_i + Q_i) + \epsilon$$

This process is repeated until 50 samples are obtained. The median response $\overline{Y}$ is then computed, and the samples are labeled according to the following rule: If $Y > \overline{Y}$, then the sample is assigned to Class 2; otherwise, the sample belongs to Class 1.

GRRF, varSelRF, and LASSO were then applied to each pathway. The parameter γ of GRRF was selected from $\{0.05, 0.25, 0.5, 0.75, 0.95\}$ resulting in five distinct models: GRRF-0.05,

GRRF-0.25, GRRF-0.50, GRRF-0.75, and GRRF-0.95. The regularization parameter of LASSO was chosen via 10-fold cross-validation. For each model, the pathways were scored and then ranked using an appropriate metric: minimum OOB error rate for the varSelRF and GRRF models; minimum cross-validated mean for LASSO models. In the event that two or more scores coincided, the ranks at each index set of ties were averaged—for instance, the list of numbers $(1, 2, 3, 1, 3, 1)$ would be ranked $(2.0, 4.0, 5.5, 2.0, 5.5, 2.0)$ under this scheme.

Model	Worst	Pathway 1	Pathway 2
GRRF-0.05	3.865	3.765	1.17
GRRF-0.25	4.205	4.13	1.26
GRRF-0.5	5.33	5.175	1.505
GRRF-0.75	5.92	5.7	2
GRRF-0.95	6.17	5.785	2.32
LASSO	4.605	4.545	1.09
varSelRF	6.46	6.19	1.94

Table 1: Pathway results for RM1.

Model	Worst	Pathway 1	Pathway 2
GRRF-0.05	3.375	2.42	2.265
GRRF-0.25	3.585	2.535	2.43
GRRF-0.5	4.515	3.29	3.1
GRRF-0.75	5.255	3.535	3.69
GRRF-0.95	5.44	3.825	3.745
LASSO	3.065	2.05	2.115
varSelRF	5.64	3.945	3.89

Table 2: Pathway results for RM2.

Model	Worst	Pathway 1	Pathway 2
GRRF-0.05	3.265	2.17	2.37
GRRF-0.25	3.64	2.43	2.59
GRRF-0.5	4.895	3.33	3.335
GRRF-0.75	5.48	3.645	3.935
GRRF-0.95	5.47	3.615	4.01
LASSO	3.015	2.09	1.975
varSelRF	5.955	3.81	4.375

Table 3: Pathway results for RM3.

In total, 100 datasets were simulated for each of RM1–RM3. The ranks of Pathways 1–2 were recorded, and, in particular, we recorded the rank of the worst performing pathway. Results are displayed in Tables 1–3. At first glance, the rankings appear surprisingly high. In particular, we see that the worst performing pathway, on average, never places in the top 3 for any of the models—and very often it does not even place in the top 5. However, the

results on the individual pathways are significantly better across the board, which suggests the inflated results are simply an artifact of the ranking procedure.

Comparing performance between models, we see that GRRF outperformed the varSelRF models in all but two cases in RM1 (GRRF-0.75 and GRRF-0.95) but tended not to fare as well against the LASSO models. However, in general, as γ of GRRF decreased—that is, as the GRRF models utilized less regularization—the differences in performance between the GRRF and LASSO models were narrowed as well. In particular, for $\gamma \in \{0.05, 0.25\}$, the GRRF models perform comparably to the LASSO models for each of RM1–RM3, generally differing by no more than half a position in the rankings of the pathways; and, in fact, within RM1, the GRRF models actually edge out the LASSO model with regard to their average rankings of Pathway 1 and the worst performing pathway.

4.2 Gene Selection

The next component of our simulation study measured the ability of GRRF, relative to varSelRF and LASSO, to identify the significant genes within predetermined important pathways.

The setup is more or less the same as in the previous section; however, we neglect to generate Pathways 3–10 as only Pathways 1–2 are significant. We then applied GRRF, varSelRF, and LASSO to each of Pathways 1–2. We allowed γ of GRRF to range over $\{0.5, 0.55, \ldots, 0.95\} \cup \{0.96, 0.97, 0.98, 0.99\}$; the regularization parameter for LASSO was selected via 10-fold cross-validation from $\{0.01, 0.02, 0.03, 0.05, 0.1, 0.2\}$. For each model, we assigned scores for each pathway based on the number of significant genes identified, minus the number of redundancies, as outlined in (Deng)—that is, if $(P_1, P_4, P_7, P_{11}, P_{13})$ are identified as significant, we would assign a score of 3, not 4, as P_1 and P_{11} are identical.

As above, this procedure was repeated 100 times for each of RM1–RM3. The results are given in Tables 4–6. Within each response model and pathway, the true positive and false positive rates of the GRRF models decreased as γ increased—which is expected, given higher γ values lead to increased regularization. Arguably, the best GRRF models were those for which the true positive and false positive rates were balanced, which roughly occurred for $\gamma \in \{0.85, 0.90, 0.95, 0.96, 0.97\}$. For γ in this range, we see that GRRF outperformed varSelRF across the board in all but a few cases—the most notable exceptions occurring in RM1, where varSelRF achieved a lower false positive rate in Pathway 1 and a higher true positive rate in Pathway 2. We also note that, though LASSO generally outperformed GRRF in terms of significant genes identified, the false positive rates of the LASSO models were typically 2–4 times higher than the GRRF models.

5 Breast Cancer Dataset

The proposed method was further assessed with an application to a breast cancer dataset consisting of three tumor classes (luminal, basal, and apocrine) for 49 breast cancer patients

(Farmer). A total of 441 pathways were associated with the dataset. Of the 441 pathways, 129 were taken from the KEGG pathway database and are responsible for metabolism, degradation, biosynthesis, and signal processing; the remaining 312 originate from BioCarta and are mostly related to signal transduction in humans and mice (Pang).

As an initial step, the pathways were separated into two groups according to whether the pathway contained less than 10 genes or not. For the first group (roughly 25% of the pathways), GRRF was ran with $\gamma = 0$ so as to provide minimal regularization; for the second, γ was selected from $\{0.05, 0.10, \ldots, 0.90, 0.95\}$ via 10-fold cross-validation. The pathways were then scored and ranked according to minimum OOB.

Table 7 displays the ten most significant pathways along with their corresponding OOB error rates, lengths, and optimal γ values. The results mesh well with recent work in the field. For instance, Palaskas et al. showed the Glycolysis-Gluconeogenesis pathway and several glycolysis-related pathways, such as Pentose-Phosphate and Carbon Fixation, were linked to the basal subtype of human breast cancer. Among the aforementioned glycolysis-related pathways, the Pentose-Phosphate pathway (PPP) has been suggested as a possible target in cancer therapies as the elevated PPP in cancer cells may help differentiate normal cells from cancer cells (Patra). It is also well known that estrogen plays an important role in breast cancer development and progression, and so it makes sense that three of the top 10 pathways (BC Downregulated of MTA-3 in ER-negative Breast Tumors, BC GATA3 Participate in Activating the Th2 Cytokine Genes Expression, and BC CARM1 and Regulation of the Estrogen Receptor) are directly linked to activities involving estrogen receptors (Kumar; Wilson; Frietze et al.). Most of the remaining pathways are associated with metabolic processes. The altered metabolisms of cancer cells (Palaskas) makes the appearance of such pathways in the top 10 at least plausible.

It is also worth noting that none of the pathways in the top 10 have less than 10 genes. In fact, the first pathway with less than 10 genes, Sulfur Metabolism, only narrowly rounds out the top 25, having placed in the 24th position in the rankings.

Model	P1 Significant	P1 Non-significant	P2 Significant	P2 Non-significant
GRRF-0.5	4.25	3.72	3.47	1.77
GRRF-0.55	3.83	3.35	3.23	1.46
GRRF-0.6	3.56	3.25	2.91	1.4
GRRF-0.65	3.39	2.8	2.65	1.19
GRRF-0.7	3.12	2.61	2.68	0.96
GRRF-0.75	2.8	2.35	2.47	1.01
GRRF-0.8	2.78	2.24	2.3	0.87
GRRF-0.85	2.59	2.04	2.23	0.75
GRRF-0.9	2.3	1.78	2.18	0.72
GRRF-0.95	2.26	1.71	2.05	0.68
GRRF-0.96	2.3	1.75	2.04	0.76
GRRF-0.97	2.29	1.73	2.2	0.79
GRRF-0.98	2.26	1.66	2.08	0.76
GRRF-0.99	2.2	1.63	2	0.7
LASSO	1.42	1.79	3.74	5.21
varSelRF	1.56	1.42	2.33	1.77

Table 4: Gene selection results for RM1.

Model	P1 Significant	P1 Non-significant	P2 Significant	P2 Non-significant
GRRF-0.5	3.97	2.63	3.88	3.07
GRRF-0.55	3.71	2.43	3.65	2.75
GRRF-0.6	3.36	2.07	3.3	2.32
GRRF-0.65	3.25	1.79	3.19	2.21
GRRF-0.7	2.96	1.59	3.07	1.81
GRRF-0.75	2.94	1.45	2.9	1.79
GRRF-0.8	2.75	1.5	2.84	1.6
GRRF-0.85	2.49	1.23	2.57	1.48
GRRF-0.9	2.57	1.1	2.48	1.48
GRRF-0.95	2.45	1.11	2.34	1.33
GRRF-0.96	2.4	1.12	2.39	1.26
GRRF-0.97	2.47	1.02	2.43	1.24
GRRF-0.98	2.4	1.07	2.31	1.26
GRRF-0.99	2.26	1.24	2.39	1.07
LASSO	3.35	4.46	3.13	4.05
varSelRF	1.95	1.49	2.04	1.6

Table 5: Gene selection results for RM2.

Model	P1 Significant	P1 Non-significant	P2 Significant	P2 Non-significant
GRRF-0.5	4.2	2.65	4.05	2.6
GRRF-0.55	3.9	2.2	3.85	2.23
GRRF-0.6	3.68	1.88	3.61	1.86
GRRF-0.65	3.45	1.63	3.39	1.76
GRRF-0.7	3.28	1.36	3.23	1.6
GRRF-0.75	3.11	1.22	3.06	1.4
GRRF-0.8	2.94	1.17	2.91	1.35
GRRF-0.85	2.74	1.17	2.81	1.1
GRRF-0.9	2.74	1.1	2.6	1.02
GRRF-0.95	2.62	0.97	2.66	1.02
GRRF-0.96	2.59	1.01	2.58	1.06
GRRF-0.97	2.56	1.01	2.54	0.97
GRRF-0.98	2.51	1	2.55	1.13
GRRF-0.99	2.6	0.86	2.59	0.97
LASSO	3.79	4.97	3.77	5
varSelRF	2.08	1.34	2.08	1.58

Table 6: Gene selection results for RM3.

Pathway	Length	Gamma	Error
Glycolysis-Gluconeogenesis	68	0.10	0.022
BC Downregulated of MTA-3 in ER-negative Breast Tumors	19	0.15	0.029
BC GATA3 Participate in Activating the Th2 Cytokine Genes Expression	21	0.05	0.031
Pentose Phosphate	22	0.05	0.051
Fructose and Mannose Metabolism	39	0.05	0.055
Carbon Fixation	25	0.05	0.071
BC CARM1 and Regulation of the Estrogen Receptor	24	0.70	0.082
Glutathione Metabolism	31	0.05	0.088
Valine, Leucine and Isoleucine Degradation	46	0.05	0.094
Purine Metabolism	112	0.10	0.100

Table 7: Breast cancer dataset results.