Foroughi pour *et al. BMC Bioinformatics* (2020) 21:156
https://doi.org/10.1186/s12859-020-3486-x

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

High dimensional model representation of log-likelihood ratio: binary classification with expression data

Ali Foroughi pour[1,2], Maciej Pietrzak[3], Lori A Dalton[1] and Grzegorz A. Rempała[2,4*]

*Correspondence:
rempala.3@osu.edu
[2]Department of Mathematics, The Ohio State University, 100 Math Tower, 31 West 18th Ave., 43210 Columbus, USA
[4]College of Public Health, 250 Cunz Hall, 1841 Neil Ave., 43210 Columbus, USA
Full list of author information is available at the end of the article

Abstract

Background: Binary classification rules based on a small-sample of high-dimensional data (for instance, gene expression data) are ubiquitous in modern bioinformatics. Constructing such classifiers is challenging due to (a) the complex nature of underlying biological traits, such as gene interactions, and (b) the need for highly interpretable glass-box models. We use the theory of high dimensional model representation (HDMR) to build interpretable low dimensional approximations of the log-likelihood ratio accounting for the effects of each individual gene as well as gene-gene interactions. We propose two algorithms approximating the second order HDMR expansion, and a hypothesis test based on the HDMR formulation to identify significantly dysregulated pairwise interactions. The theory is seen as flexible and requiring only a mild set of assumptions.

Results: We apply our approach to gene expression data from both synthetic and real (breast and lung cancer) datasets comparing it also against several popular state-of-the-art methods. The analyses suggest the proposed algorithms can be used to obtain interpretable prediction rules with high prediction accuracies and to successfully extract significantly dysregulated gene-gene interactions from the data. They also compare favorably against their competitors across multiple synthetic data scenarios.

Conclusion: The proposed HDMR-based approach appears to produce a reliable classifier that additionally allows one to describe how individual genes or gene-gene interactions affect classification decisions. Both real and synthetic data analyses suggest that our methods can be used to identify gene networks with dysregulated pairwise interactions, and are therefore appropriate for differential networks analysis.

Keywords: High dimensional model representation, Classification, Disease prediction, Log-likelihood ratio, Expression analysis

Background

The notion of a simple binary classification, as one of the corner stones of modern data analysis, has been considered in many different contexts and an abundance of algorithms have been proposed for this task. While research has recently shifted focus to classification rules in the context of big data, many bioinformatics applications deal with small-sample, high-dimensional prediction problems. Current high-throughput "omics" technologies measure tens of thousands of molecular features for each experimental unit (for instance, a patient's tissue sample); however, research data is still usually limited to small sizes, rarely more than a few hundred units, impeding reliable analysis. Additionally, data might be heavily imbalanced, which adds to the challenge of correct classification in a small-sample, high-dimensional setting, with the minimum misclassification error criteria being too unreliable for consistent feature selection across multiple datasets [1, 2].

In contrast to many applications where machine learning methods are used merely to predict and do not have to provide explicit decision rules, the bioinformatics applications demand highly interpretable glass-box models to explain how a specific decision is obtained. In many instances it is important to know which features, e.g., genes, are used by the classifier, whether these features are biologically relevant, whether the distributional differences in features across two classes indicate biological variability or are merely artifacts of the measurement/normalization process, and what is the uncertainty of prediction at a new test point?

Answers to these questions are necessary to hypothesize about biological mechanisms of complex diseases such as cancer and to evaluate clinical utility of developed decision rules for tasks such as diagnosis and prognosis. But they are also necessary to explain how certain patterns in the data might motivate different actions, such as choosing a specific treatment over another for targeted therapy, exploring alternative treatments, or how to form hypotheses on the biological mechanisms that can potentially be targeted in drug discovery applications.

The small-sample high-dimensional nature of the problem, interpretability of outputted statistics, and complex feature dependencies, force the development of methods with few degrees of freedom that place strong assumptions (e.g. distributional assumptions) on the classification problem. For example, linear discriminant analysis (LDA) assumes features are Gaussian and have the same covariance matrix in both classes, quadratic discriminant analysis (QDA) assumes features are jointly Gaussian with different class-conditioned covariances, and a logistic regression model assumes that the log-likelihood ratio is an additive function of features. The common idea behind these methods is that although the "optimal" decision rule might be very complex (e.g., a high dimensional separating surface), it can be well approximated by a low dimensional model, and an appropriate family of models should contain a point close to the "best" low dimensional representation that can be reliably approximated given the observed data. In the current paper we follow a rather similar general approach, but apply a much more flexible method for deriving classifiers that allow for more flexible classification rules.

Recent studies emphasize the importance of gene synergies and genetic interactions for reliable analysis [3]. However, two general themes of the recent method developments are leveraging big data, such as the cancer genome atlas (TCGA), or taking advantage of side information such as sets of co-mutated genes or disease protein subnetworks, e.g. [4, 5]. Such information may not be readily available or may not be

easily applicable to the current dataset, as cancer gene interactions are highly context dependent [6].

Utilizing pairwise interactions for reliable prediction aside, detecting disease-associated genetic interactions has been studied as a "gene discovery problem". To that end, mutual information based synergy scores are proposed, e.g., [7, 8]. However, reliably inferring mutual information from data is a challenging task, which can be circumvented by quantizing expression values, building dendrograms based on expressions, utilizing ranked expressions instead of raw continuous values, or defining new statistics based on gene-pair expression rankings [7–10]. In [9] it is stated that dendrogram and doublet (a specific collection of transformations merging gene pair expressions into one-dimensional values) based methods are "helpless for discovering pair-wise gene interactions". The information theoretic score of [8] cannot be easily utilized to test significance, using limited permutations of data to approximate the null, hypothesis which is computationally intensive [9]. Finally, [9] proposes a new conversion transformation, the absolute difference of ranked expressions constructing a t-statistic, which seems to balance performance and computation cost.

High dimensional model representation

Consider a set of predictors as a random vector and a dependent variable as a function of predictors, e.g., class labels as a function of observed expressions. High dimensional model representation (HDMR) is a recently proposed framework to decompose functions of a random vector, i.e., the dependent variable, into a hierarchy of low dimensional models based on partial marginals of the full joint distribution [11]. Intuitively speaking, **HDMR expansion optimally decomposes a high dimensional non-linear system into a hierarchy of lower dimensional non-linear systems**, simplifying the process of studying each high-dimensional component. It enjoys several interesting properties. For example, the d^{th} order expansion is the best representation, in mean square error (MSE) sense, to estimate the dependent variable given its marginal distribution with all subsets of the predictors with at most d elements. Additionally, higher order expansion terms are independent of the lower order terms. HDMR assumptions are mild, only requiring certain moments to exist. Unfortunately, computing the HDMR expansion requires complete knowledge of the full joint distribution, and potentially solving large families of highly complex integral equations unless simplifying assumptions are made or special cases are considered [12]. This can be a deal breaker in many practical applications, where it is not always possible to obtain the full joint distribution given the small sample size. In this work, as a workaround, we propose algorithms that aim to approximate the HDMR expansion without directly estimating the full joint distribution and solving integral equations.

Our contribution

The novelty of the proposed classification framework is three-fold. (1) Our approach provides a hierarchy of low dimensional representations of data, possibly allowing for analyzing progressively more complex interactions among features. (2) We propose a regression based approach to circumvent solving complex integral equations. (3) We can easily study the effects of any specified subsets of variables, and assess how their interactions affect the classifier output. As a side note, the proposed framework can also

easily combine different parametric and non-parametric methods for computing log-likelihood ratios, an interesting property that adds further flexibility. However, we leave this extension for future work.

The paper is organized as follows. We first briefly overview the theory of HDMR expansion, and how it can be used for binary classification, considering in particular the special case of second order HDMR expansion. We then explain the regression-based algorithm of approximating the second order HDMR expansion and perform synthetic simulations comparing our method with several other popular classification rules proposed in the literature. Finally, we provide several real data examples, studying breast cancer, leukemia, and lung cancer.

Methods

Here we describe our classification methodology based on the HDMR expansion, studied in detail in [11, 13]. We briefly review the theory, and then show how it applies to binary classification.

HDMR expansion

HDMR provides a hierarchy of functions that describe how the interactions of variables affect the output. In particular, assuming output Z as a function of input random vector $X = [X_1, \cdots, X_D]$, i.e., $Z = h(X)$, HDMR studies how $h(X)$ can be decomposed to a hierarchy of partial observations. Let $F = \{1, \cdots, D\}$. The HDMR expansion of order d is the collection of functions $h_u(X_u)$ for all $u \subseteq F$ with $|u| \leq d$ that minimize the mean square error (MSE) of estimating Z given $E(Z|X_u)$ for all u under the condition that for all $f \in u$, $E\left(h_u(X_u)|X_{u\setminus f}\right) = 0$, which is equivalent to a hierarchical orthogonality criterion [13], i.e., HDMR terms of different orders are independent of each other. From [13] we have

$$h(X) = h_0 + \sum_{\substack{u\subseteq F \\ u\neq\phi}} h_u(X_u), \tag{1}$$

$$h_0 = \int h(x)w(x)dx, \tag{2}$$

$$h_u = \int h(x)w(x_{-u})dx_{-u} - \sum_{v\subset u} h_v(x_v) - \sum_{v\neq u: v\cap u\neq\phi} \int h_v(x_v)w_{-u}dx_{-u}. \tag{3}$$

Eq. 3 suggests that in the general case of dependent variables a component function, $h_u(x_u)$, depends on all other expansion terms that have a non-empty intersection with u. However, assuming elements of X are independent, the last term of (3) equals zero. While this greatly simplifies the process of computing the HDMR expansion, the independence assumption can be heavily violated for expression data. We hereafter use $E_d(Z|X)$ to denote the d^{th} order HDMR expansion.

Approximate second order HDMR for classification

We now focus on the second order HDMR expansion for correlated features. Observe that under the independence assumption, we have

$$E_2(Z|X) = w_0 + w_f E\left(Z|X_f\right) + w_{f,f'} E\left(Z|X_{f,f'}\right), \tag{4}$$

for some $w_0, w_f, w_{ff'} \in \mathbb{R}$. In case of dependent features, we still assume the second order HDMR expansion follows a structure similar to (4), except that the coefficients $w_f, w_{ff'}$ are different than the independent case. Now, consider a binary classification problem with class labels $y = 0, 1$ and feature index set F. Let X be a random unlabeled observation with true label y_x. Given a training sample $\mathcal{S}$, it is desired to design a decision rule that assigns a label, $\hat{y}_x$ to X so that $\hat{y}_x = y_x$ with high probability. Note that given the full joint distribution of X and y_x, one could have easily computed $P(y_x = 1|X)$, or equivalently the log-likelihood ratio $L(X) = \log(P(y_x = 1|X)/P(y_x = 0|X))$, and use a decision rule $\hat{y}_x = 1_{L(X)>T}$, where 1_q is the indicator function of statement q being correct and T is a threshold.

However, the full joint distribution is typically not available, and is usually estimated using $\mathcal{S}$. Alternatively, many models assume the classification rule belongs to a family parametrized by θ, which is estimated from $\mathcal{S}$. For example, a generalized linear model (GLM) with the logit link assumes $L(X) = \beta_0 + \sum_{f\in F} \beta_f X_f$, where X_f is the value of X for feature f. Here θ is the collection of β_0 and β^f's. However, such model may be insufficient when pairwise feature interactions are of interest, and it can be difficult to train a GLM considering all potential pairwise interactions using LASSO and elastic net penalties due to the potentially large number of feature pairs. Assuming $Z = L(X)$, and $\mathcal{S}$ is the training sample, we have

$$E_2(L(X)|X, \mathcal{S}) = c_0 + \sum_{f\in F} c_f S(X_f) + \sum_{f_i f_j \in F} c_{f_i f_j} S(X_{f_i f_j}),$$

for some $c_0, c_f, c_{f_i f_j} \in \mathbb{R}$ where

$$S(X_f) = E(\log L(X)|X_f, \mathcal{S}), \tag{5}$$

$$S(X_{f_i f_j}) = E(\log L(X)|X_{f_i f_j}, \mathcal{S}) - E(\log L(X)|X_{f_i}, \mathcal{S}) - E(\log L(X)|X_{f_j}, \mathcal{S}), \tag{6}$$

It now remains to estimate coefficients c_f and $c_{f_i f_j}$, which is part of our classification algorithm discussed in the next section. Note that we also assume an external mechanism which already outputs $E(L(X)|X_f, \mathcal{S})$ and $E(L(X)|X_{f_i f_j}, \mathcal{S})$.

Identifying pairwise feature interactions

An important application is identifying feature interactions that are significantly different between the classes, i.e., identify $u = \{f_i, f_j\}$'s for which $h_u \not\equiv 0$. We have the following hypothesis test:

$$H_0 : h_u \equiv 0 \; v.s. \; H_1 : h_u \not\equiv 0. \tag{7}$$

Note this is a very difficult problem in general, and only the special case of Gaussian features is studied here. Assuming f_i and f_j are jointly Gaussian in each class, for an HDMR expansion of first order to be able to grasp the exact form of the log-likelihood ratio we need (1) all features of u to be independent given class label y, or (2) have the same covariance in both classes (assuming $\mu_0^f \neq \mu_1^f$ for all $f \in F$). In either case we have $L(X) = a_0 + \sum_f a_f L(X|X_f)$, for some coefficients $a_0, a_f \in \mathbb{R}$. It is straightforward but tedious to show that other cases result in a second order expansion. Therefore, we can reformulate the hypothesis test of Eq. 7 as

$$H_0 : \rho_0^{f_i,f_j} = \rho_1^{f_i,f_j} = 0 \text{ or} \Sigma_0^{f_i,f_j} = \Sigma_1^{f_i,f_j} \quad v.s.$$
$$H_1 : (\rho_0^{f_i,f_j} \neq 0 \text{ or } \rho_1^{f_i,f_j} \neq 0) \text{ and} \Sigma_0^{f_i,f_j} \neq \Sigma_1^{f_i,f_j}, \tag{8}$$

where $\rho_y^{f_i,f_j}$ and $\Sigma_y^{f_i,f_j}$ are the correlation coefficient and covariance matrix of feature pair f_i,f_j in class y, respectively. Figure 1 provides several examples on cases with and without pairwise feature interactions. In cases (a), (b), and (d) there are no pairwise feature interactions. Case (c) denotes a degenerate case and is studied in the Supplementary. Cases (e) and (f) depict feature pairs with pairwise interactions.

Testing conditional independence is a statistically difficult problem [14], and has been studied for certain cases in [14]. As an approximation, we adopt the following approach. Let $P_y^{f_i,f_j}$ be the p-value of an independence test performed on data in class y. We use the Pearson linear correlation test for Gaussian data. Assuming points in different classes are independent, we treat them as independent tests, and use Fisher's method for meta-analysis: $C = -2\left(\log\left(P_0^{f_i,f_j}\right) + \log\left(P_1^{f_i,f_j}\right)\right)$ follows a χ^2 distribution with 4 degrees of freedom under the null, giving us the final p-value. We use the likelihood ratio test of [15] with the χ^2 adjustment to find p-values of $\Sigma_0^{f_i,f_j} = \Sigma_1^{f_i,f_j}$. Finally, we use the union bound and add the two p-values to obtain our final p-value, which is an overestimate. We hereafter call this approach *multiple test mixing for pairwise interactions* (MTM) and note that it is appropriate for differential network analysis. Indeed, identifying feature pairs with interesting interactions, i.e., pairwise dependencies that require looking at a second order expansion, instead of a first order expansion, are a goal of co-expression and differential network analysis. Different modes of co-expression is discussed in [16], and some of its applications, such as expression analysis, functional classification, and gene-disease prediction, are described in [17, 18]. Differential network analysis is further discussed in [19, 20].

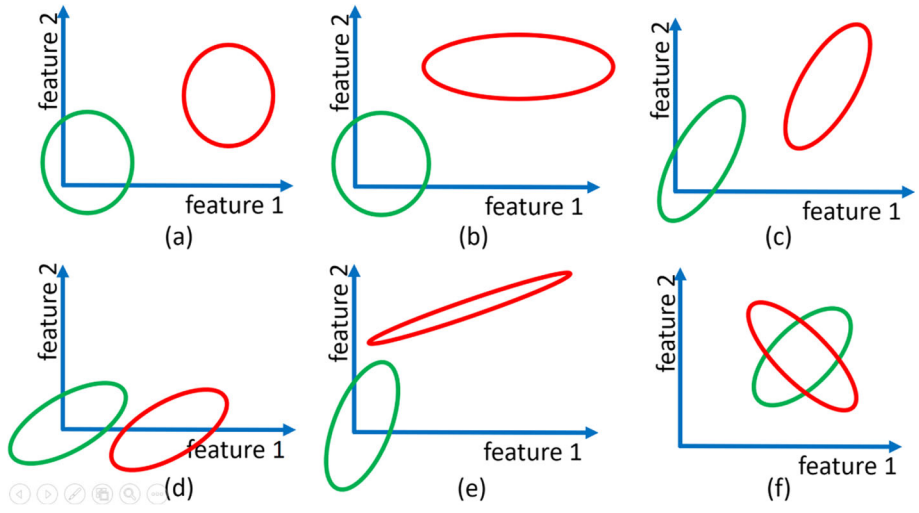

Fig. 1 An example of different cases that might occur for two Gaussian features in a binary classification problem, where the two classes are denoted by red and green: (**a**) independent and equal variances, (**b**) independent and unequal variances, (**c**) correlated and equal covariances with degenerate means, (**d**) correlated and equal covariances with generic means, (**e**) correlated, unequal covariances, and unequal mean vectors, and (**f**) correlated, unequal covariances, and equal means vectors

The classification algorithm

Here we describe our approach to build a classifier that is inspired by second order HDMR expansion of the log-likelihood ratio. Again suppose sample $\mathcal{S}$ is collected, and for each set u such that $|u| \leq 2$, we have a method that outputs the log-likelihood ratio of the test point belonging to class 1, i.e., we have a method that outputs $L(X_u|\mathcal{S})$ for all u such that $|u| \leq 2$. Given all these values, it remains to combine them into a test score. However, in a small-sample problem the number of feature pairs can be much larger than sample size. This creates an ill-posed problem as the number of equations is smaller than the number of parameters to estimate. Therefore, many classical methods for obtaining the model parameters from data, such as maximum likelihood, are not applicable. Not being able to compute the exact second order HDMR expansion by framing it as a regression problem is the main reason we need to look at alternative training methods, and is the main reason we label the concomitant classifier design approach as "HDMR expansion inspired" or as an "HDMR expansion perspective" for classification.

To circumvent the ill-posed problem, we use a variation of objective functions mostly studied in compressed sensing, e.g. [21], that estimate a sparse signal given 1-bit quantized observations. The connection between these objectives and a convex relaxation to the logistic regression problem is discussed in [22]. Heuristically speaking, given a feature vector in the form of log-likelihood ratios of partial observations x_u, here we find weights that maximize the distance between the average points of each class. The heuristic for using such objective function is as follows. On one hand, HDMR obtains the weights that result in the "best" low dimensional representation, i.e., the MSE estimate of the log-likelihood ratio, yielding low prediction errors. On the other hand, weights that maximize the distance between the projections of the center points of the two classes into a one dimensional space should also yield low prediction error. Hence, such objective function should result in a model with an error probability close to that of the HDMR expansion. Here we have used the HDMR theory to obtain the functional form of the solution, and use algorithms borrowed from 1-bit compressed sensing to estimate the HDMR coefficients.

In many high-dimensional statistics applications, it is common to favor some bias to reduce estimation variance. Examples and a detailed discussion on this issue is provided in [23]. In order to reduce variance in our setting, we remove the weakest classifiers, in forms of single feature classification rules or relatively independent feature pairs whose information is already provided in the first order expansion. We compute

$$r^f = \frac{1}{n_1} \sum_{x \in \mathcal{S}_1} L(x_f|\mathcal{S}) - \frac{1}{n_0} \sum_{x \in \mathcal{S}_0} L(x_f|\mathcal{S}), \tag{9}$$

where $\mathcal{S}_y$ is the restriction of $\mathcal{S}$ to points in class y, and n_y is sample size in class y. We remove features for which $r^f < T_1$, where T_1 is a threshold. For feature pairs we compute $r^{f_i f_j}$ defined as

$$\frac{1}{n_1} \sum_{x \in \mathcal{S}_1} L(x_{f_i f_j}|\mathcal{S}) - \sum_{x \in \mathcal{S}_0} \frac{1}{n_0} L(x_{f_i f_j}|\mathcal{S}) - r^{f_i} - r^{f_j}, \tag{10}$$

and remove feature pairs for which $|r^{f_i f_j}| < T_2$, for some threshold T_2. Feature pairs for which $r^{f_i f_j} > T_2$ are risk increasing pairs, and feature pairs for which $r^{f_i f_j} < -T_2$ are risk decreasing pairs. We now find weights that combine $S(X|X_u, \mathcal{S})$ of feature and

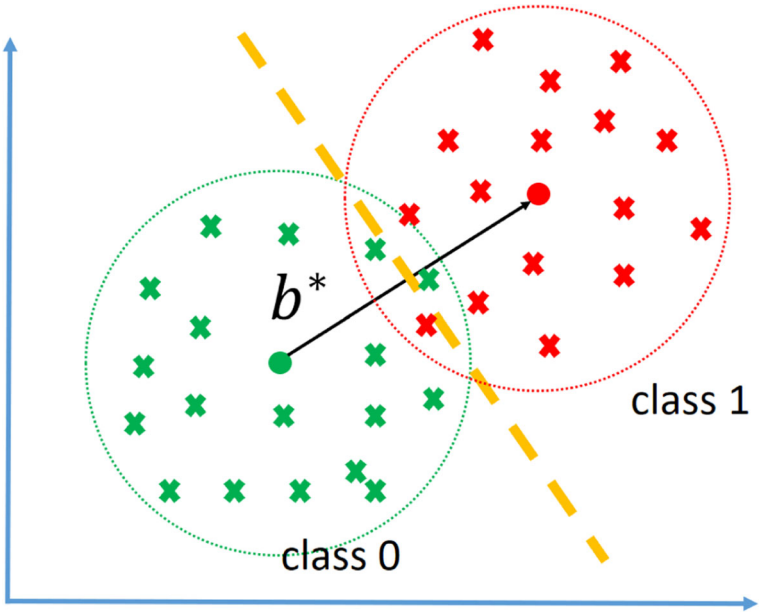

Fig. 2 The illustration of the b^* selection process. Given $V(X)$ vectors for the two classes, denoted by red and green crosses, the b^* is chosen to maximize the distance between the center of projections of $V(X)$ vectors on b

feature pairs with large r^f and $|r^{f_i f_j}|$, respectively. Although the HDMR expansion of the log-likelihood ratio is unique, the actual true weights may be complicated to derive. Since we mostly compare the HDMR expansion against a specific threshold to assign a label to a newly observed point, we may consider the following optimization problem to approximately solve for the desired weights.

$$b^* = \text{argmax}_{b:||b||_2=1} \left(\frac{1}{n_1} \sum_{x \in \mathcal{S}_1} b.V(x) - \frac{1}{n_0} \sum_{x \in \mathcal{S}_0} b.V(x) \right), \tag{11}$$

where $V(x)$ is the collection $S(X_f|\mathcal{S})$'s for all features f such that $r^f > T_1$ and $S(X_{f_i f_j}|\mathcal{S})$ of all feature pairs f_i, f_j such that $|r^{f_i f_j}| > T_2$, and "." denotes inner product. Figure 2 depicts how b^* is selected given vectors $V(X)$ for the training data. Given a new observation X, we find $R(X) = b^*.V(X)$, and we assign class label $\hat{Y} = 1_{R(X)>T}$, where T is a threshold. Note T_1, T_2, and T are parameters of the model. Here they are selected using a grid search within cross validation; however, efficient parameter tuning strategies should be explored (see the Conclusion and Future Work section). We hereafter refer to this classification rule as *linear approximation second order HDMR expansion* (LAS-HDMR). The pseudo-code of LAS-HDMR is provided in Algorithm 1. To summarize, given the data, the machinery that outputs log likelihood ratios of features and feature pairs, and the algorithm parameters, LAS-HDMR computes the risks of individual features and feature pairs, removes weak ones, and computes the weights that maximize the distance between the centers of each class. The overall pipeline is provided in Fig. 3.

The block model extension

The optimization problem in Eq. 11 maximizes the Euclidean distance between the centers of points in different classes, which might be most suitable for cases where elements

Algorithm 1 Pseudo-code of LAS-HDMR

Require: $L(x|x_u)$ for all u with $|u| \leq 2$ and all training points (x, y), thresholds T_1, T_2, and T, and new observation X.

1: For each feature f compute r^f defined in Eq. 9.
2: For each feature pair f_i, f_j compute $r^{f_i f_j}$ defined in Eq. 10.
3: Construct $V(x)$, the collection of $S(x_f)$ for features f such that $r^f > T_1$ and $S(x_{f_i f_j})$ for features fi, f_j such that $|r^{f_i f_j}| > T_2$.
4: Compute b^* using Eq. 11.
5: Compute $R(X) = b^* \cdot V(X)$ for observation X.
6: Compute $\hat{Y} = 1_{R(X)>T}$.

Ensure: $\hat{Y}$.

of $V(x)$ are independent. However, the feature pairs in LAS-HDMR can be heavily correlated. Therefore, it might be useful to merge heavily correlated feature pairs in blocks, average feature pairs of each block, and use the average log-likelihood of each block in $V(x)$. Note that almost any community detection algorithm over graphs can be used to cluster feature pairs into blocks, where each node is a feature pair and the edges measures the correlations between log-likelihood ratios of the two feature pairs (see [24, 25] for a review on graph community detection). We consider the following simple block construction scheme. For each feature f_i construct risk increasing and risk decreasing blocks P^{f_i} and N^{f_i}, respectively, as follows:

$$P^{f_i} = \{f_j : f_j \neq f_i, r^{f_i f_j} > T_2\}, \tag{12}$$

$$N^{f_i} = \{f_j : f_j \neq f_i, r^{f_i f_j} < -T_2\}. \tag{13}$$

Afterwards, for each block P^{f_i} and N^{f_i} compute $r^{P^{f_i}}$ and $r^{N^{f_i}}$, the risks of positive and negative risk feature pairs containing f_i, being the average risks of feature pairs in the block. We then remove "weak" blocks. Weak risk increasing blocks are those for which $r^{P^{f_i}} < T_3$, and weak risk decreasing blocks are those for which $r^{N^{f_i}} > -T_3$. Note T_3 is a parameter of the model that is tuned via a grid search within cross validation. Now, given observation x, $V(x)$ is comprised of the log-likelihood ratio of single features for which $r^f > T_1$ and average $S(X_{f_i f_j})$ of risk increasing and risk decreasing blocks that had absolute average risk large than T_3. We again use Eq. 11 to obtain HDMR coefficients. Finally, given new observation X, $V(X)$ is formed, $R(X) = b^*.V(X)$ is computed, and $\hat{Y} = 1_{\{R(X)>T\}}$ is

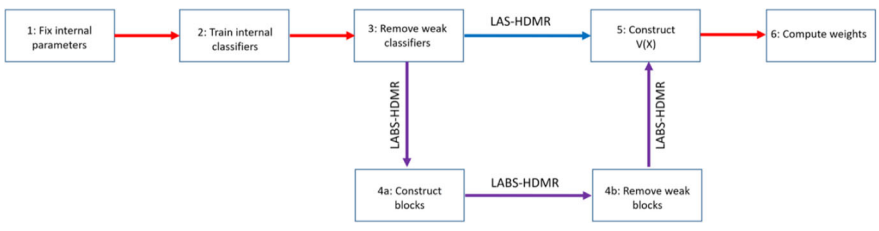

Fig. 3 The overall pipeline of LAS-HDMR and LABS-HDMR. In step 1 the algorithm parameters and the internal classification machineries are selected. In step 2 classifiers using individual features and feature pairs are trained and log likelihood ratios are saved. In step 3 weak classifiers (based on r^f and $|r^{f_i f_j}|$) are removed. LAS-HDMR moves to step 5, but LABS-HDMR moves to steps 4a and 4b, in which feature pairs construct blocks, and weak blocks are removed. In step 5 the final feature vector $V(X)$ is constructed and in step 6 the weights used to combine elements of $V(X)$ are computed

the predicted label. We hereafter call this classification algorithm *linear approximation of block second order HDMR expansion* (LABS-HDMR). The pseudo-code of LABS-HDMR is described in Algorithm 2, and the overall pipeline is provided in Fig. 3.

Algorithm 2 Pseudo-code of LABS-HDMR

Require: $L(x|x_u)$ for all u with $|u| \leq 2$ and all training points (x, y), thresholds T_1, T_2, T_3 and T, and new observation X.

1: For each feature f compute r^f defined in Eq. 9.
2: For each feature pair f_i, f_j compute $r^{f_i f_j}$ defined in Eq. 10.
3: Construct positive and negative risk blocks for each feature f.
4: Construct $V(x)$, the collection of $S(x_f)$ for features f such that $r^f > T_1$, $M_{f_k}^P$ for features f_k that $r^{Pf_k} > T_3$, and $M_{f_l}^N$ for features f_l that $r^{Nf_l} < -T_3$.
5: Compute b^* using Eq. 11.
6: Compute $R(X) = b^* \cdot V(X)$ for observation X.
7: Compute $\hat{Y} = 1_{R(X)>T}$.

Ensure: $\hat{Y}$.

Synthetic simulations

Here we perform several simulations to study LAS-HDMR and LABS-HDMR classifiers in more detail. We use a synthetic model developed to mimic microarrays and gene expression levels for data generation. The original model is proposed in [26], and has been extended in [27, 28]. Here we use the extended model of [27] in which features are markers or non-markers. Markers are either global or heterogeneous, and comprise blocks of size k, where features in the same block are dependent and features in different blocks are independent of each other. Each block of global markers in class 0 is Gaussian with zero mean and covariance $\sigma_0^2 \Sigma_0$, where diagonal elements of Σ_0 are 1 and off-diagonal elements are ρ_0. Global markers in class 1 are either synergetic or marginal. In the synergetic case mean vector of each block in class 1 is $[1, 1/2, \cdots, 1/k]$, and in the marginal case it is $[1, 0, \cdots, 0]$. The covariance matrix of the Gaussian distribution is $\sigma_1^2 \Sigma_1$. Diagonal elements of Σ_1 are 1 and off diagonal elements are ρ_1. Heterogeneous markers are similar to global markers in class 0 but comprise c subclasses in class 1, where in each subclass certain points follow a distribution similar to class 1 global markers and the remaining points are similar to class 0 markers. Non-markers are either low variance or high variance. Low variance non-markers are similar to class 0 markers. High variance non-markers are independent of each other, and each HV non-marker follows a Gaussian mixture of the form $pN(0, \sigma_0^2) + (1 - p)N(1, \sigma_1^2)$, where p is drawn uniformly at random over the interval $[0, 1]$. Note σ_0^2, σ_1^2, ρ_0, ρ_1, k, and c are parameters of the model. Figure 4, adopted from [26], shows an illustration of the feature types developed in this synthetic model.

We now study how LAS-HDMR and LABS-HDMR perform under this model. In this simulation, all features are heterogeneous markers, to create a more difficult problem. We fix $|F| = 60$, $c = 2$, $k = 10$, $\sigma_0^2 = 0.25$, and $\sigma_1^2 = 0.64$ and consider 4 scenarios: (1) synergetic markers with $\rho_0 = \rho_1 = 0.5$, (2) synergetic markers $\rho_0 = 0.1$, and $\rho_1 = 0.9$, (3) marginal markers with $\rho_0 = \rho_1 = 0.5$, and (4) marginal markers with $\rho_0 = 0.1$ and $\rho_1 = 0.9$. We generate a stratified sample of size n (to be specified below) with an equal

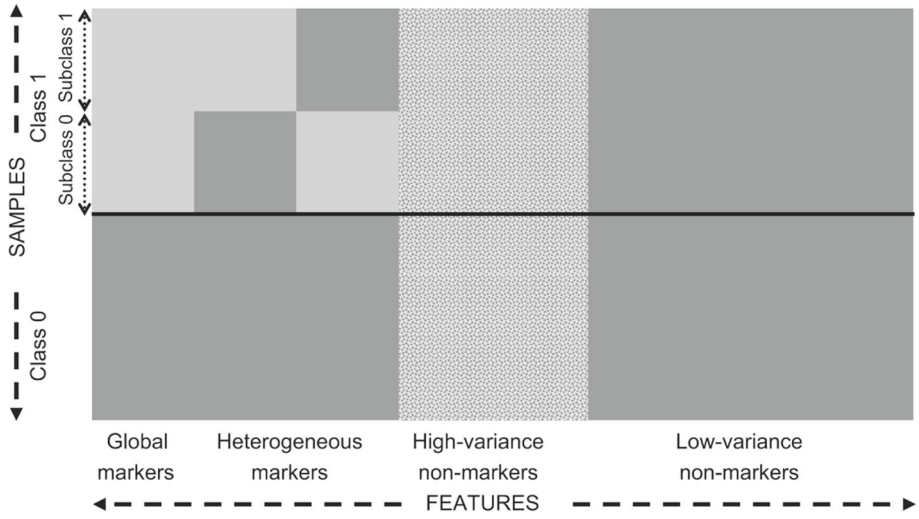

Fig. 4 An illustration of global markers, heterogeneous markers, and non-markers of the data generation model for continuous observations, adopted from [26]

number of points in each class for training, and a stratified sample of size 2000 with an equal number of points in each class for testing. Given training data, several classifiers are trained, which are then applied to the test data. We compute the receiver operator characteristic (ROC) curve and the area under curve (AUC) averaging over 100 iterations. Note large test sets were used to accurately compute prediction errors. Despite the test data being balanced, we believe AUC is a more reliable performance statistic than accuracy, as experimental data are typically imbalanced. That being said, in the current setup, for each point on the ROC curve obtained for threshold T, accuracy is $1 - 0.5(P_I + P_{II})$, where P_I and P_{II} are probabilities of Type I and Type II errors, respectively.

In addition to LAS-HDMR and LABS-HDMR we implement the following classifiers for comparison: regularized quadratic discriminant analysis (RQDA) with regularization value ranging from 0 to 1 in steps of 0.1, regularized linear discriminant analysis (RLDA) with regularization value ranging from 0 to 1 in steps of 0.1, linear support vector machine (SVM), random forest (RF) with the number of tree ranging from 10 to 100 in steps of 20, k nearest neighbors (kNN) with $k = 3, 5, \cdots, 30$, and generalized linear models with linear and quadratic probit links using LASSO and elastic net penalties ($\alpha = 0.5$) with penalty coefficients $\lambda = 0.01 : 0.01 : 0.1$. These methods are discussed in detail in [29–31].

For each family with multiple tuning parameters and for each sample size, we report the largest AUC among all tested parameter values over the test data. Figure 5 plots the AUCs over test data averaging over 100 iterations as the sample size increases from 20 to 80 in steps of 4. When features have similar correlation matrices in both classes, classical methods such as RLDA and RQDA perform best and are closely followed by LAS-HDMR and LABS-HDMR. However, when correlation coefficients differ between the two classes LAS-HDMR and LABS-HDMR outperform other tested classifiers. Here we observe little difference between LAS-HDMR and LABS-HDMR, suggesting we do not need to merge feature pairs into blocks and the number of feature pairs to consider is not too large for this problem. We leave a more thorough comparison of LAS-HDMR and LABS-HDMR for future work. ROC plots are provided in the Supplementary.

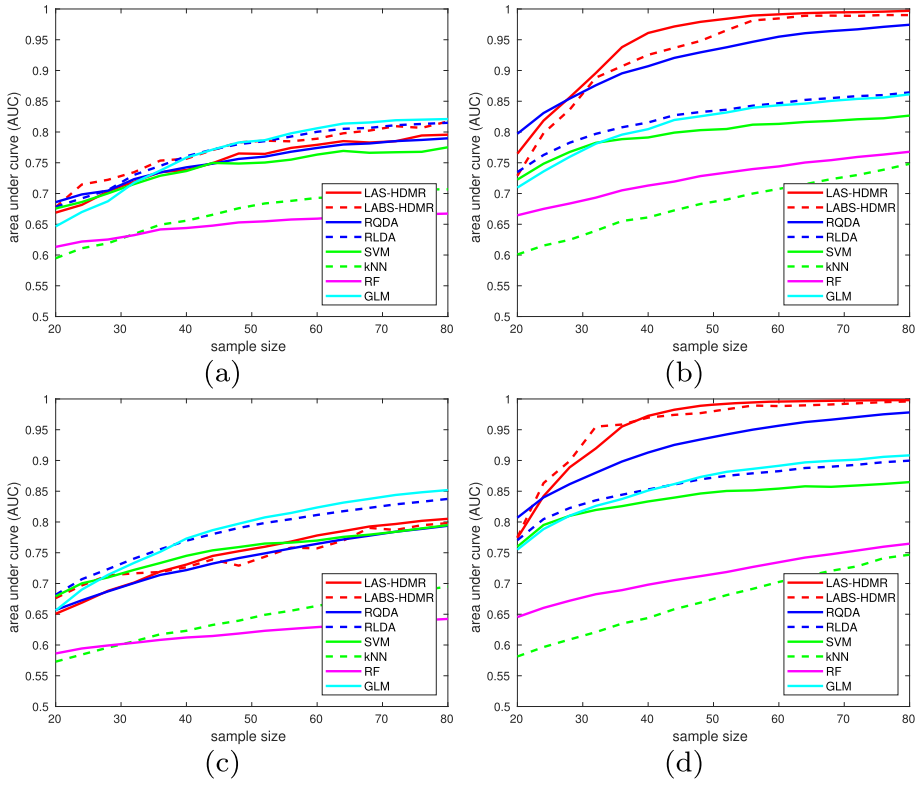

Fig. 5 AUC of different classification algorithms versus sample size for different data generation models: (**a**) synergetic features and equal correlations, (**b**) synergetic features and unequal correlations, (**c**) marginal features and equal correlations, and (**d**) marginal features and unequal correlations

Note that we started from the problem of finding the "best" second order representation of the log-likelihood ratio, but, due to computational difficulties, had to make several assumptions and approximations along the way. Therefore, it is possible that we end up mis-specifying the exact second order HDMR decomposition. In such scenarios, it can be very probable that another method outperforms LAS-HDMR and LABS-HDMR. Note LAS-HDMR and LABS-HDMR enjoy competitive overall performance in all tested scenarios, and outperform other methods when correlation coefficients differ between the classes. They are competitive methods, and hence can be suitable for a wide range of problems. Note that settings where LAS-HDMR and LABS-HDMR do not perform best correspond to those in which correlation coefficients are equal in both classes, for which RLDA and linear probit models perform best. This suggests maybe in these cases the first order HDMR expansion is more appropriate to represent the data (LDA is equivalent to a first order HDMR expansion under its modeling assumptions), although variances are slightly different between the classes. The small sample sizes used in this simulation may impede quadratic classifiers to satisfactorily estimate the distribution parameters, which may result in their poor performance. Similar patterns are observed in [32, 33]. Additionally, given that the first order expansion is sufficient to represent data, a second order HDMR model might suffer curse of dimensionality.

Identifying pairwise interactions

Here we evaluate the performance of MTM in identifying significant pairwise feature interactions. A comparison of MTM with the method of [9] is provided in the Supplementary. We fix $|F| = 5000$, with 20 global markers, 80 heterogeneous markers, and 2000 high variance non-markers. We again assume $k = 10$, $c = 2$, $\sigma_0^2 = 0.25$, $\sigma_1^2 = 0.64$, and consider the 4 scenarios of the previous section for mean types and correlation coefficients. Figure 6 provides the ROC curves of MTM for different marker and correlation values when $n = 40$, averaging over 100 iterations. MTM performs best when correlations are different between the two classes: red and blue lines denoting unequal correlations for synergetic and marginal markers, respectively, enjoy a higher probably of detection compared with black and magenta lines denoting equal correlations for synergetic and marginal markers, respectively. Note the mean types (marginal or synergetic) have little effect on the ROC curves. Figure 7 provides ROC curves of the equal correlation cases for different sample sizes, averaging over 100 iterations. As expected, with increase in the sample size it becomes easier to detect pairwise interactions. To benchmark MTM we compared it with the absolute conversion method of [9], which is proposed as a "fast" algorithm. However, we observed it is computationally more intensive than MTM. We considered marginal and synergetic markers with equal correlations, fixed $n = 40$, and reduced the number of iterations to 50. Results are provided in the Supplementary, in which MTM outperforms the method of [9].

Experimental data analysis

We apply LAS-HDMR, LABS-HDMR, and the comparison classifiers of the previous section to datasets studying relapsing breast and lung cancer patients. We also evaluate if MTM can detect significant pairwise gene interactions in realistic settings. We specifically selected datasets resulting in tasks more challenging than healthy versus normal labels. Such datasets are in particular challenging as data can be small in size and imbalanced (only a small portion of followed up patients may relapse). Additionally, breast

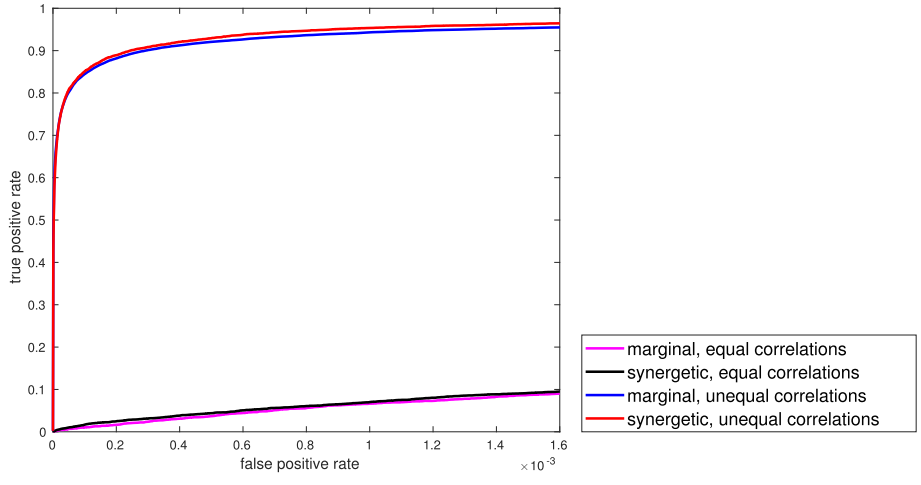

Fig. 6 ROC curves of MTM method for detecting pairwise feature interactions for different marker and correlation types when $n = 40$. Marker's class-conditioned means have almost no effect on the ROC curves as we look for certain structures on the covariance matrix; however, correlation structures have a much larger impact on ROCs. Probability of detection of much lower when correlations are equal for a fixed false alarm rate

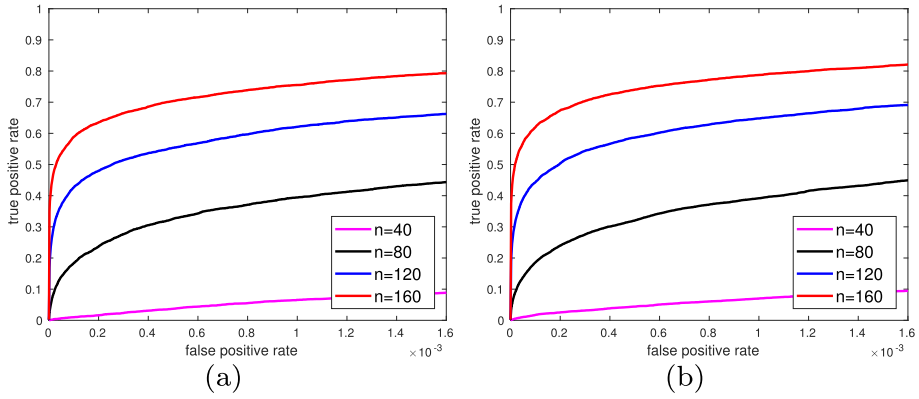

Fig. 7 ROC curves of MTM for detecting pairwise interactions for different sample sizes for (**a**) marginal and (**b**) synergetic markers with equal correlations in both classes. Marker means have almost no effect on the ROCs. MTM's ability to grasp dysregulated interactions in the challenging case of equal correlation coefficients rapidly improves as sample size increases

and lung cancers are well studied in the literature, allowing us to evaluate if the detected patterns are biologically plausible. A leukemia dataset is studied in the Supplementary.

Breast cancer

The data obtained in [34] and [35] deposited on gene expression omnibus (GEO) database [36] with accession number GSE25066, containing expression levels of 397 relapse free and 111 relapsing breast cancer patients, all of whom went through neoadjuvant taxane-anthracycline chemotherapy. Data is based on the GPL96 platform, and is already pre-processed and normalized. The dataset contains 22,283 probes, of which 20,967 map to genes. We only use probes that map to genes in our analysis. First, 100 relapsing and 360 non-relapsing patients randomly select as training data, and the remaining points are used for testing. The likelihood ratio test (LRT) statistic of [37], which is equivalent to the optimal Bayesian filter scoring function under independent Gaussian models [27, 38] under Jeffreys prior, is used to select the top 100 differentially expressed genes. We iterate 100 times.

2D and 3D t-SNE [39] plots using the cityblock distance are provided in Fig. 8, suggesting the two classes do separate. It seems each class contains a few points which may truly belong to the other class, i.e., each class is polluted with a small subpopulation truly belonging to the other class. Alternatively, larger follow-up times may be necessary to further determine if certain non-relapsing patients relapse, and hence should belong to class 1. The large number of non-relapsing patients that resemble relapsing patients reduces the measured AUC. Table 1 lists the AUC of different classifiers on this dataset (over the hold out test data). Figure 9a provides the ROC plots.

As Table 1 suggests, all methods do not enjoy a high AUC. We observed that the variant of LAS-HDMR using RQDA with $\lambda = 0.8$ achieved the highest AUC. The largest AUC

Table 1 AUC of classification algorithms for the cancer datasets

method	LAS-HDMR	LABS-HDMR	RQDA	RLDA	SVM	RF	kNN	GLM
breast cancer	64.21%	**69.95%**	62.70%	66.04%	55.34%	60.87%	58.18%	67.55%
lung cancer	66.56%	67.67%	65.47%	66.62%	63.03%	66.87%	61.60%	**68.22%**

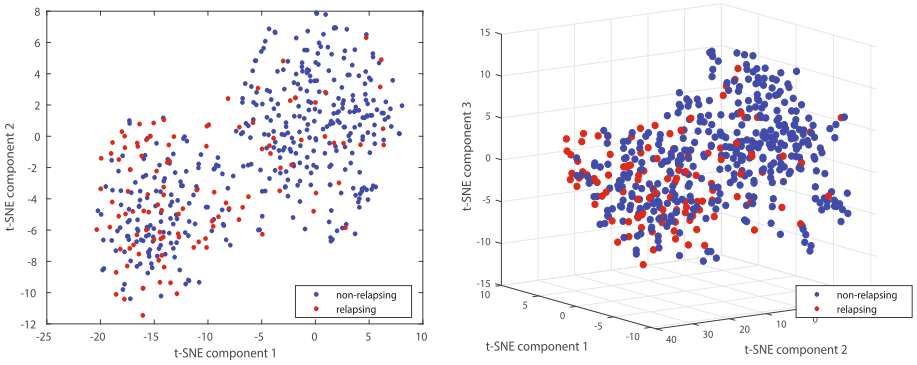

Fig. 8 t-SNE plots of the breast cancer dataset separating the relapsing (high risk) and non-relapsing (low risk) patients: (**a**) 2D and (**b**) 3D plots. Blue and red denote the non-relapsing and relapsing patients, respectively. Few relapsing patients resemble non-relapsing patients while a larger portion of non-relapsing patients resemble relapsing patients, suggesting each class may be polluted with a sub-population of points that truly belong to the other class

for the variant of LAS-HDMR using RLDA was 63.12% obtained for $\lambda = 0.9$. In contrast, LABS-HDMR seems to enjoy the highest AUC, obtained using RQDA with $\lambda = 0.1$, which is the closest tested variant to conventional QDA. This may suggest that a second order expansion is not satisfactory enough for this dataset, emphasizing the need to look at higher order expansions. Finally, Fig. 10a provides the Kaplan-Meier survivorship plots based on the assigned labels to the test data, averaging over 100 iterations for LAS-HDMR and LABS-HDMR. The figure provides extra assurance that indeed the proposed algorithms separate the two classes, the approximate second order HDMR expansion of the log-likelihood ratio, i.e., $R(X)$, is an appropriate statistic to denote the "risk" of an event, and the proposed methods can be further used in conjunction with other data analysis tools. As t-SNE plots in Fig. 8 suggest, many low risk patients resemble high risk patients, and we expect a well-designed classification rule to mislabel such points; otherwise, the separating plane (curve) should be extremely complex, raising serious concerns of overfitting. This explains why many high risk patients have not relapsed up to the follow-up time.

Although LAS-HDMR and LABS-HDMR may not yield high classification accuracy similar to other classification algorithms, their glass box nature simplifies the process of identifying genes and gene pairs that contribute the most to the classifier's prediction. We use all of the data for training, and use RQDA with $\lambda = 0.8$ to obtain the log-likelihood ratios. Table 2 lists the top 10 genes of LAS-HDMR and their risks. Many of the top genes, such as IL8 [40], also known as C-X-C motif chemokine ligand 8 (CXCL8), and growth regulating estrogen receptor binding 1 (GREB1) [41] are suggested to be affected in breast cancer. Table 3 lists the top 10 LAS-HDMR gene pairs. Comparing Tables 2 and 3 we observe that gene interactions tend to have a larger risk than individual features. Note all risks describe the average increase/decrease of the log likelihood ratio, and are hence on the same scale for all genes and gene pairs. Scatter plots of several gene pairs with interesting interactions is provided in the Supplementary). For example, we observed either GREB1 or carboxypeptidase B1 (CPB1) are over-expressed among non-relapsing patients, and under-expression of both GREB1 and CPB1 is necessary to have a high risk

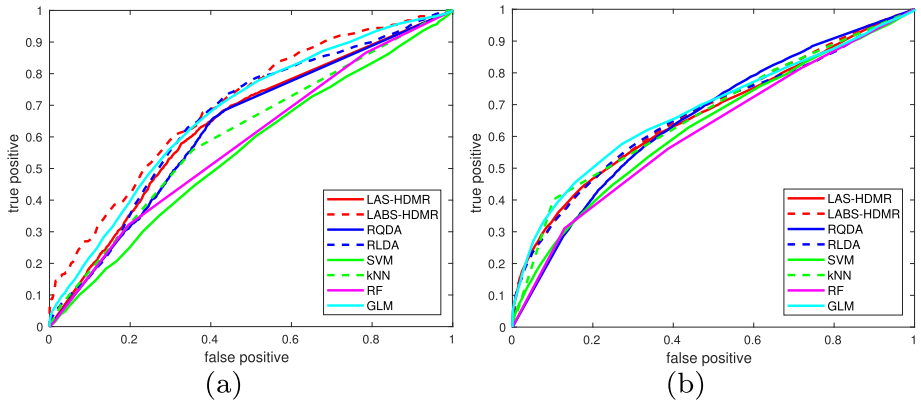

Fig. 9 ROC curve of different classifiers for the (**a**) breast and (**b**) lung cancer datasets: (**a**) In the breast cancer dataset LABS-HDMR uses RQDA to infer log-likelihood ratio of partial observations, greatly improves RQDA's accuracy, and is followed by GLMs. LAS-HDMR closely follows GLM and RLDA, uses RQDA in its design, improves RQDA's accuracy for small false positives, and performs similar to RQDA for large false alarms. (**b**) In the lung cancer dataset GLM performs best, but is closely followed by LAS-HDMR and LABS-HDMR, which both improve on the AUC of the intrinsic machinery they use for computing log-likelihood ratio of partial observations

of relapse. Finally, many of the top gene-gene interaction pairs contain GREB1, signal peptide CUB domain EGF-like 2 (SCUBE2), GATA Binding Protein 3 (GATA3), and IL8, suggesting their interaction might be key to studying breast cancer. The variant of LABS-HDMR that achieved the highest AUC used RQDA with $\lambda = 0.1$, and is studied in detail in the Supplementary.

Now we look for significant pairwise gene interactions. First, for each gene, we only consider the probe ranking highest by LRT so that probes mapping to the same genes do not disrupt the analysis, resulting in 13,211 different genes. This results in 87,258,655 tests. We observed that MTM can be heavily affected by small subpopulations, heavy tails, and outliers, and therefore used MATLAB's built in `isoutlier` function with its default settings to remove potential outliers before further analysis. Bounding the false discovery rate (FDR) by 5% using the Benjamini-Hochberg (BH) procedure [42] 1,275,351 pairwise

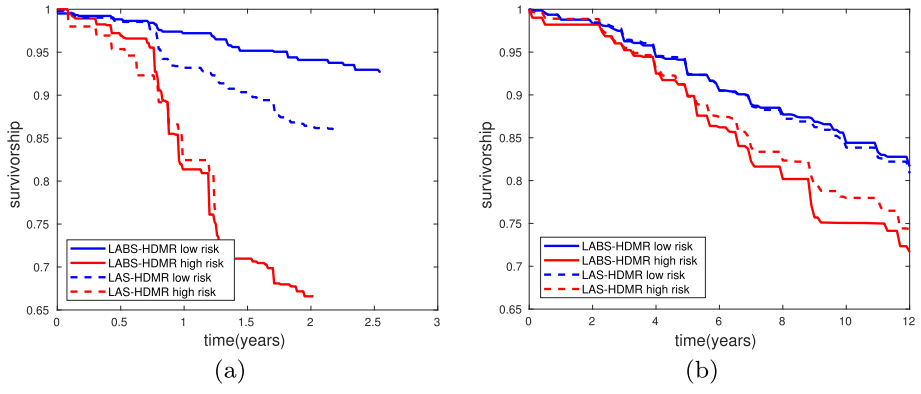

Fig. 10 Kaplan-Meier survivorship plots for the (**a**) breast and (**b**) lung cancer datasets. (**a**) In the breast cancer dataset high/low risk patients have a high/low chance of relapse withing the follow-up time. (**b**) In the lung cancer dataset high/low risk patients have a high/low chance of relapse or death due to cancer withing the follow-up time. Both LABS-HDMR and LAS-HDMR do separate the patients given input expression levels

Table 2 Top breast cancer genes used for classification by LAS-HDMR

Rank	Gene	Risk	Rank	Gene	Risk
1	ORM1, ORM2	0.92	6	ACADSB	0.75
2	IL8	0.87	7	PTOV1	0.75
3	ZNF395	0.85	8	ZNF673	0.75
4	GREB1	0.8	9	AR	0.74
5	TBC1D9	0.78	10	LGALS8	0.71

interactions are significant (about 1.46% of tested hypotheses). Table 4 lists several of the top gene pairs and their adjusted p-values. Figure 11 provides scatter plots of several gene pairs. We observe many interesting patterns that require further investigation: (1) under-expression of both SAM pointed domain containing ETS transcription factor (SPDEF) and MLPH, also known as synaptotagmin-like protein 2A (SLAC2A), increases the risk of relapse, (2) over-expression of anterior gradient protein 2 homolog (AGR2) and N-Acetyltransferase 1 (NAT1) is an indicator of low risk, over-expression of AGR2 and under-expression of NAT1 is an indicator of "medium" risk, and under-expression of both AGR2 and NAT1 is an indicator of high risk, (3) over-expression of NAT1 or DNALI1 is an indicator of low relapse risk. Comparing Fig. 11a and d suggests solute carrier family 2 member 10 (SLC2A10) and MLPH are heavily correlated with a positive correlation coefficient, which is indeed observed in the data as well.

Considering a weighted graph where nodes are genes and edge weights are $-\log p$-value of the gene pair, we observed many detected gene pairs construct highly connected clusters. To verify if the selected gene pairs are biologically relevant we (a) associated each gene with its smallest gene pair p-value, (b) selected the top 200 genes, (c) constructed their graph, (d) used community detection algorithm of [43] to identify network clusters, (e) selected the genes corresponding to the largest cluster, and (f) used Ingenuity Pathway Analysis[1] (IPA) [44] to identify the networks associated with these genes only using experimentally observed results as well as the top canonical pathways. The top canonical pathways and the largest detected network are provided in Figs. 12 and 13, respectively. Note the top ranking IPA gene network is identified with cellular development, cellular growth, and cell cycle functions. The log fold changes, computed using method of [45], were used to identify over/under-expressed genes in the network; however, as data is highly heterogeneous, these effects might not be highly pronounced. Many of the selected genes are connected directly or indirectly with only one gene in between. Literature review suggests many of the top IPA pathways are also affected in breast cancer.

We now randomly leave out one point in each class, train LAS-HDMR, and look at the genes and gene pair that yield the highest scores in absolute values for these two test points. The minimum value for HDMR terms, either $S(X_f)$ or $S(X_{f_i,f_j})$, was -3.62 and -0.59 for the points in classes 0 and 1, corresponding to gene pairs GREB1 and NAT1, and ZNF673 and ZNF391, respectively. The largest values were 3.64 and 14.32 respectively, corresponding to the ERBB2 gene, and gene pair ZNF395 and PTOV1. Figure 14 plots how the different genes and gene pairs are combined to arrive at the final log-likelihood ratio estimate.

[1] QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis

Table 3 Top LAS-HDMR gene pairs of the breast cancer dataset

Rank	Gene 1	Gene 2	Risk
1	GREB1	SCUBE2	2.24
2	GREB1	CPB1	2.23
3	ORM1, ORM2	GREB1	2.22
4	GREB1	IL8	2.22
5	ZNF395	GREB1	2.1
6	GREB1	GATA3	2.07
7	GREB1	NAT1	2.06
8	GREB1	TBC1D9	2.05
9	GREB1	ACADSB	2.03
n10	ORM1, ORM2	SCUBE2	2.02

Lung cancer

Data obtained in [46] is deposited on GEO with accession number GSE68465, containing expressions of 443 lung cancer patients. 279 patients whose cancer relapsed or died within the follow up time comprise class 1, and the remaining 164 patients comprise class 0. This dataset is based on the GPL96 platform. We again only use probes mapping to genes, perform a log-normalization step, randomly select 250 points in class 0 and 140 points in class 1 for training, use the remaining points for testing, use the top 100 LRT genes for classifier design which we evaluate on test data, and iterate 100 times. Before we train the classifiers we provide 2D and 3D t-SNE [39] plots using the cityblock distance in Fig. 15, suggesting the two classes do separate; however, many patients who relapsed or died within six year (high risk) resemble those who survived (low risk). Both t-SNE plots suggest there are at least two high risk subpopulations.

Table 1 lists the AUCs (over test data), and Fig. 9b provides the ROC curves. Again we observe that none of the classifiers enjoy a very high AUC, and both LAS-HDMR and LABS-HDMR enjoy competitive performance compared with other classifiers. This may again suggest that a quadratic model might not be enough to capture the complicate structure of data. In this dataset, the variants of LAS-HDMR and LABS-HDMR achieving the highest AUCs are RLDA with $\lambda = 0.1$ and RLDA with $\lambda = 0.2$, respectively. Figure 10b provides the Kaplan-Meier survivorship plots based on the assigned labels to the test data, averaging over 100 iterations for LAS-HDMR and LABS-HDMR, providing extra assurance that indeed the proposed algorithms separate the two classes. Here, the large ratio of high risk patients who resemble low risk ones in the t-SNE plots of Fig. 15 suggest that

Table 4 Top breast cancer gene pairs and adjusted *p*-values ($\times 10^{-25}$)

Rank	gene 1	gene 2	adj *p*-value
1	SPDEF	MLPH	$< 10^{-4}$
2	MSN	SPDEF	$< 10^{-4}$
3	SCGB2A2	SCGB1D2	$< 10^{-4}$
4	TFF3	SPDEF	$< 10^{-4}$
5	RHOB	SPDEF	$< 10^{-4}$
11	SLC44A4	SPDEF	0.0014
12	SPDEF	FAM174B	0.0064
13	FOXA1	NAT1	0.0128
14	GATA3	SPDEF	0.0477
15	TSPAN1	SPDEF	0.0477

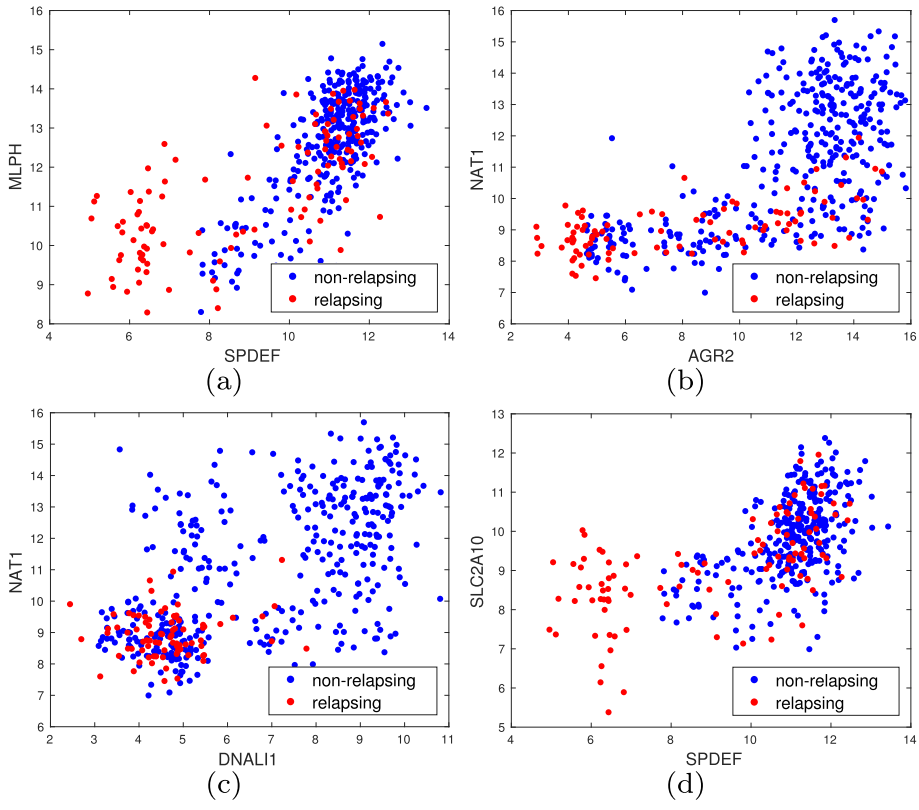

Fig. 11 Scatter plots of gene pairs with significant interactions for the breast cancer dataset. Red denotes relapsing breast cancer patient within the follow-up time and blue denotes patients who did not relapse: (**a**) decreased expression of SPDEF and MLPH increases relapse risk, (**b**) over-expression of both AGR2 and NAT1 is an indicator of low relapse risk, decreased expression of NAT1 and over-expression of AGR2 seems to result in medium risk, and patients for whom both AGR2 and NAT1 have decreased expressions are in high risk of relapse, (**c**) to have a high risk of relapse both DNALI1 and NAT1 must have decreased expression, i.e., over-expression of either gene is an indicator of low relapse risk, (**d**) decreased expression of both SPDEF and SCL2A10 is an indicator of high relapse risk

a reasonable decision rule would mislabel many high risk patients as low risk, explaining the low survivorship of the estimated low risk patients in Fig. 10b.

Table 5 lists the top 10 genes of LAS-HDMR and their associated risks. We again observe that many of the top genes, such as bromodomain PHD finger transcription factor (BPTF) [47] and LUC7 like 3 pre-MRNA splicing factor (LUC7L3) [48], are shown or

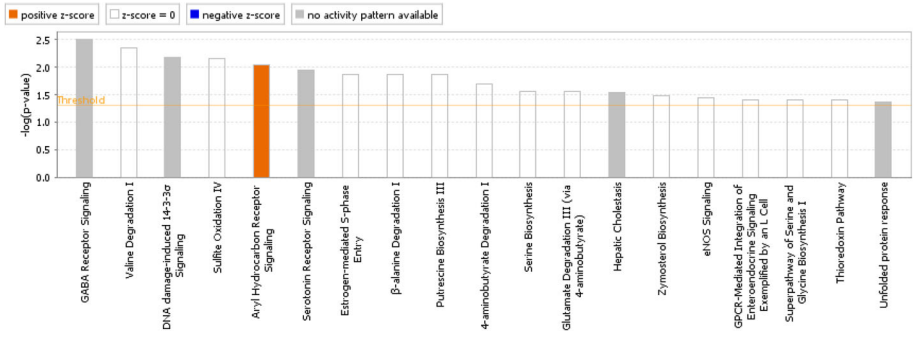

Fig. 12 Top IPA canonical pathways of the largest MTM gene cluster of the breast cancer dataset

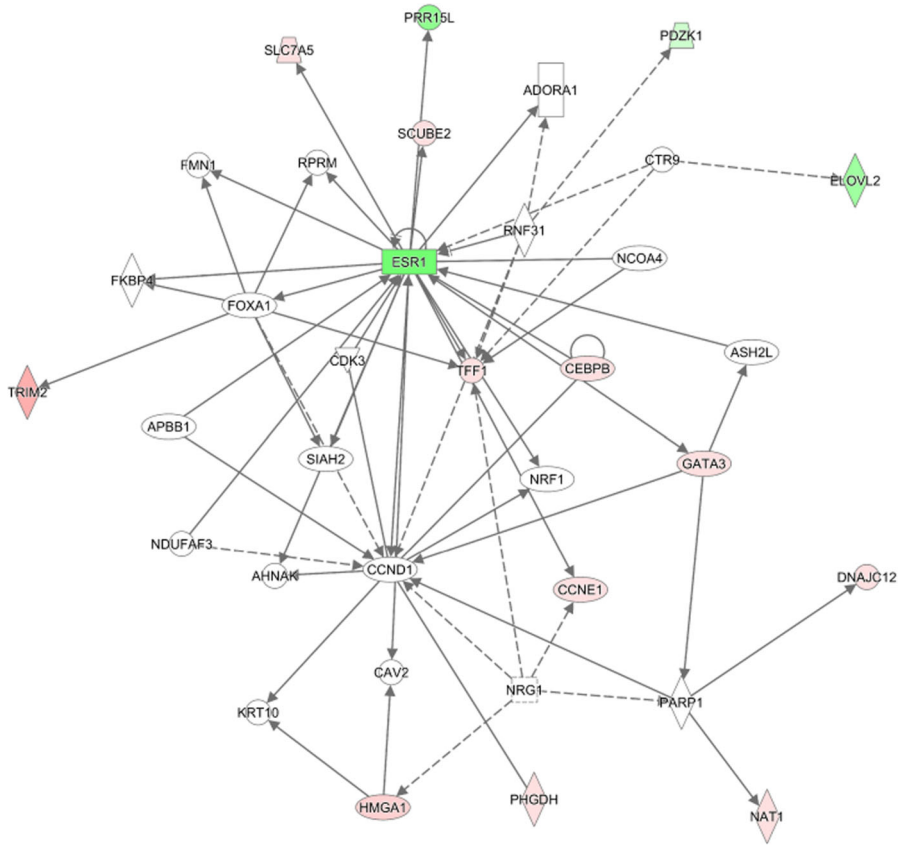

Fig. 13 The largest IPA gene network corresponding to the largest gene cluster of MTM for the breast cancer dataset. Red/Green denote genes which have increased/decreased expression among relapsing patients compared with non-relapsing patients. Solid/Dashed lines denote direct/indirect interactions. Genes highlighted with magenta borders are identified with cellular development, cellular growth, and cell cycle functions. ESR1 and CCND1 seem to be interacting with each other as well as many other genes whose interactions are affected in breast cancer. The faded colors denote smaller fold changes. As many genes involved in breast cancer relapse are heterogeneous with a small sub-population heavily affected, fold changes might fail to grasp the extent of the gene effect

suggested to be affected in lung cancer. Table 6 lists the top 10 gene pairs and their associated risks. A deeper analysis, including the results of LABS-HDMR, is provided in the Supplementary.

For each gene we only use the probe ranking highest by LRT giving us 13,211 different genes and 8,758,655 pairwise dependence tests. We also perform outlier detection to further improve identifying general gene interaction patterns. Bounding FDR by 5% using BH 701410 gene pairs are significant, about 0.8% of all tests. Table 7 lists several of the top

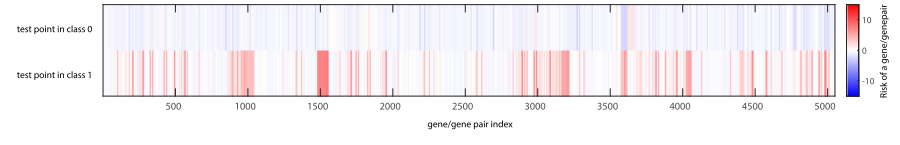

Fig. 14 The associated un-normalized risk scores, i.e., $S(X_f)$ and $S(X_{f_i,f_j})$, of two test points. Genes and gene pairs which are not used in the decision rule automatically get a risk of zero. Note most genes/gene pairs assign a larger risk to the point in class 1 (high risk patients) to the point in class 0 (low risk patient)

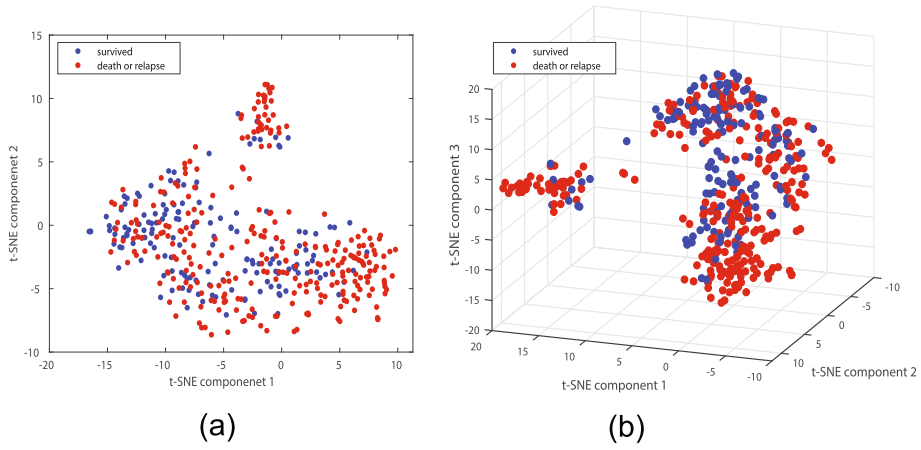

Fig. 15 t-SNE plots of the lung cancer dataset separating the relapsing or dead (high risk) and surviving (low risk) patients: (**a**) 2D and (**b**) 3D plots. In both figures blue and red denote the surviving and relapsing/dead patients, respectively. Both t-SNE plots suggest there are at least two high risk subpopulations

gene pairs and their adjusted *p*-values. Figure 16 provides scatter plots. We again observe MTM detects interesting pairwise interactions: (a) It seems there is a subpopulation of high risk lung cancer patients with poor survival for whom BCL2 associated transcription factor 1 (BCLAF1) is under-expressed and interleukin enhancer binding factor 3 (ILF3) is over-expressed. For other patients, irrespective of their label, these two genes are positively correlated. (b) Patients for whom both laminin subunit gamma 2 (LAMC2) and cadherin 3 (CDH3) are over-expressed have a high chance of relapse/death, those for whom CDH3 is over expressed and LAMC2 is under-expressed have a "medium" chance of relapse/death, and those for whom both CDH3 and LAMC2 are under-expressed have a low chance of relapse/death. (c) Under expression of either BCLAF or SOS Ras/Rho guanine nucleotide exchange factor 2 (SOS2) can be used as an indicator of poor survival. (d) Patients for whom both Annexin A1 (ANXA1) and CDH3 are over-expressed have a medium chance of relapse/death, those for whom CDH3 is over expressed and ANXA1 is under-expressed have a high chance of relapse/death, and those for whom both CDH3 and ANXA1 are under-expressed have a low chance of relapse/death. Finally, we again perform the IPA analysis similar to the breast cancer dataset, expect we include highly probable interactions in the analysis as well as experimentally observed ones. Figure 17 plots the detected network, which is associated with cell cycle, cellular assembly and organization, and cellular function and maintenance. Again observe many of the selected genes are connected with at most two genes in between. A deeper analysis is provided in the Supplementary.

Table 5 Top lung cancer genes used for classification by LAS-HDMR

Rank	Gene	Risk	Rank	Gene	Risk
1	BPTF	2.5955	6	KIAA1033	1.5576
2	SEC63	2.4024	7	LUC7L3	1.398
3	UBXN4	1.7646	8	SON	1.3294
4	SRSF2IP	1.6672	9	PPIG	1.277
5	ATRX	1.5954	10	SF3B1	1.2601

Table 6 Top LAS-HDMR gene pairs for the lung cancer dataset

Rank	Gene 1	Gene 2	Risk
1	ATRX	DDX17	0.4683
2	ZEB1	BCLAF1	0.4663
3	IQGAP1	BPTF	0.4583
4	SON	UBXN4	0.4486
5	PRMT2	RUFY3	0.4433
6	ATP6V1G2, BAT1	SEC63	0.4402
7	LUC7L3	SMC5	0.4275
8	RBL2	MLL	0.3951
9	SRSF2IP	ENC1	0.3643
10	SPIN1	UBE2W	0.356

Discussion

In the simulations and real data analyses we observed that both LAS-HDMR and LABS-HDMR enjoy competitive prediction accuracies compared with several popular classification rules. Additionally, they explicitly reveal how individual features or pairwise feature interactions motivate certain decisions, and how unique patterns of a new observation motivate its predicted label. Scatter plots of breast and lung cancer gene pairs in Figs. 11 and 16, respectively, suggest a quadratic classifier is adequate for predicting class labels from each gene pair; however, AUCs are not very high. This suggests higher order expansions, i.e., the joint interaction of three genes or more, are necessary to increase prediction accuracy. In these datasets we observe LABS-HDMR assigns larger risks to the top gene pairs compared with individual genes (see Tables 2 and 3 for breast cancer and Tables 5 and 6 for lung cancer), suggesting gene interactions are crucial to reliable prediction; not that linear classifiers miss important information in the data, but that pairwise interactions seem to carry more information about class labels than individual genes. In the leukemia dataset (studied in the Supplementary) we observed both LAS-HDMR and LAS-HDMR perform competitively, but all methods seem to enjoy high AUCs (AUC was larger than 94% for all classifiers). In particular, we observed highest AUCs for linear classifiers, suggesting there is no need to use more complex rules. In particular, quadratic rules such as RQDA perform inferior to LAS-HDMR and LABS-HDMR. Finally, in the leukemia dataset we observe gene pairs have much smaller risks compared with individual genes.

Table 7 Top lung cancer gene pairs and adjusted *p*-values ($\times 10^{-8}$)

Rank	Gene 1	Gene 2	adj *p*-value
1	BCLAF1	ILF3	$< 10^{-4}$
2	CDH3	CST6	0.0001
3	LAMC2	CDH3	0.0001
4	CDH3	PLAU	0.0006
5	S100A10	CDH3	0.0014
6	SMARCC1	BCLAF1	0.0015
7	BCLAF1	PCM1	0.004
8	ITGA3	CDH3	0.004
9	KRT19	CDH3	0.0064
10	BCLAF1	UBN1	0.0088

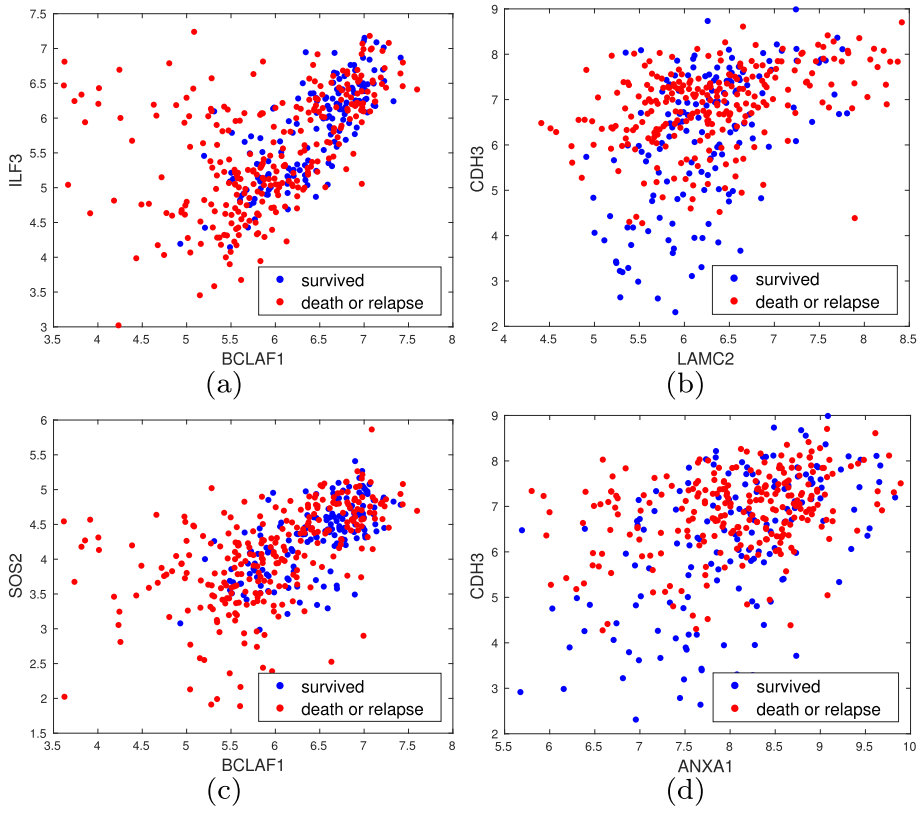

Fig. 16 Scatter plots of gene pairs with significant interactions for the lung cancer dataset. Blue denotes patients who survived for at least sic years after diagnosis (low risk) and red denotes patients who died or relapsed within six years (high risk). It seems: (**a**) for a subpopulation of high risk patients BCLAF1 has decreased expression and ILF3 is over-expressed, (**b**) decreased expression of LAMC2 and CDH3 is an indicator of low risk, decreased expression of LAMC2 and over-expression of CDH3 is an indicator of "medium" risk, and over-expression of both is an indicator of high risk, (**c**) decreased expression of BCLAF1 or SOS2 compared with each other is an indicator of high risk, and (**d**) decreased expression of CDH3 is an indicator of low risk, over-expression of CDH3 and low expression of ANXA1 is an indicator of high risk, and over-expression of both results in medium risk

MTM is an interesting test, allowing to extract co-expression patterns that differ between the two classes, i.e., identify pairwise interactions that differ between the two classes. MTM efficiently takes advantage of information in the data and is computationally fast. This is in contrast to several recent pipelines proposed for analyzing gene-pairs that are rather computationally intensive or may rely on large datasets (see [4, 5, 7–10] for examples).

Conclusion and future work

Glass box models for binary classification open many avenues of research for analyzing genomic data as they enable us to make meaningful hypotheses about the underlying biological dysfunctions that are involved in a disease. To that end, HDMR seems to be an interesting theory for studying low dimensional glass-box models. The limitation of this approach in its current form is the assumption of known mechanisms that output the log-likelihood ratios given any partial observation. On the other hand, different feature sets can use different classification rules tailored to their joint distribution. This provides

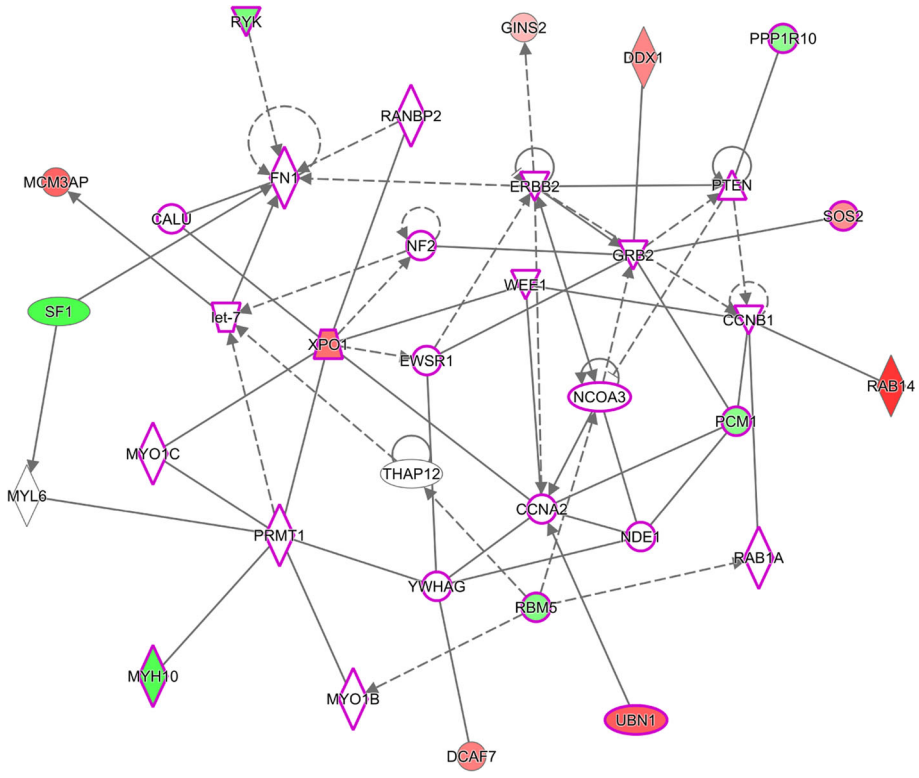

Fig. 17 The largest IPA gene network corresponding to the largest gene cluster of MTM for the lung cancer dataset. Green/Red denote decreased/increased expression among high risk patients compared with low risk patients. Solid/Dashed lines denote direct/indirect interactions. Genes highlighted with magenta borders are associated with cell cycle, cellular assembly and organization, and cellular function and maintenance. Many dysregulated gene pairs seem to be affected via a mediating gene. Note the effect of each individual gene is more pronounced than the breast cancer dataset

HDMR with tremendous flexibility, for instance, we may use QDA for some feature while a GLM is used for another feature pair. While this is a very exciting potential benefit of HDMR, we did not really exploit it in our current analysis and leave its careful consideration for possible future work. Another future direction for improving LAS-HDMR and LABS-HDMR is the development of more efficient methods of parameter selection to reduce computation cost of the currently implemented grid search approach.

The ability of MTM to identify pairwise interactions containing information not encoded in the likelihood function of each feature is a very practical contribution of our work. Here, instead of determining if two features are dependent, the goal is to verify if the pairwise interactions contain additional information about the classes which no model can extract by observing features individually. To do so, we developed a test where the null assumes the first order HDMR expansion is sufficient to explain the class differences. In the case of Gaussian features this translates into testing specific structures on the covariance matrices. Synthetic simulations and experimental data analyses suggest that MTM is indeed a powerful tool to extract dysregulated pairwise gene co-expression patterns that motivate new hypotheses about cancer biology. In the real data analyses we observed many pairs might have a gene in common, for instance, SPDEF in the breast cancer dataset and BCLAF1 and CDH3 in the lung cancer dataset. These patterns motivate hypotheses

not only about gene pairs, but more generally about heavily correlated marker families. Graph representations are interesting tools for analyzing families of pairwise interactions, and graph community detection algorithms can infer marker interaction blocks given pairwise interaction graphs. However, future work should investigate the use of HDMR for directly inferring such structures. To that end, LABS-HDMR seems to be an ideal first step, where its constructed blocks can be potential first approximations to marker families of interest.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3486-x.

Additional file 1: Supplementary. The supplementary contains extra information regarding the synthetic simulations and real data analysis. It also studies a Leukemia dataset not discussed in the main manuscript.

Abbreviations

AGR2: Anterior gradient protein 2 homolog; ANXA1: Annexin A1; AUC: Area under the ROC curve; BCLAF1: BCL2 associated transcription factor 1; BH: Benjamini-Hochberg step-up procedure; BPTF: Bromodomain PHD finger transcription factor; CDH3: Cadherin 3; CPB1: Carboxypeptidase B1 ; CXCL8: C-X-C motif chemokine ligand 8; FDR: False discovery rate; GATA3: GATA binding protein 3; GEO: Gene expression omnibus; GLM: Generalized linear model; GREB1: Growth regulating estrogen receptor binding 1; HDMR: High dimensional model representation; ILF3: Interleukin enhancer binding factor 3; IPA: Ingenuity pathway analysis; kNN: k nearest neighbors; LABS-HDMR: Linear approximation of block second order HDMR expansion; LAMC2: Laminin subunit gamma 2; LAS-HDMR: Linear approximation second order HDMR expansion; LASSO: Least absolute shrinkage and selection operator; LDA: Linear discriminant analysis; LRT: Likelihood ratio test; LUC7L3: LUC7 like 3 pre-MRNA splicing factor; MSE: Mean squared error; MTM: Multiple test mixing for pairwise interactions; NAT1: N-Acetyltransferase 1; QDA: Quadratic discriminant analysis; RF: Random forest; ROC: Receiver operator characteristic; RLDA: Regularized linear discriminant analysis; RQDA: Regularized quadratic discriminant analysis; SLC2A10: Solute carrier family 2 member 10; SOS2: Ras/Rho guanine nucleotide exchange factor 2; SPDEF: SAM pointed domain containing ETS transcription factor; SCUBE2: Signal peptide CUB domain EGF-like 2; SLAC2A: Synaptotagmin-like protein 2A; SVM: Support vector machine; TCGA: The cancer genome atlas

Acknowledgements

Not applicable.

Authors' contributions

AF and GR developed the proposed approach and performed analyses. AF drafted the initial manuscript. LAD, AF, MP, and GR provided comments on the draft and helped edit the manuscript. All authors reviewed and approved the final manuscript. AF is currently at the Jackson laboratory for genomic medicine.

Funding

This work was supported by the National Science Foundation (DMS 1853587 and DMS 1923038 to GR). The funding body did not play any role in the design of the study, analysis and interpretation of data, or in writing the manuscript.

Availability of data and materials

All cancer data used in this work is publicly available via GEO database accession numbers GSE25066 and GSE68465. The synthetic datasets are available upon request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Electrical and Computer Engineering, The Ohio State University, 205 Dreese laboratories, 2015 Neil Ave., 43210 Columbus, USA. [2]Department of Mathematics, The Ohio State University, 100 Math Tower, 31 West 18th Ave., 43210 Columbus, USA. [3]Department of Biomedical Informatics, The Ohio State University, 1585 Neil Ave, 43210 Columbus, USA. [4]College of Public Health, 250 Cunz Hall, 1841 Neil Ave., 43210 Columbus, USA.

Received: 27 November 2019 Accepted: 8 April 2020
Published online: 25 April 2020

References

1. Sima C, Dougherty ER. What should be expected from feature selection in small-sample settings. Bioinformatics. 2006;22(19):2430–6.
2. Sima C, Dougherty ER. The peaking phenomenon in the presence of feature-selection. Pattern Recognit Lett. 2008;29(11):1667–74.
3. Tutuncuoglu B, Krogan NJ. Mapping genetic interactions in cancer: a road to rational combination therapies. Genome Med. 2019;11(1):62.
4. Regan-Fendt KE, Xu J, DiVincenzo M, Duggan MC, Shakya R, Na R, Carson WE, Payne PR, Li F. Synergy from gene expression and network mining (syngenet) method predicts synergistic drug combinations for diverse melanoma genomic subtypes. NPJ Systs Biol Appl. 2019;5(1):1–15.
5. Deng X, Das S, Valdez K, Camphausen K, Shankavaram U. Sl-biodp: Multi-cancer interactive tool for prediction of synthetic lethality and response to cancer treatment. Cancers. 2019;11(11):1682.
6. Henkel L, Rauscher B, Boutros M. Context-dependent genetic interactions in cancer. Curr Opin Genet Dev. 2019;54:73–82.
7. Chen Y, Cao D, Gao J, Yuan Z. Discovering pair-wise synergies in microarray data. Sci Rep. 2016;6:30672.
8. Watkinson J, Wang X, Zheng T, Anastassiou D. Identification of gene interactions associated with disease from gene expression data using synergy networks. BMC Syst Biol. 2008;2(1):10.
9. Xing P, Chen Y, Gao J, Bai L, Yuan Z. A fast approach to detect gene–gene synergy. Sci Rep. 2017;7(1):1–8.
10. Chopra P, Lee J, Kang J, Lee S. Improving cancer classification accuracy using gene pairs. Plos ONE. 2010;5(12):.
11. Li G, Rabitz H. General formulation of HDMR component functions with independent and correlated variables. J Math Chem. 2012;50(1):99–130.
12. Lu R, Wang D, Wang M, Rempała GA. Estimation of Sobol's sensitivity indices under generalized linear models. Commun Stat-Theory Methods. 2018;47(21):5163–95.
13. Hooker G. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. J Comput Graph Stat. 2007;16(3):709–32.
14. Shah RD, Peters J. The hardness of conditional independence testing and the generalised covariance measure. arXiv preprint arXiv:1804.07203. 2018.
15. Gupta AK, Tang J. Distribution of likelihood ratio statistic for testing equality of covariance matrices of multivariate Gaussian models. Biometrika. 1984;71(3):555–9.
16. Crow M, Paul A, et al. Exploiting single-cell expression to characterize co-expression replicability. Genome Biol. 2016;17(1):101.
17. van Dam S, Võsa U, et al. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinforma. 2017;19(4):575–92.
18. Ruan J, Dean AK, et al. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst Biol. 2010;4(1):8.
19. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012;8(1):565.
20. Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics. 2010;11(1):95. https://doi.org/10.1186/1471-2105-11-95.
21. Plan Y, Vershynin R. One-bit compressed sensing by linear programming. Commun Pure Appl Math. 2013;66(8):1275–97.
22. Plan Y, Vershynin R. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. IEEE Trans Inf Theory. 2013;59(1):482–94.
23. Wasserman L. All of Nonparametric Statistics, 1st edn. New York: Springer; 2010.
24. Fortunato S. Community detection in graphs. Phys Rep. 2010;486(3):75–174.
25. Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis. Phys Rev E. 2009;80(5):056117.
26. Hua J, Tembe WD, et al. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recog. 2009;42(3):409–24.
27. Foroughi pour A, Dalton LA. Heuristic algorithms for feature selection under Bayesian models with block-diagonal covariance structure. BMC Bioinformatics. 2018;19(3):70.
28. Foroughi pour A, Dalton LA. Robust feature selection for block covariance Bayesian models. In: Proceedigns of IEEE International Conference on Acoustics, Speech and Signal Processing; 2017. p. 2696–700.
29. Fukunaga K. Introduction to Statistical Pattern Recognition, 2nd edn. Boston, MA: Academic Press; 1990. https://doi.org/10.1016/B978-0-08-047865-4.50005-3. http://www.sciencedirect.com/science/article/pii/B9780080478654500053.
30. Theodoridis S, Koutroumbas K. Pattern Recognition, 4th edn. Boston, MA: Academic Press; 2009. https://doi.org/10.1016/B978-1-59749-272-0.50005-0. http://www.sciencedirect.com/science/article/pii/B9781597492720500050.
31. Bishop CM. Pattern Recognition and Machine Learning, 1st edn. New York: Springer; 2006.
32. Lu J, Plataniotis KN, et al. Regularized discriminant analysis for the small sample size problem in face recognition. Pattern Recog Lett. 2003;24(16):3079–87.
33. Wu B, Abbott T, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics. 2003;19(13):1636–43.
34. Itoh M, Iwamoto T, et al. Estrogen receptor (ER) mRNA expression and molecular subtype distribution in ER-negative/progesterone receptor-positive breast cancers. Breast Cancer Res Treat. 2014;143(2):403–9.
35. Hatzis C, Pusztai L, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA. 2011;305(18):1873–81.
36. Edgar R, Domrachev M, et al. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
37. Pearson ES, Neyman J. On the problem of two samples. In: Neyman J, Pearson ES, editors. Joint Statistical Papers 1967; 1930. p. 99–115.
38. Foroughi pour A, Dalton LA. Optimal bayesian filtering for biomarker discovery: Performance and robustness. IEEE/ACM Trans Comput Biol Bioinforma (to appear). 2018.

39. Maaten Lvd, Hinton G. Visualizing data using t-sne. J Mach Learn Res. 2008;9(Nov):2579–605.
40. Finak G, Bertos N, et al. Stromal gene expression predicts clinical outcome in breast cancer. Nat Med. 2008;14(5):518–27.
41. Rae JM, Johnson MD, et al. GREB1 is a critical regulator of hormone dependent breast cancer growth. Breast Cancer Res Treat. 2005;92(2):141–9.
42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol). 1995289–300.
43. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):10008.
44. Krämer A, Green J, et al. Causal analysis approaches in ingenuity pathway analysis. Bioinformatics. 2013;30(4):523–30.
45. Zhou W, Wang Y, et al. A standardized fold change method for microarray differential expression analysis used to reveal genes involved in acute rejection in murine allograft models. FEBS Open Bio. 2018;8(3):481–90.
46. Shedden K, Taylor JMG, et al. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med. 2008;14(8):822–7.
47. Dai M, Lu J-J, et al. BPTF promotes tumor growth and predicts poor prognosis in lung adenocarcinomas. Oncotarget. 2015;6(32):33878–92.
48. Lu Y, Wang L, et al. Gene-expression signature predicts postoperative recurrence in stage I non-small cell lung cancer patients. PloS ONE. 2012;7(1):30880.

Publisher's Note