The Annals of Applied Statistics
2020, Vol. 14, No. 4, 1676–1698
https://doi.org/10.1214/20-AOAS1389
© Institute of Mathematical Statistics, 2020

FEATURE SELECTION FOR DATA INTEGRATION WITH MIXED MULTIVIEW DATA

BY YULIA BAKER[1], TIFFANY M. TANG[2] AND GENEVERA I. ALLEN[3]

[1]*Department of Statistics, Rice University,* yulia.baker@rice.edu
[2]*Department of Statistics, University of California,* tiffany.tang@berkeley.edu
[3]*Department of Electrical and Computer Engineering, Rice University,* gallen@rice.edu

Data integration methods that analyze multiple sources of data simultaneously can often provide more holistic insights than can separate inquiries of each data source. Motivated by the advantages of data integration in the era of "big data," we investigate feature selection for high-dimensional multiview data with mixed data types (e.g., continuous, binary, count-valued). This heterogeneity of multiview data poses numerous challenges for existing feature selection methods. However, after critically examining these issues through empirical and theoretically-guided lenses, we develop a practical solution, the Block Randomized Adaptive Iterative Lasso (B-RAIL) which combines the strengths of the randomized Lasso, adaptive weighting schemes and stability selection. B-RAIL serves as a versatile data integration method for sparse regression and graph selection, and we demonstrate the effectiveness of B-RAIL through extensive simulations and a case study to infer the ovarian cancer gene regulatory network. In this case study, B-RAIL successfully identifies well-known biomarkers associated with ovarian cancer and hints at novel candidates for future ovarian cancer research.

1. Introduction. As the amount of data grows in volume and variety, data integration, or the analysis of multiple sources of data simultaneously, is becoming increasingly necessary in numerous disciplines. For example, in genomics, scientists can gather data from many related, yet distinct sources, including gene expression, miRNA expression, point mutations and DNA methylation. Since all of these genomic sources interact within the same biological system, it can be advantageous to analyze them together via data integration. Ultimately, the abundance and diversity of information captured by integrated data offers an invaluable opportunity to gain a better and more holistic understanding of the phenomena at hand.

In this work we aim to perform feature selection for a common family of integrated data sets called high-dimensional *multiview* data. Multiview data refers to data collected on the same set of samples but with features from multiple sources of potentially mixed types (e.g., categorical, binary, count, proportion, continuous and skewed continuous values). More formally, suppose we observe multiview data with K high-dimensional views (or sources), $\mathbf{X}_1 \in \mathbb{R}^{n\times p_1}, \dots, \mathbf{X}_K \in \mathbb{R}^{n\times p_K}$, which are measured on the same n samples but with features of mixed types. We seek to recover a sparse set of features from each $\mathbf{X}_k$ associated with the response $\mathbf{y} \in \mathbb{R}^n$ by considering

$$\underset{\alpha,\boldsymbol{\beta}_1,\dots\boldsymbol{\beta}_K}{\text{minimize}} -\frac{1}{n}\ell\left(\mathbf{y}; \alpha\mathbf{1}_n + \sum_{k=1}^{K}\mathbf{X}_k\boldsymbol{\beta}_k\right) \quad \text{subject to } \sum_{k=1}^{K}\|\boldsymbol{\beta}_k\|_0 \le \nu. \tag{1.1}$$

Here, $\boldsymbol{\beta}_k \in \mathbb{R}^{p_k}$ are the coefficients associated with view k, $\nu > 0$ is a tuning parameter, which regulates the sparsity level, and $\ell()$ is the generalized linear model (GLM) log-likelihood

Received March 2019; revised January 2020.
Key words and phrases. Data fusion, multimodal data, integrative genomics, variable selection, Lasso/GLM Lasso, stability selection, mixed graphical models.

associated with $\mathbf{y}$. Note, we not only consider continuous (Gaussian) responses but also the broader class of GLMs including the Poisson (log-linear) and Bernoulli (logistic) families.

While there are many applications for multiview feature selection in genomics, imaging, national security, economics and other fields, major difficulties, stemming from the heterogeneity of features and how to appropriately integrate such differences, have prevented the successful use of multiview feature selection in practice. To our knowledge, no one has proposed an effective practical solution to perform feature selection with multiview data. A plethora of works have studied feature selection in the high-dimensional setting via the Lasso or GLM Lasso (Tibshirani (1996, 2013), Yuan and Lin (2007), Zhao and Yu (2006)), and others have studied various data integration problems (Hall and Llinas (1997), Shen, Olshen and Ladanyi (2009), Acar, Kolda and Dunlavy (2011)). However, there is limited research at the intersection of the two fields.

The one area that touches on multiview feature selection is in the context of mixed graphical models which estimate sparse graphs between features in multiview data (Cheng et al. (2017), Lee and Hastie (2013), Yang et al. (2014a, 2014b), Haslbeck and Waldorp (2015)). Using the nodewise neighborhood estimation approach of Meinshausen and Bühlmann (2006), mixed graphical models estimate the neighborhood of each node (i.e., feature) separately via a penalized regression model (typically based on the Lasso or GLM Lasso) and combine neighborhoods using an "AND" or "OR" rule. Though mixed graphical models perform well in idealized settings for which theoretical guarantees have been proven, we will demonstrate in Section 2 that there are severe limitations with these approaches in realistic settings with correlated, heterogeneous features commonly found in multiview data.

To facilitate more effective integrative analyses in practice, we investigate the understudied problem of high-dimensional multiview feature selection, and we propose a practical solution. Our work is the first to identify and to critically examine the fundamental challenges of multiview feature selection, and we leverage this deep understanding of the challenges to develop a new high-dimensional multiview selection method, the Block Randomized Adaptive Iterative Lasso (B-RAIL). B-RAIL is a practical tool for multiview feature selection with its roots grounded in theory, and it builds upon adaptive ℓ_1 penalties, the randomized Lasso and stability selection (Meinshausen and Bühlmann (2010)) to overcome the issues incurred by existing methods. Our method can be used for both regression and mixed graphical selection, thus lending itself to a host of important applications.

In Section 2 we investigate the major challenges of multiview feature selection and highlight the literature gaps relating to these issues. We also show that the culmination of these challenges lead to poor feature recovery in existing Lasso-type methods and mixed graphical models. In Section 3 we introduce our proposed method, B-RAIL, which takes steps to address the challenges from Section 2. In Section 4 we showcase the strong empirical performance of B-RAIL through simulations and contrast it to existing methods. In Section 5 we further demonstrate the effectiveness of B-RAIL in a novel integrative genomics case study for ovarian cancer, and we provide concluding remarks in Section 6.

2. Challenges.

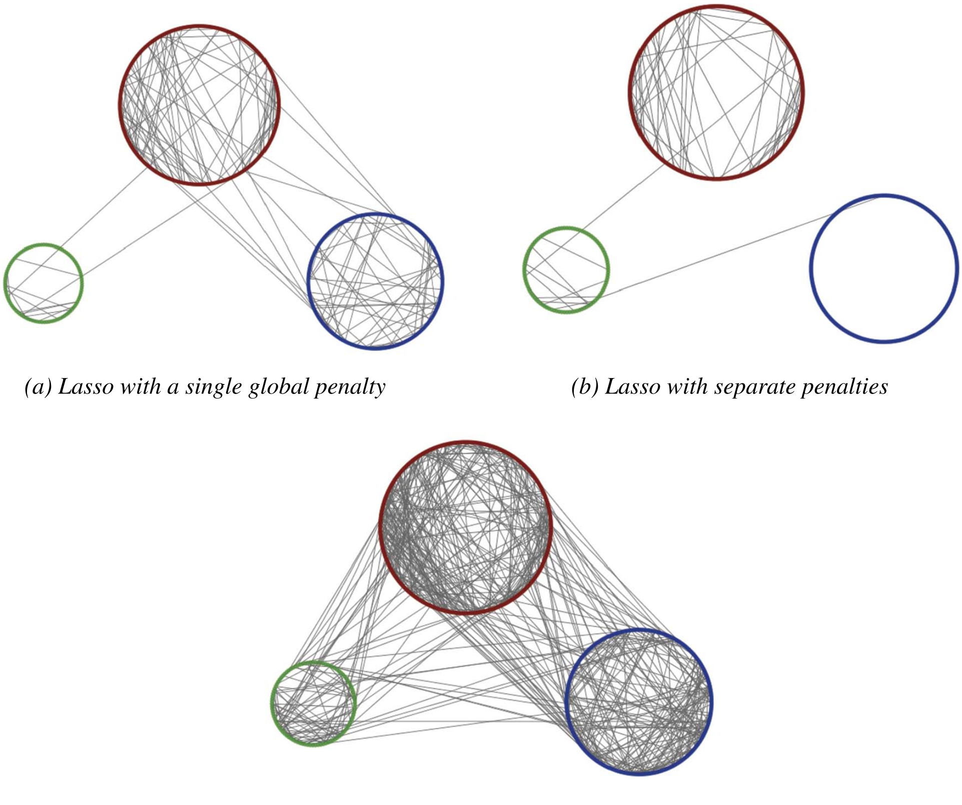
Before introducing our proposed method, it is instructive to understand the challenges posed by feature selection in the multiview setting. These challenges have been overlooked in previous methods and thus contribute to many of their shortcomings. In this section we focus on the challenges faced by linear models with Lasso-type penalties due to their overwhelming popularity and desirable statistical properties (Meinshausen and Yu (2009), Tibshirani (1996, 2013), Yuan and Lin (2007), Zhang and Huang (2008), Zhao and Yu (2006)). Given data $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response $\mathbf{y} \in \mathbb{R}^n$, recall that the (GLM) Lasso solves

$$\hat{\alpha}, \hat{\boldsymbol{\beta}} = \underset{\alpha \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} - \frac{1}{n} \ell(\mathbf{y}; \alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1, \tag{2.1}$$

where $\lambda > 0$ is a regularization parameter and $\ell()$ is the GLM log-likelihood associated with the response. For clarity, we use the term "Lasso" to refer to the ℓ_1-penalized model with continuous (Gaussian) responses, "GLM Lasso" to mean the ℓ_1-penalized model with non-Gaussian GLM responses (e.g., binary, Poisson) and "Lasso-type" methods to mean either the Lasso or GLM Lasso with some form of ℓ_1 penalty (e.g., a global penalty, separate penalties, adaptive penalties).

Our focus here is not on deriving new theoretical guarantees for the Lasso in multiview settings. Rather, we highlight deep practical concerns which are rooted in theory and commonly arise in feature selection for data integration. By identifying these practical challenges, we open up numerous avenues for future theoretical research and set the stage for the construction of a new method which overcomes the identified issues.

2.1. *Motivating example.* To first illustrate the current challenges and motivate the need for a solution, we present in Figure 1 the estimated graphs from common Lasso-type methods and our proposed method when applied to real ovarian cancer genomics data. Here, there are $n = 293$ samples and $p = 836$ features from three views: count-valued RNASeq data ($p_{\text{RNASeq}} = 408$), continuous miRNA data ($p_{\text{miRNA}} = 307$) and proportion-valued methylation data ($p_{\text{Methyl}} = 301$) (refer to Section 5 for data collection and preprocessing details). As in several previous graphical models and mixed graphical models (Meinshausen and



(a) *Lasso with a single global penalty*

(b) *Lasso with separate penalties*

(c) *The B-RAIL algorithm*

FIG. 1. *We compare three different graph selection methods when applied to real ovarian cancer genomics data. The data is comprised of three blocks*: *RNASeq*, *miRNA and methylation*, *with* $n = 293$ *and* $p = 836$. *For all three methods we use stability selection with the threshold* 0.98 *to select stable edges*, *and*, *hence*, *we can directly compare the number of selected edges across the various methods. The Lasso with a single global penalty and the Lasso with separate penalties select few edges within blocks and almost no edges between blocks*, *indicating that these methods are highly unstable to small perturbations of the data.*

Bühlmann (2006), Ravikumar, Wainwright and Lafferty (2010), Jalali et al. (2011)), we estimated the graphs using nodewise neighborhood selection. We then combined neighborhoods using the "AND" rule and applied stability selection (Meinshausen and Bühlmann (2010), Liu, Roeder and Wasserman (2010)) with the threshold 0.98 to select stable edges.

Figure 1 specifically compares three types of estimation schemas at each node: (a) GLM Lasso with one global penalty, (b) GLM Lasso with separate penalties for each view and (c) our proposed B-RAIL algorithm (introduced in Section 3). The first two methods have been proposed in several mixed graphical models (Chen, Witten and Shojaie (2015), Yang et al. (2014a), Haslbeck and Waldorp (2015)) and satisfy strong theoretical guarantees in idealized settings. However, in the real data example the Lasso-type methods are unstable (illustrated by the fewer edges), favor feature selection within one view and select only a few edges between views. This overall instability indicates that the Lasso-type methods are not robust to small perturbations of the data and raises serious concerns about the reproducibility and reliability of the results (Yu (2013)). Our proposed B-RAIL algorithm, in contrast, avoids these issues and exhibits greater stability as well as balance, selecting a larger number of within and between block edges under the same thresholding value. We will later see through extensive simulations in Section 4 that the issues with existing Lasso-type methods observed here are recurring problems in very general multiview scenarios.

To begin understanding why existing Lasso-type methods struggle in practice, we identify and study four major challenges of feature selection for high-dimensional multiview data: (1) scaling, (2) ultra-high-dimensionality, (3) signal interference and (4) domain-specific beta-min. These issues stem from a combination of domain differences, signal differences and the high dimensionality of each view. Together, these challenges can have a significant adverse effect on feature recovery for data integration. We next examine each of these challenges in greater detail.

2.2. *Scaling.* The first and most obvious challenge with integrative analyses revolves around scaling. That is, each view in a multiview data set is often measured on a different scale, and it is unclear how to most effectively integrate such differences. Many believe that normalizing all features to mean 0 and variance 1 remedies the scaling differences, but this is not always the case. Even after centering and scaling, data views remain distinct if they differ in ways beyond the first and second moments. This issue is especially problematic with binary and count-valued data blocks, two common types in multiview data, since they are defined by much higher moments. We thus highly discourage using the ordinary (GLM) Lasso with a single penalty (2.1) on normalized multiview data.

Now, while one can use different regularization parameters for each view to help alleviate the scaling differences, this generates another set of issues that are complicated by the following challenges. We will revisit the scaling issue in light of these complications later in this section.

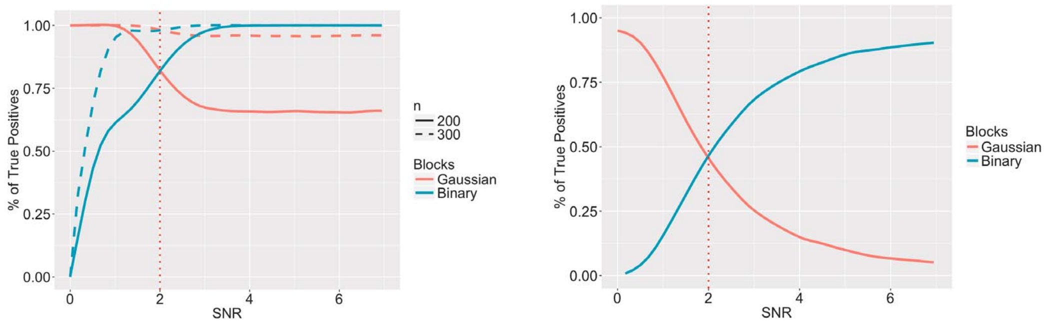
2.3. *Ultrahigh dimensionality.* In addition to the scaling issue, performing exact feature selection with the Lasso is already difficult in the ordinary high-dimensional setting. For exact feature selection, the number of samples n must be above a theoretical minimum known as the *sample complexity*. In the highly idealized scenario of an i.i.d. standard Gaussian design and a Gaussian response, Wainwright (2009) showed that the sample complexity scales at approximately $2s\log(p-s)$, where p is the number of features and s is the number of nonzero features. This idealized lower bound can be difficult to attain in many applications including genomics, where typical values of $p = 1000$ and $s = 30$ demand $n \approx 400$ patients—a large and highly expensive study. We informally refer to the regime where $p \gg n \geq 2s\log(p-s)$ as "high dimensional" and $n < 2s\log(p-s)$ as "ultrahigh dimensional." Roughly, the Lasso can never perform exact feature selection in the ultrahigh-dimensional regime.

For non-Gaussian responses and more realistic designs, such as correlated, heterogeneous views in multiview data, the sample complexity is significantly higher than the idealized Gaussian bound (Chen, Witten and Shojaie (2015), Ravikumar, Wainwright and Lafferty (2010)). As an example, the Poisson GLM's sample complexity scales at approximately $s^2 \log(p(\log p)^2)$ (Yang et al. (2015)), so if $p = 1000$ and $s = 30$, we require $n \approx 10{,}000$ samples. This problem is further exacerbated in multiview settings since combining multiple high-dimensional views for data integration almost always results in an ultrahigh-dimensional problem.

2.4. *Signal interference.* The third challenge we identify stems from a problem with the Lasso known as shrinkage noise. Su, Bogdan and Candès (2017) showed that with high probability, no matter how strong the effect sizes, false discoveries appear early on the Lasso path due to pseudo noise introduced by shrinkage in the high- and ultrahigh-dimensional regimes. When the Lasso selects its first few features using large regularization parameters, the residuals still contain much of the signal associated with the selected features, and it is this extra noise which Su, Bogdan and Candès (2017) calls shrinkage noise.

In the multiview context, shrinkage noise becomes a very complex and serious issue due to the different signals across blocks. Since the Lasso naturally selects features from the block with the highest signal first, the resulting shrinkage noise will mask the weaker signals from other blocks and compromise our ability to select from these weaker blocks. We refer to this adverse consequence of shrinkage noise as *signal interference*.

The problem of shrinkage noise has not been widely studied beyond the i.i.d. Gaussian design in Su, Bogdan and Candès (2017), but we provide strong empirical evidence in Figure 2 that confirms the existence of shrinkage noise and signal interference in non-Gaussian multiview settings. In the case of an i.i.d. Gaussian and an i.i.d. binary block, shown in Figure 2, the Lasso achieves perfect recovery in the Gaussian block when the signal-to-noise ratio (SNR) in the binary block is 0, but, as the SNR of the binary block increases, it interferes with our ability to recover the Gaussian features in the small sample scenario of $n = 200$. This signal interference is especially disastrous in the GLM Lasso with binary responses, where support recovery in the Gaussian block tends to 0. However, when we increase the sample size to $n = 300$, there is no decline in the recovery of the Gaussian block in Figure 2(a). This agrees with the known result from Su, Bogdan and Candès (2017) that shrinkage noise occurs when



(a) Gaussian response, Gaussian and Binary predictors

(b) Binary response, Gaussian and Binary predictors

FIG. 2. *We illustrate signal interference for both Gaussian and binary responses given i.i.d. Gaussian* $\mathbf{X}_1$ *and i.i.d. binary* $\mathbf{X}_2$ *predictors. We simulate* $n = 200$, $p_1 = p_2 = 1000$ *and* 10 *true features in each block. We fix the SNR for the Gaussian block at* 2 *and let the SNR of the binary block vary between* 0 *and* 7. *The dotted vertical line highlights the point at which* $SNR_1 = SNR_2 = 2$. *As the SNR of the binary block increases, it interferes with the ability to recover the true Gaussian features. This signal interference is even more severe for binary responses.*

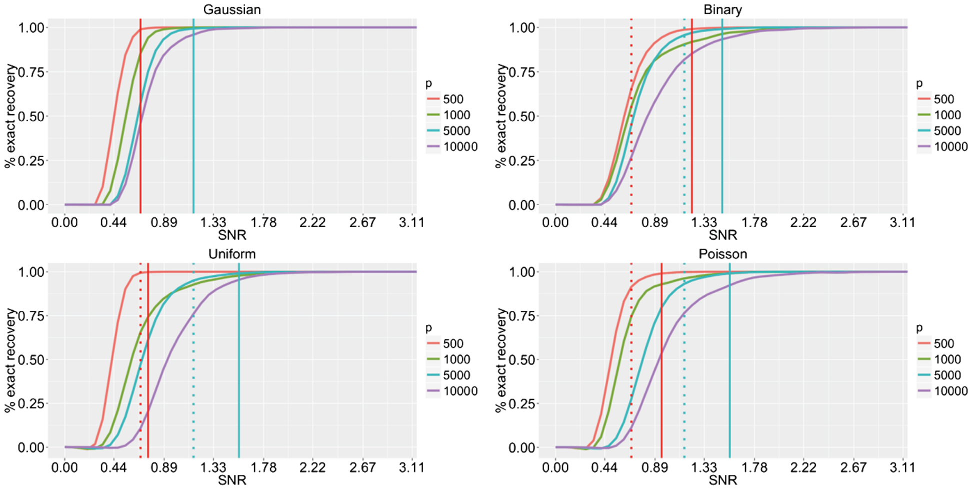


FIG. 3. *We simulate Gaussian responses given four types of predictors* (*Gaussian*, *binary*, *uniform*, *Poisson*) *and compare our ability to recover the true features under the four designs*. *There are* $n = 200$ *samples*, p *features and* 10 *true features*. *The dashed vertical lines indicate the minimum SNR required to achieve* 99% *recovery for* $p = 500$ (*darker line*) *and* $p = 5000$ (*lighter line*). *In the case of non-Gaussian predictors*, *dotted vertical lines are overlayed to compare the minimum SNR requirements to those of Gaussian predictors*. *These results show that different data types can tolerate different minimum SNRs*.

the Lasso's theoretical conditions are violated and, in particular, when n is not sufficiently large.

2.5. *Domain-specific beta-min condition*. Finally, analogous to how signal differences can exacerbate the Lasso's shrinkage noise issue, domain differences in multiview problems can complicate the Lasso's *beta-min condition*, which establishes a lower bound for the minimum amount of signal (i.e., SNR) required for feature recovery (Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Bühlmann (2013)).

In Figure 3 we report our ability to recover the true features for a simple simulation with i.i.d. features from four data types (Gaussian, binary, uniform and Poisson). In each subplot the dashed lines indicate the minimum SNR needed to recover 99% of all true features when $p = 500$ (darker line) and $p = 5000$ (lighter line). We observe that the minimum SNR requirement varies based upon the domain of the features and that the Gaussian predictors can tolerate the lowest SNR. These empirical results reveal that if two blocks have the same amount of signal but are from different domains, we can only recover the features that pass the minimum signal threshold dictated by the domains. Put concretely, if we were to perform feature selection on our simulated multiview data with $p = 500$ and SNR $= 0.66$, we would be able to recover 99% of the true features in the Gaussian block but only about $3/4$ of the true features in the binary block. This observed phenomenon agrees with previous work which has shown that an increase in the sparsity in $\mathbf{X}$ effectively reduces the SNR in the high-dimensional setting (Wang, Wainwright and Ramchandran (2010)). Beyond this, however, the beta-min condition has been relatively unexplored for the GLM Lasso and domain differences and remains a ripe area for future theoretical work.

2.6. *Additional challenges*. It is important to note that the four challenges above do not act independently from one another. In fact, the main source of difficulty with multiview feature selection is arguably the interactions between challenges. For instance, consider the problem of selecting features from high-dimensional discrete blocks with weak signals. The ultrahigh-dimensionality issue can exacerbate the already existing problem of signal interference which can then worsen scaling issues, increase minimum SNR requirements and amplify the overall difficulty of the problem.

In conjunction with these complex interactions, the need to select an appropriate amount of regularization λ through model selection methods can also increase the difficulty of multiview feature selection. We will compare three common selection methods, namely, stability selection (Meinshausen and Bühlmann (2010), Liu, Roeder and Wasserman (2010)), cross-validation (Allen (1974), Stone (1974), Shao (1993)) and extended BIC (Chen and Chen (2012)), and discuss their additional challenges in Section 4.

We lastly note that the majority of our discussion has been focused on the Lasso. Feature selection is even more challenging for the GLM Lasso. Chen, Witten and Shojaie (2015) investigated this for mixed graphical models and concluded that the predictors associated with Gaussian responses are easier to recover than those with responses from other exponential families. Specifically, Gaussian responses require fewer samples, allow for a wider range of tuning parameters and, generally, have a higher probability of success.

2.7. *Challenges with existing methods.* Having identified a host of challenges, we return to address why common Lasso-type methods are not well suited for multiview feature selection.

To begin with its most simple form, the *Lasso with a single global penalty* (2.1) uses the same penalty for all views and does not alleviate the scaling issues or signal interference issues in multiview data. The consequences of these problems are evident, especially in the case of non-Gaussian blocks with weak signals, in Figure 1(a), where fewer edges are selected within the proportion-valued methylation block.

By employing the *Lasso with separate penalties* for each data view, we can mitigate the issue of scaling. Nevertheless, model selection becomes more challenging with multiple penalties, and signal interference remains a driver of poor recovery. In fact, having a separate penalty for each view exacerbates signal interference and encourages selection from the block with the strongest signal and no selection from the blocks with weak signals. This signal interference is exemplified in Figure 1 by the extreme selection imbalance among views, with almost no selection in the miRNA block and heavy selection in the RNASeq block.

In the *Adaptive Lasso* (Zou (2006)), the amount of ℓ_1-regularization associated with β_j is typically $\lambda/|\hat{\beta}_j^{\mathrm{OLS}}|^\gamma$ for some constants $\gamma, \lambda > 0$. This adaptive penalty mitigates the scaling issue by adjusting for signal differences through $\hat{\beta}_j^{\mathrm{OLS}}$, but it also encourages selection of features with higher signals and penalizes features with weaker signals. The Adaptive Lasso hence complicates signal interference by treating weaker signals as noise and results in little to no selection in the blocks with weak signals.

While the previous methods all struggle with signal interference, one simple way to reduce the signal interference between blocks is to perform *separate Lassos* for each data view. Since independently-estimated blocks cannot possibly interfere with one another, this method addresses both scaling and signal interference issues. It also avoids the problem of ultrahigh-dimensionality. However, each view by itself usually does not contain sufficient information to explain much of the variability in the response, and we lose the advantages of data integration.

Beyond the Lasso-type methods, there are selection methods with nonconvex penalties such as SCAD and MCP (Fan and Li (2001), Zhang (2010)). These nonconvex penalties tend to scale better than the Lasso-type penalties but are still not variable selection consistent in the ultrahigh-dimensional regime, especially for non-Gaussian responses and highly correlated data. We investigate MCP/SCAD feature selection in Table 6 in the Supplementary Material (Baker, Tang and Allen (2020)), but our primary focus in this paper is on the more commonly used Lasso-type penalties.

Algorithm 1 Outline of Block—Randomized Adaptive Iterative Lasso

```
Initialize $t = 0$ and $\hat{\boldsymbol{\beta}}_k^{(0)}$ to have a fixed proportion of sparsity for $k = 1, \dots K$.

Do until Supp($\hat{\boldsymbol{\beta}}^{(t)}$) stops changing:
  • Set $t = t + 1$.
  • For $k = 1, \dots, K$, estimate $\hat{\boldsymbol{\beta}}_k^{(t)}$ blockwise, holding $\hat{\boldsymbol{\beta}}_l^{(t)}$ $(l < k)$ and $\hat{\boldsymbol{\beta}}_l^{(t-1)}$ $(l > k)$ fixed:
      1. Estimate the support $\hat{S}_k^{(t)}$ of block $k$:
         – Use stability selection with the randomized Lasso and adaptive penalties
      2. Given $\hat{S}_k^{(t)}$, estimate $[\hat{\boldsymbol{\beta}}_k^{(t)}]_{\hat{S}_k^{(t)}}$, the estimated nonzero coefficient values of block $k$

Output $\hat{\boldsymbol{\beta}}_1, \dots \hat{\boldsymbol{\beta}}_K$.
```

3. Block randomized adaptive iterative lasso.

Driven by the many challenges and the lack of effective tools, we propose a new method for multiview feature selection, the Block Randomized Adaptive Iterative Lasso (B-RAIL). For the sake of notation, suppose we observe the response vector $\mathbf{y}$ and multiview data $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_K]$ with K views of potentially mixed types, n samples and p total features. We will assume $p \gg n$ and typically $p_k \gg n$ for each view. Let S denote the indices of the support, and let $[\mathbf{X}]_S$ denote the columns of $\mathbf{X}$ indexed by S. We will introduce B-RAIL in the context of regression and later discuss its extension to graph selection.

Under the regression framework the goal of B-RAIL can be viewed as two-fold: 1) to select features from each view $\mathbf{X}_k$ that are associated with the response $\mathbf{y}$, and 2) to do so while avoiding the challenges discussed in Section 2. With this goal in mind, we briefly outline the B-RAIL algorithm in Algorithm 1 and summarize the key steps taken to overcome the current challenges.

At a high level, B-RAIL iterates across the data blocks $k = 1, \dots, K$ and estimates $\hat{\boldsymbol{\beta}}_k$ for each data block $\mathbf{X}_k$ separately while holding all other blocks fixed. This iterative procedure is motivated by the advantages of performing separate Lassos, namely, that it mitigates the ultrahigh-dimensionality and signal interference issues. Then, within each of the individual block estimations, B-RAIL first estimates the block's support and, subsequently, the coefficient values given the support. Here, B-RAIL leverages ideas from adaptive weighting schemes, stability selection and the randomized Lasso in an attempt to reduce the scaling discrepancies and domain-specific beta-min issues.

We next provide the full B-RAIL algorithm in Algorithm 2 and proceed to discuss each step of the B-RAIL algorithm in greater detail.

3.1. *Initialization.*

In our proposed B-RAIL algorithm the coefficients are first initialized to a prespecified sparsity level (e.g., $0.2p_k$ nonzero features per block) by fitting separate Lasso path regressions for each block. As long as the algorithm is initialized to an overselection of the support of $\boldsymbol{\beta}$, we have found in all of our empirical simulations that the B-RAIL algorithm tends to perform well and is very robust to the exact choice of initialization.

After initializing $\boldsymbol{\beta}$, we must specify the order of the blocks to iterate over. This ordering can slightly alter the estimation results of B-RAIL, as accurate estimation of previous blocks makes subsequent estimations of other blocks easier, but, in most cases, we have found that the block ordering is not as important to B-RAIL's performance as initializing the coefficients to an overselection of the support. Nevertheless, for best practices, since previous Lasso results guarantee a high probability of support recovery when n is sufficiently large compared

Algorithm 2 Block—Randomized Adaptive Iterative Lasso (B-RAIL)

```
Initialization:
  • Set $t = 0$.
  • Initialize $\hat{\boldsymbol{\beta}}^{(0)} = [\, \hat{\beta}_1^{(0)} \cdots \hat{\beta}_K^{(0)} \,]$, where $\|\hat{\boldsymbol{\beta}}_k^{(0)}\|_0 \approx 0.2 p_k$ for $k = 1, \dots K$.
  • Re-order blocks in the data $\boldsymbol{X}$, if necessary.

Do:
  • Set $t = t + 1$.
  • For $k = 1, \dots, K$, estimate $\hat{\boldsymbol{\beta}}_k^{(t)}$ blockwise, holding $\hat{\boldsymbol{\beta}}_l^{(t)}$ $(l < k)$ and $\hat{\boldsymbol{\beta}}_l^{(t-1)}$ $(l > k)$ fixed:
    1. Update $\hat{S}_k^{(t)}$, the estimated support for block $k$:
      (a) Set adaptive regularization:
(3.1)       $\lambda_{k,j}^{(t)} = \begin{cases} \eta_k^{(t)} & \text{if } \hat{\beta}_{k,j}^{(t-1)} \neq 0, \\ 2\eta_k^{(t)} & \text{otherwise,} \end{cases}$
          where
(3.2)       $\eta_k^{(t)} = \frac{\Lambda_{\max}(\widehat{\boldsymbol{\Theta}}^{(t-1)})}{\Lambda_{\max}(\boldsymbol{X}^T \boldsymbol{X})} \frac{1}{\sqrt{n}} \|\hat{\boldsymbol{\beta}}_k^{(t-1)}\|_2 \sqrt{\frac{\log(p_k)}{n} \|\hat{\boldsymbol{\beta}}_k^{(t-1)}\|_0}$
          and $\widehat{\boldsymbol{\Theta}}^{(t-1)} = \boldsymbol{X}^T \boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t-1)})\boldsymbol{X}$ is the estimated Fisher information matrix.
      (b) Perform stability selection:
         i. Take $B$ bootstrap samples: $\{\boldsymbol{y}^{*b}, \boldsymbol{X}^{*b}\}_{b=1}^{B}$.
        ii. Solve the randomized Lasso: For each $b = 1, \dots, B$,
(3.3)       $\hat{\boldsymbol{\beta}}_k^{(t)}(b) = \arg\min_{\alpha, \boldsymbol{\beta}} -\frac{1}{n} \ell(\boldsymbol{y}^{*b}; \alpha + \boldsymbol{X}_k^{*b}\boldsymbol{\beta} + \boldsymbol{\Phi}_k^{(t)}(b)) + \sum_{j=1}^{p_k} \gamma_j \lambda_{k,j}^{(t)} |\beta_j|$
            where $\gamma_j \overset{\text{IID}}{\sim} \mathcal{U}([0.5, 1.5])$ and $\boldsymbol{\Phi}_k^{(t)}(b) = \sum_{l<k} \boldsymbol{X}_l^{*b} \hat{\boldsymbol{\beta}}_l^{(t)} + \sum_{l>k} \boldsymbol{X}_l^{*b} \hat{\boldsymbol{\beta}}_l^{(t-1)}$.
       iii. Select features at stability level $\tau$:
(3.4)       $\hat{S}_k^{(t)} = \left\{ j : \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\hat{\beta}_{k,j}^{(t)}(b) \neq 0) \geq \tau \right\}.$
    2. Update $[\hat{\boldsymbol{\beta}}_k^{(t)}]_{\hat{S}_k^{(t)}}$, the estimated nonzero coefficients for block $k$:
(3.5)       $[\hat{\boldsymbol{\beta}}_k^{(t)}]_{\hat{S}_k^{(t)}} = \arg\min_{\alpha, \boldsymbol{\beta}} -\frac{1}{n} \ell(\boldsymbol{y}; \alpha + [\boldsymbol{X}_k]_{\hat{S}_k^{(t)}} \boldsymbol{\beta} + \boldsymbol{\Phi}_k^{(t)}) + \epsilon \|\boldsymbol{\beta}\|_2^2,$
       where $\boldsymbol{\Phi}_k^{(t)} = \sum_{l<k} \boldsymbol{X}_l \hat{\boldsymbol{\beta}}_l^{(t)} + \sum_{l>k} \boldsymbol{X}_l \hat{\boldsymbol{\beta}}_l^{(t-1)}$.

Until: $\text{Supp}(\hat{\boldsymbol{\beta}}^{(t)}) = \text{Supp}(\hat{\boldsymbol{\beta}}^{(t-1)})$, where $\text{Supp}(\cdot)$ denotes the signed support of a vector.
Output: $\hat{\boldsymbol{\beta}}_{\text{B-RAIL}} = [\, \hat{\beta}_1^{(t)} \cdots \hat{\beta}_K^{(t)} \,]$.
```

to p, we advise estimating the blocks with the smallest p first, especially if $p \leq n$. If dimensions of all the blocks are of similar sizes or much larger than n, we recommend starting with Gaussian blocks which tend to have better support recovery than non-Gaussian blocks.

3.2. *Estimating support* ($\hat{S}_k^{(t)}$). After initialization, we repeatedly iterate across the K data blocks and estimate the support of each block separately, holding the estimates of all other blocks fixed. This blockwise estimation avoids the ultrahigh-dimensionality issue, and, because shrinkage noise is mainly a problem in the ultrahigh-dimensional regime, the signal interference issue is also mitigated as a direct biproduct.

Furthermore, to effectively handle correlated features in practice, we incorporate stability selection with the randomized Lasso (Meinshausen and Bühlmann (2010)) to estimate each block. As given by step 1(b) in Algorithm 2, we solve the Lasso B times using the bootstrap and randomized penalty terms, and we threshold the stability score at τ (3.4) to select the most stable features. Though $\tau \in (0, 1)$ is a user-specified hyperparameter, the B-RAIL algorithm is insensitive to choices of τ within reasonable ranges. This insensitivity to τ has also been observed in previous work on stability selection (Meinshausen and Bühlmann (2010)). Ultimately, by utilizing randomized penalties and stability selection when estimating the support of each block, B-RAIL leverages the key property that the randomized Lasso is feature selection consistent, even when the Lasso's irrepresentable condition is violated (Meinshausen and Bühlmann (2010)), and hence can effectively handle correlated features.

3.3. *Adaptive regularization* (λ). Now, looking more closely at the penalty term in (3.3) of the randomized Lasso, the penalty term includes a random weight γ like the original randomized Lasso. However, in order to account for the scaling discrepancies, signal variability and domain differences between blocks, we introduce a block-specific adaptive penalty λ in (3.3) as well. For feature j in block k, we define the adaptive weight

$$\lambda_{k,j}^{(t)} = \begin{cases} \eta & \text{if } \hat{\beta}_{k,j}^{(t-1)} \neq 0, \\ 2\eta & \text{otherwise,} \end{cases} \tag{3.1}$$

where

$$\eta = \underbrace{\frac{\Lambda_{\max}(\widehat{\mathbf{\Theta}}^{(t-1)})}{\Lambda_{\max}(\mathbf{X}^T\mathbf{X})}}_{\substack{(a)\text{ domain}\\ \text{correction}}} \underbrace{\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{\beta}}_k^{(t-1)}\|_2}_{\substack{(b)\text{ signal}\\ \text{correction}}} \underbrace{\sqrt{\frac{\log(p_k)}{n}\|\hat{\boldsymbol{\beta}}_k^{(t-1)}\|_0}}_{\substack{(c)\text{ Lasso}\\ \text{penalty}}}. \tag{3.2}$$

Here, $\widehat{\mathbf{\Theta}}^{(t-1)}$ is the Fisher information matrix corresponding to the GLM of the response $\mathbf{y}$, and $\Lambda_{\max}$ denotes the maximum eigenvalue.

In this definition of λ, there are two moving parts. First, the multiplicative scheme in (3.1) encourages previously selected features to remain selected while still allowing all features to freely enter or exit the model. Second, η accounts for the heterogeneity of multiview data and helps to mitigate the challenges of Section 2.

Though the exact form of η was derived experimentally, η can be interpreted as the product of three factors, each of which is rooted in solid theoretical foundations. Namely, part (c) of (3.2) is closely related to the theoretical bound on the regularization parameter needed for selection consistency of the Lasso (Zhao and Yu (2006), Meinshausen and Bühlmann (2006)). The ratio of eigenvalues in part (a) (i.e., the domain correction term) is motivated by the theoretical conditions imposed on the Fisher information matrix for exponential family distributions (Yang et al. (2015)), and part (b) of (3.2) (i.e., the signal correction term) can be viewed as the average signal in block k since $\frac{1}{\sqrt{n}}$ is derived from the theoretical sparsity level within each block (Bunea, Tsybakov and Wegkamp (2007)).

By constructing the adaptive penalty η in this way, B-RAIL accounts for different block sizes through the $\frac{\log(p_k)}{n}$ term and automatically penalizes non-Gaussian blocks less heavily than Gaussian blocks since $\Lambda_{\max}(\hat{\Theta})$ is larger for Gaussian blocks. This helps to balance the inherently different beta-min conditions. In addition, because $\|\hat{\boldsymbol{\beta}}_k\|_2$ captures information about both the signal and scale of the kth block, η addresses the scaling differences and penalizes the stronger signal blocks more heavily to allow for the possibility of selection from weaker blocks.

While, in theory, this specific combination of weights should correct for scaling and domain-specific beta-min differences across views, we reinforce our choice of η through strong empirical results in Section 4. We also note that even if the form of η is slightly misspecified, stability selection is known to be fairly robust to the exact amount of regularization as long as the amount of regularization is within reason (Meinshausen and Bühlmann (2010)). Incorporating stability selection with the randomized Lasso thus serves as a built-in check within B-RAIL which is advantageous in practice.

3.4. *Coefficient estimations.* After estimating the blockwise support using the randomized Lasso with adaptive weights, we seek to estimate the coefficients of the support as accurately as possible since these values are used in future block estimations and iterations of B-RAIL. We hence refit a penalized regression model with a small ridge penalty (e.g., $\epsilon \approx 10^{-4}$) in (3.5) to avoid the known bias issues with the Lasso. The only reason to include the tiny ridge penalty is to ensure that we can still estimate coefficients when the selected support is greater than n. The exact choice of ϵ has a negligible impact in practice because it is chosen to be so small.

3.5. *Convergence.* We finally declare convergence of B-RAIL's iterative block estimation procedure when the estimated support remains unchanged. Our empirical analysis indicates that B-RAIL has quick support convergence, and we provide one example of this fast convergence in Figure 6 in the Supplementary Material (Baker, Tang and Allen (2020)). Using the ovarian cancer simulation (see Section 5) for three different responses (Gaussian, binary and Poisson), we report that the average number of iterations until convergence is between four and five with the maximum number of iterations reaching 15 (over 100 runs). These ranges are similar for all designs, empirically demonstrating B-RAIL's fast convergence.

Though convergence of the full-blown B-RAIL algorithm is currently limited to empirical analysis, B-RAIL can be viewed as a block coordinate descent algorithm which can be studied theoretically under some simplifications to gain additional insights into the full B-RAIL algorithm. Namely, if we omit the adaptive regularization parameter and apply the ordinary (GLM) Lasso in each block of the algorithm (as detailed in Algorithm 3 in the Supplementary Material (Baker, Tang and Allen (2020))), we call the resulting algorithm the blockwise (GLM) Lasso and discuss its convergence below.

PROPOSITION 1. *Consider the optimization problem*

$$\hat{\alpha}, \hat{\boldsymbol{\beta}} = \underset{\substack{\alpha_1,\dots,\alpha_K \in \mathbb{R}, \\ \boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_K \in \mathbb{R}^p}}{\arg\min} -\frac{1}{n}\sum_{k=1}^{K} \ell(\mathbf{y}; \alpha_k \mathbf{1}_n + \mathbf{X}_k \boldsymbol{\beta}_k) + \sum_{k=1}^{K}\sum_{j=1}^{p_k} \lambda_{k,j} |\boldsymbol{\beta}_{k,j}|, \tag{3.6}$$

where $\hat{\alpha} = [\hat{\alpha}_1, \dots, \hat{\alpha}_K]$, $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_K]$ *and* $\lambda_{k,j} \geq 0$, *and suppose that the objective function in* (3.6) *is bounded below. Then, the blockwise* (*GLM*) *Lasso converges to a global solution of*(3.6).

To prove Proposition 1, we leverage the block coordinate descent view of the blockwise Lasso and apply Theorem 4.1 from Tseng (2001) since the objective function is convex and separable with respect to each block $\boldsymbol{\beta}_k$. The detailed proof is provided in the Supplementary Material (Baker, Tang and Allen (2020)).

The two main differences between the blockwise Lasso and B-RAIL are the adaptive regularization parameter η and the use of stability selection with the randomized Lasso in each block's update. If the estimated support from stability selection converges to a common support across the many block iterations in B-RAIL, then B-RAIL reduces to the blockwise Lasso algorithm, for which we have shown convergence. While there is empirical evidence to believe that stability selection applied iteratively with the adaptive regularization parameter, as in B-RAIL, converges to a common support, proving this is theoretically challenging due to the purely algorithmic and random nature of stability selection. In addition, existing optimization-theoretic frameworks cannot handle adaptive parameters (η) that are dependent on previous iterates ($\hat{\boldsymbol{\beta}}^{(t-1)}$). Developing such a framework to handle both of these issues is beyond the scope of this work, but we plan to further investigate it in the future. For now, due to these serious difficulties, we rely on our empirical analysis to demonstrate B-RAIL's quick convergence.

3.6. *B-RAIL summary.* While we have introduced B-RAIL under the regression framework, B-RAIL can be naturally extended to estimate mixed graphical models via a penalized nodewise regression approach (Meinshausen and Bühlmann (2006)). As in the motivating example in Section 2, we can use B-RAIL to estimate the neighborhood of each node separately via penalized regressions and then combine the neighborhoods using an"AND" or "OR" rule to obtain the graph.

In either the regression or graph selection setting, our B-RAIL algorithm deliberately takes steps to exploit the practical advantages of existing Lasso-type methods while avoiding the drawbacks described in Section 2. For instance, by performing iterative block-by-block estimations, B-RAIL inherits the advantages of performing separate Lassos and avoids the issue of ultrahigh-dimensionality. This, in turn, reduces signal interference between blocks since shrinkage noise is only a concern when n is not sufficiently large relative to p. Furthermore, we mitigate the scaling and beta-min problems by engineering adaptive ℓ_1 penalties in B-RAIL to correct for domain and signal differences between blocks. In this construction, slightly weaker non-Gaussian blocks are penalized less heavily and thus not completely overshadowed by Gaussian blocks. Still, selecting an appropriate amount of ℓ_1 regularization is challenging in practice, especially due to highly correlated data. B-RAIL thus incorporates randomized stability selection, which is known to be feature selection consistent under stronger and more complex dependencies than can be handled by the Lasso. This boosts the support estimation of correlated features, and together, with the previous components, B-RAIL effectively overcomes the many practical challenges of multiview feature selection and lends itself to a plethora of data integration applications.

4. Numerical studies. We next reinforce the theoretically-guided choices in our B-RAIL construction and demonstrate its effectiveness through extensive simulations. In these simulations we evaluate B-RAIL against four common Lasso-based parametric methods: (i) Lasso with a global penalty for all blocks, (ii) Lasso with separate penalties for each block, (iii) separate Lassos for each block and (iv) Adaptive Lasso. For the Adaptive Lasso we use ridge weights, as they are better adapted to handle correlated features. Moreover, to avoid biases from penalty selection methods, we use oracle information to select features in the Lasso-based models. That is, if k is the number of true features in the simulation, we fit the full path of the Lasso and select the first k features. In the case of the Lasso with separate penalties, we select the k features with the largest number of true positives. We do not, however, use oracle information for B-RAIL. Instead, B-RAIL internally selects the number of features using stability selection with the threshold $\tau = 0.8$, as outlined in Algorithm 2, and we set $\epsilon = 0.001/p$.

To systematically compare these methods, we simulate from various designs of $\mathbf{X}$ with three blocks—namely, a Gaussian $\mathbf{X}_1$, Bernoulli $\mathbf{X}_2$ and Poisson $\mathbf{X}_3$ block—and various types of GLM responses $\mathbf{y}$. Due to the popular use of the Gaussian, Bernoulli and Poisson GLMs, we run simulations with responses $\mathbf{y}$ from each of these families. For the Gaussian response we fit the linear model $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $\epsilon \sim N(0, 1)$. For the binary and Poisson responses, we use copula transformations (Nelsen (1999)) to simulate $\mathbf{y}$.

In addition to these response models for $\mathbf{y}$, we consider four simulation designs for $\mathbf{X}$ to understand model behavior under different assumptions. The four simulations designs are: (i) i.i.d. features, (ii) independent features with nonconstant variance, (iii) correlated features with covariance structure from a Block Directed Markov Random Field and (iv) a real data-inspired simulation with features from The Cancer Genome Atlas (TCGA) ovarian cancer study. We elaborate on each of these designs below.

Note, in all of the simulations we set the number of true features in each covariate block to 10, and the magnitudes of the true features are drawn from Unif(4, 10) with random sign assignment. However, for the Gaussian block in the non-i.i.d. simulations below, we artificially lowered the SNR since we know that recovering continuous features is easier than recovering noncontinuous features (see Figure 3). Unless stated otherwise, we simulate $n = 200$ samples and $p_1 = p_2 = p_3 = 300$ features. We also center and scale the design matrix $\mathbf{X}$ before estimation.

I.i.d. design. For each of the three covariate blocks, we simulate $n = 200$ samples from i.i.d. features. Here, $p_1 = p_2 = p_3 = 300$ for the high-dimensional design and $p_1 = p_2 = p_3 = 100$ for the low-dimensional design.

Heteroscedasticity design. In this design we assume that the features are independent but have nonconstant variance. For the Gaussian block the entries in each column are simulated from the normal distribution $N(0, \sigma^2)$, where $\sigma \sim \Gamma(3, 0.6)$. In the Bernoulli block each column is simulated independently with entries drawn from Bern(p) with $p \sim$ Unif(0.2, 0.8). Similarly, in the Poisson block the mean λ of each column is drawn from the Gamma distribution $\Gamma(4, 0.6)$ (using the shape/scale parameterization).

Block directed graph design. We next drop the independence assumption and use a Block Directed Markov Random Field (BDMRF) (Yang et al. (2014a)) graph to simulate correlated features. In this case, $\mathbf{X}$ is simulated via Gibbs sampling with the partial ordering of the underlying mixed graph given by $P[X_1, X_2, X_3] = P[X_1|X_2, X_3]P[X_2|X_3]P[X_3]$, where $P[X_1|X_2, X_3]$ is a pairwise Gaussian conditional random field (CRF), $P[X_2|X_3]$ is a pairwise Ising CRF (Ravikumar, Wainwright and Lafferty (2010)) and $P[X_3]$ is a pairwise Poisson Markov Random Field (MRF) (Yang et al. (2013, 2012)). We set high correlations for the Gaussian and Poisson blocks and low correlations for the binary block and between block structure.

Ovarian cancer inspired simulation design. In an attempt to simulate data closest to real-world scenarios, we take the continuous-valued miRNA data, proportion-valued methylation data and the count-valued RNASeq data from The Cancer Genome Atlas (TCGA) ovarian cancer database (The Cancer Genome Atlas Research Network (2011)) to be our covariates. After merging and preprocessing the TCGA ovarian cancer data (refer to Section 5 for details), we arrive at $n = 293$ samples and $p_{\text{RNASeq}} = 408$, $p_{\text{miRNA}} = 307$ and $p_{\text{Methyl}} = 301$ features.

Under each of these simulation scenarios, we evaluate the performance of B-RAIL and the oracle Lasso-type methods by reporting the true positive rate (TPR) and false discovery proportion (FDP) for overall feature recovery and individual block recoveries. Due to the large number of features, we use FDP, defined as the number of false positives divided by total the number of recovered nonzero features, instead of the false discovery rate.

We summarize the results of our simulations with Gaussian responses in Table 1 and those with binary and Poisson responses in Table 2. Note that, for the binary and Poisson responses, we show the block directed graph results here and provide the other simulation results in the Supplementary Material (Baker, Tang and Allen (2020)). We also highlight in bold the TPR/FDP combination with the highest TPR*(1-FDP) value for overall recovery. In almost all scenarios the results in Table 1 and Table 2 indicate that B-RAIL (with no oracle information) is able to achieve a higher TPR and lower FDP than its competitive Lasso-type methods with oracle information.

When oracle information is unavailable, model selection techniques can introduce additional errors and further complicate feature selection. Table 3 shows one such case and compares the block directed graph simulation performance of B-RAIL against the Lasso-type methods using five-fold cross-validation, extended BIC and stability selection to select the penalty parameters. We also include the oracle estimators for the same set of simulations to emphasize the large decrease in performance when the Lasso-type methods do not have oracle information. These simulations indicate that cross-validation tends to overselect the number of features in the model while extended BIC underselects, and stability selection performs the best but pales in comparison to oracle selection. In contrast, B-RAIL, when initialized to an over-selection using the prespecified sparsity level of $0.2p_k$, outperforms the Lasso-type methods even when oracle selection is used for these competitive methods. Additional simulations, confirming the strong empirical performance of B-RAIL, are provided in the Supplementary Material (Baker, Tang and Allen (2020)).

5. Case study: Integrative genomics of ovarian cancer.

One promising practical application for our research on multiview feature selection lies in integrative cancer genomics. Here, scientists seek to integrate data from multiple sources of high-throughput genomic data to more holistically model the genomic systems in cancer cells, leading to a better understanding of disease mechanisms and possible therapies.

In this case study we seek to integrate three different types of genomic data to study how epigenetics and short RNAs influence the gene regulatory system in ovarian cancer. Specifically, we are interested in discovering miRNAs and CpG sites which affect the gene expression of well-known oncogenes in ovarian cancer and hence can serve as potential drug targets for blocking or decreasing the expression of these oncogenes. Driven by this goal of discovering potential drug targets, we use our proposed B-RAIL method to estimate the integrative ovarian cancer gene regulatory network with the specific intention of identifying miRNAs and CpG sites that are directly linked to known oncogenes of ovarian cancer.

In this investigation we integrate the following three data sets: (1) count-valued gene expression measured via RNASeq, (2) continuous (Gaussian) miRNA expression and (3) proportion-valued DNA methylation data from The Cancer Genome Atlas (TCGA) ovarian cancer study which is publicly available (The Cancer Genome Atlas Research Network (2011)). The TCGA data originally contained 19,990 genes, 27,578 CpG sites and 799 miRNAs but only $n = 293$ common patients across all three data sets of interest. We hence reduced the number of features to manageable sizes by first filtering features according to their association with several important clinical outcomes—*survival* via a univariate cox model, *chemo-resistance* via a univariate logistic model and *recurrence* via a univariate logistic model. In addition, we transformed the RNASeq data using the Kolmogorov–Smirnov test

TABLE 1

We compare various selection methods under five different simulation scenarios. For each scenario we simulate **X** *with three blocks (continuous, binary, counts) and a Gaussian response* **y**. *We report the true positive rate (TPR) and false discovery proportion (FDP) for overall feature recovery and individual block recoveries, averaged across* 200 *runs with standard errors in parentheses. We bold the best overall TPR* $*$ $(1 - \mathit{FDP})$ *values for each simulation scenario. Note that we used oracle information for the Lasso-type methods*

	Total		Continuous		Binary		Counts	
	TPR	FDP	TPR	FDP	TPR	FDP	TPR	FDP
i.i.d. Case, $p = 300$								
B-RAIL	**1.00 (1.1e−3)**	**0.00 (0.0e−0)**	1.00 (1.7e−3)	0.00 (0.0e−0)	1.00 (1.4e−3)	0.00 (0.0e−0)	1.00 (1.4e−3)	0.00 (0.0 0)
Lasso-λ (oracle)	0.87 (1.2e−3)	0.12 (1.9e−3)	0.90 (1.4e−3)	0.12 (5.3e−3)	0.81 (2.9e−3)	0.20 (5.2e−4)	0.90 (0.0 0)	0.03 (4.5e−3)
Lasso-λ_k (oracle)	0.92 (1.6e−3)	0.08 (1.6e−3)	0.96 (4.8e−3)	0.00 (0.0e−0)	0.90 (0.0e−0)	0.15 (4.5e−3)	0.90 (0.0e−0)	0.08 (4.4e−3)
Separate Lasso (oracle)	0.75 (1.6e−3)	0.24 (4.9e−4)	0.80 (0.0e−0)	0.20 (0.0e−0)	0.70 (1.7e−3)	0.30 (5.7e−4)	0.77 (4.8e−3)	0.21 (1.4e−3)
Adaptive Lasso (oracle)	**1.00 (0.0e−0)**	**0.00 (0.0e−0)**	1.00 (0.0e−0)	0.00 (0.0e−0)	1.00 (0.0e−0)	0.00 (0.0e−0)	1.00 (0.0e−0)	0.00 (0.0e−0)
i.i.d. Case, $p = 900$								
B-RAIL	**0.94 (7.3e−3)**	**0.12 (1.4e−2)**	0.98 (5.0e−3)	0.14 (1.7e−2)	0.95 (9.9e−3)	0.08 (1.1e−2)	0.90 (9.2e−3)	0.12 (1.5e−2)
Lasso-λ (oracle)	0.63 (3.3e−4)	0.37 (5.0e−4)	0.80 (0.0e−0)	0.20 (0.0e−0)	0.50 (1.0e−3)	0.50 (1.4e−3)	0.60 (0.0e−0)	0.40 (0.0e−0)
Lasso-λ_k (oracle)	0.74 (1.7e−3)	0.26 (1.7e−3)	0.98 (4.4e−3)	0.19 (3.3e−3)	0.63 (7.4e−3)	0.36 (6.5e−3)	0.62 (3.9e−3)	0.20 (7.4e−3)
Separate Lasso (oracle)	0.53 (9.6e−4)	0.46 (1.3e−3)	0.60 (0.0e−0)	0.38 (4.4e−3)	0.49 (2.9e−3)	0.51 (1.6e−3)	0.50 (0.0e−0)	0.50 (0.0e−0)
Adaptive Lasso (oracle)	0.66 (9.6e−4)	0.34 (6.9e−4)	0.79 (3.1e−3)	0.28 (1.3e−3)	0.50 (3.5e−3)	0.44 (1.7e−3)	0.70 (0.0e−0)	0.30 (9.0e−4)
Nonconstant Variance (Heteroscedasticity)								
B-RAIL	**0.97 (3.3e−4)**	**0.01 (1.3e−3)**	0.90 (1.0e−3)	0.00 (0.0e−0)	1.00 (0.0e−0)	0.00 (0.0e−0)	1.00 (0.0e−0)	0.02 (3.7e−3)
Lasso-λ (oracle)	0.77 (2.2e−3)	0.21 (2.2e−3)	0.88 (3.9e−3)	0.19 (2.9e−3)	0.83 (4.7e−3)	0.06 (5.4e−3)	0.60 (0.0e−0)	0.37 (3.6e−3)
Lasso-λ_k (oracle)	0.83 (0.0e−0)	0.17 (0.0e−0)	0.90 (0.0e−0)	0.11 (2.8e−3)	1.00 (0.0e−0)	0.18 (2.4e−3)	0.60 (0.0e−0)	0.22 (4.9e−3)
Separate Lasso (oracle)	0.56 (1.4e−3)	0.43 (1.1e−3)	0.58 (4.3e−3)	0.41 (1.9e−3)	0.80 (0.0e−0)	0.19 (2.3e−3)	0.30 (0.0e−0)	0.70 (1.4e−3)
Adaptive Lasso (oracle)	0.86 (1.3e−3)	0.13 (1.2e−3)	0.90 (1.0e−3)	0.10 (1.8e−3)	1.00 (0.0e−0)	0.11 (3.4e−3)	0.69 (3.4e−3)	0.20 (5.6e−3)

TABLE 1
(*Continued*)

	Total		Continuous		Binary		Counts	
	TPR	FDP	TPR	FDP	TPR	FDP	TPR	FDP
				Block Directed Graph Structure				
B-RAIL	**0.88 (4.8e−3)**	**0.16 (1.1e−2)**	0.79 (4.8e−3)	0.12 (1.4e−2)	0.96 (6.8e−3)	0.12 (7.0e−3)	0.89 (5.2e−3)	0.22 (1.4e−2)
Lasso-λ (oracle)	0.72 (1.9e−3)	0.27 (1.9e−3)	0.88 (5.8e−3)	0.17 (3.2e−3)	1.00 (0.0e−0)	0.28 (3.5e−3)	0.30 (0.0e−0)	0.42 (4.2e−3)
Lasso-λ_k (oracle)	0.77 (0.0e−0)	0.23 (0.0e−0)	1.00 (0.0e−0)	0.20 (4.0e−3)	1.00 (0.0e−0)	0.24 (5.4e−3)	0.30 (0.0e−0)	0.26 (1.5e−2)
Separate Lasso (oracle)	0.51 (2.0e−3)	0.46 (1.9e−3)	0.79 (3.1e−3)	0.18 (4.6e−3)	0.55 (5.0e−3)	0.42 (2.8e−3)	0.20 (0.0e−0)	0.79 (1.3e−3)
Adaptive Lasso (oracle)	0.70 (0.0e−0)	0.30 (3.4e−4)	0.80 (0.0e−0)	0.20 (0.0e−0)	1.00 (0.0e−0)	0.29 (1.3e−3)	0.30 (0.0e−0)	0.49 (3.0e−3)
				OV Data				
B-RAIL	**0.95 (2.9e−3)**	**0.11 (4.7e−3)**	0.85 (8.7e−3)	0.23 (4.9e−3)	1.00 (0.0e−0)	0.09 (8.2e−3)	1.00 (1.0e−3)	0.03 (4.4e−3)
Lasso-λ (oracle)	0.57 (8.5e−4)	0.43 (1.2e−3)	0.80 (0.0e−0)	0.38 (0.0e−0)	0.41 (2.6e−3)	0.31 (6.2e−3)	0.50 (0.0e−0)	0.54 (1.5e−3)
Lasso-λ_k (oracle)	0.65 (1.6e−3)	0.35 (1.6e−3)	0.80 (1.0e−3)	0.37 (2.3e−3)	0.51 (3.6e−3)	0.00 (1.7e−3)	0.63 (4.6e−3)	0.48 (4.6e−3)
Separate Lasso (oracle)	0.40 (1.6e−3)	0.60 (1.3e−3)	0.58 (4.1e−3)	0.41 (1.9e−3)	0.30 (0.0e−0)	0.70 (0.0e−0)	0.30 (2.0e−3)	0.69 (2.3e−3)
Adaptive Lasso (oracle)	0.93 (1.2e−3)	0.09 (1.1e−3)	0.80 (0.0e−0)	0.12 (2.8e−3)	0.98 (3.6e−3)	0.00 (1.0e−3)	1.00 (0.0e−0)	0.09 (1.7e−3)

TABLE 2

We compare various selection methods under the block directed graph simulation design with binary responses and with Poisson responses. We report the TPR and FDP for feature recovery, averaged across 200 *runs with standard errors in parentheses. We bold the best overall TPR* $*$ $(1 - FDP)$ *values for each simulation scenario*

	Total		Continuous		Binary		Counts	
	TPR	FDP	TPR	FDP	TPR	FDP	TPR	FDP
				Binary Response Block Directed Graph Structure				
B-RAIL	**0.81 (8.6e−3)**	**0.07 (5.2e−3)**	0.80 (0.0e−0)	0.04 (5.4e−3)	0.97 (5.7e−3)	0.09 (6.2e−3)	0.68 (2.1e−2)	0.07 (1.2e−2)
Lasso-λ (oracle)	0.70 (0.0e−0)	0.29 (1.3e−3)	0.90 (0.0e−0)	0.23 (4.1e−3)	0.90 (0.0e−0)	0.37 (2.0e−3)	0.30 (0.0e−0)	0.18 (1.3e−2)
Lasso-λ (oracle)	0.70 (8.5e−4)	0.30 (8.5e−4)	0.89 (2.6e−3)	0.21 (4.3e−3)	0.90 (0.0e−0)	0.27 (3.6e−3)	0.30 (0.0e−0)	0.52 (7.4e−3)
Separate Lasso (oracle)	0.50 (1.3e−3)	0.46 (2.3e−3)	0.80 (0.0e−0)	0.20 (0.0e−0)	0.51 (3.9e−3)	0.41 (7.0e−3)	0.20 (0.0e−0)	0.79 (1.7e−3)
Adaptive Lasso (oracle)	0.70 (9.6e−4)	0.29 (1.4e−3)	0.81 (2.9e−3)	0.20 (1.3e−3)	1.00 (0.0e−0)	0.36 (2.5e−3)	0.30 (0.0e−0)	0.27 (5.7e−3)
				Poisson Response Block Directed Graph Structure				
B-RAIL	**0.81 (8.6e−3)**	**0.07 (5.2e−3)**	0.80 (0.0e−0)	0.04 (5.4e−3)	0.97 (5.7e−3)	0.09 (6.2e−3)	0.68 (2.1e−2)	0.07 (1.2e−2)
Lasso-λ (oracle)	0.70 (0.0e−0)	0.29 (1.3e−3)	0.90 (0.0e−0)	0.23 (4.1e−3)	0.90 (0.0e−0)	0.37 (2.0e−3)	0.30 (0.0e−0)	0.18 (1.3e−2)
Lasso-λ (oracle)	0.70 (8.5e−4)	0.30 (8.5e−4)	0.89 (2.6e−3)	0.21 (4.3e−3)	0.90 (0.0e−0)	0.27 (3.6e−3)	0.30 (0.0e−0)	0.52 (7.4e−3)
Separate Lasso (oracle)	0.50 (1.3e−3)	0.46 (2.3e−3)	0.80 (0.0e−0)	0.20 (0.0e−0)	0.51 (3.9e−3)	0.41 (7.0e−3)	0.20 (0.0e−0)	0.79 (1.7e−3)
Adaptive Lasso (oracle)	0.70 (9.6e−4)	0.29 (1.4e−3)	0.81 (2.9e−3)	0.20 (1.3e−3)	1.00 (0.0e−0)	0.36 (2.5e−3)	0.30 (0.0e−0)	0.27 (5.7e−3)

TABLE 3

We compare feature recovery for B-RAIL and Lasso-type methods with various model selection methods. Here, we simulate from the block directed graph simulation design with Gaussian responses and report the TPR and FDP, averaged across 200 *runs with standard errors in parentheses. We highlight the best overall TPR* $*$ $(1 - \mathit{FDP})$ *values in bold. Note, for stability selection, we initialize* λ *using the* λ *selected by CV*

	Total		Continuous		Binary		Counts	
	TPR	FDP	TPR	FDP	TPR	FDP	TPR	FDP
Block Directed Graph Structure								
B-RAIL	**0.86 (1.2e−2)**	**0.20 (2.8e−2)**	0.78 (9.2e−3)	0.15 (3.6e−2)	0.93 (1.9e−2)	0.16 (1.8e−2)	0.88 (1.2e−2)	0.29 (4.3e−2)
Lasso-λ (oracle)	0.72 (3.5e−3)	0.27 (4.8e−3)	0.87 (1.1e−2)	0.20 (5.0e−3)	1.00 (0.0e−0)	0.40 (1.6e−2)	0.30 (0.0e−0)	0.21 (5.0e−3)
Lasso-λ_k (oracle)	0.77 (0.0e−0)	0.23 (0.0e−0)	1.00 (0.0e−0)	0.24 (1.4e−2)	1.00 (0.0e−0)	0.32 (2.7e−2)	0.30 (0.0e−0)	0.14 (2.2e−2)
Separate Lasso (oracle)	0.51 (5.0e−3)	0.44 (5.6e−3)	0.79 (8.2e−3)	0.18 (1.2e−2)	0.55 (1.1e−2)	0.41 (5.0e−3)	0.20 (0.0e−0)	0.73 (1.1e−2)
Adaptive Lasso (oracle)	0.70 (0.0e−0)	0.30 (2.3e−3)	0.80 (0.0e−0)	0.20 (0.0e−0)	1.00 (0.0e−0)	0.42 (9.2e−3)	0.30 (0.0e−0)	0.27 (1.1e−2)
Five-Fold Cross-Validation								
Lasso-λ	0.98 (3.0e−3)	0.62 (4.1e−3)	1.00 (0.0e−0)	0.58 (6.3e−3)	1.00 (0.0e−0)	0.63 (4.0e−3)	0.95 (8.9e−3)	0.63 (3.9e−3)
Lasso-λ_k	0.60 (2.3e−3)	0.17 (2.2e−3)	0.80 (0.0e−0)	0.13 (3.6e−3)	0.69 (6.8e−3)	0.07 (6.3e−3)	0.30 (0.0e−0)	0.40 (1.0e−3)
Separate Lasso	0.37 (8.1e−3)	0.28 (1.5e−2)	0.58 (9.9e−3)	0.05 (9.9e−3)	0.54 (1.8e−2)	0.38 (2.0e−2)	0.00 (0.0e−0)	0.47 (5.0e−2)
Adaptive Lasso	0.73 (3.0e−3)	0.37 (7.8e−3)	0.87 (6.8e−3)	0.23 (6.4e−3)	1.00 (2.0e−3)	0.39 (8.1e−3)	0.31 (2.9e−3)	0.55 (9.5e−3)
Extended BIC								
Lasso-λ	0.03 (0.0e−0)	0.00 (0.0e−0)	0.10 (0.0e−0)	0.00 (0.0e−0)	0.00 (0.0e−0)	0.00 (0.0e−0)	0.00 (0.0e−0)	0.00 (0.0e−0)
Lasso-λ_k	0.58 (2.0e−3)	0.18 (3.1e−3)	0.80 (0.0e−0)	0.15 (5.3e−3)	0.64 (5.9e−3)	0.04 (6.1e−3)	0.30 (0.0e−0)	0.42 (4.0e−3)
Separate Lasso	0.18 (1.7e−3)	0.07 (7.9e−3)	0.50 (1.0e−3)	0.00 (0.0e−0)	0.04 (4.9e−3)	0.00 (0.0e−0)	0.00 (0.0e−0)	0.47 (5.0e−2)
Adaptive Lasso	0.30 (0.0e−0)	0.00 (0.0e−0)	0.50 (0.0e−0)	0.00 (0.0e−0)	0.40 (0.0e−0)	0.00 (0.0e−0)	0.00 (0.0e−0)	0.00 (0.0e−0)
Stability Selection								
Lasso -λ	0.67 (2.4e−3)	0.01 (1.8e−3)	0.80 (0.0e−0)	0.02 (4.2e−3)	0.92 (6.9e−3)	0.00 (9.1e−4)	0.29 (2.4e−3)	0.00 (0.0e−0)
Lasso-λ_k	0.58 (2.7e−3)	0.04 (2.3e−3)	0.80 (0.0e−0)	0.09 (4.7e−3)	0.66 (6.5e−3)	0.00 (0.0e−0)	0.28 (4.3e−3)	0.00 (2.5e−3)
Separate Lasso	0.42 (3.9e−3)	0.28 (6.6e−3)	0.69 (2.7e−3)	0.13 (7.2e−3)	0.55 (1.1e−2)	0.39 (8.8e−3)	0.00 (0.0e−0)	0.29 (4.6e−2)
Adaptive Lasso	0.65 (1.7e−3)	0.08 (2.6e−3)	0.80 (0.0e−0)	0.11 (0.0e−0)	0.84 (5.1e−3)	0.07 (6.0e−3)	0.30 (0.0e−0)	0.01 (4.3e−3)

($\alpha = 0.262$) to alleviate the problem of very large counts (up to 20,000). This preprocessing yielded $p_1 = 408$ genes, $p_2 = 301$ CpG sites and $p_3 = 307$ miRNAs in the RNASeq, methylation and miRNA data sets, respectively. Lastly, per the recommendation of scientists we included 20 additional highly mutated genes that were experimentally identified as important in ovarian cancer, resulting in $p_1 = 428$ genes in the RNASeq data set.

To estimate the integrated ovarian cancer network, we fit a Block Directed Markov Random Field (BDMRF) model (Yang et al. (2014a)) using B-RAIL to estimate the neighborhood of each node in the graph. Note that since miRNAs and methylation are both gene regulatory mechanisms, miRNAs and methylation can affect expression levels (measured via RNASeq), but the converse is not possible. To agree with this known physical mechanism, we set the partial ordering of the mixed graph underlying BDMRF as $P[X_1, X_2, X_3] = P[X_1|X_2, X_3]P[X_2]P[X_3]$, where $P[X_2]$ is a pairwise Ising MRF for the proportion-valued methylation data, $P[X_3]$ is a pairwise Gaussian MRF for the continuous miRNA data and $P[X_1|X_2, X_3]$ is a pairwise Poisson CRF for the count-valued RNASeq data. However, we recall that only negative conditional dependencies are permitted in the Poisson MRF and CRF models. Since this constraint is unrealistic for genomics data, we fit a sublinear Poisson CRF, in lieu of the usual Poisson CRF, to allow for both positive and negative conditional dependencies (Yang et al. (2013)). Under this specified BDMRF model, we employ nodewise neighborhood selection (Meinshausen and Bühlmann (2006), Yang et al. (2015)) using B-RAIL to learn the edge structure of the integrated network.

Our overall BRAIL-estimated network is presented in Figure 4, and in Figure 5 we more closely examine the relationships between the oncogenes, miRNAs and CpG sites by zooming in on the subnetworks for the well-known oncogene, BRCA1, and its direct neighbor, miRNA23b. Both BRCA1 and miRNA23b are well-known biomarkers and have been implicated in several ovarian cancer studies (Antoniou et al. (2003), BRCA (1994), Geng et al. (2012), King et al. (2003), Li et al. (2014), Yan et al. (2016)). Moreover, miRNA23b is

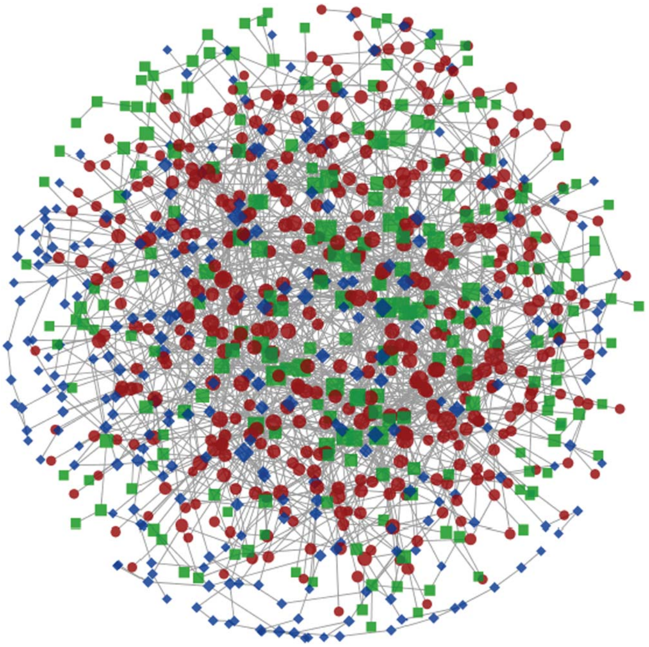
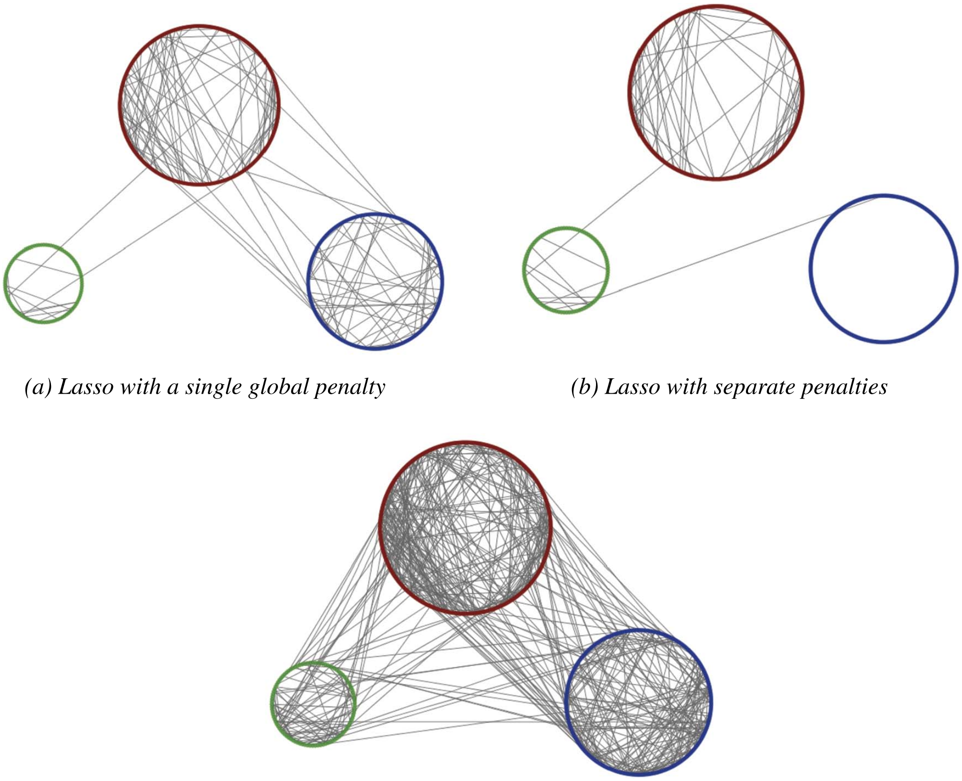


FIG. 4. *We present the integrated ovarian cancer genetic network estimated by the B-RAIL algorithm. The diamond nodes denote miRNAs, square nodes denote CpG sites, circle nodes denote gene expression via RNASeq and the size of each node is proportional to the number of connected first neighbors.*



(a) Lasso with a single global penalty

(b) Lasso with separate penalties

(c) The B-RAIL algorithm

FIG. 5. *We zoom in on the subnetworks for two known biomarkers which have been previously implicated in ovarian cancer studies. Key mutated cancer biomarkers, such as miR23b and BRCA*1, *are found to have many interconnections to biomarkers, which are circled, that are consistent with the cancer literature* (*Buchholtz et al.* (2014), *Freier* (2016), *Gao et al.* (2009), *Giannakakis et al.* (2008), *Obermayr et al.* (2010), *Tong et al.* (2017), *Toyama et al.* (1999)).

known to play a key role in p53 signaling (via TP53) (Boren et al. (2009)), agreeing with the estimated edge between the TP53 oncogene and miRNA23b in Figure 5(b).

Aside from this link, however, the estimated edges between genes, CpG sites and miRNAs in Figure 4 are largely unexplored and unknown by researchers since B-RAIL is one of the first practical approaches for multiview feature selection. Nevertheless, we can partially validate our B-RAIL-estimated network by highlighting the many genes with verified connections in the ovarian cancer and cancer proliferation/suppression literatures. In Figure 5 we circle this collection of implicated genes which includes LDOC1, SGCB and miRNA210 (Buchholtz et al. (2014), Obermayr et al. (2010), Giannakakis et al. (2008)).

As we have noted, there is substantial evidence in the scientific literature, suggesting that our proposed B-RAIL algorithm successfully identified promising candidates as well as known biomarkers involved in ovarian cancer. By taking into account the relationships between genes, miRNAs and CpG sites, our integrative analysis via B-RAIL leads to valuable insights beyond a single biomarker type and to novel discoveries of direct connections between miRNAs, CpG sites and known oncogenes which may aid the development of targeted drug therapies for ovarian cancer. This is the first integrative analysis of its kind, and future experiments studying the connections between known ovarian cancer oncogenes and candidate miRNAs and CpG sites would be of great value to validate our findings.

6. Discussion. Though we have primarily focused on applications to integrative genomics in this work, B-RAIL is not limited to this context. B-RAIL can be applied to any field that yields high-dimensional multiview data, and, with the rapid advances in technologies, we expect B-RAIL to have a growing and far-reaching impact in fields such as imaging genetics, national security, climate studies, spatial statistics, Internet data, marketing and economics. B-RAIL is also a versatile tool that can be used to for any sparse regression or graph selection problem in this multiview context.

In addition to developing an effective data integration tool for multiview feature selection, our work addresses the many difficulties of performing multiview feature selection in practice. These practical challenges were severely understudied prior to this work, but we partially resolve this gap, identifying four root challenges which interact with one another to impede recovery. Throughout our investigation of these practical challenges, we provide strong empirical evidence of the existence as well as the adverse consequences of such challenges. However, the theoretical underpinnings of these issues are still unknown. Understanding exactly how challenges such as shrinkage noise and the beta-min condition are influenced by varying domains and signals would be of great benefit to the field of data integration as a whole. We also highlight that, while the Lasso has been well studied under Gaussianity and idealized assumptions, the increasing abundance of correlated non-Gaussian data in multiview settings requires a greater push for theoretical studies on feature selection with heterogeneous data and the GLM Lasso.

Overall, we have demonstrated many challenges posed by multiview feature selection, and, in our investigation of these challenges, we opened up new avenues for future theoretical work to rigorously understand how the heterogeneity of multiview data complicates feature selection. Driven by these challenges and the ineffectiveness of existing methods, we developed a practical solution to overcome the current challenges. Our method, B-RAIL, is one of the first practical tools for multiview feature selection and is grounded in deep theoretical foundations. With its versatility and strong empirical performance, B-RAIL facilitates impactful integrative analyses across a broad spectrum of fields.

Acknowledgments. Y. Baker acknowledges support from the NIH/NCI T32 CA096520 training program in Biostatistics for Cancer Research, grant number 5T32CA09652011. T. Tang acknowledges support from the NSF Graduate Research Fellowship Program DGE-1752814 and ONR N00014-16-1-2664. G. Allen acknowledges support from NSF DMS-1554821, NSF NeuroNex-1707400 and NSF DMS-1264058. We thank Zhandong Liu, Ying-Wooi Wan and Matthew L. Anderson at Baylor College of Medicine for thoughtful discussions related to this work. We also thank the Editor and reviewers for their constructive feedback which helped to improve the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Feature selection for data integration with mixed multiview data" (DOI: 10.1214/20-AOAS1389SUPP; .pdf). We provide additional plots to further support the strong empirical performance of B-RAIL and its quick convergence. We also provide the proof of Proposition 1.

REFERENCES

ACAR, E., KOLDA, T. G. and DUNLAVY, D. M. (2011). All-at-once optimization for coupled matrix and tensor factorizations. ArXiv Preprint. Available at arXiv:1105.3422.

ALLEN, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16** 125–127. MR0343481 https://doi.org/10.2307/1267500

ANTONIOU, A., PHAROAH, P., NAROD, S., RISCH, H. A., EYFJORD, J. E., HOPPER, J., LOMAN, N., OLSSON, H., JOHANNSSON, O. et al. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. *Am. J. Hum. Genet.* **72** 1117–1130.

BAKER, Y., TANG, T. M and ALLEN, G. I (2020). Supplement to "Feature Selection for Data Integration with Mixed Multiview Data." https://doi.org/10.1214/20-AOAS1389SUPP

BOREN, T., XIONG, Y., HAKAM, A., WENHAM, R., APTE, S., CHAN, G., KAMATH, S. G., CHEN, D.-T., DRESSMAN, H. et al. (2009). MicroRNAs and their target messenger RNAs associated with ovarian cancer response to chemotherapy. *Gynecol. Oncol.* **113** 249–255.

BRCA, S. G. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266** 7.

BUCHHOLTZ, M.-L., BRÜNING, A., MYLONAS, I. and JÜCKSTOCK, J. (2014). Epigenetic silencing of the LDOC1 tumor suppressor gene in ovarian cancer cells. *Arch. Gynecol. Obstet.* **290** 149–154.

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549 https://doi.org/10.3150/12-BEJSP11

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149 https://doi.org/10.1214/07-EJS008

CHEN, J. and CHEN, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statist. Sinica* **22** 555–574. MR2954352 https://doi.org/10.5705/ss.2010.216

CHEN, S., WITTEN, D. M. and SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64. MR3335095 https://doi.org/10.1093/biomet/asu051

CHENG, J., LI, T., LEVINA, E. and ZHU, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. MR3640193 https://doi.org/10.1080/10618600.2016.1237362

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FREIER, C. (2016). Role of regulatory T cells and associated chemokines in human gynecological tumors Ph.D. thesis LMU.

GAO, J.-Q., TSUDA, Y., HAN, M., XU, D.-H., KANAGAWA, N., HATANAKA, Y., TANI, Y., MIZUGUCHI, H., TSUTSUMI, Y. et al. (2009). NK cells are migrated and indispensable in the anti-tumor activity induced by CCL27 gene therapy. *Cancer Immunology, Immunotherapy* **58** 291.

GENG, J., LUO, H., PU, Y., ZHOU, Z., WU, X., XU, W. and YANG, Z. (2012). Methylation mediated silencing of miR-23b expression and its role in glioma stem cells. *Neurosci. Lett.* **528** 185–189.

GIANNAKAKIS, A., SANDALTZOPOULOS, R., GRESHOCK, J., LIANG, S., HUANG, J., HASEGAWA, K., LI, C., O'BRIEN-JENKINS, A., KATSAROS, D. et al. (2008). miR-210 links hypoxia with cell cycle regulation and is deleted in human epithelial ovarian cancer. *Cancer Biol. Ther.* **7** 255–264.

HALL, D. L. and LLINAS, J. (1997). An introduction to multisensor data fusion. *Proc. IEEE* **85** 6–23.

HASLBECK, J. and WALDORP, L. J. (2015). mgm: Structure estimation for time-varying mixed graphical models in high-dimensional data. ArXiv Preprint. Available at arXiv:1510.06871.

JALALI, A., RAVIKUMAR, P., VASUKI, V. and SANGHAVI, S. (2011). On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 378–387.

KING, M.-C., MARKS, J. H., MANDELL, J. B. et al. (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302** 643–646.

LEE, J. and HASTIE, T. (2013). Structure learning of mixed graphical models. In *Artificial Intelligence and Statistics* 388–396.

LI, W., LIU, Z., CHEN, L., ZHOU, L. and YAO, Y. (2014). MicroRNA-23b is an independent prognostic marker and suppresses ovarian cancer progression by targeting runt-related transcription factor-2. *FEBS Lett.* **588** 1608–1615.

LIU, H., ROEDER, K. and WASSERMAN, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in Neural Information Processing Systems* 1432–1440.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523 https://doi.org/10.1111/j.1467-9868.2010.00740.x

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351 https://doi.org/10.1214/07-AOS582

NELSEN, R. B. (1999). *An Introduction to Copulas. Lecture Notes in Statistics* **139**. Springer, New York. MR1653203 https://doi.org/10.1007/978-1-4757-3076-0

OBERMAYR, E., SANCHEZ-CABO, F., TEA, M.-K. M., SINGER, C. F., KRAINER, M., FISCHER, M. B., SEHOULI, J., REINTHALLER, A., HORVAT, R. et al. (2010). Assessment of a six gene panel for the molecular detection of circulating tumor cells in the blood of female cancer patients. *BMC Cancer* **10** 666.

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 https://doi.org/10.1214/09-AOS691

SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494. MR1224373

SHEN, R., OLSHEN, A. B. and LADANYI, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25** 2906–2912.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147. MR0356377

SU, W., BOGDAN, M. and CANDÈS, E. (2017). False discoveries occur early on the Lasso path. *Ann. Statist.* **45** 2133–2150. MR3718164 https://doi.org/10.1214/16-AOS1521

THE CANCER GENOME ATLAS RESEARCH NETWORK (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474** 609.

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

TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. MR3066375 https://doi.org/10.1214/13-EJS815

TONG, M., WONG, T. L., LUK, S. T.-C., CHE, N., WONG, K. Y., FUNG, T. M., GUAN, X.-Y., LEE, N. P., YUAN, Y.-F. et al. (2017). Down-regulation of 4-hydroxyphenylpyruvate dioxygenate (HPD) contributes to the pathogenesis of hepatocellular carcinoma (HCC) through ERK/BCL-2 signalling activation.

TOYAMA, T., IWASE, H., WATSON, P., MUZIK, H., SAETTLER, E., MAGLIOCCO, A., DIFRANCESCO, L., FORSYTH, P., GARKAVTSEV, I. et al. (1999). Suppression of ING1 expression in sporadic breast cancer. *Oncogene* **18**.

TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069 https://doi.org/10.1023/A:1017501703105

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018

WANG, W., WAINWRIGHT, M. J. and RAMCHANDRAN, K. (2010). Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inf. Theory* **56** 2967–2979. MR2683451 https://doi.org/10.1109/TIT.2010.2046199

YAN, J., JIANG, J.-Y., MENG, X.-N., XIU, Y.-L. and ZONG, Z.-H. (2016). MiR-23b targets cyclin G1 and suppresses ovarian cancer tumorigenesis and progression. *J. Exp. Clin. Cancer Res.* **35** 31.

YANG, E., ALLEN, G., LIU, Z. and RAVIKUMAR, P. K. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems* 1358–1366.

YANG, E., RAVIKUMAR, P. K., ALLEN, G. I. and LIU, Z. (2013). On Poisson graphical models. In *Advances in Neural Information Processing Systems* 1718–1726.

YANG, E., RAVIKUMAR, P., ALLEN, G. I., BAKER, Y., WAN, Y.-W. and LIU, Z. (2014a). A general framework for mixed graphical models. ArXiv Preprint. Available at arXiv:1411.0288.

YANG, E., BAKER, Y., RAVIKUMAR, P., ALLEN, G. and LIU, Z. (2014b). Mixed graphical models via exponential families. In *Artificial Intelligence and Statistics* 1042–1050.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. MR3450553

YU, B. (2013). Stability. *Bernoulli* **19** 1484–1500. MR3102560 https://doi.org/10.3150/13-BEJSP14

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824 https://doi.org/10.1093/biomet/asm018

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735