Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Simulation Study on the Stability and Recovery Behaviour of a Relative Belief Ratio Filter for Binary Feature Selection

Maher Emarly

b00097022@alumni.aus.edu

American University of Sharjah

Ayman Alzaatreh

American University of Sharjah

Luai Al-Labadi

University of Toronto

Firuz Kamalov

Canadian University of Dubai

Research Article

Keywords: Bayesian feature selection, Relative belief ratio, Filter methods, Feature selection stability, Synthetic data, Simulation study, Binary classification, Benchmarking

Posted Date: September 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10883318/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Simulation Study on the Stability and Recovery Behaviour of a Relative Belief Ratio Filter for Binary Feature Selection

Abstract

The relative belief ratio (RBR) filter is a Bayesian feature selection method for binary classification in which each continuous feature is scored by the evidence against equality of its class-conditional means. Although predictive utility has been demonstrated on real gene-expression data, recovery and selection stability under known feature relevance had not been evaluated. A controlled simulation study was conducted with informative, redundant, and noise features fixed by construction. Five data-generating scenarios were examined: a Gaussian mean shift, a block-correlated Gaussian variant, and heavy-tailed, right-skewed, and bimodal within-class distributions that depart from the filter's normality assumption while retaining a class-location difference. RBR was compared with Information Gain, Symmetrical Uncertainty, mRMR, ReliefF, and the Boruta wrapper on 300 independent datasets per condition. A one factor at a time design was used, with sample sizes from $n = 50$ to $n = 200$ and dimensionalities from $p = 500$ to $p = 1000$. From $n = 100$ upward, higher informative-feature recovery and selection stability were generally obtained with RBR under the mean-shift, correlated, and skewed scenarios. At $n = 200$, complete recovery was achieved under the mean-shift and skewed scenarios, with near-exact reproducibility of the selected sets. Little change in recovery was observed when the dimensionality was doubled under these two scenarios. Lower recovery was obtained with RBR at $n = 50$ and under the heavy-tailed and mixture scenarios, for which higher recovery was generally obtained with Information Gain or Boruta. Because distributional shape and standardised marginal signal strength were not varied independently, the lower performance in these scenarios cannot be attributed to non-normality alone. Under the conditions examined, the relative performance of RBR was strongest when the marginal location signal was clear and at least 100 observations were available.

Keywords: Bayesian feature selection; Relative belief ratio; Filter methods; Feature selection stability; Synthetic data; Simulation study; Binary classification; Benchmarking

1 Introduction

Feature selection is used to identify, from a set of candidate variables, the subset relevant to a learning problem (Saeys, Inza, & Larrañaga, 2007; Kamalov et al., 2025). When informative variables are retained and irrelevant or redundant variables are removed, model dimensionality, storage requirements, and computational cost can be reduced. Predictive performance and interpretability may also be improved. These considerations are especially relevant to high-dimensional classification with quantitative predictors and a categorical response, where the number of candidate features can exceed the number of observations. Such settings occur in gene-expression analysis, chemometrics, and signal processing.

Feature selection methods are commonly classified as filters, wrappers, or embedded methods according to their interaction with the learning algorithm (Saeys et al.,

2007). In a filter, feature relevance is evaluated without fitting a downstream classifier, so the resulting procedure is usually inexpensive and can be applied to large feature spaces. The criteria used by individual filters differ. Information Gain (IG) and Symmetrical Uncertainty (SU) measure the entropy reduction in the class label induced by a discretised feature (Roobaert, Karakoulas, & Chawla, 2006; Chai, Li, & Li, 2023). Minimum Redundancy Maximum Relevance (mRMR) balances marginal relevance against pairwise redundancy (Peng, Long, & Ding, 2005; Bugata & Drotár, 2020), whereas ReliefF assigns importance through nearest-neighbour comparisons and can respond to local feature interactions (Urbanowicz et al., 2018). Performance therefore depends on the structure of the signal, and the same filter is not expected to be optimal for every dataset (Bolón-Canedo et al., 2014).

Many filter methods return a feature ordering without a formal variable-level significance decision (Kamalov et al., 2025). A selected subset must then be defined through an external cutoff, such as the top k features or a fixed proportion. This requirement is consequential in high-dimensional, small-sample settings, where spurious associations may be difficult to distinguish from informative features. An explicit decision rule is provided by the Boruta wrapper: a random forest is fitted, and a feature is confirmed only when its importance significantly exceeds that of permuted "shadow" features (Kursa & Rudnicki, 2010). Compared with a univariate filter, however, a wrapper generally requires more computation and its decisions depend on the fitted learning algorithm. A filter that also provides a model-based evidential decision may therefore be useful.

Recent benchmarks have assessed selection accuracy, redundancy, predictive performance, computational cost, and stability across multiple datasets (Bommert et al., 2020; Barbieri, Grisci, & Dorn, 2024). High predictive accuracy does not imply that the same variables will be selected after a small change in the training data. Selection *stability* describes this sensitivity: a procedure is unstable when replacement or resampling of a small number of observations produces a large change in the selected subset (Kalousis, Prados, & Hilario, 2007). Low stability weakens the reproducibility of conclusions based on feature identity. This issue is particularly relevant when selected variables are interpreted as candidate biomarkers, because a classifier may remain accurate while a different set of variables is selected (Kuncheva, 2007). Instability is common in high-dimensional, small-sample settings, where several subsets may fit the observed data similarly well (Kalousis et al., 2007).

Selection stability has been quantified through measures such as the consistency index of Kuncheva (2007) and the Jaccard index, both of which compare subsets obtained after repeated perturbations of the data. Set-overlap measures can be affected by highly correlated or near-duplicate features, and adjusted measures have been proposed to account for similarity among features (Bommert & Rahnenführer, 2020). Reviews of stability concepts and proposed remedies are given by Khaire and Dhanalakshmi (2022) and Büyükkeçeci and Okur (2023). The effect of stability on conclusions from high-dimensional genetic data has also been examined empirically (Łukaszuk & Krawczuk, 2024).

Recovery and stability cannot usually be evaluated against known feature relevance in observed data. With synthetic data, the informative, redundant, and irrelevant variables can be specified before generation. Recovery of the informative variables, admission of noise, treatment of redundant surrogates, and agreement across indepen-

dently generated datasets can then be measured directly. Controlled simulations with known feature relevance have been used to evaluate Sure Independence Screening (Fan & Lv, 2008), features-annealed independence rules (Fan & Fan, 2008), distance-correlation screening (Li, Zhong, & Zhu, 2012), and selection stability (Kuncheva, 2007).

Designed synthetic benchmarks include the continuous predictors with prescribed functional relationships proposed by Friedman (1991) and the `MADELON` generator from the NIPS 2003 variable-selection challenge (Guyon, 2003). In `MADELON`, informative features are placed on the vertices of a multidimensional hypercube, so signal can be multivariate without marginal class information. Artificial datasets were used by Bolón-Canedo, Sánchez-Maroño, and Alonso-Betanzos (2013) to examine increasing numbers of irrelevant features, input and label noise, feature redundancy, attribute interaction, and small sample-to-feature ratios. In that review, selection of a redundant feature was penalised in the same manner as selection of an irrelevant feature. The same scoring convention is used here, and one design factor is varied at a time where the design permits.

Toolkits are now available for generating prescribed numbers of relevant, redundant, and irrelevant features from Gaussian, gamma, exponential, and other marginal families (Mitra et al., 2024, 2022). Synthetic datasets motivated by engineering applications have also been provided (Kamalov, Sulieman, & Cherukuri, 2023). In broader empirical benchmarks, stability has been evaluated together with predictive accuracy, runtime, and redundancy (Bommert et al., 2020, 2022; Barbieri et al., 2024). These studies provide the simulation components and evaluation measures used in the present design.

For a marginal location criterion, a direct reference design is the two-class location model, in which informative features differ between classes through a shift in their conditional means and the remaining features are non-informative (Fan & Fan, 2008). Correlated and redundant extensions of this model have also been used in screening studies (Fan & Lv, 2008). Because the signal is placed in the first class-conditional moment, the target of a location-based test is specified explicitly. The corresponding boundary is also clear: signal confined to higher moments or to joint dependence with null marginals is not targeted by such a test.

Filter methods are often evaluated in one fixed configuration or on observed datasets for which the informative features are unknown. Less evidence is available on selection stability across different data-generating scenarios and on its sensitivity to sample size and dimensionality. For a marginal location filter, both recovery under its reference model and changes in recovery when the distribution or sampling regime is altered require evaluation.

In the present study, a binary feature selection filter based on the relative belief ratio (RBR) was evaluated. RBR is a Bayesian criterion derived from the relative-belief framework for measuring statistical evidence (Evans, 2015, 2016). For each candidate feature, observations are partitioned by class and equality of the class-conditional means is tested under a within-class normal model. The estimated strength of the evidence is then used to rank the features. The filter was proposed for binary classification by Alzaatreh et al. (2025) and was later extended to multiclass classification (Emarly, Alzaatreh, & Al-Labadi, 2025); performance on high-dimensional cancer

gene-expression datasets was reported in both settings. Unlike rank-only filters such as IG, SU, mRMR, and ReliefF, RBR also provides a model-based evidential criterion: a relative belief ratio below one is interpreted as evidence that a feature discriminates between the classes. Its recovery and stability under known feature relevance, and their response to sample size and dimensionality, had not been systematically evaluated.

The study was designed to characterise the recovery and stability of RBR under known feature relevance; uniform superiority was not assumed. Five data-generating scenarios were used: a Gaussian mean shift (Fan & Fan, 2008), correlated and redundant feature structures (Fan & Lv, 2008; Bolón-Canedo et al., 2013; Mitra et al., 2024), and heavy-tailed, right-skewed, and bimodal within-class distributions. Recovery of informative features, admission of noise, selection of redundant surrogates, and agreement between independently generated datasets were evaluated with fixed-cardinality selected sets and established stability indices (Kuncheva, 2007; Bommert & Rahnenführer, 2020). RBR was compared with Information Gain, Symmetrical Uncertainty, mRMR, ReliefF, and the Boruta wrapper (Kursa & Rudnicki, 2010). The assessment was repeated in a one-factor-at-a-time design over sample sizes from $n = 50$ to $n = 200$ at fixed dimensionality and dimensionalities from $p = 500$ to $p = 1000$ at fixed sample size. These comparisons were used to identify the conditions under which recovery and stability were higher or lower for RBR than for the comparison methods.

The RBR filter is summarised in Section 2, with the full Bayesian development cited there. The five data-generating scenarios are specified in Section 3. The experimental design and evaluation measures are given in Section 4, and the results are reported in Section 5.

2 The Relative Belief Ratio Filter

The relative belief ratio (RBR) filter introduced by Alzaatreh et al. (2025) was assessed. Whereas a rank-only filter orders features without testing them for significance, RBR assigns each feature a measure of evidence for or against discrimination between the two classes and a calibrated measure of evidential strength. The procedure is based on the relative-belief framework of Evans (2015, 2016) and the Bayesian test for equality of means developed by Abdelrazeq, Al-Labadi, and Alzaatreh (2020) and Al-Labadi and Alzaatreh (2026). Only the components required for the simulation study are summarised here. The full derivation, prior elicitation, Monte Carlo computation, and pseudocode are provided by Alzaatreh et al. (2025).

2.1 The relative belief ratio and its strength

Consider a statistical model with density $f_\theta(x)$ indexed by a parameter θ in a parameter space Θ, and let $\pi(\theta)$ denote the prior density of θ. After observing the data x, the posterior density is obtained by Bayes' rule,

$$\pi_\theta(\theta \mid x) = \frac{f_\theta(x)\,\pi_\theta(\theta)}{l(x)}, \qquad l(x) = \int_\Theta f_\theta(x)\,\pi_\theta(\theta)\,\mathrm{d}\theta, \tag{1}$$

where $l(x)$ is the prior predictive density of the data. For a hypothesised value $H_0 : \theta = \theta_0$, and assuming the prior and posterior densities are continuous at θ_0,

the relative belief ratio is defined as the ratio of the posterior to the prior density evaluated at θ_0,

$$RB_\theta(\theta_0 \mid x) = \frac{\pi_\theta(\theta_0 \mid x)}{\pi_\theta(\theta_0)}. \tag{2}$$

The quantity in (2) measures how the data have changed the belief that θ_0 is the true value in passing from the prior to the posterior (Evans, 2015). Its interpretation is direct: if $RB_\theta(\theta_0 \mid x) > 1$, the data have increased the belief in θ_0 and thus provide evidence *in favour of* θ_0; if $RB_\theta(\theta_0 \mid x) < 1$, the data have decreased that belief and thus provide evidence *against* θ_0; and if $RB_\theta(\theta_0 \mid x) = 1$, the data provide no evidence either way.

The relative belief ratio records the *direction* of the evidence, but not, on its own, how strong that evidence is. To calibrate it, the strength of the evidence is measured by the posterior probability that the relative belief ratio is no larger than the value attained at θ_0 (Evans, 2015; Al-Labadi & Evans, 2017),

$$\mathrm{Str}(\theta_0 \mid x) = \Pi(RB_\theta(\theta \mid x) \le RB_\theta(\theta_0 \mid x) \,|\, x) = \int_{\{\theta:\, RB_\theta(\theta|x) \le RB_\theta(\theta_0|x)\}} \pi_\theta(\theta \mid x)\, \mathrm{d}\theta. \tag{3}$$

When $RB_\theta(\theta_0 \mid x) < 1$, a small value of $\mathrm{Str}(\theta_0 \mid x)$ indicates strong evidence against θ_0, whereas when $RB_\theta(\theta_0 \mid x) > 1$, a large value of $\mathrm{Str}(\theta_0 \mid x)$ indicates strong evidence in favour of θ_0. Sensitivity of inferences based on (2) and (3) to the choice of prior, and their relationship with prior–data conflict, were examined by Al-Labadi and Evans (2017).

2.2 Per-feature testing problem

Let $\mathbb{S} \equiv (\boldsymbol{X}, Y)$ denote a dataset with p features and n observations, where the observation matrix is $\boldsymbol{X} = [x_{wj}]_{n\times p} \in \mathbb{R}^{n\times p}$ and the response Y is binary with classes labelled 1 and 2. The features of $\boldsymbol{X}$ are denoted by the set $\{V_j \mid j = 1, \ldots, p\}$, and each feature V_j, assumed continuous, is partitioned by the class label into two subsets: $\mathbb{V}_j^1 = (x_{j1}^1, \ldots, x_{jn_1}^1)$, comprising the observations of V_j for which $Y = 1$, and $\mathbb{V}_j^2 = (x_{j1}^2, \ldots, x_{jn_2}^2)$, comprising those for which $Y = 2$. The two subsets are treated as independent samples from normal populations,

$$\mathbb{V}_j^i \sim N(\mu_i, \sigma_i^2), \qquad i = 1, 2, \qquad (\mu_i, \sigma_i^2) \in \mathbb{R} \times \mathbb{R}^+. \tag{4}$$

The relevance of feature V_j is then assessed through the hypotheses

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_a : \mu_1 \neq \mu_2. \tag{5}$$

A feature whose class-conditional means coincide carries no marginal mean-location information of the type targeted by the present criterion, whereas differing class-conditional means provide such information. Accordingly, the relative belief ratio for (5) is converted into a selection rule: if the data provide evidence against H_0 (that is, $RB < 1$), the feature V_j is judged significant for discriminating between the classes and becomes eligible for selection; if the data provide evidence in favour of H_0 (that is, $RB > 1$), the feature is judged uninformative and may be discarded. In the simulation study, this evidence criterion determines eligibility for the RBR candidate set but does not alone determine membership in the final fixed-size selected set.

2.3 Prior, posterior, and the KL divergence

To construct the relative belief ratio for (5), conjugate priors are placed on the class-conditional parameters (Al-Labadi & Alzaatreh, 2026). For $i = 1, 2$,

$$\frac{1}{\sigma_i^2} \sim \text{gamma}_{\text{rate}}(\alpha_{0i}, \beta_{0i}), \tag{6}$$

$$\mu_i \mid \sigma_i^2 \sim N\big(\mu_{0i},\, \lambda_{0i}^2 \sigma_i^2\big)\,, \tag{7}$$

where $\text{gamma}_{\text{rate}}$ denotes the gamma distribution in shape–rate parameterisation, and μ_{0i}, λ_{0i}, and $(\alpha_{0i}, \beta_{0i})$ are hyperparameters fixed through an elicitation procedure. The elicitation assumes, with high prior probability, that μ_i lies in a specified interval and that the class-conditional spread lies within specified bounds; the resulting closed-form expressions for the hyperparameters, and the specific values adopted in this study, are given in Alzaatreh et al. (2025) and in the experimental design (Section 4), respectively. Under (6)–(7), the posterior distribution of (μ_i, σ_i^2) is again normal–gamma,

$$\frac{1}{\sigma_i^2} \,\Big|\, \mathbb{V}_j^i \sim \text{gamma}_{\text{rate}}\big(\alpha_{0i} + \tfrac{n_i}{2},\, \beta_{x_i}\big)\,, \tag{8}$$

$$\mu_i \mid \sigma_i^2, \mathbb{V}_j^i \sim N\left(\mu_{x_i},\, \left(n_i + \tfrac{1}{\lambda_{0i}^2}\right)^{-1} \sigma_i^2\right), \tag{9}$$

where

$$\mu_{x_i} = \left(n_i + \tfrac{1}{\lambda_{0i}^2}\right)^{-1}\left(\frac{\mu_{0i}}{\lambda_{0i}^2} + n_i \bar{x}_i\right), \qquad \beta_{x_i} = \beta_{0i} + \frac{(n_i - 1)S_i^2}{2} + \frac{n_i\,(\bar{x}_i - \mu_{0i})^2}{2\,(n_i \lambda_{0i}^2 + 1)}, \tag{10}$$

with $\bar{x}_i = \frac{1}{n_i}\sum_{w=1}^{n_i} x_{jw}^i$ and $S_i^2 = \frac{1}{n_i - 1}\sum_{w=1}^{n_i}(x_{jw}^i - \bar{x}_i)^2$ the class sample mean and variance of the feature V_j in class i.

The hypothesis in (5) is assessed by comparing the concentration of the Kullback–Leibler (KL) divergence between the reference (null) model and the prior with its concentration between the reference model and the posterior (Abdelrazeq et al., 2020). Under the normal model (4), this KL divergence reduces to a single standardised KL divergence between the class-conditional means. Evaluated at the prior draws of (μ_i, σ_i^2) from (6)–(7), the prior KL divergence is

$$D = \frac{(\mu_1 - \mu_2)^2}{2\,(\sigma_1^2 + \sigma_2^2)}, \tag{11}$$

and evaluated at the posterior draws from (8)–(9), the posterior KL divergence D_x takes the same form,

$$D_x = \frac{(\mu_1 - \mu_2)^2}{2\,(\sigma_1^2 + \sigma_2^2)}. \tag{12}$$

Although (11) and (12) are formally identical, they are computed under different distributions: (11) under the prior and (12) under the posterior. The null hypothesis $\mu_1 = \mu_2$ corresponds to the value $D = 0$, and the relative belief ratio $RB(0 \mid x)$ of (2), applied to the KL divergence at 0, measures the evidence for or against equality of the class means. A value $RB(0 \mid x) < 1$ therefore constitutes evidence that the feature discriminates between the classes.

2.4 Computation and feature ranking

Because the prior and posterior densities of D are not available in closed form, the relative belief ratio $RB(0 \mid x)$ and its strength are estimated by Monte Carlo (Al-Labadi & Alzaatreh, 2026). Independent samples of (μ_i, σ_i^2) are drawn from the prior (6)–(7) and from the posterior (8)–(9), the KL divergences (11) and (12) are evaluated for each draw, and the relative belief ratio at 0 is estimated from the prior and posterior empirical distribution functions of D, with the strength estimated as the corresponding posterior tail probability defined in (3). The number of Monte Carlo draws used in the estimation is reported in Section 4, and the full discretised estimators are given in Alzaatreh et al. (2025).

The estimated strength of the relative belief ratio is used as the feature-importance score. Among features satisfying $RB(0 \mid x) < 1$, the features are ranked in ascending order of their estimated strength. Writing $\ell_j \in [0, 1]$ for the strength obtained for feature V_j, a smaller strength indicates stronger evidence against equality of the class means and hence a more important feature, so that the most informative features occupy the top of the ranking. The complete procedure, comprising the per-feature relative belief computation and the subsequent ranking, is given as Algorithms I and II of Alzaatreh et al. (2025).

The criterion is univariate and marginal. Through (11), the evidence assigned to a feature depends only on its within-class means and variances, so features with different class-conditional means are targeted. The rule $RB(0 \mid x) < 1$ also provides a model-based significance decision rather than an ordering alone. The extent to which these properties yield informative-feature recovery, noise avoidance, and stable selected sets was examined in the simulation study.

3 Simulation Study

Five data-generating scenarios were constructed for the marginal, location-based RBR criterion (Bolón-Canedo et al., 2013; Mitra et al., 2024). A Gaussian location model and a block-correlated variant were included, together with heavy-tailed, right-skewed, and bimodal within-class distributions. In each non-Gaussian scenario, a difference between the class locations was retained while the within-class normality assumption was violated. The scenarios are specified below; the sampling configurations, comparison methods, and evaluation measures are given in Section 4.

Structure common to all scenarios
Each synthetic dataset is denoted by $\mathbb{S} = (\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X} = [x_{wj}]$ contains n observations on p features and $y_w \in \{1, 2\}$ is the class label for observation w. Exactly $n/2$ observations were assigned to each class. The values of n and p are specified in Section 4.1. Three feature groups were defined: $p_{\text{inf}} = 25$ informative features associated with the label, $p_{\text{red}} = 5$ redundant features, and $p_{\text{noise}} = p - 30$ noise features independent of the label. Thus, 30 predictors were relevant, of which 25 were informative and 5 were redundant; the other $p - 30$ predictors were irrelevant. The informative and redundant counts were fixed across configurations, so the proportion of relevant features decreased as p increased. Recovery was scored against the 25 informative features, the redundant features were recorded separately, and the noise features formed the irrelevant set. These labels supplied the ground truth

for the measures in Section 4.

The redundant features were generated as noisy linear copies of informative features. Each redundant feature was therefore associated with the label but contained no information beyond that in its informative source. Writing $\iota(r) \in \{1, \ldots, p_{\text{inf}}\}$ for the informative feature copied by redundant feature r, with the informative features recycled in order,

$$x_{wr}^{\text{red}} \;=\; \rho_{\text{red}}\, x_{w,\iota(r)}^{\text{inf}} \;+\; \sqrt{1-\rho_{\text{red}}^2}\,\varepsilon_{wr}, \qquad \varepsilon_{wr} \overset{\text{iid}}{\sim} N(0,1), \qquad \rho_{\text{red}} = 0.90, \tag{13}$$

where ρ_{red} is the mixing coefficient. A strongly dependent noisy surrogate of the corresponding informative feature is produced by this construction. The induced Pearson correlation equals ρ_{red} only when the informative feature has unit variance and differs from it in the non-Gaussian scenarios. Noise features were generated independently of the class label, with their marginal distributions specified within each scenario.

The same informative, redundant, and noise feature counts were used in all five scenarios, while the predictor distributions and dependence structures were varied. The short names *mean shift*, *correlated*, *heavy-tailed*, *skewed*, and *mixture* are used in the tables and figures.

3.1 Independent Gaussian mean shift

In the first scenario, informative features were generated from mutually independent Gaussian distributions with class-conditional means separated by δ (Fan & Fan, 2008). A common variance was used in both classes, and the noise features were standard Gaussian,

$$x_{wj}^{\text{inf}} \mid y_w \;\sim\; N\big(\delta \mathbb{1}\{y_w = 2\},\, 1\big), \qquad x_{wj}^{\text{noise}} \overset{\text{iid}}{\sim} N(0,1), \qquad \delta = 1.0, \tag{14}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. In (14), all discriminative information is contained in the class-conditional means, with equal variances and independent coordinates. The marginal location criterion is therefore correctly specified in this scenario, and the KL divergence used by RBR is determined by the standardised mean difference. This scenario was used as the reference condition.

3.2 Block-correlated Gaussian signal

In the second scenario, the Gaussian location signal was retained while informative features were generated in correlated blocks through a shared latent factor (Fan & Lv, 2008). The informative features were divided into blocks of five. Within block b, a latent variable $z_{wb} \sim N(0,1)$ was shared by all members, and the class-2 mean was shifted by δ,

$$x_{wj}^{\text{inf}} \;=\; \sqrt{\rho}\, z_{wb} \;+\; \sqrt{1-\rho}\, e_{wj} \;+\; \delta \mathbb{1}\{y_w = 2\}, \qquad e_{wj} \overset{\text{iid}}{\sim} N(0,1), \tag{15}$$

with within-block correlation $\rho = 0.65$ and shift $\delta = 0.80$. Noise features were generated in weakly correlated blocks with shared-factor correlation 0.30 and block size 20. The marginal mean shift of each informative feature was retained, while strong dependence was introduced among informative coordinates and weaker dependence among noise coordinates. Recovery was thereby evaluated in the presence of correlated informative and irrelevant features.

3.3 Heavy-tailed mean shift

The normal–gamma model used by RBR assumes Gaussian within-class observations. In the third scenario, this assumption was violated while a class-location difference was retained. The Gaussian marginals in (14) were replaced by Student-t marginals with three degrees of freedom,

$$x_{wj}^{\mathrm{inf}} \;=\; \delta \mathbb{1}\{y_w = 2\} \;+\; t_{wj}, \qquad t_{wj} \sim t_3, \qquad x_{wj}^{\mathrm{noise}} \sim t_3, \qquad \delta = 1.0, \tag{16}$$

where t_3 denotes the Student-t distribution with three degrees of freedom. Extreme observations occur more frequently under this distribution than under a Gaussian distribution of comparable scale. The location shift was kept equal to that in the mean-shift scenario, so the signal remained in the class means. However, the class variances in the denominator of the KL divergence were estimated from heavy-tailed data and were more sensitive to individual extreme observations.

3.4 Right-skewed mean shift

In the fourth scenario, asymmetric within-class distributions were generated from a log-normal model, with the class-2 log-mean displaced by δ (Fan & Fan, 2008; Mitra et al., 2024),

$$x_{wj}^{\mathrm{inf}} \mid y_w \;\sim\; \mathrm{LN}\big(\delta \mathbb{1}\{y_w = 2\},\, \sigma_{\log}^2\big), \quad x_{wj}^{\mathrm{noise}} \sim \mathrm{LN}\big(0, \sigma_{\log}^2\big), \quad \delta = 0.60,\; \sigma_{\log} = 0.50, \tag{17}$$

where $\mathrm{LN}(\mu, \sigma_{\log}^2)$ denotes the log-normal distribution whose logarithm is $N(\mu, \sigma_{\log}^2)$. The shift on the log scale also produces different raw-scale means, so a raw-scale location difference was retained. The class distributions were positively skewed, and the sample mean entering the KL divergence was influenced by the right tail. This scenario was included to evaluate the Gaussian criterion under asymmetric observations resembling the marginal shape of expression measurements.

3.5 Within-class mixture mean shift

The fifth scenario was constructed to attenuate the standardised mean difference targeted by the filter (Fan & Fan, 2008). Each informative feature had a bimodal within-class distribution formed by two Gaussian components separated by s, and the full distribution was shifted by δ in class 2. With a latent component indicator $c_{wj} \sim \mathrm{Bernoulli}(1/2)$,

$$x_{wj}^{\mathrm{inf}} \;=\; (2c_{wj} - 1)\,\frac{s}{2} \;+\; \delta \mathbb{1}\{y_w = 2\} \;+\; \eta_{wj}, \qquad \eta_{wj} \sim N(0, \sigma_c^2), \tag{18}$$

with mode separation $s = 2.5$, class shift $\delta = 1.0$, and component standard deviation $\sigma_c = 0.70$; the noise features were standard Gaussian. The informative class-conditional means differed by δ, but within-class bimodality increased the variance in the denominator of the KL divergence and reduced the standardised mean difference. Location detection was therefore evaluated when a specified mean shift was accompanied by a large within-class spread.

4 Experimental Design

The simulation study was designed to assess the recovery and selection stability of the RBR filter under known ground truth and to identify the conditions under which

reliable feature selection is obtained. Five data-generating scenarios (Section 3) were evaluated under four (n, p) configurations (Table 1), with $B = 300$ independently generated datasets for each condition. The target set consisted of 25 known informative features, while the 5 redundant features were evaluated separately and were not counted as informative recoveries. RBR, Information Gain, Symmetrical Uncertainty, mRMR, ReliefF, and the Boruta wrapper were each applied to the same 300 datasets within every condition. Performance was evaluated using the numbers of informative, redundant, and noise features in the top 25 selected set, informative-feature precision, and Jaccard and Kuncheva selection stability. Monte Carlo standard errors or bootstrap confidence intervals were reported where appropriate.

Monte Carlo uncertainty was quantified for the reported mean performance measures. Because all methods were evaluated on the same independently generated datasets within each condition, between-method comparisons were conducted using paired mean differences with corresponding confidence intervals rather than comparisons of marginal means. These paired mean differences and their confidence intervals were treated as the primary inferential summaries. Two-sided paired Wilcoxon signed-rank tests were also conducted as a supplementary analysis, with p-values adjusted by the Holm procedure across the full family of 100 comparisons comprising five competitors, five scenarios, and four configurations.

Datasets were generated according to the scenarios described in Section 3. Before the feature-selection methods were applied, each predictor was standardised to zero mean and unit variance so that differences in raw measurement scale could not influence the comparison. The resulting standardised predictors were processed according to the requirements of each method: the entropy- and mutual-information-based filters operated on discretised representations, whereas RBR, ReliefF, and Boruta used the continuous values directly.

For each scenario and configuration, $B = 300$ independent datasets were generated. Each method was applied to every dataset, and the $k = 25$ top-ranked predictors, matching the number of informative features, constituted the selected set. For RBR, features satisfying $RB(0 \mid x) < 1$ formed the candidate set; these significant features were ranked in ascending order of their estimated strength, and the 25 highest-ranked candidates were retained. At least 25 features satisfied $RB(0 \mid x) < 1$ in every replicate across all scenarios and configurations, so every feature included in an evaluated RBR top 25 set satisfied the specified evidence criterion. Accordingly, the reported RBR recovery and stability results evaluate the evidence-screened top 25 ranking rather than the variable-size subset obtained by retaining every feature satisfying $RB(0 \mid x) < 1$.

The adequacy of $B = 300$ replications was assessed through the Monte Carlo standard error (Morris, White, & Crowther, 2019). Writing $\hat{\theta}$ for a mean over the B datasets and s for the standard deviation across replicates,

$$\mathrm{MCSE}(\hat{\theta}) = \frac{s}{\sqrt{B}}, \qquad 95\%\ \text{interval} = \hat{\theta} \pm 1.96\,\mathrm{MCSE}(\hat{\theta}). \tag{19}$$

Across all methods, scenarios, and configurations, the largest observed replicate standard deviations were 3.0 features for informative recovery and 3.5 features for noise admission. With $B = 300$, the corresponding Monte Carlo standard errors were no greater than 0.18 and 0.21 features, respectively.

The RBR filter was applied as described in Section 2. For each predictor, an estimated strength was obtained from the relative-belief test of equality of the class-conditional means, and the predictors were ranked in ascending order of strength. The internal Monte Carlo computation used $N_{\mathrm{MC}} = 5000$ posterior draws per feature and a 100-bin discretisation of the KL divergence. A single prior was elicited on the standardised scale and reused for all features and configurations. Following Alzaatreh et al. (2025), the prior was specified by requiring the class-conditional mean to lie in $(-1, 1)$ and, with prior probability 0.999, the class-conditional variance to lie in $(1/3, 3)$. The resulting hyperparameters, common to both classes, were $\mu_{0i} = 0$, $\lambda_{0i} = 0.419$, $\alpha_{0i} = 9.82$, and $\beta_{0i} = 7.83$.

4.1 Simulation configurations

Four configurations were examined around a baseline of $n = 100$ observations and $p = 500$ predictors (Table 1). At $p = 500$, sample size was varied over $n = 50$, 100, and 200. Dimensionality was examined by increasing p from 500 to 1000 while holding $n = 100$. The informative and redundant feature counts were fixed at 25 and 5, so the proportion of relevant predictors decreased from 6% at $p = 500$ to 3% at $p = 1000$. The dimensionality comparison therefore also represents a sparser screening problem. Exactly half of the observations were assigned to each class in every configuration, and $B = 300$ independent datasets were generated for each scenario and configuration.

Table 1: The four simulation configurations.

Configuration	n	p	Informative	Redundant	Noise	Datasets B
Baseline	100	500	25	5	470	300
Reduced sample size	50	500	25	5	470	300
Enlarged sample size	200	500	25	5	470	300
Increased dimensionality	100	1000	25	5	970	300

4.2 Comparison methods

RBR was compared with four filter methods and one wrapper. Information Gain (IG) ranks each predictor by the reduction in class entropy it induces (Roobaert et al., 2006), and Symmetrical Uncertainty (SU) normalises that quantity by the summed entropies to correct its bias toward many-valued features (Chai et al., 2023). Minimum Redundancy Maximum Relevance (mRMR) balances marginal relevance against redundancy with already-selected features (Peng et al., 2005), and ReliefF weights each predictor according to its separation of instances from their nearest same- and different-class neighbours (Urbanowicz et al., 2018). These four methods produce rankings. Boruta was included as a wrapper reference and confirms a predictor only when its random-forest importance consistently exceeds that of permuted shadow features (Kursa & Rudnicki, 2010).

The implementations and settings are reported to make the comparison reproducible (Nießl et al., 2022). IG and SU were computed with the `FSelector` package, which applies supervised discretisation to continuous predictors; mRMR was computed with `praznik`, with a complete ranking of all p predictors requested; and ReliefF

was computed with `CORElearn`, using the `ReliefFexpRank` estimator and 100 Relief iterations. Boruta was run with `maxRuns = 100`, and tentative attributes were resolved using `TentativeRoughFix`. Features were ordered by decision status—Confirmed, Tentative, and then Rejected—and by decreasing `meanImp` within each group, with feature names used to resolve ties. The first 25 features formed the selected set. Apart from the stated iteration counts, package defaults were used and no competitor was tuned to a data-generating scenario.

4.3 Evaluation measures

Recovery quantifies how well the top-k selected set matches the ground truth. Recovery is summarised by the composition of the selected set: the number of informative features recovered, the number of redundant features selected, and the number of noise features admitted. Because every selected set contains exactly 25 features, these three quantities sum to 25. Informative-feature precision is defined as the number of informative features recovered divided by 25. The five redundant features are counted separately and are not credited as informative recoveries (Bolón-Canedo et al., 2013). Every mean is reported with its Monte Carlo standard error from (19).

Repeated-sampling selection stability is defined here as the agreement of selected sets across independently generated datasets from the same data-generating mechanism. For two selected sets S_a, S_b of size k drawn from p features, the Jaccard index and the Kuncheva consistency index are reported (Kuncheva, 2007),

$$\text{Jaccard}(S_a, S_b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}, \qquad \text{Kuncheva}(S_a, S_b) = \frac{|S_a \cap S_b| - k^2/p}{k - k^2/p}, \tag{20}$$

each averaged over all $\binom{300}{2} = 44{,}850$ pairs of datasets. The Kuncheva index corrects for the overlap expected by chance and is therefore the measure used for comparisons *across* configurations with different p; the Jaccard index, which carries no such correction, is reported alongside it within configurations. Because the 44,850 pairwise values are mutually dependent, with each dataset appearing in 299 of them, their standard deviation divided by the number of pairs is not a valid standard error. Interval estimates are therefore obtained by resampling the 300 *datasets* with replacement, recomputing the mean pairwise index over the pairs of distinct resampled datasets, and taking the 2.5th and 97.5th percentiles of 1000 such replicates.

5 Results

5.1 Recovery analysis

An overview of the mean number of informative features included in the top 25 selected set is given in Figure 1 for the six methods across the four configurations and five scenarios. All 120 means are shown, and no Monte Carlo standard error exceeds 0.18 features. The top 25 selected set is decomposed into informative, redundant, and noise features at the baseline configuration in Table 2, together with the corresponding informative-feature precision.

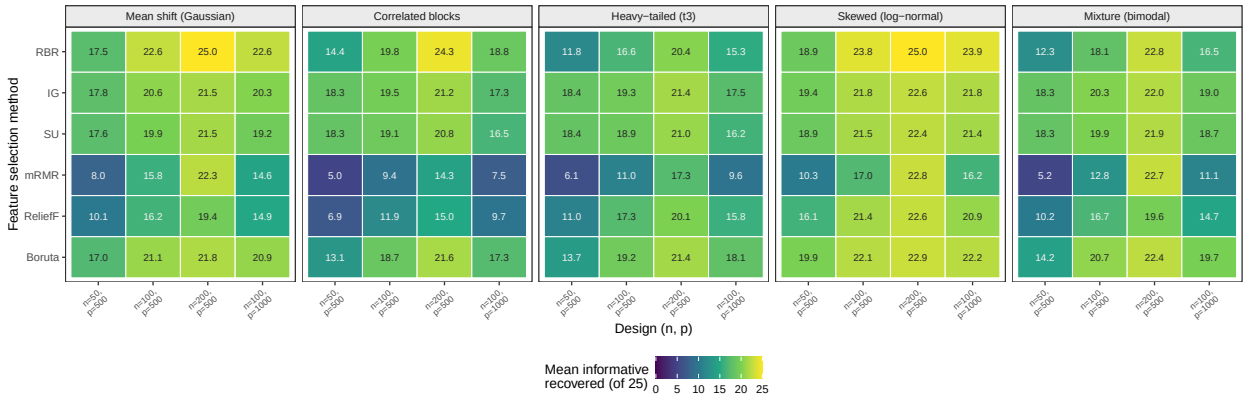


Figure 1: Mean number of informative features over 300 independent datasets per condition.

Across the conditions, higher informative-feature recovery was generally obtained with RBR, IG, SU, and Boruta than with ReliefF and mRMR, although the ordering depended on the scenario and configuration. Greater between-scenario variation was observed for RBR at the baseline configuration: mean recovery ranged from 16.55 features under the heavy-tailed scenario to 23.84 under the skewed scenario, a range of 7.29 features. By comparison, the baseline recovery of IG ranged from 19.30 to 21.80 features, and that of Boruta ranged from 18.71 to 22.14. Mean ranks across all conditions are reported in Figure 2.

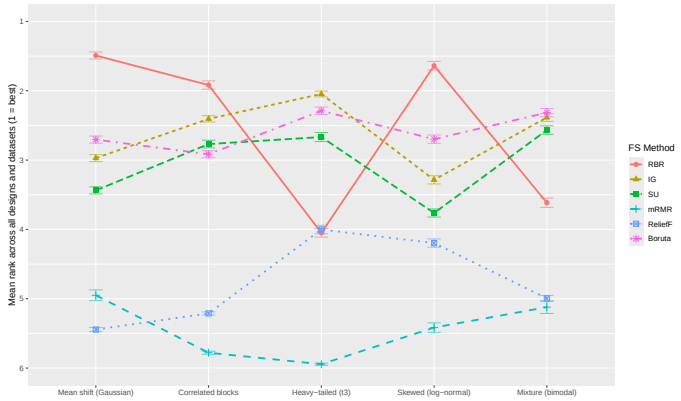


Figure 2: Mean rank of the six feature selection methods by informative-feature recovery, across all four configurations and five scenarios.

The recovery measures in Table 2 are consistent with this pattern at the baseline configuration ($n = 100$, $p = 500$). Under the mean-shift scenario, averages of 22.64 informative, 1.85 redundant, and 0.51 noise features were included in the RBR top 25 set. The highest informative-feature recovery was obtained with RBR under the mean-shift (22.64), correlated (19.75), and skewed (23.84) scenarios. Under the heavy-tailed and mixture scenarios, mean RBR recovery was 16.55 and 18.07 features, compared with 19.30 for IG and 20.71 for Boruta, respectively, and more noise features were admitted by RBR. The replicate distributions are reported in Figures A1–A4.

The effect of sample size was evaluated by holding $p = 500$ fixed and varying n over 50, 100, and 200 (Figure 3). Informative-feature recovery increased with sample size

Table 2: Composition of the top 25 selected set and informative-feature precision at $n = 100$, $p = 500$. Entries are means over 300 datasets with Monte Carlo standard errors in parentheses.

Method	Mean shift	Correlated	Heavy-tailed	Skewed	Mixture
Informative (correct)					
RBR	**22.64** (0.085)	**19.75** (0.102)	16.55 (0.103)	**23.84** (0.057)	18.07 (0.084)
IG	20.64 (0.063)	19.48 (0.105)	**19.30** (0.097)	21.80 (0.052)	20.26 (0.079)
SU	19.87 (0.074)	19.09 (0.112)	18.89 (0.107)	21.52 (0.054)	19.94 (0.086)
mRMR	15.85 (0.111)	9.45 (0.134)	11.02 (0.108)	17.04 (0.078)	12.85 (0.104)
ReliefF	16.19 (0.097)	11.85 (0.158)	17.33 (0.098)	21.44 (0.064)	16.66 (0.101)
Boruta	21.05 (0.057)	18.71 (0.113)	19.21 (0.075)	22.14 (0.051)	**20.71** (0.061)
Redundant					
RBR	1.85 (0.082)	3.22 (0.078)	3.03 (0.056)	0.84 (0.054)	3.32 (0.051)
IG	3.46 (0.051)	2.71 (0.086)	2.80 (0.059)	3.06 (0.050)	2.78 (0.055)
SU	3.35 (0.051)	2.62 (0.084)	2.72 (0.059)	3.08 (0.053)	2.72 (0.053)
mRMR	1.92 (0.060)	1.15 (0.054)	1.14 (0.055)	1.17 (0.051)	0.67 (0.042)
ReliefF	2.77 (0.059)	1.83 (0.079)	2.77 (0.059)	2.43 (0.058)	1.56 (0.057)
Boruta	3.47 (0.051)	2.93 (0.075)	3.18 (0.052)	2.81 (0.050)	2.95 (0.054)
Noise					
RBR	0.51 (0.040)	2.03 (0.101)	5.42 (0.119)	0.32 (0.031)	3.61 (0.093)
IG	0.90 (0.052)	2.81 (0.099)	2.90 (0.092)	0.14 (0.021)	1.97 (0.078)
SU	1.78 (0.073)	3.30 (0.100)	3.39 (0.099)	0.40 (0.034)	2.34 (0.083)
mRMR	7.23 (0.112)	14.40 (0.134)	12.84 (0.112)	6.79 (0.083)	11.48 (0.104)
ReliefF	6.03 (0.107)	11.31 (0.179)	4.89 (0.108)	1.13 (0.057)	6.77 (0.111)
Boruta	0.47 (0.039)	3.37 (0.124)	2.61 (0.085)	0.05 (0.013)	1.34 (0.063)
Informative-feature precision					
RBR	**0.906** (0.0034)	**0.790** (0.0041)	0.662 (0.0041)	**0.954** (0.0023)	0.723 (0.0034)
IG	0.826 (0.0025)	0.779 (0.0042)	**0.772** (0.0039)	0.872 (0.0021)	0.810 (0.0032)
SU	0.795 (0.0030)	0.764 (0.0045)	0.756 (0.0043)	0.861 (0.0022)	0.798 (0.0034)
mRMR	0.634 (0.0044)	0.378 (0.0054)	0.441 (0.0043)	0.682 (0.0031)	0.514 (0.0042)
ReliefF	0.648 (0.0039)	0.474 (0.0063)	0.693 (0.0039)	0.858 (0.0026)	0.666 (0.0040)
Boruta	0.842 (0.0023)	0.748 (0.0045)	0.768 (0.0030)	0.886 (0.0020)	**0.828** (0.0024)

for all methods, and a steeper increase was observed for RBR than for the comparison methods.

Because all methods were applied to the same datasets, for concise presentation, Tables 3 and 4 compare RBR with the best observed competitor in each condition and are interpreted as descriptive summaries; the complete pre-specified pairwise comparisons of RBR with all competitors are reported in Table A2. Paired mean differences and their 95% Monte Carlo confidence intervals, calculated as $\bar{d} \pm 1.96\,\mathrm{MCSE}(\bar{d})$, are the primary summaries. Two-sided paired Wilcoxon signed-rank tests were used as a supplementary analysis, with p-values Holm-adjusted across the full family of 100 comparisons. The intervals summarise the mean paired differences, whereas the Wilcoxon tests assess the location of the paired-difference distributions. At $n = 50$, higher recovery was not obtained with RBR under any scenario. Under the Gaussian mean-shift scenario, no statistically detectable difference was found between RBR and IG ($\bar{d} = -0.24$, 95% CI -0.51 to $+0.03$); under the remaining scenarios, between 1.00 and 6.55 fewer informative features were recovered with RBR than with the best observed competitor. This pattern is consistent with lower precision in the class-moment estimates at this sample size. At $n = 200$, higher recovery was obtained with RBR under the mean-shift, correlated, and skewed scenarios, with paired mean differences of 2.72, 2.74, and 2.09 features, respectively. No statistically detectable

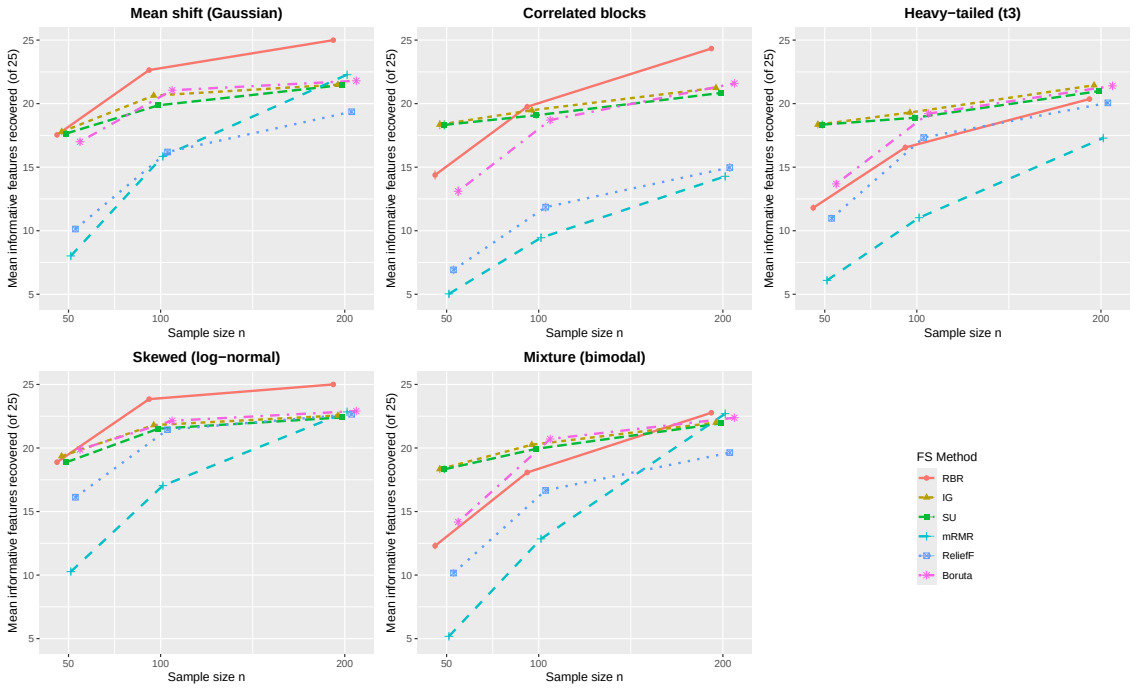


Figure 3: Mean number of informative features recovered across sample sizes $n = 50, 100, 200$ at $p = 500$.

difference was found under the mixture scenario ($\bar{d} = 0.07$, 95% CI -0.11 to 0.26), whereas lower recovery remained under the heavy-tailed scenario ($\bar{d} = -1.09$, 95% CI -1.25 to -0.93).

At fixed $n = 100$, the number of predictors was increased from $p = 500$ to $p = 1000$ (Figure 4). Because the informative and redundant feature counts were fixed, the proportion of relevant predictors decreased from 6% to 3%. Mean RBR recovery remained approximately 22.6 features under the mean-shift scenario and 23.9 under the skewed scenario.

Larger reductions were observed under the heavy-tailed and mixture scenarios, where recovery decreased from 16.6 to 15.3 and from 18.1 to 16.5 features. At $p = 1000$, mean recovery with RBR exceeded that of the best observed competitor by 1.68, 1.44, and 1.69 features under the mean-shift, correlated, and skewed scenarios, respectively. It was lower by 2.73 and 3.25 features under the heavy-tailed and mixture scenarios (Table 4).

The between-scenario differences in RBR recovery were consistent with differences in the marginal class-location signal. Higher recovery was observed under the mean-shift, correlated, and skewed scenarios, where class separation was greater relative to within-class variability. Lower recovery was observed under the heavy-tailed and mixture scenarios, where within-class dispersion was greater. Distributional shape and standardised signal strength were not varied independently, so the effects of non-normality and reduced class separation cannot be distinguished from these comparisons.

Table 3: Paired difference in informative features recovered, RBR minus the best performing competitor in each condition

Scenario	Best observed competitor	$\bar{d}$	95% CI	p
n=50, p=500				
Mean shift	IG	−0.24	[−0.51, +0.03]	0.762
Correlated	IG	−3.96	[−4.37, −3.56]	$< 10^{-16}$
Heavy-tailed	IG	−6.55	[−6.90, −6.20]	$< 10^{-16}$
Skewed	Boruta	−1.00	[−1.18, −0.82]	$< 10^{-16}$
Mixture	IG	−6.03	[−6.37, −5.70]	$< 10^{-16}$
n=100, p=500				
Mean shift	Boruta	+1.59	[+1.41, +1.76]	$< 10^{-16}$
Correlated	IG	+0.27	[+0.03, +0.51]	0.043
Heavy-tailed	IG	−2.75	[−3.01, −2.49]	$< 10^{-16}$
Skewed	Boruta	+1.70	[+1.56, +1.84]	$< 10^{-16}$
Mixture	Boruta	−2.64	[−2.80, −2.47]	$< 10^{-16}$
n=200, p=500				
Mean shift	mRMR	+2.72	[+2.61, +2.83]	$< 10^{-16}$
Correlated	Boruta	+2.74	[+2.61, +2.88]	$< 10^{-16}$
Heavy-tailed	IG	−1.09	[−1.25, −0.93]	$< 10^{-16}$
Skewed	Boruta	+2.09	[+1.99, +2.20]	$< 10^{-16}$
Mixture	mRMR	+0.07	[−0.11, +0.26]	1.000

The confidence intervals describe mean paired differences. The p-values are from two-sided paired Wilcoxon signed-rank tests and are Holm-adjusted across the full family of 100 comparisons.

Table 4: Paired difference in informative features recovered, RBR minus the best observed competitor, at $n = 100$, $p = 1000$. Positive values favour RBR.

Scenario	Best observed competitor	$\bar{d}$	95% CI	p
Mean shift	Boruta	+1.68	[+1.51, +1.84]	$< 10^{-16}$
Correlated	Boruta	+1.44	[+1.24, +1.64]	$< 10^{-16}$
Heavy-tailed	Boruta	−2.73	[−2.95, −2.51]	$< 10^{-16}$
Skewed	Boruta	+1.69	[+1.53, +1.84]	$< 10^{-16}$
Mixture	Boruta	−3.25	[−3.45, −3.05]	$< 10^{-16}$

5.2 Stability analysis

Feature-level selection frequencies were examined as a complementary measure of reproducibility across the 300 independently generated datasets (Figure 5). Informative features were generally selected at similar frequencies within each method, whereas greater variation was observed among the redundant features. Under the mean-shift and skewed scenarios, the informative features were selected more consistently by RBR than their redundant surrogates.

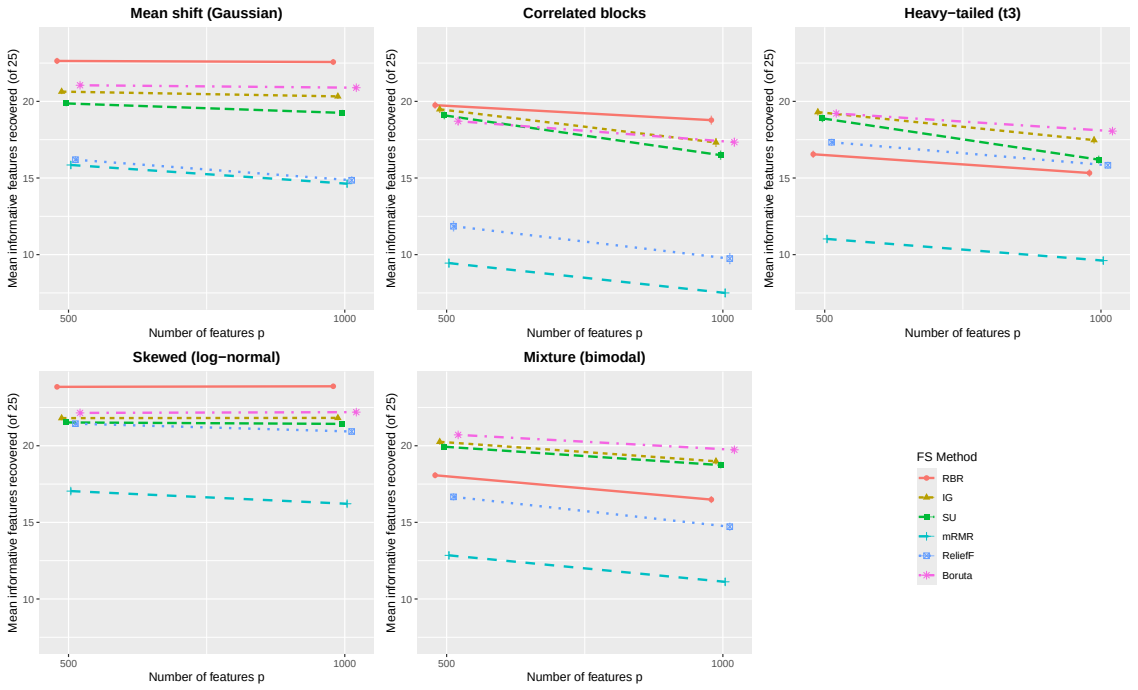


Figure 4: Mean informative features recovered at $p = 500$ versus $p = 1000$, at $n = 100$.

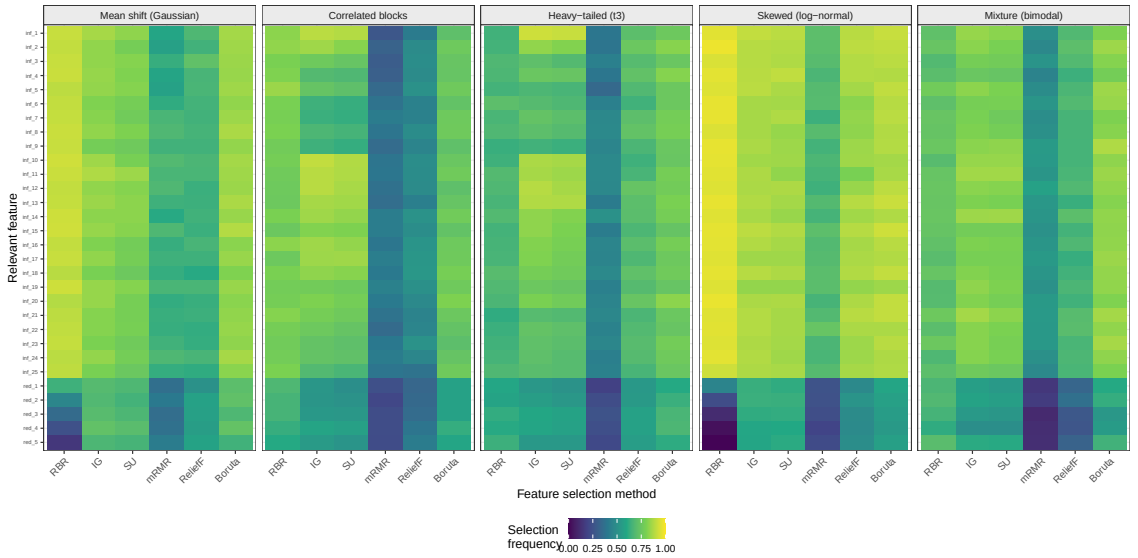


Figure 5: Selection frequency of the 30 relevant features at the baseline configuration over 300 independently generated datasets.

Set-level stability was assessed from the agreement of the top 25 selected features across independently generated datasets using the chance-corrected Kuncheva index. The baseline results are presented in Table 5. Higher stability was observed for RBR under the mean-shift, correlated, and skewed scenarios, with Kuncheva indices of 0.845, 0.692, and 0.916, respectively. Lower stability was obtained under the heavy-tailed and mixture scenarios, where more consistent selected sets were produced by Information Gain and Boruta. Thus, the three baseline scenarios with the highest RBR recovery were also those with the highest RBR stability.

The effect of the sampling regime on stability is shown in Figure 6. Stability generally

Table 5: Mean pairwise Kuncheva index at the baseline configuration, with 95% bootstrap intervals.

Method	Mean shift	Correlated	Heavy-tailed	Skewed	Mixture
RBR	0.845 [0.837, 0.854]	0.692 [0.680, 0.705]	0.488 [0.476, 0.501]	0.916 [0.909, 0.923]	0.591 [0.581, 0.602]
IG	0.765 [0.758, 0.772]	0.656 [0.644, 0.668]	0.647 [0.635, 0.658]	0.826 [0.823, 0.830]	0.704 [0.694, 0.714]
SU	0.707 [0.697, 0.716]	0.626 [0.614, 0.639]	0.617 [0.604, 0.629]	0.807 [0.802, 0.811]	0.680 [0.669, 0.690]
mRMR	0.407 [0.396, 0.419]	0.128 [0.121, 0.137]	0.178 [0.171, 0.185]	0.452 [0.444, 0.461]	0.241 [0.232, 0.250]
ReliefF	0.457 [0.446, 0.468]	0.224 [0.213, 0.237]	0.520 [0.508, 0.532]	0.772 [0.765, 0.779]	0.440 [0.429, 0.451]
Boruta	0.795 [0.790, 0.801]	0.610 [0.596, 0.624]	0.655 [0.645, 0.665]	0.839 [0.836, 0.843]	0.743 [0.735, 0.751]

increased with sample size, and the largest increases for RBR were observed under the mean-shift and skewed scenarios. At $n = 200$, Kuncheva indices approached 1.0 in both settings. The value of 1.0 under the skewed scenario coincided with complete recovery: because all selected sets contained the same 25 informative features, perfect stability followed from the fixed-cardinality design. When dimensionality was increased from $p = 500$ to $p = 1000$ at fixed $n = 100$, stability was reduced for most methods, although the highest values under the mean-shift and skewed scenarios were still obtained with RBR. The within-configuration ordering was also reproduced by the Jaccard results in Figure A5.

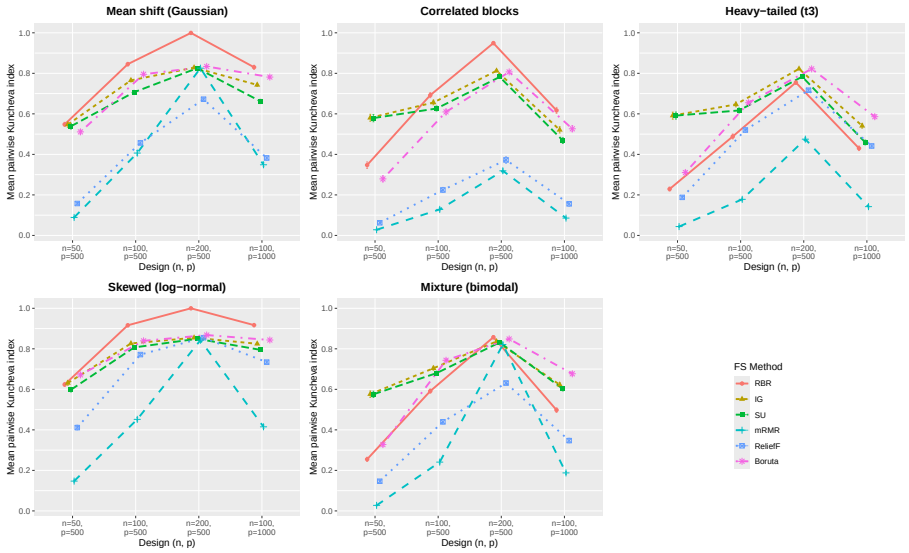


Figure 6: Mean pairwise Kuncheva index across the four configurations, with 95% bootstrap intervals.

6 Discussion and Conclusion

The recovery and stability of the relative belief ratio (RBR) filter were examined across five data-generating scenarios and four sampling configurations for which feature relevance was known. Performance was found to depend on the data-generating regime and sample size, and uniform superiority was not observed. When class information was represented by a sufficiently strong marginal difference in the class-conditional means and at least $n = 100$ observations were available, more informative features were generally recovered and more reproducible top 25 sets were obtained with RBR than with the comparison methods. The highest recovery and stability were observed under the independent mean-shift, correlated, and right-

skewed scenarios. Under the heavy-tailed and within-class mixture scenarios, lower recovery and stability were obtained with RBR, while higher values were obtained with IG or Boruta. RBR is therefore interpreted as a marginal location-based filter whose performance depends on the data-generating regime.

RBR performance varied with sample size. At $n = 50$,chigher recovery was not obtained with RBR under any of the five scenarios. This pattern is consistent with the lower precision of the class-conditional mean and variance estimates when few observations are available within each class. As the sample size increased, the separation between informative and noise features improved. At $n = 200$, complete recovery was achieved by RBR under the mean-shift and skewed scenarios, and a recovery advantage was obtained under the correlated scenario. Under the mixture scenario, no statistically detectable difference was found between RBR and the best-performing comparison method; RBR remained less competitive under the heavy-tailed scenario.

Similar patterns were observed for recovery and stability. Under the fixed-cardinality design, perfect stability is implied when all 25 informative features are recovered in a selected set of size 25, because the selected set must equal the ground-truth set. When recovery was incomplete, instability reflected competition near the selection boundary among weaker informative features, redundant features, and noise features. Under the mean-shift, correlated, and right-skewed scenarios, high RBR stability was observed together with consistent recovery of the informative features rather than repeated selection of irrelevant variables.

Poorer RBR performance cannot be attributed to non-normality alone. High recovery was observed under the right-skewed scenario despite the violation of normality, whereas lower recovery was observed under the heavy-tailed and mixture scenarios. Because distributional shape and effective signal strength were not varied independently, their individual effects cannot be separated. No causal effect of non-normality was identified by the present design.

Little reduction in RBR recovery was observed when the number of predictors was doubled from $p = 500$ to $p = 1000$ under the mean-shift and skewed scenarios. Under the correlated scenario, mean recovery with RBR exceeded that of the best observed competitor by 1.44 features at $p = 1000$. Larger reductions were observed under the heavy-tailed and mixture scenarios. The reduction associated with increased dimensionality was smaller when the marginal location signal was clear and larger when informative features were difficult to distinguish from noise. Because the proportion of relevant predictors decreased from 6% to 3%, dimensionality and sparsity were changed together.

For RBR, features satisfying $RB(0 \mid x) < 1$ were treated as significant and ranked according to their estimated strength. The 25 highest-ranked features were then retained to match the known number of informative features. Across all scenarios and configurations, the evidence criterion was satisfied by every feature included in the evaluated RBR top 25 sets. Recovery and stability were evaluated for this evidence-screened top 25 ranking. The size and composition of the variable-cardinality set obtained by retaining every feature satisfying $RB(0 \mid x) < 1$ were not evaluated.

The simulation was restricted to balanced binary-classification scenarios based primarily on marginal location differences. Settings involving pure scale differences,

interactions without marginal effects, class imbalance, label noise, and missing observations were not examined. The comparison methods were applied using the reported package settings, and predictive performance and computational cost were not evaluated. Within this scope, higher recovery and stability were obtained with RBR at moderate sample sizes when the marginal location signal was clear; the same advantage was not observed at $n = 50$ or under the heavy-tailed and mixture scenarios.

References

Abdelrazeq, I., Al-Labadi, L., & Alzaatreh, A. (2020). On one-sample Bayesian tests for the mean. *Statistics*, *54*(2), 424–440. doi: 10.1080/02331888.2020.1726918

Al-Labadi, L., & Alzaatreh, A. (2026). k-sample Bayesian test for the mean. *Annals of Data Science*. doi: 10.1007/s40745-025-00668-y

Al-Labadi, L., & Evans, M. (2017). Optimal robustness results for relative belief inferences and the relationship to prior-data conflict. *Bayesian Analysis*, *12*(3), 705–728. doi: 10.1214/16-BA1024

Alzaatreh, A., Emarly, M., Mitra, R., & Al-Labadi, L. (2025). Feature selection in binary classification problems using relative belief ratio. *Physica Scripta*, *100*(8), 086010.

Barbieri, M. C., Grisci, B. I., & Dorn, M. (2024). Analysis and comparison of feature selection methods towards performance and stability. *Expert Systems with Applications*, *249*, 123667. doi: 10.1016/j.eswa.2024.123667

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, *34*(3), 483–519. doi: 10.1007/s10115-012-0487-8

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, *282*, 111–135. doi: 10.1016/j.ins.2014.05.042

Bommert, A., & Rahnenführer, J. (2020). Adjusted measures for feature selection stability for data sets with similar features. In *Machine learning, optimization, and data science (LOD 2020), lecture notes in computer science* (Vol. 12565, pp. 203–214). Springer. doi: 10.1007/978-3-030-64583-0_19

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, *143*, 106839. doi: 10.1016/j.csda.2019.106839

Bommert, A., Welchowski, T., Schmid, M., & Rahnenführer, J. (2022). Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings in Bioinformatics*, *23*(1), bbab354. doi: 10.1093/bib/bbab354

Bugata, P., & Drotár, P. (2020). On some aspects of minimum redundancy maximum relevance feature selection. *Science China Information Sciences*, *63*(1), 112103. doi: 10.1007/s11432-019-2701-x

Büyükkeçeci, M., & Okur, M. C. (2023). A comprehensive review of feature selection and feature selection stability in machine learning. *Gazi University Journal of Science*, *36*(4), 1506–1520. doi: 10.35378/gujs.993763

Chai, Z., Li, W., & Li, Y. (2023). Symmetric uncertainty based decomposition multi-objective immune algorithm for feature selection. *Swarm and Evolutionary Computation*, *78*, 101286. doi: 10.1016/j.swevo.2023.101286

Emarly, M., Alzaatreh, A., & Al-Labadi, L. (2025). A Bayesian feature selection method for multiclass classification with application to cancer gene expression data. (Manuscript under review)

Evans, M. (2015). *Measuring statistical evidence using relative belief*. Boca Raton, FL: Chapman and Hall/CRC.

Evans, M. (2016). Measuring statistical evidence using relative belief. *Computational and Structural Biotechnology Journal*, *14*, 91–96. doi: 10.1016/j.csbj.2015.12.001

Fan, J., & Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, *36*(6), 2605–2637. doi: 10.1214/07-AOS504

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, *70*(5), 849–911. doi: 10.1111/j.1467-9868.2008.00674.x

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67. doi: 10.1214/aos/1176347963

Guyon, I. (2003). *Design of experiments for the NIPS 2003 variable selection benchmark*. (NIPS 2003 Workshop on Feature Extraction (MADELON dataset))

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowledge and Information Systems*, *12*(1), 95–116. doi: 10.1007/s10115-006-0040-8

Kamalov, F., Sulieman, H., & Cherukuri, A. K. (2023). Synthetic data for feature selection. In *Lecture notes in computer science* (pp. 291–301). Springer.

Kamalov, F., Sulieman, H., Alzaatreh, A., Emarly, M., Chamlal, H., & Safaraliev, M. (2025). Mathematical methods in feature selection: A review. *Mathematics*, *13*(6), 996. doi: 10.3390/math13060996

Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University – Computer and Information Sciences*, *34*(4), 1060–1073. doi: 10.1016/j.jksuci.2019.06.012

Kuncheva, L. I. (2007). A stability index for feature selection. In *Proceedings of the 25th IASTED international multi-conference: Artificial intelligence and applications* (pp. 421–427).

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*(11), 1–13. doi: 10.18637/jss.v036.i11

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*(499), 1129–1139. doi: 10.1080/01621459.2012.695654

Łukaszuk, T., & Krawczuk, J. (2024). Importance of feature selection stability in the classifier evaluation on high-dimensional genetic data. *PeerJ*, *12*, e18405. doi: 10.7717/peerj.18405

Mitra, R., Varam, D., Ali, E., Sulieman, H., & Kamalov, F. (2022). Development of synthetic data benchmarks for evaluating feature selection algorithms. In *Proceedings of the 2022 2nd international seminar on machine learning, optimization, and data science (ISMODE)* (pp. 47–52). IEEE.

Mitra, R., Ali, E., Varam, D., Sulieman, H., & Kamalov, F. (2024). Variable selection in data analysis: A synthetic data toolkit. *Mathematics*, *12*(4), 570. doi: 10.3390/math12040570

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. doi: 10.1002/sim.8086

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., & Boulesteix, A. L. (2022). Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *WIREs Data Mining and Knowledge Discovery*, *12*(2), e1441. doi: 10.1002/widm.1441

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238.

doi: 10.1109/TPAMI.2005.159

Roobaert, D., Karakoulas, G., & Chawla, N. V. (2006). Information gain, correlation and support vector machines. In *Feature extraction: Foundations and applications* (pp. 463–470). Springer.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. doi: 10.1093/bioinformatics/btm344

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, *85*, 189–203. doi: 10.1016/j.jbi.2018.07.014

A1 Appendix

This appendix contains the software implementations and settings used in the study, the complete matrix of paired comparisons between RBR and all five competitors, the replicate distributions of informative-feature recovery, and the supplementary stability and selected-set composition figures. The software settings are reported in Table A1, and the complete paired comparisons are reported in Table A2. Replicate distributions are shown in Figures A1–A4; the uncorrected Jaccard stability results are shown in Figure A5; and the selected-set compositions are shown in Figures A6–A11.

Table A1: Software implementations and settings used in the simulation study.

Method	Implementation	Principal settings
RBR	custom (this study)	$N_{\mathrm{MC}} = 5000$ posterior draws per feature; $M = 100$ bins; single prior elicited once on the standardised scale.
IG	`FSelector`	`information.gain`; package-internal supervised discretisation
SU	`FSelector`	`symmetrical.uncertainty`; package-internal supervised discretisation
mRMR	`praznik`	`MRMR`; complete ranking of all p predictors
ReliefF	`CORElearn`	`attrEval`, estimator `ReliefFexpRank`, 100 Relief iterations
Boruta	`Boruta`	`maxRuns` = 100; `TentativeRoughFix`; decision order Confirmed, Tentative, Rejected; decreasing `meanImp`; feature-name tie-breaking

All simulations were carried out in R, with each competitor invoked through the function and settings listed in Table A1. A base seed of 101 was offset by configuration and scenario, and the seed for dataset i within each condition was set before generation. Each dataset can therefore be reproduced independently of the order in which the conditions are run. A Monte Carlo sample of $N_{\mathrm{MC}} = 5000$ draws from (6)–(7) was used to construct the prior distribution of the KL divergence. The same sample size was used for the posterior distribution of each feature. The prior sample was generated once with a separate fixed seed and was used under all conditions. The boundaries of the $M = 100$ bins were defined by its empirical quantiles. Thus, the

bin boundaries and the prior probability assigned to the first bin were held constant across conditions.

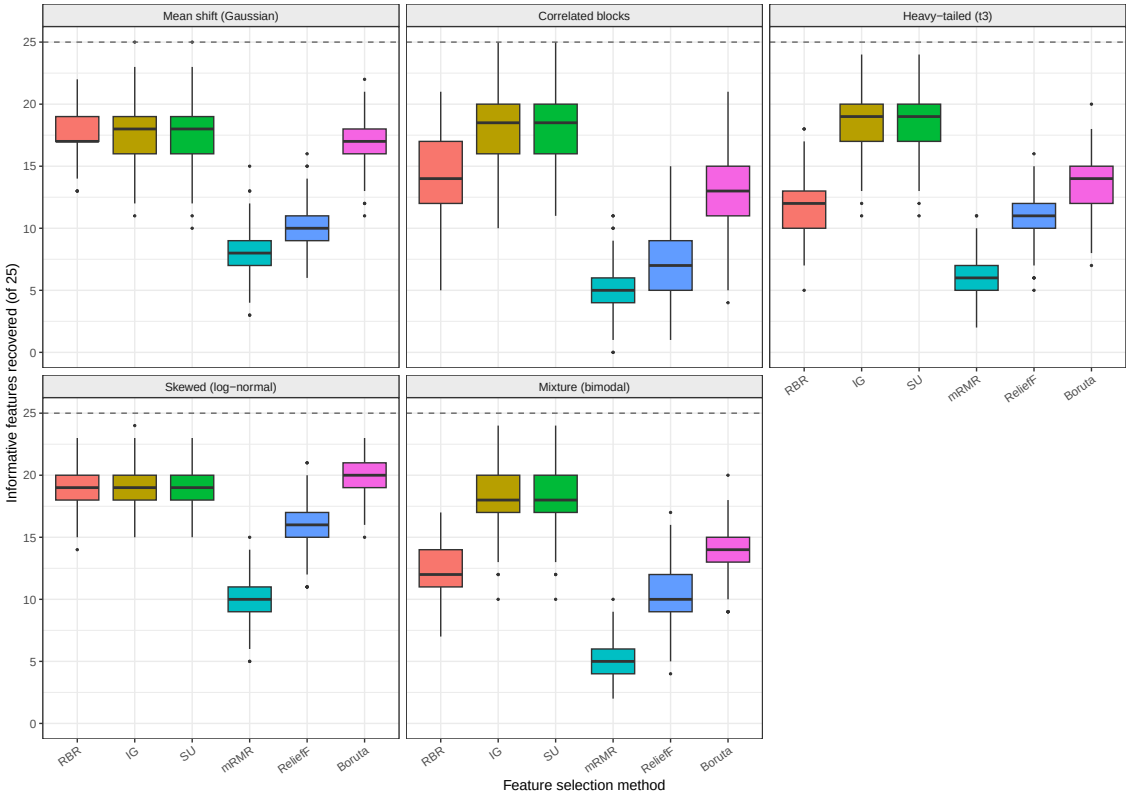


Figure A1: Distribution of informative features recovered at $n = 50$, $p = 500$.

Table A2: Paired difference in informative features recovered, RBR minus each competitor, in every condition. Positive values favour RBR.

Scenario	Competitor	$\bar{d}$	95% CI	p
n=50, p=500				
Mean shift	Boruta	+0.54	[+0.34, +0.74]	< 0.001
Mean shift	IG	−0.24	[−0.51, +0.03]	0.762
Mean shift	ReliefF	+7.41	[+7.17, +7.65]	$< 10^{-16}$
Mean shift	SU	−0.10	[−0.39, +0.18]	1.000
Mean shift	mRMR	+9.53	[+9.28, +9.78]	$< 10^{-16}$
Correlated	Boruta	+1.28	[+1.08, +1.48]	$< 10^{-16}$
Correlated	IG	−3.96	[−4.37, −3.56]	$< 10^{-16}$
Correlated	ReliefF	+7.46	[+7.17, +7.76]	$< 10^{-16}$
Correlated	SU	−3.94	[−4.35, −3.53]	$< 10^{-16}$
Correlated	mRMR	+9.35	[+9.06, +9.65]	$< 10^{-16}$
Heavy-tailed	Boruta	−1.88	[−2.09, −1.68]	$< 10^{-16}$
Heavy-tailed	IG	−6.55	[−6.90, −6.20]	$< 10^{-16}$
Heavy-tailed	ReliefF	+0.82	[+0.55, +1.09]	< 0.001
Heavy-tailed	SU	−6.55	[−6.90, −6.20]	$< 10^{-16}$
Heavy-tailed	mRMR	+5.71	[+5.41, +6.00]	$< 10^{-16}$
Skewed	Boruta	−1.00	[−1.18, −0.82]	$< 10^{-16}$
Skewed	IG	−0.47	[−0.67, −0.27]	< 0.001
Skewed	ReliefF	+2.76	[+2.51, +3.00]	$< 10^{-16}$
Skewed	SU	+0.01	[−0.21, +0.23]	1.000
Skewed	mRMR	+8.61	[+8.37, +8.85]	$< 10^{-16}$
Mixture	Boruta	−1.87	[−2.09, −1.65]	$< 10^{-16}$
Mixture	IG	−6.03	[−6.37, −5.70]	$< 10^{-16}$
Mixture	ReliefF	+2.14	[+1.84, +2.44]	$< 10^{-16}$
Mixture	SU	−6.02	[−6.36, −5.68]	$< 10^{-16}$
Mixture	mRMR	+7.12	[+6.85, +7.39]	$< 10^{-16}$
n=100, p=500				
Mean shift	Boruta	+1.59	[+1.41, +1.76]	$< 10^{-16}$
Mean shift	IG	+2.00	[+1.83, +2.18]	$< 10^{-16}$
Mean shift	ReliefF	+6.45	[+6.23, +6.66]	$< 10^{-16}$
Mean shift	SU	+2.77	[+2.58, +2.96]	$< 10^{-16}$
Mean shift	mRMR	+6.79	[+6.55, +7.03]	$< 10^{-16}$
Correlated	Boruta	+1.05	[+0.88, +1.22]	$< 10^{-16}$
Correlated	IG	+0.27	[+0.03, +0.51]	0.043
Correlated	ReliefF	+7.90	[+7.62, +8.18]	$< 10^{-16}$
Correlated	SU	+0.67	[+0.40, +0.94]	< 0.001
Correlated	mRMR	+10.31	[+10.06, +10.55]	$< 10^{-16}$
Heavy-tailed	Boruta	−2.66	[−2.86, −2.46]	$< 10^{-16}$
Heavy-tailed	IG	−2.75	[−3.01, −2.49]	$< 10^{-16}$
Heavy-tailed	ReliefF	−0.78	[−1.03, −0.54]	< 0.001
Heavy-tailed	SU	−2.34	[−2.63, −2.05]	$< 10^{-16}$
Heavy-tailed	mRMR	+5.53	[+5.24, +5.81]	$< 10^{-16}$
Skewed	Boruta	+1.70	[+1.56, +1.84]	$< 10^{-16}$
Skewed	IG	+2.04	[+1.90, +2.18]	$< 10^{-16}$
Skewed	ReliefF	+2.40	[+2.26, +2.55]	$< 10^{-16}$
Skewed	SU	+2.32	[+2.18, +2.46]	$< 10^{-16}$
Skewed	mRMR	+6.80	[+6.62, +6.98]	$< 10^{-16}$
Mixture	Boruta	−2.64	[−2.80, −2.47]	$< 10^{-16}$
Mixture	IG	−2.18	[−2.38, −1.99]	$< 10^{-16}$
Mixture	ReliefF	+1.41	[+1.18, +1.64]	$< 10^{-16}$
Mixture	SU	−1.86	[−2.07, −1.65]	$< 10^{-16}$
Mixture	mRMR	+5.23	[+5.00, +5.45]	$< 10^{-16}$
n=200, p=500				
Mean shift	Boruta	+3.20	[+3.11, +3.29]	$< 10^{-16}$
Mean shift	IG	+3.50	[+3.41, +3.59]	$< 10^{-16}$
Mean shift	ReliefF	+5.63	[+5.47, +5.79]	$< 10^{-16}$
Mean shift	SU	+3.53	[+3.44, +3.62]	$< 10^{-16}$
Mean shift	mRMR	+2.72	[+2.61, +2.83]	$< 10^{-16}$
Correlated	Boruta	+2.74	[+2.61, +2.88]	$< 10^{-16}$
Correlated	IG	+3.09	[+2.94, +3.24]	$< 10^{-16}$

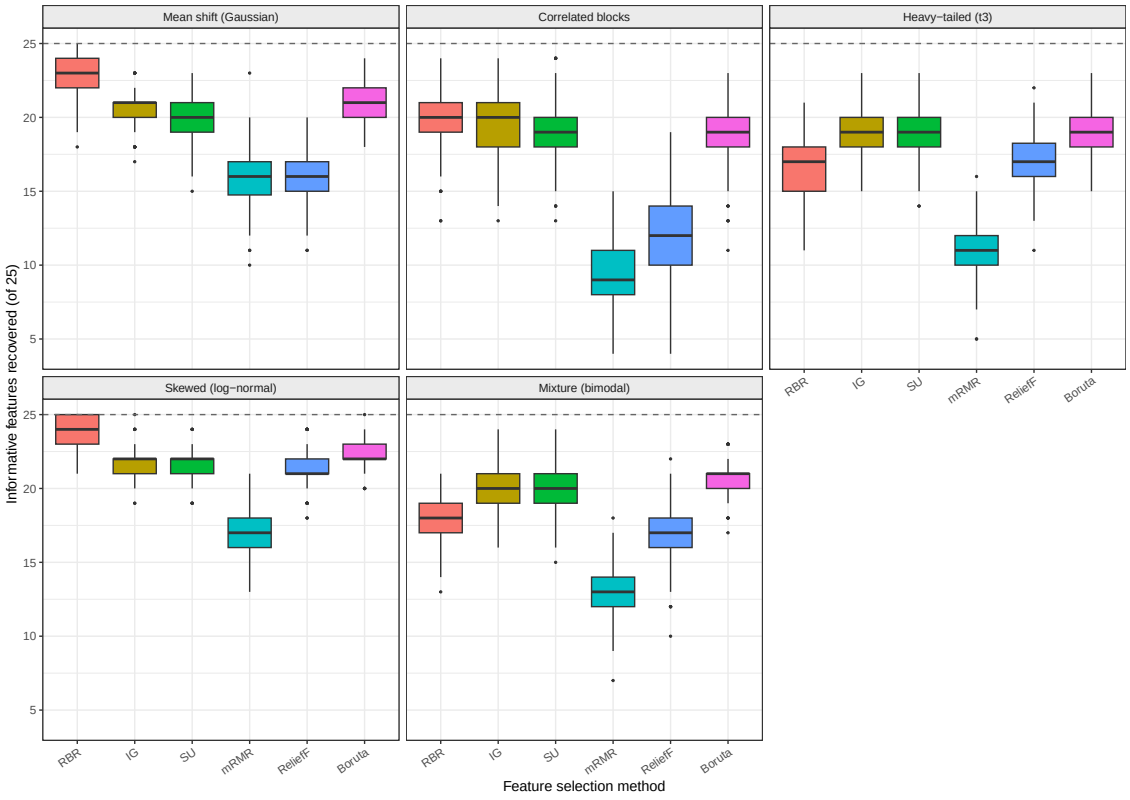


Figure A2: Distribution of informative features recovered at the baseline configuration ($n = 100$, $p = 500$); the dashed line marks perfect recovery.

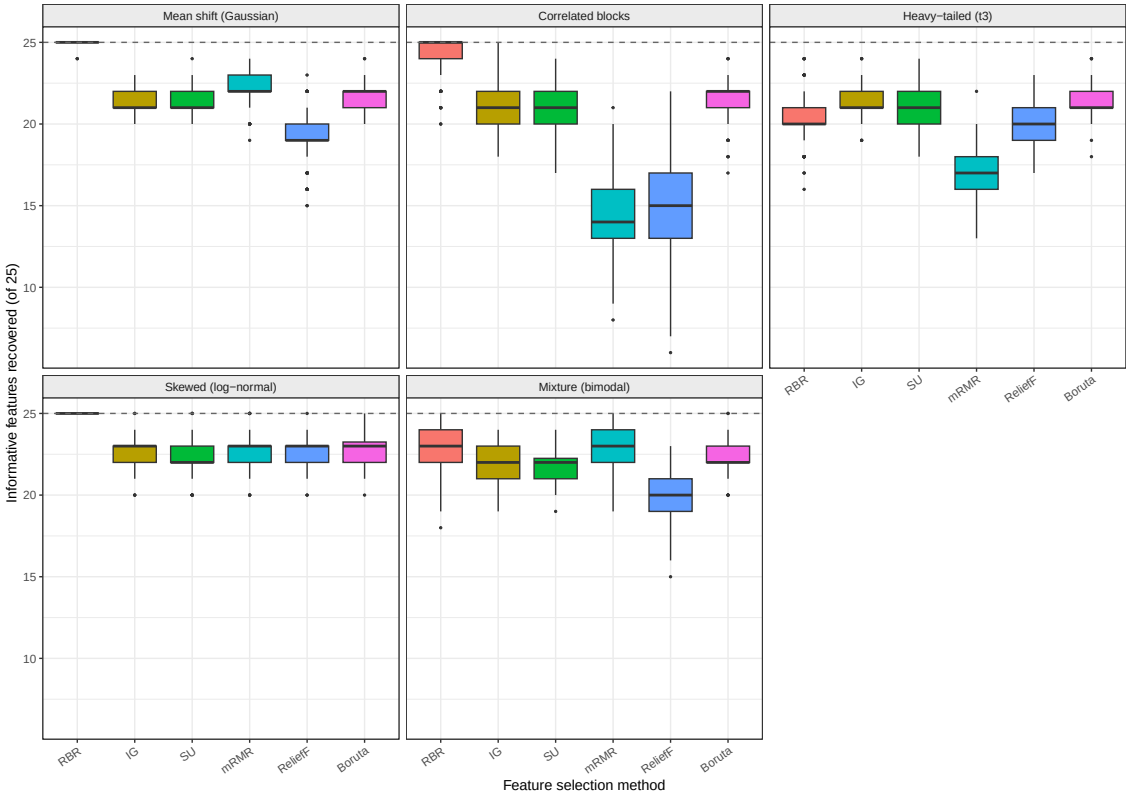


Figure A3: Distribution of informative features recovered at $n = 200$, $p = 500$.

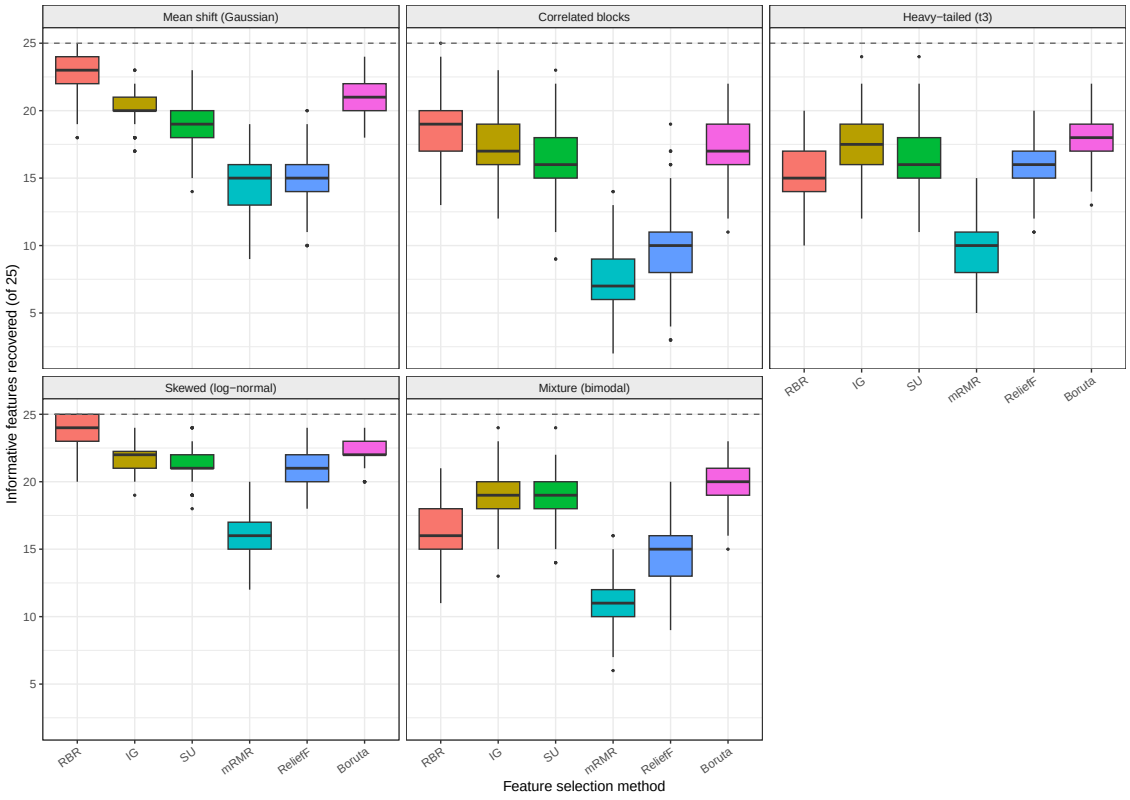


Figure A4: Distribution of informative features recovered at $n = 100$, $p = 1000$.

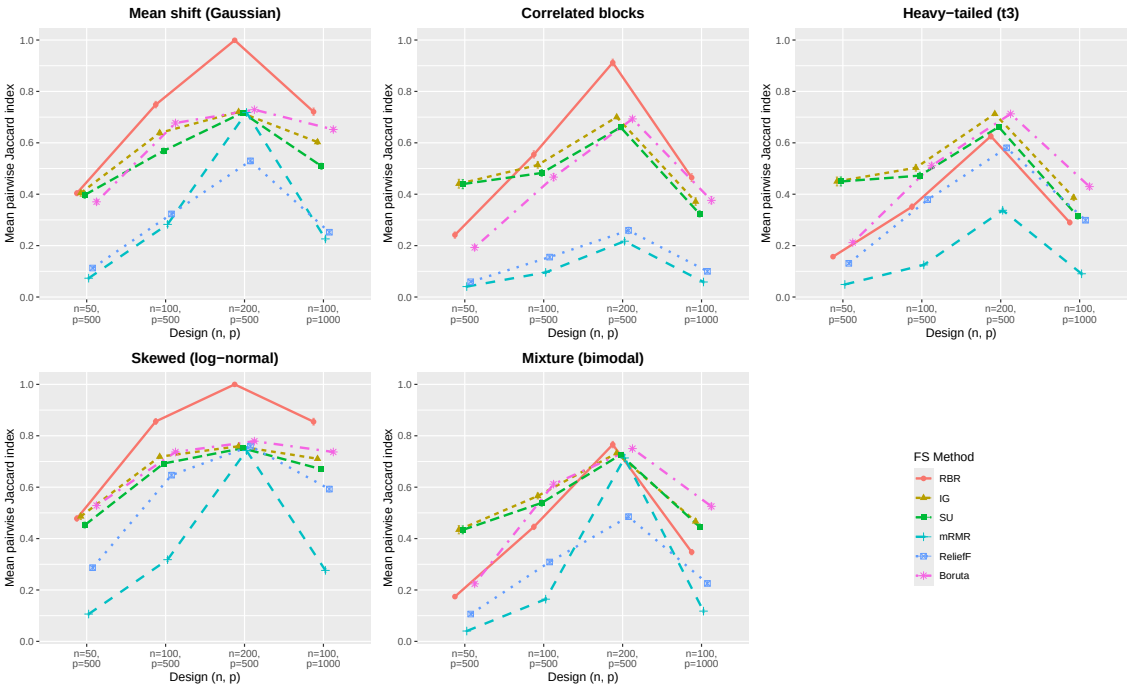


Figure A5: Mean pairwise Jaccard index across the four configurations, with 95% bootstrap intervals.

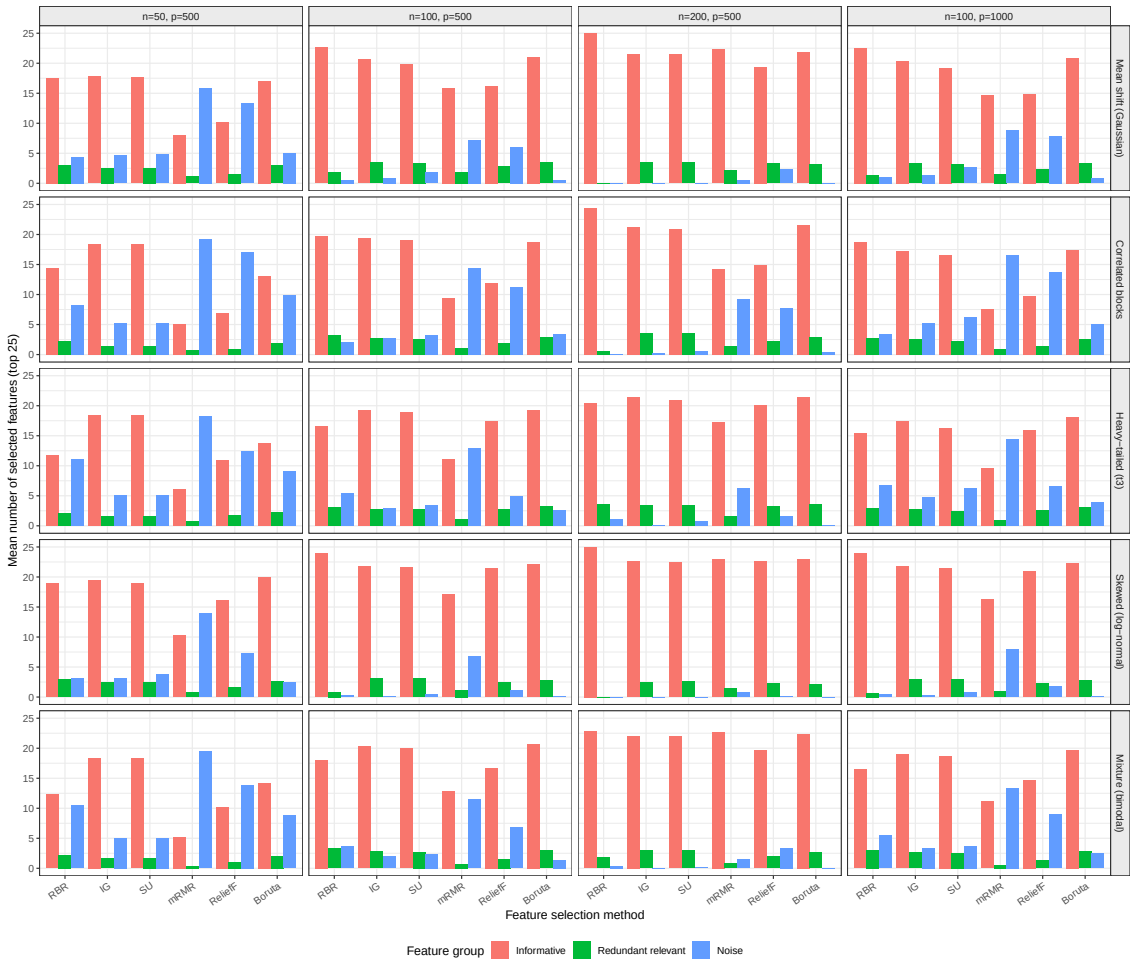


Figure A6: Mean composition of the top 25 selected set, for every scenario (rows) and configuration (columns).

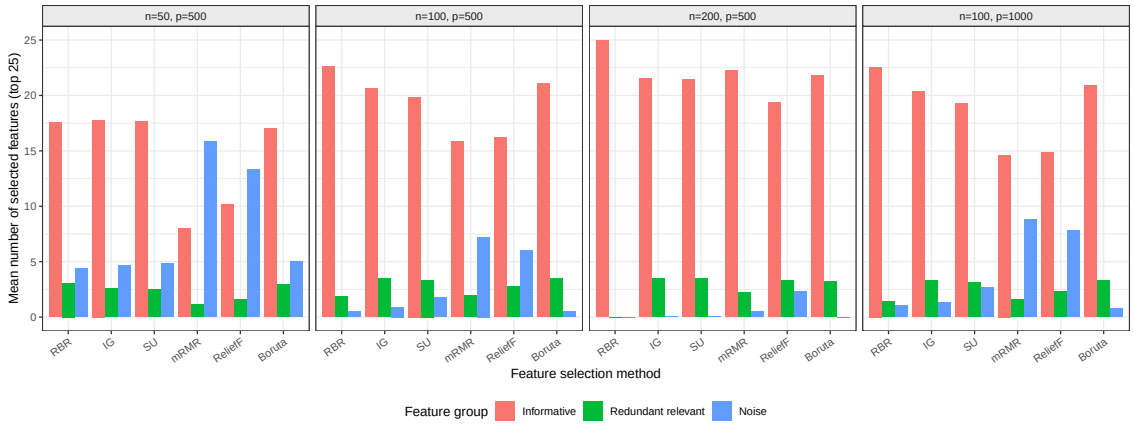


Figure A7: Top 25 composition across the four configurations: independent Gaussian mean shift.

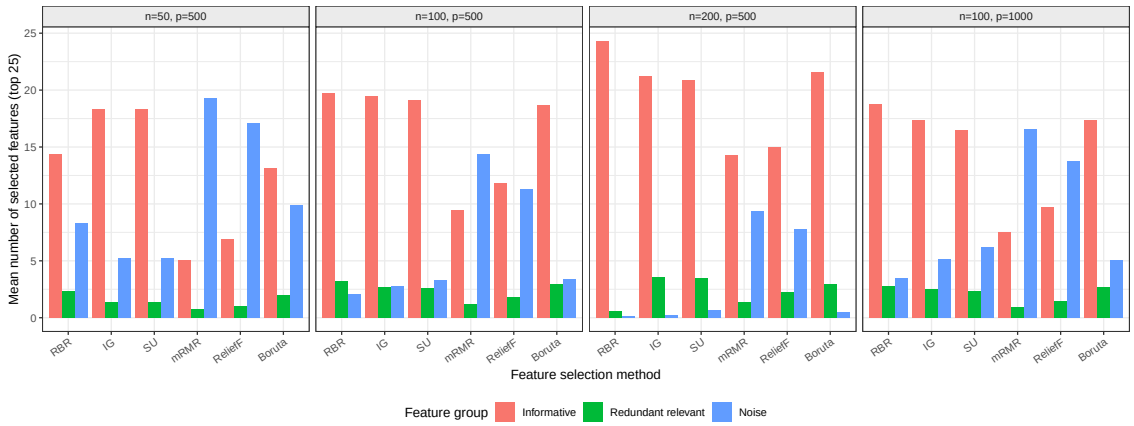


Figure A8: Top 25 composition across the four configurations: block-correlated Gaussian signal.

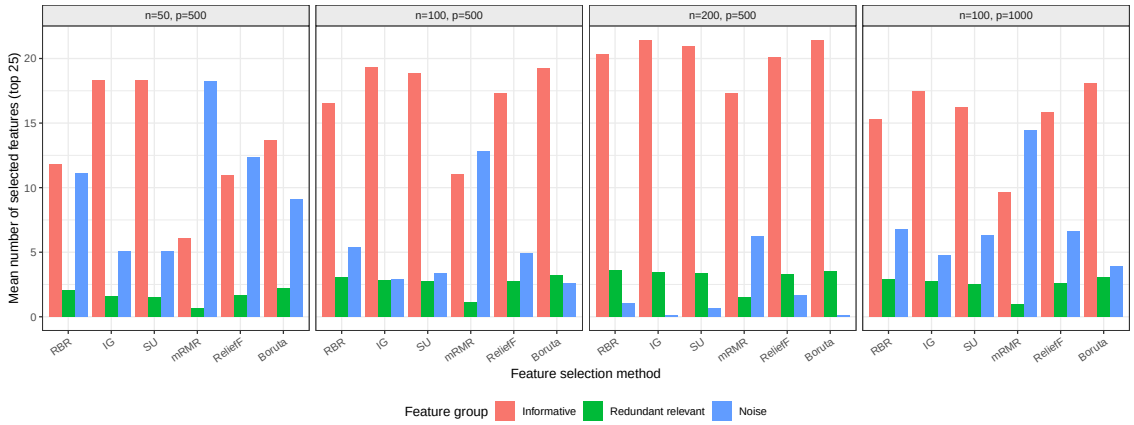


Figure A9: Top 25 composition across the four configurations: heavy-tailed mean shift.

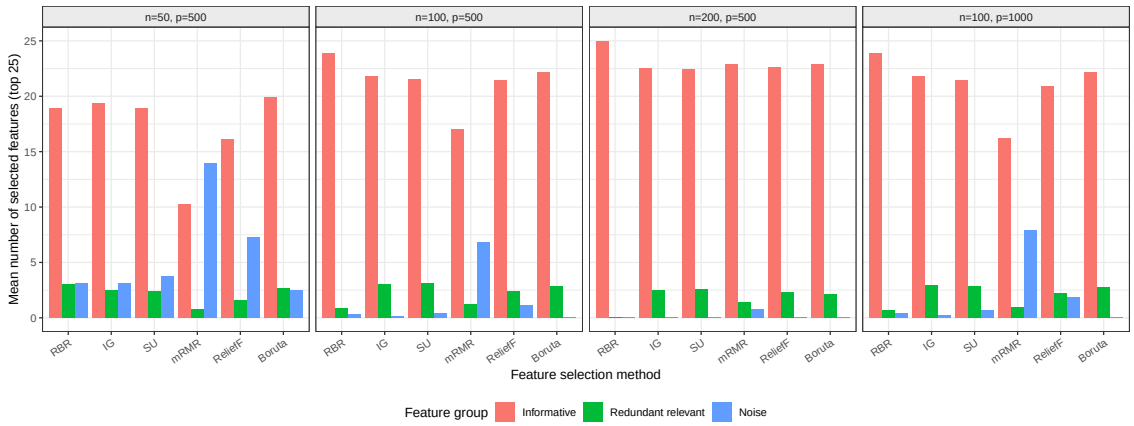


Figure A10: Top 25 composition across the four configurations: right-skewed mean shift.

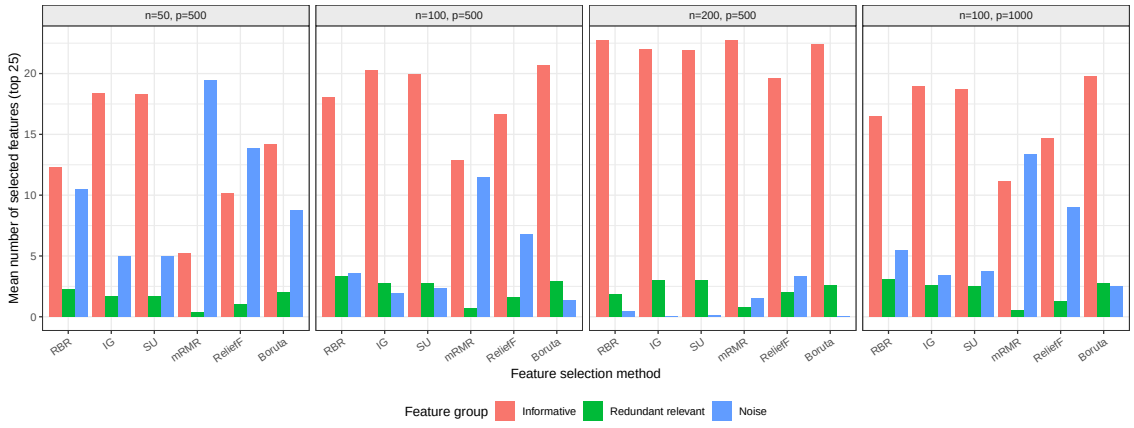


Figure A11: Top 25 composition across the four configurations: within-class mixture mean shift.