Foroughi pour and Dalton *BMC Bioinformatics* 2018, **19**(Suppl 3):70
https://doi.org/10.1186/s12859-018-2059-8

BMC Bioinformatics

RESEARCH Open Access

Heuristic algorithms for feature selection under Bayesian models with block-diagonal covariance structure

Ali Foroughi pour[1*] and Lori A. Dalton[1,2]

From The Fourth International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2017)
Boston, MA, USA. 20 August 2017

Abstract

Background: Many bioinformatics studies aim to identify markers, or features, that can be used to discriminate between distinct groups. In problems where strong individual markers are not available, or where interactions between gene products are of primary interest, it may be necessary to consider combinations of features as a marker family. To this end, recent work proposes a hierarchical Bayesian framework for feature selection that places a prior on the set of features we wish to select and on the label-conditioned feature distribution. While an analytical posterior under Gaussian models with block covariance structures is available, the optimal feature selection algorithm for this model remains intractable since it requires evaluating the posterior over the space of all possible covariance block structures and feature-block assignments. To address this computational barrier, in prior work we proposed a simple suboptimal algorithm, 2MNC-Robust, with robust performance across the space of block structures. Here, we present three new heuristic feature selection algorithms.

Results: The proposed algorithms outperform 2MNC-Robust and many other popular feature selection algorithms on synthetic data. In addition, enrichment analysis on real breast cancer, colon cancer, and Leukemia data indicates they also output many of the genes and pathways linked to the cancers under study.

Conclusions: Bayesian feature selection is a promising framework for small-sample high-dimensional data, in particular biomarker discovery applications. When applied to cancer data these algorithms outputted many genes already shown to be involved in cancer as well as potentially new biomarkers. Furthermore, one of the proposed algorithms, SPM, outputs blocks of heavily correlated genes, particularly useful for studying gene interactions and gene networks.

Keywords: Feature selection, Bayesian learning, Biomarker discovery, Heuristic search algorithms

Background

Many bioinformatics studies aim to identify predictive biomarkers that can be used to establish diagnosis or prognosis, or to predict a drug response [1–3]. This problem can often be framed as a feature selection task, where the goal is to identify a list of features (molecular biomarkers) that can discriminate between groups of interest based on high-dimensional data from microarray, RNA-seq, or other high-throughput technologies.

Initially, exploratory studies are often conducted on small samples to generate a shortlist of biomarker candidates before a large-sample validation study is performed [4]. However, such studies have too often been unsuccessful at producing reliable and reproducible biomarkers [5]. Biomarker discovery is inherently difficult, given the large number of features, highly complex interactions between genes and gene products, enormous variety of dysfunctions that can occur, and many sources of error in the data. As a result, feature selection algorithms are

*Correspondence: foroughipour.1@osu.edu
[1]Department of Electrical and Computer Engineering, The Ohio State University, 2015 Neil Avenue, 43210 Columbus, Ohio, USA
Full list of author information is available at the end of the article

often implemented without much consideration of the particular demands of the problem. For instance, variants of t-test are perhaps the most widely implemented selection strategies in bioinformatics, but can only detect strong individual features, and fail to take correlations into account.

Given that molecular signaling is often inherently multivariate, there is a need for methods that can account for correlations and extract combinations of features as a marker family. Wrapper methods do this by ranking sets of features according to some objective function, usually the error of a classifier. However, methods based on classifier error are computationally expensive, and may not necessarily produce the best markers; indeed, strong features can be excluded if they are correlated with other strong features. Furthermore, analysis downstream from feature selection may include gene set enrichment analysis, where the hope is to identify known pathways or other biological mechanisms that contain a statistically significant number of genes in the reported gene set, or may involve the development of new pathways and gene networks. We are thus motivated to develop methods that not only select markers useful for discrimination, but select all relevant markers, even individually weak ones.

To address this, in prior work we proposed a hierarchical Bayesian framework for feature selection, labeling features as "good" or "bad", where good features are those we wish to select, i.e., biomarkers. This framework places a prior on the set of good features and the underlying distribution parameters. Three Gaussian models have been considered. Under independent features, Optimal Bayesian Filtering reports a feature set of a given size with a maximal expected number of truly good features (CMNC-OBF) [6]. Assuming fully dependent good features and independent bad features, 2MNC-DGIB is a fast suboptimal method that ranks features by evaluating all sets of size 2 [7]. Finally, assuming good and bad features are separately dependent, 2MNC-Robust proposes an approximation of the posterior on good features and uses a ranking strategy similar to 2MNC-DGIB to select features [8].

While 2MNC-DGIB has outstanding performance when its assumptions are satisfied [7], it performs poorly when bad features are dependent [9]. On the other hand, CMNC-OBF and 2MNC-Robust have been shown to have robust performance across Bayesian models with block-diagonal covariances [9]. CMNC-OBF is extremely fast and enjoys particularly excellent performance when markers are individually strong with low correlations, but, like all filter methods, may miss weak features that are of interest due to high correlations with strong features [6, 9]. 2MNC-Robust is computationally very manageable and generally improves upon CMNC-OBF in the presence of correlations.

Although CMNC-OBF and 2MNC-Robust are robust to different block-diagonal covariance structures, they do not attempt to detect these underlying structures, and their assumptions and approximations constrain performance. Thus, in this work we propose three new feature selection algorithms that: (1) use an iterative strategy to update the approximate posterior used in 2MNC-Robust, (2) use a novel scoring function inspired by Bayes factors to improve overall rankings, and (3) attempt to actually detect the underlying block structure of the data. We show that these algorithms have comparable computation time to 2MNC-Robust, while outperforming 2MNC-Robust and many other popular feature selection algorithms on a synthetic Bayesian model assuming block-diagonal covariance matrices, and a synthetic microarray data model. Finally, we apply the proposed algorithms and CMNC-OBF to breast cancer, colon cancer, and AML datasets, and perform enrichment analysis on each to address validation.

Feature selection model

We review a hierarchical Bayesian model that serves as a reference for the approximate posterior developed in 2MNC-Robust [8, 9] and will be used in the algorithms we present in the next section.

Consider a binary feature selection problem with class labels $y = 0, 1$. Let F be the set of feature indices. Assume features are partitioned into blocks, where features in each block are dependent, but features in different blocks are independent. Assume each block is either good or bad. A good block has different class-conditioned distributions between the two classes, while a bad block has the same distribution in both classes. We denote a partitioning of F to good and bad blocks by $P = (P_G, P_B)$, and hereafter call it a feature partition, where $P_G = \{G_1, \cdots, G_u\}$ is the set of u good blocks and $P_B = \{B_1, \cdots, B_v\}$ is the set of v bad blocks. Furthermore, denote the set of all features in good blocks as good features, $G = \cup_{i=1}^{u} G_i$, and denote all features in bad blocks as bad features, $B = \cup_{j=1}^{v} B_j$. Denote the random feature partition by $\bar{P} = (\bar{P}_G, \bar{P}_B)$, the random set of good features by $\bar{G}$, and the random set of bad features by $\bar{B}$.

We define $\pi(P) = \mathrm{P}(\bar{P} = P)$ to be the prior distribution on $\bar{P}$. Let P be fixed. Let θ^P be the parameter describing the joint feature distribution of P. Since blocks are independent of each other we can write $\theta^P = \left[\theta_0^{G_1}, \cdots, \theta_0^{G_u}, \theta_1^{G_1}, \cdots, \theta_1^{G_u}, \theta^{B_1}, \cdots, \theta^{B_v}\right]$, where $\theta_y^{G_i}$ parametrizes class-y features in G_i, and θ^{B_j} parametrizes features in B_j. Assume $\theta_y^{G_i}$ and θ^{B_j}'s are independent given P, i.e., $\pi\left(\theta^P\right) = \prod_{i=1}^{u} \pi\left(\theta_0^{G_i}\right) \pi\left(\theta_1^{G_i}\right) \prod_{j=1}^{v} \pi\left(\theta^{B_j}\right)$.

Given a training set, S, of n independent and identically distributed (i.i.d.) points, with n_y points in each class,

we have $\pi^*\left(\theta_y^{G_i}\right) \propto \pi\left(\theta_y^{G_i}\right) f\left(\mathcal{S}_y^{G_i}|\theta_y^{G_i}\right)$ and $\pi^*\left(\theta^{B_j}\right) \propto \pi\left(\theta^{B_j}\right) f\left(\mathcal{S}^{B_j}|\theta^{B_j}\right)$, where $\pi^*(.)$ denotes posterior, $\mathcal{S}_y^{G_i}$ and $\mathcal{S}^{B_j}$ are class-y points in G_i and points in B_j, respectively, and $f\left(\mathcal{S}_y^{G_i}|\theta_y^{G_i}\right) = \prod_{x\in\mathcal{S}_y^{G_i}} f\left(x|\theta_y^{G_i}\right)$ and $f\left(\mathcal{S}^{B_j}|\theta^{B_j}\right) = \prod_{x\in\mathcal{S}^{B_j}} f\left(x|\theta^{B_j}\right)$ are the likelihoods. Following steps in [7, 10], we have

$$\begin{aligned}\pi^*(P) \propto \pi(P) &\prod_{i=1}^{u}\int \pi\left(\theta_0^{G_i}\right) f\left(\mathcal{S}_0^{G_i}|\theta_0^{G_i}\right) d\theta_0^{G_i} \\ &\times \prod_{i=1}^{u}\int \pi\left(\theta_1^{G_i}\right) f\left(\mathcal{S}_1^{G_i}|\theta_1^{G_i}\right) d\theta_1^{G_i} \\ &\times \prod_{j=1}^{v}\int \pi\left(\theta^{B_j}\right) f\left(\mathcal{S}^{B_j}|\theta^{B_j}\right) d\theta^{B_j}. \qquad (1)\end{aligned}$$

In addition, the marginal posterior of a feature set G is $\pi^*(G) = \mathrm{P}(\bar{G} = G|\mathcal{S}) = \sum_{P:G=\cup P_G} \pi^*(P)$, and marginal posterior of a feature f is $\pi^*(f) = \mathrm{P}(f \in \bar{G}|\mathcal{S}) = \sum_{P:f\in\cup P_G} \pi^*(P)$. Note $\pi^*(f) = \mathrm{P}(f \in \bar{G}|\mathcal{S})$ is different than $\pi^*(\{f\}) = \mathrm{P}(\bar{G} = \{f\}|\mathcal{S})$.

Gaussian model

Here we solve Eq. (1) for jointly Gaussian features. We assume for a block A, $\theta_y^A = \left[\mu_y^A, \Sigma_y^A\right]$ and $\theta^A = \left[\mu^A, \Sigma^A\right]$, where μ_y^A and μ^A are the mean vectors, and Σ_y^A and Σ^A are the covariance matrices.

Let P be a feature partition. Suppose A is a good block of P. Assume $\pi(\theta_y^A)$ is Normal-Inverse-Wishart (NIW). Hence, $\pi\left(\theta_y^A\right) = \pi\left(\Sigma_y^A\right)\pi\left(\mu_y^A|\Sigma_y^A\right)$, where

$$\pi\left(\Sigma_y^A\right) = K_y^A|\Sigma_y^A|^{-\frac{\kappa_y^A+|A|+1}{2}}\exp\left(-0.5\mathrm{Tr}\left(S_y^A\left(\Sigma_y^A\right)^{-1}\right)\right),$$

$$\begin{aligned}\pi\left(\mu_y^A|\Sigma_y^A\right) &= L_y^A|\Sigma_y^A|^{-0.5} \\ &\times \exp\left(-0.5\nu_y^A\left(\mu_y^A - m_y^A\right)^T\left(\Sigma_y^A\right)^{-1}\left(\mu_y^A - m_y^A\right)\right),\end{aligned}$$

where for a matrix $|.|$ denotes determinant. S_y^A, κ_y^A, m_y^A, and ν_y^A are hyperparameters, which are assumed given and fixed. S_y^A is an $|A| \times |A|$ matrix, where for a set $|.|$ denotes cardinality. For a proper prior S_y^A is symmetric and positive-definite, and $\kappa_y^A > |A| - 1$. If $\kappa_y^A > |A| + 1$, then $E\left(\Sigma_y^A\right) = S_y^A/\left(\kappa_y^A - |A| - 1\right)$. Furthermore, m_y^A is an $|A| \times 1$ vector describing the average mean of features and for a proper prior we need $\nu_y^A > 0$. K_y^A and L_y^A represent the relative weights of each distribution. For a proper distribution we have $K_y^A = |S_y^A|^{0.5\kappa_y^A} 2^{-0.5\kappa_y^A|A|}/\Gamma_{|A|}\left(0.5\kappa_y^A\right)$ and $L_y^A = \left(2\pi/\nu_y^A\right)^{-0.5|A|}$, where Γ_d denotes the multivariate gamma function.

Since NIW is a conjugate prior of Gaussian distribution, given sample, $\pi^*\left(\theta_y^A\right)$ is again NIW with updated hyperparameters: $\kappa_y^{A^*} = \kappa_y^A + n_y$, $\nu_y^{A^*} = \nu_y^A + n_y$, $m_y^{A^*} = \frac{\nu_y^A m_y^A + n_y\hat{\mu}_y^A}{\nu_y^{A^*}}$, and

$$S_y^{A^*} = S_y^A + (n_y - 1)\hat{\Sigma}_y^A + \frac{\nu_y^A n_y}{\nu_y^A + n_y}\left(\hat{\mu}_y^A - m_y^A\right)\left(\hat{\mu}_y^A - m_y^A\right)^T,$$

where $\hat{\mu}_y^A$ and $\hat{\Sigma}_y^A$ are class-conditioned sample mean and covariance of $\mathcal{S}_y^A$, respectively [11]. Now suppose A is a bad block. We assume the prior on θ^A is NIW with hyperparameters S^A, κ^A, m^A, and ν^A, and relative weights K^A and L^A. Given sample, $\pi^*\left(\theta^A\right)$ is NIW with $\kappa^{A^*} = \kappa^A + n$, $\nu^{A^*} = \nu^A + n$, $m^{A^*} = \frac{\nu^A m^A + n\hat{\mu}^A}{\nu^{A^*}}$, and

$$S^{A^*} = S^A + (n - 1)\hat{\Sigma}^A + \frac{\nu^A n}{\nu^A + n}\left(\hat{\mu}^A - m^A\right)\left(\hat{\mu}^A - m^A\right)^T,$$

where $\hat{\mu}^A$ and $\hat{\Sigma}^A$ are sample mean and covariance of $\mathcal{S}^A$, respectively [11]. As long as $\pi^*(P)$ is proper, using the normalization constant of NIW distribution to compute the integrals in Eq. (1) we have

$$\begin{aligned}\pi^*(P) \propto \pi(P) &\prod_{i=1}^{u} Q_0^{G_i} Q_1^{G_i} \left|S_0^{G_i^*}\right|^{-0.5\kappa_0^{G_i^*}} \left|S_1^{G_i^*}\right|^{-0.5\kappa_1^{G_i^*}} \\ &\times \prod_{j=1}^{v} Q^{B_j}|S^{B_j^*}|^{-0.5\kappa^{B_j^*}},\end{aligned}$$

where

$$Q_y^A = K_y^A L_y^A 2^{0.5\kappa_y^{A^*}|A|}\Gamma_{|A|}\left(0.5\kappa_y^{A^*}\right)\left(2\pi/\nu_y^{A^*}\right)^{0.5|A|},$$

$$Q^A = K^A L^A 2^{0.5\kappa^{A^*}|A|}\Gamma_{|A|}\left(0.5\kappa^{A^*}\right)\left(2\pi/\nu^{A^*}\right)^{0.5|A|}.$$

Assuming: (1) $\pi(P)$ is such that the block structure, i.e., the number and size of good and bad blocks, is fixed, (2) for each good block A, K_y^A, L_y^A, κ_y^A, and ν_y^A do not depend on the features indices in A, and (3) for each bad block A, K^A, L^A, κ^A, and ν^A do not depend on the features indices in A,

$$\pi^*(P) \propto \pi(P)\left(\prod_{i=1}^{u}\left|S_0^{G_i^*}\right|^{\kappa_0^{G_i^*}}\left|S_1^{G_i^*}\right|^{\kappa_1^{G_i^*}}\prod_{j=1}^{v}\left|S^{B_j^*}\right|^{\kappa^{B_j^*}}\right)^{-0.5}.$$

Methods

Here we describe the set selection methods used. Note we aim to find the set of true good features, rather than the true underlying feature partition. The Maximum Number Correct (MNC) criterion [7] outputs the set maximizing the expected number of correctly labeled features and the Constrained MNC (CMNC) criterion outputs the set with maximum expected number of correctly labeled features constrained to having exactly D selected features, where D

is a parameter of the optimization problem [9]. The solution of MNC is $\{f \in F : \pi^*(f) > 0.5\}$ [7] and the solution of CMNC is picking the top D features with largest $\pi^*(f)$ [9]. Therefore, both MNC and CMNC require computing $\pi^*(f)$ for all $f \in F$, which is not computationally feasible for an arbitrary block structure unless $|F|$ is very small. We review two previously proposed algorithms, OBF and 2MNC-Robust, and then present three new algorithms.

Optimal Bayesian filter

Optimal Bayesian Filter (OBF) assumes all blocks have size one, i.e., all features are independent, and assumes the events $\{f \in \bar{G}\}$ are independent a priori. In this case $\pi^*(f)$ can be found in closed form with little computation cost [6, 9]. OBF is optimal under its modeling assumptions. As argued in [9], in the presence of correlation OBF is a robust suboptimal algorithm that can detect individually strong good features, i.e., those whose mean and/or variance is very different between the two classes, but cannot take advantage of correlations to correctly label individually weak good features, those whose mean and variance are similar in both classes.

2MNC-Robust

The 2MNC algorithm [7] suggests approximating $\pi^*(f)$ using $\pi^*(G)$ for all sets G such that $|G| = 2$, and picking the top D features. Since finding $\pi^*(G)$ for all feature partitions where $| \cup P_G| = 2$ is typically infeasible, an approximate posterior, $\tilde{\pi}^*(G)$, is proposed [8], where for all $G \subseteq F$ of size 2,

$$\tilde{\pi}^*(G) \propto \tilde{\pi}(G) \frac{\int \pi\left(\theta_0^G\right) f\left(\mathbb{S}_0^G | \theta_0^G\right) d\theta_0^G \int \pi\left(\theta_1^G\right) f\left(\mathbb{S}_1^G | \theta_1^G\right) d\theta_1^G}{\int \pi\left(\theta^G\right) f\left(\mathbb{S}^G | \theta^G\right) d\theta^G}.$$

The normalization constant is found such that $\sum_{G \subseteq F:|G|=2} \tilde{\pi}^*(G) = 1$. $\tilde{\pi}(G)$ mimics the role of $\pi(G) = P(\bar{G} = G) = \sum_{P:\cup P_G = G} \pi(P)$. Using some suboptimal method might affect one's decision of the value used as the prior of a feature set, replacing $\pi^*(G)$ with $\tilde{\pi}^*(G)$. For example, knowing $|\bar{G}| > 2$ implies $\pi(G) = 0$ for all sets of size 2; however, 2MNC-Robust only evaluates such sets. In this case, $\tilde{\pi}(G) = P(G \subseteq \bar{G})$ might be a suitable choice to replace $\pi(G)$. $\tilde{\pi}^*(f) = \sum_{G:f \in G} \tilde{\pi}^*(G)$ is the approximate marginal posterior of $f \in F$. For the Gaussian model, if the number of good features is fixed and hyperparameters do not depend on the feature indices,

$$\tilde{\pi}^*(G) \propto \tilde{\pi}(G) \left(|S_0^{G^*}|^{\kappa_0^{G^*}} |S_1^{G^*}|^{\kappa_1^{G^*}} / |S^{G^*}|^{\kappa^{G^*}} \right)^{-0.5}. \quad (2)$$

2MNC-Robust is implementing 2MNC with $\tilde{\pi}^*(f)$. As mentioned before, 2MNC-Robust does not tune itself to the underlying block structure of data.

Recursive marginal posterior inflation

It is easy to show that $\sum_{f \in F} \tilde{\pi}^*(f) = 2$ when only sets of size 2 are used to find $\tilde{\pi}^*(f)$. Hence, under MNC criterion one would at most pick 4 good features, implying we underestimate $\pi^*(f)$ by only using sets of size 2 when $|\bar{G}| >> 2$. REcursive MArginal posterior INflation (REMAIN) aims to sequentially detect good features by rescaling $\tilde{\pi}^*(f) = \sum_{G:f \in G, |G|=2} \tilde{\pi}^*(G)$. We initialize REMAIN with the set of all features, $F_r = F$. Then, REMAIN uses the MArginal posterior INflation (MAIN) algorithm to identify several features as good, removes them from F_r, and feeds MAIN with the truncated F_r to select additional features. This process iterates until MAIN does not output any features. REMAIN is nothing but repetitive calls to MAIN with shrinking feature sets, making MAIN the heart of this algorithm.

Algorithm 1 Pseudo-code of MAIN: $\tilde{G} = MAIN\ (F_t, T_1, T_2)$

Require: feature index set F_t, and threshold values T_1 and T_2.

$\tilde{G} := \phi$.
$\tilde{\pi}^*(f) := \sum_{G:f \in G, |G|=2} \tilde{\pi}^*(G)$ for all $f \in F_t$.
do
  $G_s := \{f \in F_t : \tilde{\pi}^*(f) > T_1\}$.
  $\tilde{G} := \tilde{G} \cup G_s$.
  Update $F_t := F_t \backslash G_s$.
  For all $f \in G_s$ set $\tilde{\pi}^*(f) := 0$.
  $sum := \sum_{f \in F_t} \tilde{\pi}^*(f)$.
  For all $f \in F_t$ set $\tilde{\pi}^*(f) := 2\tilde{\pi}^*(f)/sum$.
while $G_s \neq \phi$ **&** $\max_{G \subseteq F_t, |G|=2} H(G) > T_2$.

Ensure: $\tilde{G}$.

Pseudo-code of MAIN is provided in Algorithm 1, where $H(G)$ is the right hand side of Eq. (2). Inputted with a feature set F_t, MAIN finds $\tilde{\pi}^*(f)$ using sets of size 2, and finds the set $G_s = \{f \in F_t : \tilde{\pi}^*(f) > T_1\}$. MAIN adds G_s to $\tilde{G}$, the set of features in F_t already labeled as good. It then updates F_t to $F_t \backslash G_s$, and rescales $\tilde{\pi}^*(f)$ of features $f \in F_t$ so that $\sum_{f \in F_t} \tilde{\pi}^*(f) = 2$. Note features in $\tilde{G}$ are used to compute $\tilde{\pi}^*(f)$ for features $f \in F_t$, but $\tilde{\pi}^*(f)$ of features $f \in \tilde{G}$ are not used in the scaling of $\sum_{f \in F_t} \tilde{\pi}^*(f) = 2$. MAIN iterates until $\tilde{G} = \phi$, or $H(G) \leq T_2$ for all $G \subseteq F_t$ with $|G| = 2$.

Not finding new features in MAIN might be due to the remaining good features being weaker and independent of $\tilde{G}$. Hence, REMAIN removes $\tilde{G}$ from F_r, and feeds MAIN with the updated F_r. This way, features in $\tilde{G}$ are not used to compute $\tilde{\pi}^*(f)$ anymore for any feature $f \in F_r$, thus making it easier to detect weaker good features that are independent of features already selected by REMAIN. Pseudo-code of REMIAN is provided in Algorithm 2.

T_1 mimics the role of the threshold used in the MNC criterion. Hence, $T_1 \in [0, 1]$. Recall that by evaluating sets

of size 2 we underestimate $\tilde{\pi}^*(f)$ when $|\bar{G}| >> 2$. Therefore, when confident $|\bar{G}| >> 2$, one might opt for smaller values for T_1 rather than values close to 1. As T_2 is a threshold over un-normalized posteriors, $H(G)$, extra care must be taken when setting T_2. We suggest $T_2 = n$ for high-dimensional feature selection applications, which is a good rule of thumb based on our simulation results and asymptotic analysis of $H(.)$.

Algorithm 2 Pseudo-code of REMAIN

Require: feature index set F, and threshold values T_1 and T_2.

$\hat{G} := \phi$.
$F_r := F$.
do
  $\tilde{G} := MAIN(F_r, T_1, T_2)$.
  $\hat{G} := \hat{G} \cup \tilde{G}$.
  $F_r := F_r \backslash \tilde{G}$.
while $\tilde{G} \neq \phi$.

Ensure: $\hat{G}$.

Note the number of features reported by REMAIN is variable; however, one can easily obtain close to a desired number of selected features by tuning T_1. To illustrate, we provide an example based on the data generation model used in the "Synthetic microarray simulations" section, where we assume there are 100 markers, i.e., good features, and 4900 non-markers, i.e., bad features. We use the synergetic model with block size $k = 5$ and correlation coefficients $\rho_0 = \rho_1 = 0.9$. Fixing $T_2 = n$, Table 1 lists the average number of markers and non-markers selected over 1000 iterations for $n = 20$ and 100 across different values of T_1. REMAIN outputs very few features when T_1 is large, and too many features when T_1 is extremely low. The best choice for T_1 can vary greatly from case to case, but one strategy is to choose T_1 so that REMAIN selects close to a given number of features. For example, $T_1 = 0.05$ is a good choice in this simulation if one desires approximately 100 selected features. Another strategy is based on the number of features selected by REMAIN across various values of T_1. When T_1 is large, reducing T_1 only slightly increases the output feature size, for instance when $T_1 > 0.05$ in this simulation. However, one might observe a rapid increase in the output size by slightly reducing T_1, for instance T_1 changing from 0.05 to 0.01 in this simulation. For such observed patterns, the value for which this phenomenon occurs might be a desirable choice.

Posterior factor

Feature selection can be construed as a model selection problem where each model is a set of good features. Let f be a feature. If f is a good feature, we expect that if we add f to any model G, i.e., a set of good features, then $\tilde{\pi}^*(G \cup \{f\})/\tilde{\pi}^*(G) >> 1$. If f is a bad feature, we expect $\tilde{\pi}^*(G \cup \{f\})/\tilde{\pi}^*(G)$ to be much smaller. Hence, if we average this ratio over a family of models that do not contain f, and denote it by $\beta(f)$, we expect $\beta(f) >> 1$ if f is a good feature, and comparable to or smaller than 1 if f is a bad feature. $\tilde{\pi}^*(G \cup \{f\})/\tilde{\pi}^*(G)$ is similar, but not identical, to the *Bayes factor* encountered in model selection [12], where here we compare a model with good feature set $G \cup \{f\}$ versus a model with good feature set G excluding f. The posterior factor, $\beta(f)$, averages the approximate posterior ratio across all feature sets $G \in F \setminus \{f\}$, i.e.,

$$\beta(f) = \frac{1}{|\Omega^f|} \sum_{G \in \Omega^f} \frac{\tilde{\pi}^*(G \cup \{f\})}{\tilde{\pi}^*(G)}, \tag{3}$$

where $\Omega^f = \{G \subseteq F : f \notin G\}$. As the summation over Ω^f is computationally infeasible, we propose *approximate posterior factor*, hereafter denoted by $\tilde{\beta}$.

$$\tilde{\beta}(f) = \frac{1}{|F| - 1} \sum_{f' \in F \setminus \{f\}} \frac{\tilde{\pi}^*(\{f, f'\})}{\tilde{\pi}^*(\{f'\})}. \tag{4}$$

We propose *approximate POsterior FActor-Constrained* (POFAC) algorithm as follows: Use $\tilde{\beta}(f)$ to rank features, and pick the top D features. Note that D is a parameter of the algorithm.

Sequential partition mustering

Sequential Partition Mustering (SPM) aims to improve feature selection performance by sequentially detecting good blocks, and adding them to the set of previously selected features. To find a good block, we start with the most significant feature, i.e., the feature with largest $\tilde{\beta}$, and find the block containing it. Note we do not aim to find the structure of bad blocks, and as soon as we declare no more good features remain the algorithm terminates.

Suppose u_0 is the current most significant feature. In order to find the block containing u_0 we propose the Good

Table 1 Performance of REMAIN for various values of T_1

Objective	n	$T_1 = 0.3$	$T_1 = 0.2$	$T_1 = 0.1$	$T_1 = 0.05$	$T_1 = 0.01$	$T_1 = 0.005$
Marker found	20	0.9	1.4	2.5	4.6	15.3	26.2
Non-marker found	20	2.0	3.7	9.4	23.3	164.5	403.0
Marker found	100	52.5	56.2	58.2	60.0	68.9	76.1
Non-marker found	100	1.6	3.3	8.8	20.2	124.7	288.5

Seed Grower (GSG) algorithm, which can be construed as a seed growing algorithm with u_0 as the seed. Pseudo-code of GSG is presented in Algorithm 3, where for any two non-empty disjoint sets $G_1, G_2 \subset F$,

$$C_1(G_1,G_2) = \frac{\tilde{\pi}(G_1,G_2)}{1-\tilde{\pi}(G_1,G_2)} \frac{Q_0^{G_{12}} Q_1^{G_{12}}}{Q_0^{G_1} Q_1^{G_1} Q_0^{G_2} Q_1^{G_2}} \times \left(\frac{\left|S_0^{G_{12}^*}\right|^{\kappa_0^{G_{12}^*}} \left|S_1^{G_{12}^*}\right|^{\kappa_1^{G_{12}^*}}}{\left|S_0^{G_1^*}\right|^{\kappa_0^{G_1^*}} \left|S_1^{G_1^*}\right|^{\kappa_1^{G_1^*}} \left|S_0^{G_2^*}\right|^{\kappa_0^{G_2^*}} \left|S_1^{G_2^*}\right|^{\kappa_1^{G_2^*}}} \right)^{-0.5},$$

$G_{12} = G_1 \cup G_2$, and $\tilde{\pi}(G_1,G_2)$ approximates $\pi(G_1,G_2)$, the prior probability that at least one of the features in G_2 is not independent of G_1. Note $\pi(G_1,G_2) = \sum_{P\in\mathcal{P}} \pi(P)$, where $\mathcal{P}$ is the family of feature partitions that contain a block U such that $U \cap G_1 \neq \phi$ and $U \cap G_2 \neq \phi$. At each iteration, GSG finds the feature u^* that maximizes $C_1(U,\{u^*\})$, where U is the currently detected sub-block of the block containing u_0. GSG declares u^* and U belong to the same block if $C_1(U,\{u^*\}) > T_3$, and adjoins u^* to U; otherwise, it terminates and declares U as the block containing u_0. Here we assume $T_3 = t_1 n^{t_2|U|}$, where $t_1, t_2 > 0$ are parameters of GSG. While we have only considered one possible family of thresholds, we expect this family to be large enough for most practical purposes.

Algorithm 3 Pseudo-code of GSG: $U = GSG(u_0, F_s, t_1, t_2)$

Require: initial seed u_0, feature set index F_s, parameters t_1 and t_2.

$U := [u_0]$.
$check := 1$.
while $check = 1$ **do**
  $u^* := \text{argmax}_{u\in F_s\setminus U} C_1(U,\{u\})$.
  $T_3 := t_1 n^{t_2|U|}$.
  if $C_1(U,\{u^*\}) > T_3$ **then** $U := U \cup \{u^*\}$,
  else $check := 0$.

Ensure: U.

Pseudo-code of SPM is explained in Algorithm 4. Let F_t be the feature set used by SPM initialized to $F_t = F$. We start with the most significant feature u_0 and find the block containing it, U. We then update F_t to $F_t \setminus U$. If $\tilde{\beta}(f) < T_4$ for all $f \in F_t$, then SPM declares F_t does not contain any good features and terminates; otherwise, it picks the most significant feature of F_t and iterates. Similar to REMAIN, SPM cannot be forced to output a fixed number of features, but T_4 can be used to tune SPM to output close to a desired number of features. In addition, t_1 and t_2 can be used to avoid picking large blocks, which also affects the output feature set size.

Algorithm 4 Pseudo-code of SPM

Require: feature set F, parameters t_1 and t_2, threshold T_4.

$\hat{G} := \phi$.
$F_t := F$.
$check := 1$.
while $check = 1$ **do**
  $u_0 := \text{argmax}_{u\in F_t} \tilde{\beta}(u)$.
  if $\tilde{\beta}(u_0) > T_4$ **then**
    $U := GSG(u_0, F_t, t_1, t_2)$,
    $\hat{G} := \hat{G} \cup U$.
    $F_t := F_t \setminus U$.
  else $check := 0$.

Ensure: $\hat{G}$.

We again provide an example based on the simulation we did for REMAIN. We let $t_1 = 10^1, 10^2, \cdots, 10^5$, $t_2 = 1, 2, 3, 4, 5$, and $T_4 = 10^2, 10^4, 10^6$, and consider all combinations to construct a very wide range of parameters. Figures 1 and 2 illustrate how parameters of SPM affect its performance for $n = 20$ and 100, respectively. While low thresholds mislabel more non-markers as good features, they correctly label more markers compared with large thresholds. When $n = 20$, in order to correctly label at least 10 markers on average, at least 50 non-markers are mislabeled, and to mislabel at most 5 non-markers on average, one cannot correctly detect more than 5 markers. On the other hand, when $n = 100$, one can simultaneously correctly label at least 80 markers and mislabel at most 10 non-markers for almost all parameters. Moving from lowest parameter values to highest, we observe the average outputted feature size varies from approximately 400 to less than 5 when $n = 20$, while it only varies from 120 to 70 when $n = 100$. Thereby, $n = 100$ is less sensitive to the choice of parameters than $n = 20$. Suppose $n = 100$ and one aims to find most markers and few non-markers. Many parameters can achieve this goal. In addition, within the range of such parameters, the number of features labeled as good does not change very much by slightly varying the parameters. Hence, for a fixed sample, one can implement SPM over a wide range of parameters, find the range where the output feature size does not vary much by slightly changing the parameters, and pick a value in that region that outputs a feature set with close to a reasonable number of features.

Simulations

We compare the performance of proposed algorithms with many popular feature selection algorithms over a Bayesian setting, and a synthetic microarray model introduced in [13] and extended in [8, 9].

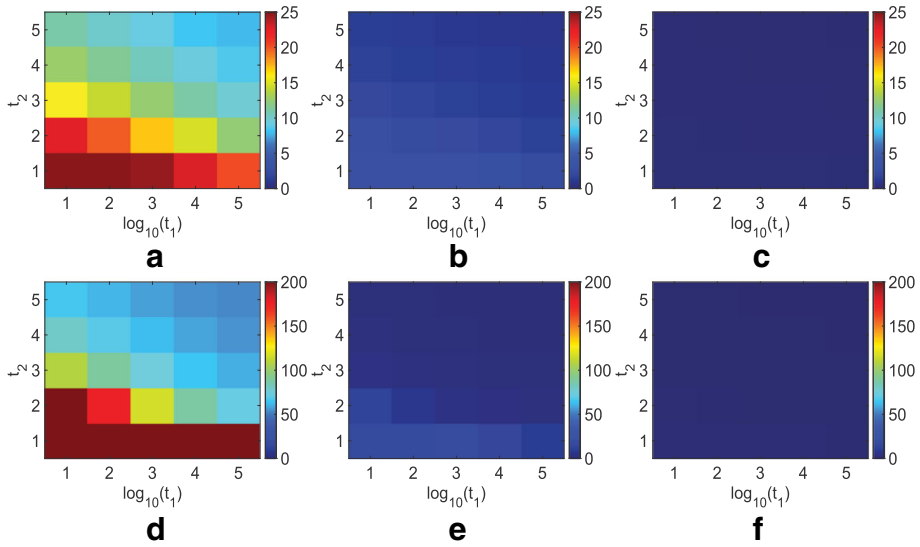

Fig. 1 Performance of SPM for various values of t_1, t_2, T_4, and $n = 20$. Average number of markers labeled as good for **a** $T_4 = 10^2$, **b** $T_4 = 10^4$, and **c** $T_4 = 10^6$. Average number of non-markers labeled as good for **d** $T_4 = 10^2$, **e** $T_4 = 10^4$, and **f** $T_4 = 10^6$

Bayesian simulation

In this simulation we assume $|F| = 4100$ and $|\tilde{G}| = 100$. We assume there is 1 good block for each of the following sizes: 10, 20, 30, and 40. We also assume there are 20 bad blocks for each of the following sizes: 5, 10, 15, 20, 50, and 100. We first randomly assign each feature to a block such that the assumed block structure is satisfied, effectively constructing $\bar{P}$. Afterwards, distribution parameters are randomly drawn from the following NIW prior. For each good block, A, we have $S_0^A = S_1^A = 0.5 \times I_{|A|\times|A|}, \kappa_0^A = \kappa_1^A = |A| + 2, m_0^A = m_1^A = 0$, and $\nu_0^A = \nu_1^A = 4$, where I is the identity matrix. Also, for a bad block, A, we have $S^A = 0.5 \times I_{|A|\times|A|}, \kappa^A = |A| + 2, m^A = 0$, and $\nu^A = 4$. Given distribution parameters, a stratified sample of size n with equal points in each class is drawn. The following feature selection methods declare the set of good features: t-test, Bhatacharrya Distance (BD), Mutual Information (MI) using the non-parameter method of [14] with spacing parameter $m = 1$, Sequential Forward Search using the bolstered error estimate [15] of Regularized Linear Discriminant Analysis applied to the top 300 features of BD (SFS-RLDA), FOrward selection

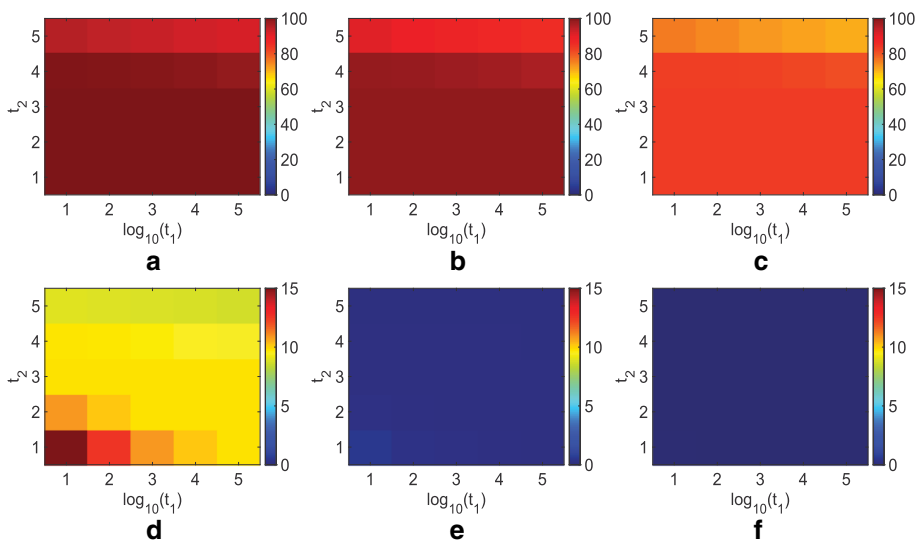

Fig. 2 Performance of SPM for various values of t_1, t_2, T_4, and $n = 100$. Average number of markers labeled as good for **a** $T_4 = 10^2$, **b** $T_4 = 10^4$, and **c** $T_4 = 10^6$. Average number of non-markers labeled as good for **d** $T_4 = 10^2$, **e** $T_4 = 10^4$, and **f** $T_4 = 10^6$

using Hilbert-Schmidt Independence Criterion (FOHSIC) [16] applied to the top 300 features of BD, CMNC-OBF, 2MNC-Robust, REMAIN, POFAC, and SPM. Note t-test, MI, BD, and CMNC-OBF are filter methods. All methods except REMAIN and SPM output $|\bar{G}|$ features. CMNC-OBF assumes the events $\{f \in \bar{G}\}$ are independent and $P(f \in \bar{G})$ is constant for all $f \in F$. 2MNC-Robust and REMAIN assume $\tilde{\pi}(G)$ is uniform over all sets of size 2, and zero otherwise. POFAC assumes $\tilde{\pi}(G)$ is uniform over all sets of size 1 and 2. Finally, SPM assumes $\tilde{\pi}(G_1, G_2) = 0.5$ for all sets $G_1, G_2 \subseteq F$, and uses the same $\tilde{\pi}(G)$ of POFAC to compute $\tilde{\beta}(f)$. Bayesian algorithms use proper priors with hyperparameters of the same form given previously (PP), and Jeffreys non-informative prior (JP), where for each set, A, S_y^A and S^A are zero matrices, $K_y^A = K^A = L_y^A = L^A = 1$, and $\kappa_y^A = \kappa^A = \nu_y^A = \nu^A = 0$. With $\nu_y^A = \nu^A = 0$ we do not need to specify m_y^A and m^A. We use $T_1 = 0.3$ and $T_2 = n$ for REMAIN using both PP and JP. For SPM-PP we set $t_1 = 100$, $t_2 = 0.5$, and $T_4 = 100n^2$, which resulted in adequate performance among all sample sizes. When using SPM-JP we use the same t_1 and T_4, but set $t_2 = 1$ to avoid picking large blocks. This process iterates 1000 times.

Figure 3 plots the average number of correctly labeled features as sample size increases from 10 to 100 in steps of 10. SPM-PP has the best performance; however, SPM-JP experiences a sharp drop under small sample sizes. For larger sample sizes, POFAC-PP performs second only to SPM-PP. However, POFAC-JP outperforms SPM-JP. REMAIN adequately balances performance across all sample sizes. All proposed algorithms, except SPM-JP, outperform 2MNC-Robust, CMNC-OBF, and other feature selection algorithms. SPM-JP outperforms previous algorithms if sample size is not very small.

In this simulation filter methods were the fastest with comparable computation time, and FOHSIC was the most computationally intensive method. A comparison of run-times for this specific simulation is provided in Table 2 assuming the run-time of 2MNC-Robust is the unit of time. Parallel processing can be used to speed up these algorithms, for instance, in the 4th step of GSG, and to compute $\tilde{\pi}^*(G)$ in 2MNC-Robust and POFAC. Although SPM is a sequential algorithm, its bottle-neck is step 4 of GSG, making parallel processing a good strategy to extensively speed up SPM.

Synthetic microarray simulations

Here an extended version of a synthetic model developed to mimic microarrays is used to generate data. The original model is introduced in [13], and has been extended in [8, 9]. In these models features are markers or non-markers. Markers are either global or heterogeneous. Global markers (GM) are homogeneous within each class. Heterogeneous Markers (HM) compromise c subclasses, where for each specific set of heterogeneous markers, a specific subset of the training sample has a different distribution than markers in class 0, and the remaining sample points have the same distribution as class 0. Markers comprise blocks of size k, where each block in class y is Gaussian with mean μ_y and covariance $\sigma_y \Sigma_y$. Diagonal elements of Σ_y are 1 and non-diagonal elements are ρ_y. The original model of [13] forced $\rho_0 = \rho_1$. We also have $\mu_0 = [0, \cdots, 0]$. There are three types of markers according to their mean in class 1: redundant, synergetic, and marginal, with μ_1 being $[1, \cdots, 1]$, $[1, 1/2, \cdots, 1/k]$, and $[1, 0, \cdots, 0]$, respectively. Non-markers are either Low Variance (LV) or High Variance (HV). In the original model LV non-markers are independent, each with a Gaussian distribution, $N(0, \sigma_0)$. However, in the extended model of [8, 9], similar to markers in class 0, LV non-markers comprise blocks of size k, where in each block features are jointly Gaussian with mean μ_0 and covariance $\sigma_0 \Sigma_0$. HV non-markers are independent with marginal

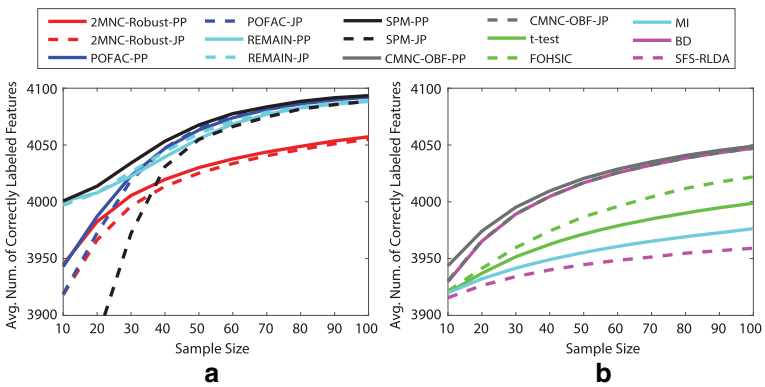

Fig. 3 Performance of various feature selection algorithms under Gaussian data. Average number of correctly labeled features versus sample size for randomly generated parameters for (**a**) 2MNC-Robust-PP, 2MNC-Robust-JP, POFAC-PP, POFAC-JP, REMAIN-PP, REMAIN-JP, SPM-PP, SPM-JP, (**b**) CMNC-OBF-PP, CMNC-OBF-JP, t-test, FOHSIC, MI, BD, and SFS-RLDA

Table 2 Run-time comparison of Bayesian simulation

Alg.	Filter	2MNC-Robust	REMAIN	POFAC	SPM	SFS-RLDA	FOHSIC
Time	$<10^{-3}$	1	2	1.05	1.5	10	15

distribution $pN(0,\sigma_0)+(1-p)N(1,\sigma_1)$, where p is drawn from the uniform distribution over $[0,1]$.

We assume $|F|=5000, |GM|=20, |HM|=80, |HV|=2000$, and $c=2$. We consider all possible combinations of the following parameters: all 3 mean types, $k=5,10,20$, and $\rho_0,\rho_1=0.1,0.5,0.9$. We also consider the "large and unequal variance" setting of Table 1 in [13], which sets $\sigma_0=0.25$ and $\sigma_1=0.64$. Given each set of distribution parameters, we randomly assign features to blocks of global markers, heterogeneous markers, and LV non-markers. The remaining features comprise the independent HV non-markers. We generate a stratified sample of size n with equal points in each class. The following algorithms are used to declare the set of good features: t-test, BD, MI, SFS-RLDA, CMNC-OBF, 2MNC-Robust, REMAIN, POFAC, and SPM. We removed FOHSIC due to its computation cost. All Bayesian algorithms use JP. We use thresholds of the Bayesian simulation, except we set $T_1=0.05$. One can tune T_3 and T_4 for one of the 81 possible settings, or a specific sample size, but it can affect the performance of other settings. We picked the thresholds of the Bayesian simulation as they provided satisfactory performance among large sample sizes. This process iterates 500 times. For each set of distribution parameters we define performance as the average number of markers identified as good plus the average number of non-markers identified as bad. Figure 4 plots the average and worst case performance for each fixed mean type across other distribution parameters as sample size increases from 10 to 100 in steps of 10. Bayesian methods tend to outperform non-Bayesian methods. While simpler methods such as CMNC-OBF outperform more

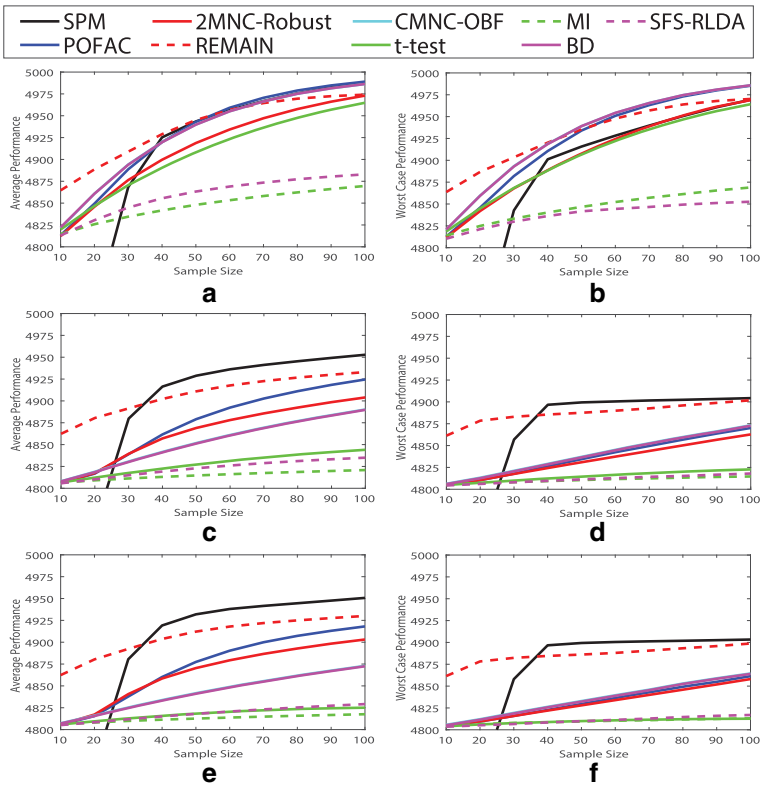

Fig. 4 Average and worst case performance of feature selection algorithms. Average and worst case performance are obtained using 27 combinations of the synthetic microarray model parameters k, ρ_0 and ρ_1 with fixed mean type, where performance is defined to be the average number of markers identified as good plus the average number of non-markers identified as bad over 500 iterations: **a** average performance for redundant means, **b** worst case performance for redundant means, **c** average performance for synergetic means, **d** worst case performance for synergetic means, **e** average performance for marginal means, and **f** worst case performance for marginal means

complicated methods when sample size is small, complicated methods such as SPM have superior performance when sample size is large.

For small sample sizes, or cases where correctly labeling good features is more difficult, REMAIN tends to output very few features resulting in very good performance, in contrast to methods that are forced to output $|\bar{G}| = 100$ features. OBF can be implemented with the MNC objective instead of CMNC to enjoy this characteristic of REMAIN. SPM seems to have the most diverse behavior. While it performs inferior to all feature selection algorithms when sample size is very small, it tends to outperform all other methods for larger sample sizes. In order for the quantities used in SPM to be well-defined under JP, sample size must be larger than the block size. Hence, under small samples SPM with JP tends to break good blocks into smaller blocks, thereby losing some of its ability to identify weak good features with strong dependencies, and making it more prone to detecting blocks incorrectly. Also note that we have used the same parameters for SPM across all data models and sample sizes, and performance is expected to improve if t_1, t_2 and T_4 are calibrated each time it is run.

POFAC is an interesting option, enjoying competitive performance across all sample sizes. It outperforms 2MNC-Robust while its computation cost is only slightly larger. CMNC-OBF tends to select individually strong markers, i.e., markers with class 1 mean far from 0. CMNC-OBF performs very similar to BD in this simulation, with their performance graphs almost overlapping.

Figure 5 plots average performance for fixed class-conditioned correlation coefficients across other distribution parameters. Simpler methods outperform complicated algorithms when sample size is small, and REMAIN enjoys outstanding performance for small sample sizes by reporting very few features. REMAIN has difficulty detecting weak markers, i.e., heterogeneous markers with class 1 mean close to 0, as for larger sample sizes its performance increment is very little for a 10 point increase in sample size. Average performance with respect to sample size for each of the 81 possible data generation settings is provided in the supplementary [see Additional file 1].

While correctly labeling more features tends to result in lower classification error, maximizing the average number of correctly labeled features does not necessarily minimize classification error [see Additional file 1]. An example can be seen in the Supplementary, where we examine the prediction error of several popular classifiers with feature selection on the synthetic microarray model [see Additional file 1].

Results

We apply CMNC-OBF, POFAC, REMAIN, and SPM with the same priors used for synthetic microarray simulations to cancer microarray datasets, select the top genes, and perform enrichment analysis. We list the top 5 genes

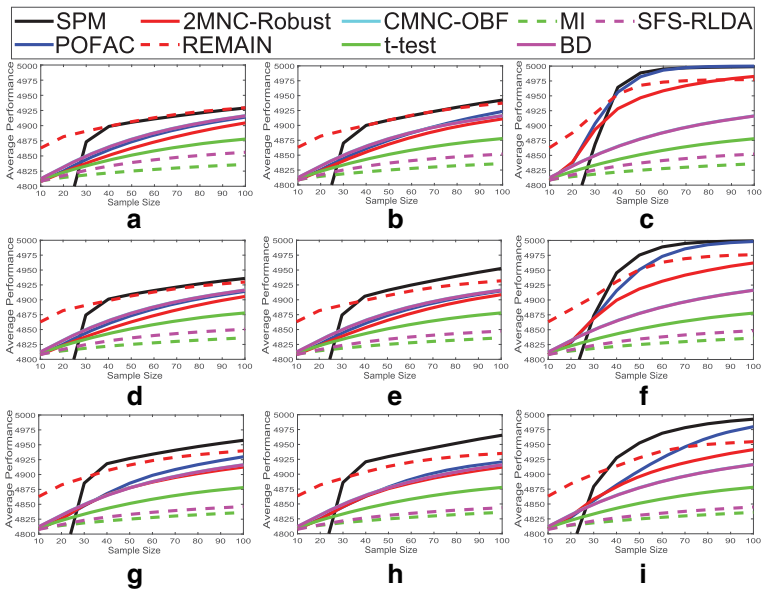

Fig. 5 Average performance of feature selection algorithms. Average performance is obtained using 9 combinations of the synthetic microarray model parameters k and mean type with fixed ρ_0 and ρ_1, where performance is defined to be the average number of markers identified as good plus the average number of non-markers identified as bad over 500 iterations: **a** $\rho_0 = 0.1, \rho_1 = 0.1$, **b** $\rho_0 = 0.5, \rho_1 = 0.1$, **c** $\rho_0 = 0.9, \rho_1 = 0.1$, **d** $\rho_0 = 0.1, \rho_1 = 0.5$, **e** $\rho_0 = 0.5, \rho_1 = 0.5$, **f** $\rho_0 = 0.9, \rho_1 = 0.5$, **g** $\rho_0 = 0.1, \rho_1 = 0.9$, **h** $\rho_0 = 0.5, \rho_1 = 0.9$, and **i** $\rho_0 = 0.9, \rho_1 = 0.9$

selected by CMNC-OBF, POFAC, and REMAIN. The top 100 genes are provided in the supplementary [see Additional file 1]. REMAIN ranks genes as follows. In each call to MAIN, we rank genes of $\tilde{G}$ by the order they are added to $\tilde{G}$, and if several genes are added at once in step 5 of MAIN, they are ranked based on $\tilde{\pi}^*(f)$. In addition, $\tilde{G}$'s are ranked by the order they are obtained using consecutive calls to the MAIN subroutine. Note SPM outputs a set of feature blocks, not a feature ranking. Studying blocks of SPM might provide invaluable information about the underlying biological mechanisms of the disease under study, but we leave this for future work.

We perform enrichment analysis using PANTHER [17, 18]. The top 20 enriched pathways are reported in the supplementary [see Additional file 1]. We list their names, number of known genes in each pathway, number of selected genes that belong to the pathway, and the corresponding p-value. Here we only list the top 3 pathways and their p-values. We study if among the top genes and pathways any are already suggested to be involved in the cancer under study. The complete analysis, with references that suggest involvement of the top reported genes and pathways involved in cancer, is provided in the supplementary [see Additional file 1]. Here, we only report the conclusions made in the supplementary based on our literature review [see Additional file 1].

CMNC-OBF tends to find individually strong genes, which are typically those already known to be involved in cancer. Hence, CMNC-OBF tends to give the best enrichment analysis results, but it might not be the best option to find biomarkers that are individually weak, but heavily correlated to strong biomarkers. POFAC and REMAIN tend to find genes that are individually strong or highly correlated to individually strong biomarkers. Hence, they might be very useful for many practical applications, particularly for those where it is desired to target genes directly involved in cancer, or genes directly interacting with them. SPM is specifically designed to find all genes correlated to individually strong biomarkers. Hence, it tends to report large gene sets. Thereby, this algorithm is particularly useful for identifying and hypothesizing which biological functions are affected in the cancer under study.

POFAC and CMNC-OBF require the user to specify the number of genes to select, which we fix to 2000 so that a reasonable number of genes are identified by the pathway enrichment analysis database. On the other hand, REMAIN and SPM cannot take a predetermined number of genes to select. We adjust their thresholds for each dataset so that a reasonable number of genes are selected. We fix $T_2 = n$, and tune T_1, t_1, t_2, and T_4.

The following process is used on each dataset. We first remove probes that are not mapped to any genes. We then use OBF and POFAC to rank probes, and use REMAIN to select a subset of probes. If multiple probes are mapped to the same genes, only the probe with the highest rank is retained. This gives the selected genes of REMAIN, and final gene rankings of OBF and POFAC. $D = 2000$ is used to obtain gene sets of CMNC-OBF and POFAC. SPM uses the gene ranking obtained by POFAC with the corresponding $\tilde{\beta}(.)$, where among probes mapped to the same genes only the probe ranking highest is retained. Running all algorithms, using MATLAB2015b, on a server with 4 XEON E5-4650L processors and 512GB of RAM took about 20 minutes for the breast cancer dataset, and about 70 minutes for each of the colon cancer and AML datasets. For all datasets REMAIN and SPM took about 55% and 25% of the total run-time, respectively.

Breast cancer

Data obtained in [19] is curated on Gene Expression Omnibus (GEO) [20] with accession number GSE1456, containing 159 points. 119 breast cancer relapse free patients comprise class 0 and 40 patients with breast cancer relapses comprise class 1. In this dataset, "the raw expression data were normalized using the global mean method" [19]. Feature selection algorithms pick the top genes, and enrichment analysis is performed using PANTHER. Here we implement REMAIN with $T_1 = 0.005$, and obtain 1413 genes, and SPM with $T_4 = 1000n^2, t_1 = 1000$, and $t_2 = 1$, and obtain 101 blocks containing 1048 genes. Top genes and pathways are listed in Tables 3 and 4, respectively. PANTHER pathways recognize 358, 254, 328, and 183 of the genes selected by CMNC-OBF, REMAIN, POFAC, and SPM, respectively. Many of the top genes and pathways are suggested to be involved in breast cancer. For instance, PHTF1, ZNF192, and MUC5AC are already shown to be involved in breast cancer. Furthermore, DCT and ZP2 are high-profile biomarkers, and their role in breast cancer requires further investigation. Among pathways, the gonadotropin-releasing hormone receptor pathway, ubiquitin proteasome pathway, CCKR signaling map, and integrin signalling pathway are shown to be involved in breast cancer.

Due to different properties of these algorithms, different types of biomarkers they tend to pick, and our limited knowledge of cancer pathways, it is natural to

Table 3 Top genes of breast cancer

Rank	CMNC-OBF	REMAIN	POFAC
1	DCT	DCT	DCT
2	PHTF1	ZNF192	PHTF1
3	ZNF227	ZP2	MUC5AC
4	ZP2	PCSK6	HUWE1
5	CEACAM7	CEACAM7	MLANA

Table 4 Top pathways of breast cancer

Algorithm	Pathway	*P*-value
CMNC-OBF	Gonadotropin-releasing hormone receptor pathway	$4.93E-05$
	p53 pathway	$8.34E-05$
	Ubiquitin proteasome pathway	$1.26E-04$
REMAIN	Ubiquitin proteasome pathway	$3.33E-07$
	Angiogenesis	$3.40E-04$
	FAS signaling pathway	$7.23E-04$
POFAC	CCKR signaling map	$3.82E-05$
	p53 pathway	$2.61E-04$
	Ubiquitin proteasome pathway	$4.18E-04$
SPM	Ubiquitin proteasome pathway	$6.25E-07$
	Integrin signalling pathway	$1.72E-06$
	Pyrimidine Metabolism	$2.87E-05$

obtain different gene sets, *p*-values, and pathway rankings. However, there is reasonable consistency between the enrichment analysis results. For instance, the ubiquitin proteasome pathway, which is in the top 3 pathways of all algorithms, is shown to be involved in breast cancer. Many of the top 20 pathways are in common between at least 3 algorithms and are shown to be involved in breast cancer. For instance, the gonadotropin-releasing hormone receptor pathway, FAS signaling pathway, P53 pathway, CCKR signaling map, de novo purine biosynthesis, TCA cycle, Cytoskeletal regulation by Rho GTPase, and cell cycle are involved in breast cancer. In addition, many of the top 20 genes are in common between algorithms ranking features. For instance, PHTF1, MUC5AC, ZNF192, PCSK6, and HDGFRP3 are shown to be involved in breast cancer, and some common genes such as DCT, ZP2, and CEACAM7 might be involved in breast cancer.

Colon cancer

Data obtained in [21, 22] is curated on GEO [20] with accession number GSE17538, containing gene expression levels of 238 patients in stages 1-4 of colon cancer. Twenty eight stage 1 patients comprise class 0 and the remaining patients comprise class 1. Bioconductor's affy package with its default settings has normalized the data. Feature selection algorithms pick the top genes, and enrichment analysis is performed using PANTHER. Here we implement REMAIN with $T_1 = 0.01$ to obtain 1289 genes, and SPM with $T_4 = 10^7 n^4$, $t_1 = 10^6$, and $t_2 = 4$, to obtain 159 blocks containing 1560 genes. Top genes and pathways are listed in Tables 5 and 6, respectively. PANTHER pathways recognize 312, 159, 208, and 174 of the genes selected by CMNC-OBF, REMAIN, POFAC, and SPM, respectively. Many of the top genes and pathways are suggested to be involved in colon cancer. For instance, CPNE4 and EPHA7 are already shown to be involved in colon cancer. Among pathways, the cadherin signaling pathway and ionotropic glutamate receptor pathway are shown to be involved in colon cancer.

Table 5 Top genes of colon cancer

Rank	CMNC-OBF	REMAIN	POFAC
1	CPNE4	EPHA7	EPHA7
2	GAGE1,12,4,5,6,7	NBLA00301	CPNE4
3	GAGE1,12,2,4,5,6,7,8	LOC100133920 LOC286297	LOC100133920 LOC286297
4	S100A7	PDK4	SCN7A
5	EPHA7	MYH11	NBLA00301

In the supplementary (Additional file 1) we show: (1) CPNE4, EPHA7, and LOC286297, which are among the top 20 genes of all three algorithms that rank genes, are shown to be involved in colon cancer, (2) many of the top 20 genes in common between two of the gene ranking algorithms, such as the GAGE genes, RYR3, PDK4, and MYH11, are suggested to be involved in colon cancer, and (3) among the common top 20 enriched pathways, the plasminogen activating cascade, blood coagulation, and the beta1 adrenergic receptor signaling pathway are suggested to be involved in colon cancer.

Acute myeloid leukemia

Data obtained in [23–25] is deposited on GEO with accession number GSE13204, containing gene expression levels of 2096 points. 74 points belong to healthy people, 542 points belong to Acute Myeloid Leukemia (AML) patients, and the remaining points are other subtypes of leukemia. Healthy points comprise class 0 and

Table 6 Top pathways of colon cancer

Algorithm	Pathway	*P*-value
CMNC-OBF	Cadherin signaling pathway	$1.83E-20$
	Wnt signaling pathway	$7.25E-14$
	Plasminogen activating cascade	$7.67E-05$
REMAIN	Cadherin signaling pathway	$2.21E-10$
	Plasminogen activating cascade	$7.95E-05$
	Wnt signaling pathway	$9.27E-05$
POFAC	Ionotropic glutamate receptor pathway	$1.72E-05$
	Metabotropic glutamate receptor group III pathway	$1.04E-02$
	Nicotinic acetylcholine receptor signaling pathway	$1.70E-02$
SPM	Ionotropic glutamate receptor pathway	$2.98E-03$
	Heterotri. G-prot. sig. P.W., Gi alpha & Gs alpha med. P.W.	$4.73E-03$
	Axon guidance mediated by Slit/Robo	$5.79E-03$

AML patients comprise class 1. The data is already preprocessed, including a summarization and quantile normalization step. Feature selection algorithms pick the top genes, and enrichment analysis is performed using PANTHER. Here we implement REMAIN with $T_1 = 0.05$ to obtain 957 genes, and SPM with $T_4 = 10^7 n^{10}$, $t_1 = 10^6$, and $t_2 = 4$, to obtain 522 blocks containing 5172 genes. Although the thresholds of SPM are chosen to be very large, we still pick very many genes. This might imply that many of the genes involved in AML might be individually weak, but highly correlated.

Top genes and pathways are listed in Tables 7 and 8, respectively. PANTHER pathways recognize 276, 141, 266, and 671 of the genes selected by CMNC-OBF, REMAIN, POFAC, and SPM, respectively. Many of the top genes and pathways are suggested to be involved in AML. For instance, ORM1 and ORM2 are already shown to be involved in AML, LTF is a high-profile gene whose role in AML needs further investigation, and S100A12 is shown to be involved in similar subtypes of leukemia, such as ALL, and is suggested to be involved in AML as well. Among pathways, heme biosynthesis, the interferon-gamma signaling pathway, pentose phosphate pathway and ubiquitin proteasome pathway have suggested involvement in AML.

Studying the top 20 genes and pathways in the supplementary (Additional file 1) we see that ORM1, ORM2, LTF, CAMP, LCN2, MMP9, CYP4F3, WT1, and CRISP3 are among the top 20 genes of all gene ranking algorithms, and are shown or suggested to be involved in AML. Among the top pathways in common between all methods, the interferon signaling pathway, and the inflammation mediated by chemokine and cytokine signaling pathway are involved in AML. Many of the top pathways picked by at least 3 methods, such as heme biosynthesis, denovo purine biosynthesis, and T-cell activation are also suggested to be involved in AML.

Table 8 Top pathways of AML

Algorithm	Pathway	*P*-value
CMNC-OBF	Heme biosynthesis	$2.78E-04$
	Pentose phosphate pathway	$7.34E-03$
	De novo purine biosynthesis	$8.40E-03$
REMAIN	Interferon-gamma signaling pathway	$5.86E-04$
	Alzheimer disease-presenilin pathway	$6.11E-03$
	Inflammation med. by chemokine & cytokine sig. P.W.	$6.22E-03$
POFAC	Pentose phosphate pathway	$1.15E-03$
	Formyltetrahydroformate biosynthesis	$1.15E-03$
	Heme biosynthesis	$1.77E-03$
SPM	Ubiquitin proteasome pathway	$1.72E-08$
	T cell activation	$2.07E-05$
	Inflammation med. by chemokine & cytokine sig. P.W.	$3.00E-05$

Discussion

Here we proposed several suboptimal feature selection algorithms outperforming many popular algorithms. However, the ability to correctly detect weaker biomarkers via these suboptimal methods comes at the expense of less intuitive objective functions compared with the optimal solutions. Although the proposed algorithms are more computationally intensive than OBF and 2MNC-Robust, they are still much faster than many popular feature selection algorithms.

Table 7 Top genes of AML

Rank	CMNC-OBF	REMAIN	POFAC
1	ORM1 ORM2	LTF	S100A12
2	LTF	CRISP3	S100A9
3	CRISP3	ORM1 ORM2	ORM1 ORM2
4	CHIT1	CHIT1	CRISP3
5	DNAH10	DNAH10	LTF

While the previously introduced OBF is suitable to find individually strong biomarkers, REMAIN and POFAC find individually strong biomarkers as well as individually weak biomarkers heavily correlated to strong biomarkers, which is useful for many practical applications. For instance, in drug development it might be desirable to target genes directly involved in biological mechanisms of the cancer under study, or target genes strongly correlated to them to indirectly control the behavior of genes directly involved in cancer.

When sample size is small or correlations are not very strong one could use REMAIN to find a small set of high-profile biomarkers. REMAIN cannot be forced to output a predetermined number of features, but for a given fixed dataset its parameters can be tuned to output close to a desired number of features. Note the output of REMAIN is a feature ranking that greatly depends on T_1. The larger T_1 is, the smaller is the output feature ranking. However, the user would have more confidence that declared features are biomarkers. While $\pi^*(f)$ is a very intuitive quantity to evaluate the quality of a feature, $\tilde{\pi}^*(f)$ obtained by finding $\tilde{\pi}^*(G)$ for sets of size 2 is not as easy to work with.

On the other hand, if sample size is large or correlations are strong, POFAC is a suitable option. POFAC provides a feature ranking based on $\tilde{\beta}(f)$ and the user specifies how many features to select, similar to CMNC-OBF and 2MNC-Robust. However, $\tilde{\beta}(f)$ is not as intuitive as $\pi^*(f)$.

SPM outputs a family of good blocks, which is very useful in studying the interactions between biomarkers, and is very useful to hypothesize about biological mechanisms that are involved in the disease under study. As SPM is designed to pick all biomarkers correlated to strong ones, it typically reports larger feature sets. SPM is very desirable when correlations are large; however, it should only be used when sample size is relatively large. Finally, parameters of SPM can be used to adjust the trade-off between the size of detected blocks, and the minimum desired dependence between biomarkers. However, one cannot intuitively determine what values should be used to achieve a certain point in the trade-off. Trial-and-error can be used on a fixed sample to find the desired parameters, as we did in this paper. One of the main reasons it is not easy to predetermine SPM parameters is their dependence on sample size and the underlying distribution parameters.

Conclusion

The proposed Bayesian framework is indeed promising for biomarker discovery applications. The objective of finding the posterior probability that a feature set is the set of good features does not suffer many of the drawbacks of heuristics used in biomarker discovery. For instance, while t-test only captures differences in the means, OBF can capture differences in variances, 2MNC-Robust, REMAIN, and POFAC take pairwise dependencies into account, and SPM looks at the joint distribution of good blocks. While the proposed suboptimal methods can efficiently take advantage of dependencies to find the set of good features, many heuristics proposed to consider dependencies do not perform well under small samples [13], which is observed in our simulations as well.

While the optimal solution under the general block model is computationally infeasible, the success of proposed suboptimal algorithms shows the Bayesian framework can serve as a foundation to model biomarker discovery problems and develop efficient suboptimal methods. Future work includes studying the properties of the proposed algorithms, for instance their asymptotic properties, further analyzing outputs of the proposed algorithms on real datasets, and exploring the specific applications suitable for each algorithm in greater detail. In addition, prior construction may be used to design prior distributions for SPM to boost its performance under small samples. Finally, while here we have studied SPM's ability to select all relevant features, it actually outputs blocks of features that appear highly correlated and differentially expressed. In future work we will examine SPM as a block detection algorithm, which has important applications in gene network modeling.

Additional file

Additional file 1: Supplementary. Additional detail on the synthetic simulations and a comparison of classification error of the selected features of each algorithm is provided. The supplementary contains the top 100 selected genes and top 20 enriched pathways of each of the proposed algorithms as well as limma. (PDF 507 kb)

Acknowledgements
Not applicable.

Funding
The publication costs of this article was funded by the National Science Foundation (CCF-1422631 and CCF-1453563).

Availability of data and materials
The breast cancer, colon cancer, and Leukemia datasets used in this paper are publicly available on Gene Expression Omnibus (GEO) [20] with accession numbers, GSE1456, GSE17538, and GSE13204, respectively. A MATLAB implementation of algorithms is available on Github.

About this supplement
This article has been published as part of *BMC Bioinformatics* Volume 19 Supplement 3, 2018: Selected original research articles from the Fourth International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2017): bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-3.

Authors' contributions
AF proposed the main idea and worked on the simulations and manuscript. LAD contributed to the formulation of main idea, and revised the manuscript. Both authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Department of Electrical and Computer Engineering, The Ohio State University, 2015 Neil Avenue, 43210 Columbus, Ohio, USA. [2]Department of Biomedical Informatics, The Ohio State University, 250 Lincoln Tower, 1800 Cannon Drive, 43210 Columbus, Ohio, USA.

Published: 21 March 2018

References

1. Ramachandran N, Srivastava S, LaBaer J. Applications of protein microarrays for biomarker discovery. Proteomics Clin Appl. 2008;2(10-11):1444–59.
2. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: The long and uncertain path to clinical utility. Nat Biotechnol. 2006;24(8):971–83.
3. Ilyin SE, Belkowski SM, Plata-Salamán CR. Biomarker discovery and validation: Technologies and integrative approaches. Trends Biotechnol. 2004;22(8):411–6.
4. Feng Z, Prentice R, Srivastava S. Research issues and strategies for genomic and proteomic biomarker discovery and validation: A statistical perspective. Pharmacogenomics. 2004;5(6):709–19.

5. Diamandis EP. Cancer biomarkers: Can we turn recent failures into success? J Natl Cancer Inst. 2010;102(19):1462–7.
6. Foroughi pour A, Dalton LA. Optimal Bayesian feature filtering. In: Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics. Atlanta: ACM; 2015. p. 651–2.
7. Foroughi pour A, Dalton LA. Optimal Bayesian feature selection on high dimensional gene expression data. In: Proceedings of the 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Atlanta: IEEE; 2014. p. 1402–5.
8. Foroughi pour A, Dalton LA. Multiple sclerosis biomarker discovery via Bayesian feature selection. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Seattle: ACM; 2016. p. 540–1.
9. Foroughi pour A, Dalton LA. Robust feature selection for block covariance bayesian models. In: Proceedigns of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE; 2017. p. 2696–700.
10. Dalton LA. Optimal Bayesian feature selection. In: Proceedings of 2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Austin: IEEE; 2013. p. 65–8.
11. Murphy KP. Conjugate Bayesian analysis of the Gaussian distribution. Canada: Technical report, University of British Columbia; 2007.
12. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130(12):1005–13.
13. Hua J, Tembe WD, Dougherty ER. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recog. 2009;42(3):409–24.
14. Beirlant J, Dudewicz EJ, Györfi L, Van der Meulen EC. Nonparametric entropy estimation: An overview. Int J Math Stat Sci. 1997;6(1):17–39.
15. Braga-Neto U, Dougherty ER. Bolstered error estimation. Pattern Recognit. 2004;37(6):1267–81.
16. Song L, Smola A, Gretton A, Borgwardt KM, Bedo J. Supervised feature selection via dependence estimation. In: Proceedings of the 24th International Conference on Machine Learning. 2007. p. 823–30.
17. Mi H, Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. In: Nikolsky Y, Bryant J, editors. Protein Networks and Pathway Analysis. Totowa: Humana Press; 2009. p. 123–40. https://doi.org/10.1007/978-1-60761-175-2_7.
18. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. Panther version 11: Expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017;45(1):183–9.
19. Pawitan Y, Bjöhle J, Amler L, Borg A-L, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedrén S, Bergh J. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res. 2005;7(6):953–64.
20. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
21. Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, Lu P, Johnson JC, Schmidt C, Bailey CE, Eschrich S, Kis C, Levy S, Washington MK, Heslin MJ, Coffey RJ, Yeatman TJ, Shyr Y, Beauchamp RD. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology. 2010;138(3):958–68.
22. Freeman TJ, Smith JJ, Chen X, Washington MK, Roland JT, Means AL, Eschrich SA, Yeatman TJ, Deane NG, Beauchamp RD. Smad4-mediated signaling inhibits intestinal neoplasia by inhibiting expression of β-catenin. Gastroenterology. 2012;142(3):562–71.
23. Kohlmann A, Kipps TJ, Rassenti LZ, Downing JR, Shurtleff SA, Mills KI, Gilkes A, Hofmann W-K, Basso G, Dell'Orto MC, Foà R, Chiaretti S, De Vos J, Rauhut S, Papenhausen PR, Hernández JM, Lumbreras E, Yeoh AE, Koay ES, Li R, Liu W-m, Williams PM, Wieczorek L, Haferlach T. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the microarray innovations in leukemia study prephase. Br J Haematol. 2008;142(5):802–7.
24. Haferlach T, Kohlmann A, Wieczorek L, Basso G, Kronnie GT, Béné M-C, Vos JD, Hernández JM, Hofmann W-K, Mills KI, Gilkes A, Chiaretti S, Shurtleff SA, Kipps TJ, Rassenti LZ, Yeoh AE, Papenhausen PR, Liu W-M, Williams PM, Foà R. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the international microarray innovations in leukemia study group. J Clin Oncol. 2010;28(15):2529–37.
25. Kühnl A, Gökbuget N, Stroux A, Burmeister T, Neumann M, Heesch S, Haferlach T, Hoelzer D, Hofmann W-K, Thiel E, Baldus CD. High BAALC expression predicts chemoresistance in adult B-precursor acute lymphoblastic leukemia. Blood. 2010;115(18):3737–44.