Quantitative Biology 2017, 5(2): 159–172
DOI 10.1007/s40484-017-0092-7

RESEARCH ARTICLE

Elastic restricted Boltzmann machines for cancer data analysis

Sai Zhang[1], Muxuan Liang[2], Zhongjun Zhou[1], Chen Zhang[1], Ning Chen[3], Ting Chen[3,4] and Jianyang Zeng[1,*]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China
[2] Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706-1685, USA
[3] Bioinformatics Division, TNLIST, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[4] Program in Computational Biology and Bioinformatics, Univesity of Southern California, Los Angeles, CA 90089, USA
* Correspondence: zengjy321@tsinghua.edu.cn

Received September 14, 2016; Revised November 24, 2016; Accepted November 26, 2016

***Background*: Restricted Boltzmann machines (RBMs) are endowed with the universal power of modeling (binary) joint distributions. Meanwhile, as a result of their confining network structure, training RBMs confronts less difficulties when dealing with approximation and inference issues. But little work has been developed to fully exploit the capacity of these models to analyze cancer data, e.g., cancer genomic, transcriptomic, proteomic and epigenomic data. On the other hand, in the cancer data analysis task, the number of features/predictors is usually much larger than the sample size, which is known as the "$p \gg N$" problem and is also ubiquitous in other bioinformatics and computational biology fields. The "$p \gg N$" problem puts the *bias-variance trade-off* in a more crucial place when designing statistical learning methods. However, to date, few RBM models have been particularly designed to address this issue.**

***Methods*: We propose a novel RBMs model, called *elastic restricted Boltzmann machines (eRBMs)*, which incorporates the elastic regularization term into the likelihood function, to balance the model complexity and sensitivity. Facilitated by the classic contrastive divergence (CD) algorithm, we develop the elastic contrastive divergence (eCD) algorithm which can train eRBMs efficiently.**

***Results*: We obtain several theoretical results on the rationality and properties of our model. We further evaluate the power of our model based on a challenging task — predicting dichotomized survival time using the molecular profiling of tumors. The test results show that the prediction performance of eRBMs is much superior to that of the state-of-the-art methods.**

***Conclusions*: The proposed eRBMs are capable of dealing with the "$p \gg N$" problems and have superior modeling performance over traditional methods. Our novel model is a promising method for future cancer data analysis.**

Keywords: RBMs; regularization; cancer data analysis; survival time prediction

INTRODUCTION

In the past decade, with the advent of high-throughput sequencing techniques, comprehensive efforts have been made to collect molecular profiling (e.g., genomic, transcriptomic, epigenomic and proteomic) data of tumor samples. The availability of these large amounts of cancer molecular profiles facilitates extensive cancer disease studies and further novel biological discoveries, in which appropriate computational and statistical tools are needed to perform the data analysis task [1−3], e.g., exploring data features (co-expression genes) and influential elements (genes or mRNAs), integrating various data types and predicting certain biological responses (survival time or cancer subtypes) we are interested in.

However, it is not straightforward to apply extant machine learning methods for the cancer data analysis, mainly due to the specific properties of cancer molecular profiles: unlike other data analysis tasks, such as image classification and automatic speech recognition, the

dimension of cancer data features greatly outnumbers the sample size. For example, a typical cancer genomic dataset downloaded from The Cancer Genome Atlas (TCGA) [4] contains only hundreds (denoted by N) of samples, but each sample contains more than $10k$ (denoted by p) measurements of genomic profiling, such as gene expression and copy number variation. This well-known "$p \gg N$" problem [5,6] requires the designed model to be able to address the *bias-variance trade-off* issue [5] effectively: on one hand, as the feature dimension p is large, the model should be equipped with enough potential/complexity for discovering complicated statistical characteristics across input features; on the other hand, we need the model to perform well in sensitivity/generalization since the observed sample size N is extremely limited compared with p, which will lead to overfitting easily. In statistics and machine learning fields, one popular principle to address the "$p \gg N$" problem is *variable selection* [7], also referred to as *feature selection*, which integrates additional penalty/regularization terms into the original objective function, and performs estimation as well as selects pivotal features simultaneously. LASSO [8] and its descendant the *elastic net* [9] are two of the most classic methods realizing feature selection. Nevertheless, most of these models are linear in nature and lack the ability to model complicated statistical characteristics, especially when p is large.

Restricted Boltzmann machines (RBMs) [10] have been widely studied and used in the machine learning fields. For example, they are commonly the fundamental building blocks of the deep learning frameworks [11−14], like deep belief networks (DBNs) [12] and deep Boltzmann machines (DBMs) [15]. In particular, RBMs have been proven to own the universal potential of approximating discrete distributions [16]. In addition, Hinton's training algorithm, i.e., the *contrastive divergence* (CD) algorithm [17], can be used to train RBMs efficiently. Both these aspects make the RBM an ideal candidate for modeling complex statistical characteristics of the input data. In fact, based on the original RBMs, various modifications have been made in different applications, for instance, Gaussian RBMs are proposed to model images [10], replicated softmax model are used to model word distributions and extract latent topics in documents [18], RBMs can also be made to perform collaborative filtering tasks [19]. Note that in the bioinformatics field, Wang and Zeng proposed an RBM-liked model to predict drug-target interactions effectively [20]. Though RBMs have been applied successfully in numerous applications, for the cancer data analysis, it is far from making use of RBMs directly, mainly due to the aforementioned "$p \gg N$" problem.

The overall goal of this study is to solve the above problems, i.e., to fully exploit the modeling power of RBMs in the "$p \gg N$" scenario, to analyze high-dimensional cancer data, which fills the gap between the model complexity and sensitivity. More specifically, we develop a new RBMs model, called *elastic restricted Boltzmann machines* (eRBMs), which extends the traditional RBMs model by adding an *elastic regularization* term to the likelihood function. Under the maximum likelihood estimation (MLE), our eRBMs can be trained using the standard CD method efficiently, with only a few modifications. We have also derived several theoretical conclusions to demonstrate that the regularized optimization problem of eRBMs own nice properties that satisfy our complexity-vs-generalization demand. To evaluate the power of our model empirically, we further perform a challenging task, i.e., to predict dichotomized survival time using the molecular profiling of tumors. Test results show that the prediction performance of eRBMs are much superior to the state-of-the-art methods.

RESULTS

Elastic restricted Boltzmann machines

We proposed a novel model to solve the "$p \gg N$" problem, which is to solve the following optimization problem,

$$\underset{\mathbf{W}}{\text{minimize}} -\frac{1}{|\mathcal{S}|}\ln\mathcal{L}(\mathcal{S}) + \lambda_1\|\mathbf{W}\|_1 + \lambda_2\|\mathbf{W}\|_2^2, \quad (1)$$

where λ_1 and λ_2 are two fixed coefficients measuring the contributions of corresponding regularization terms, $\|\mathbf{W}\|_1 := \Sigma_{ij}|w_{ij}|$, $\|\mathbf{W}\|_2^2 := \Sigma_{ij}w_{ij}^2$ and S is the training set and w_{ij} is the weight associated with edge. Let $\alpha = \lambda_2 / (\lambda_1 + \lambda_2)$, then the above problem is further equivalent to the problem

$$\begin{aligned} &\underset{\mathbf{W}}{\text{minimize}} -\frac{1}{|\mathcal{S}|}\ln\mathcal{L}(\mathcal{S}) \\ &\text{subject to } (1-\alpha)\|\mathbf{W}\|_1 + \alpha\|\mathbf{W}\|_2^2 \leqslant t \text{ for some } t, \end{aligned} \quad (2)$$

where we call the function $(1-\alpha)\|\mathbf{W}\|_1 + \alpha\|\mathbf{W}\|_2^2$ the *elastic regularization term*, which is a convex combination of the l_1- and l_2-norms. The corresponding regularization technique is referred to as the *elastic regularization*, and the resulting RBMs with Problem (2) is called the *elastic restricted Boltzmann machines* (eRBMs). Note that when $\alpha = 0$, Problem (2) degenerates to Problem (15), while when $\alpha = 1$, Problem (2) loses its l_1 regularization term and is called the *weight decay* method in neural networks [21]. In this paper, we only consider $\alpha \in [0,1)$. Also note that the elastic regularization term is strictly convex as the l_2 regularization is considered, i.e., $\alpha > 0$. More detailed derivation of our model are given in Section of Materials and Methods.

Theoretical analysis

To characterize the effects of the elastic regularization term of Problem (1) (or equivalently Problem (2)), here we provide several theoretical results on both the extreme situation in which several visible variables are exactly positively correlated, namely, they are identical, as well as the general case. Note that in this paper, we can only consider the local minima of Problem (1) because of its non-convexity. For brevity, we denote the objective function of Problem (1) as

$$O(\mathbf{W}) := L(\mathbf{W}) + R(\mathbf{W}),$$

where

$$L(\mathbf{W}) = -\frac{1}{\mathcal{S}}\ln \mathcal{L}(\mathcal{S}), R(\mathbf{W}) = \lambda_1\|\mathbf{W}\|_1 + \lambda_2\|\mathbf{W}\|_2^2.$$

Furthermore, let us denote a local minimizer of Problem (1) in the neighbourhood U as $\widehat{\mathbf{W}}$, which means that $\forall \mathbf{W} \in U$, we have $O(\mathbf{W}) \geqslant O\left(\widehat{\mathbf{W}}\right)$.

Lemma 1. *Assume that for the training examples, $V_i = V_j$, where $i, j \in \{1, ..., m\}$. Then we have the following conclusions*:

$$(i)\ \textit{If}\ \lambda_1, \lambda_2 > 0, \textit{then}\ \hat{w}_{ik} = \hat{w}_{jk}, \forall k \in \{1, ..., n\}.$$

$$(ii)\ \textit{If}\ \lambda_2 = 0, \textit{then}\ \hat{w}_{ik}\hat{w}_{jk} \geqslant 0, \forall k \in \{1, ..., n\}.$$

For fixed k, $\widehat{\mathbf{W}}^*$ is another minimizer of Problem (1), where

$$\hat{w}^*_{gh} = \begin{cases} \hat{w}_{gh} & \forall h \in \{1, ..., n\}, \textit{if}\ g \neq i, j, \\ \hat{w}_{gh} & \textit{if}\ g = i\ \textit{or}\ g = j, \textit{but}\ h \neq k, \\ s(\hat{w}_{ik} + \hat{w}_{jk}) & \textit{if}\ g = i\ \textit{and}\ h = k, \\ (1-s)(\hat{w}_{ik} + \hat{w}_{jk}) & \textit{if}\ g = j\ \textit{and}\ h = k, \end{cases}$$

for $|s - \hat{w}_{ik}/(\hat{w}_{ik} + \hat{w}_{jk})|$ is small enough.

Proof. (i) First, let us fix $\lambda_1, \lambda_2 > 0$. If $\hat{w}_{ik} \neq \hat{w}_{jk}$ for some k, let us consider another weight vector $\widehat{\mathbf{W}}^*$ as follows:

$$\hat{w}^*_{gh} = \begin{cases} \hat{w}_{gh} & \forall h \in \{1, ..., n\}, \textit{if}\ g \neq i, j, \\ \hat{w}_{gh} & \textit{if}\ g = i\ \textit{or}\ g = j, \text{ but } h \neq k, \\ (1-\varepsilon)\hat{w}_{ik} + \varepsilon\hat{w}_{jk} & \textit{if}\ g = i\ \textit{and}\ h = k, \\ \varepsilon\hat{w}_{ik} + (1-\varepsilon)\hat{w}_{jk} & \textit{if}\ g = j\ \textit{and}\ h = k, \end{cases}$$

where $\varepsilon<1$ is small enough such that for some $\delta > 0$, $B\left(\widehat{\mathbf{W}}, \varepsilon|\hat{w}_{ik} - \hat{w}_{jk}| + \delta\right) \subset U$. Here the open ball $B(\mathrm{x}_0, r)$ is defined by $B(\mathrm{x}_0, r) := \{\mathrm{x} : |\mathrm{x} - \mathrm{x}_0| < r, \mathrm{x} \in \mathbb{R}^{m\times n}\}$. Hence, $\widehat{\mathbf{W}}^*$ is also located in the neighbourhood U. Since $V_i = V_j$, it is evident that $L(\widehat{\mathbf{W}}) = \mathrm{L}(\widehat{\mathbf{W}}^*)$. However, since $\lambda_1, \lambda_2 > 0$, the elastic regularization term $R(\mathbf{W})$ is strictly convex, which yields $R(\widehat{\mathbf{W}}^*) < R(\widehat{\mathbf{W}})$. Here comes a contradiction.

(ii) As $\lambda_2 = 0$ and the regularization term $R(\mathbf{W})$ degenerates to the l_1 penalty, it is not strictly convex now. Suppose that $\hat{w}_{ik}\hat{w}_{jk}<0$ for some k, consider the same $\widehat{\mathbf{W}}^*$ as that in (i). Without loss of generality, we assume that $\hat{w}_{ik} > 0$ and $\hat{w}_{jk}<0$, and at the same time, we set ε to be small enough (without contradicting the constraints in (i)) such that $\varepsilon<1/2$, $(1-\varepsilon)\hat{w}_{ik} + \varepsilon\hat{w}_{jk} > 0$ while $\varepsilon\hat{w}_{ik} + (1-\varepsilon)\hat{w}_{jk}<0$. Thus, we have $|(1-\varepsilon)\hat{w}_{ik} + \varepsilon\hat{w}_{jk}| + |\varepsilon\hat{w}_{ik} + (1-\varepsilon)\hat{w}_{jk}|$ $= (1-2\varepsilon)(\hat{w}_{ik} - \hat{w}_{jk}) < |\hat{w}_{ij}| + |\hat{w}_{jk}|$, which yields $R(\widehat{\mathbf{W}}^*)$ $< R(\widehat{\mathbf{W}})$ contradicting to the assumption. The case of $\hat{w}_{ik}<0$ and $\hat{w}_{jk} > 0$ can be discussed in the same manner.

Since we have validated that $\hat{w}_{ik}$ and $\hat{w}_{jk}$ cannot own the reverse signs, we have $R(\widehat{\mathbf{W}}^*) = R(\widehat{\mathbf{W}})$, and further $O(\widehat{\mathbf{W}}^*) = O(\widehat{\mathbf{W}})$. After replacing the above ε by s, we note that s shall satisfy that $d\left(\widehat{\mathbf{W}}, \widehat{\mathbf{W}}^*\right) = |s - \hat{w}_{ik}/(\hat{w}_{ik} + \hat{w}_{jk})|$ is small enough to make $\widehat{\mathbf{W}}^* \in U$. □

Lemma 1. provides us with nice properties of the eRBMs. First, it guarantees the same weight solutions for two exactly correlated variables. Though empirical data rarely get this extreme correlations, Lemma 1 presents the potential of the eRBMs to model variable correlations explicitly, which satisfies our requirement discussed in Section of Elastic Regularization. In addition, without the l_2 regularization, the problem (1) may have infinite solutions around some local minimum, which makes it less stable. Note that this is a concrete illustration of the regularization technique to solve the ill-posed problem.

Let us consider the Pearson's correlation coefficient defined by

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\mathbb{E}((X-\mu_X)(Y-\mu_Y))}{\sigma_X\sigma_Y}, \qquad (3)$$

where cov denotes the covariance, σ is the standard deviation, and μ represents the mean. For the reverse direction, we have the following lemma.

Lemma 2. *Suppose that in an RBM, for the visible variables V_i and V_j, their weights are equal to each other, i.e., $w_{ik} = w_{jk}, \forall k \in \{1, ..., n\}$, then their Pearson's correlation coefficient equals to 1.*

Proof. Since the weights of V_i and V_j are equal to each other, we can treat them as the identical variable. Thus based on the definition of Pearson's correlation coefficient, the conclusion of this lemma follows after simple computations, which are omitted here. □

The complexity of the RBMs model lies in that the

same distribution may have various model configurations, which implies that though two visible variables are correlated, their weights can be vastly distinct. However, this ill-posed problem can be solved by our eRBMs model perfectly. In fact, we have the following somewhat anti-intuitive theorem.

Theorem 1. *For the solutions of eRBMs, we have $V_i = V_j$, $i, j \in \{1, \ldots, m\}$ if and only if $w_{ik} = w_{jk}, \forall k \in \{1, \ldots, n\}$.*

Proof. This is a direct result of Lemma 1 and Lemma 2. □

Next, we consider the general case, in which the assumption of two identical variables is dropped. We have the following general result that correlates/bounds the difference between the parameter paths of variables to/by their correlations quantitatively.

Theorem 2. *Suppose that the regularization coefficients λ_1, λ_2 are both positive in the eRBMs, and let $\widehat{\mathbf{W}}$ be a local minimizer of Problem (1). Assume that the empirical distribution $q(\mathbf{v})$ and the model distribution $p(\mathbf{v})$ fit well, i.e., there is a constant $C > 0$ such that $p(\mathbf{v})/q(\mathbf{v}) < C$, $\forall \mathbf{v} \in \{0,1\}^m$. Also assume that $\hat{w}_{ik}\hat{w}_{jk} > 0$ for some $k \in \{1, \ldots, m\}$. Following [9], we define*

$$D_{\lambda_1,\,\lambda_2}(i, j, k) := |\hat{w}_{ik}(\lambda_1, \lambda_2) - \hat{w}_{jk}(\lambda_1, \lambda_2)|.$$

Then $D_{\lambda_1,\,\lambda_2}(i, j, k)$ is bounded above by the empirical correlation of variables V_i and V_j. Precisely, we have

$$D_{\lambda_1,\,\lambda_2}(i, j, k) = \mathcal{O}(\mathbb{P}_e(v_i \neq v_j)), \tag{4}$$

where we use $\mathbb{P}$ and $\mathbb{P}_e$ to denote the model and empirical probabilities, respectively.

Proof. First, under the distribution regularity assumption, we claim that the probabilities $\mathbb{P}(V_i \neq V_j)$ and $\mathbb{P}_e(V_i \neq V_j)$ also fit well. Indeed, for $\mathbb{P}(V_i = 0, V_j = 1)/\mathbb{P}_e(V_i = 0, V_j = 1)$, we have

$$\begin{aligned}\frac{\mathbb{P}(V_i=0,\,V_j=1)}{\mathbb{P}_e(V_i=0,\,V_j=1)} &= \frac{\sum_{\ldots,\,v_i=0,\,v_j=1,\,\ldots} p(\mathbf{v})}{\sum_{\ldots,\,v_i=0,\,v_j=1,\,\ldots} q(\mathbf{v})} \\ &= \sum \frac{p(\mathbf{v})}{\sum_{\ldots,\,v_i=0,\,v_j=1,\,\ldots} p(\mathbf{v})} \\ &\leqslant \sum_{\ldots, v_i=0, v_j=1, \ldots} \frac{p(\mathbf{v})}{q(\mathbf{v})} \\ &\leqslant 2^{m-2}C.\end{aligned}$$

As for $\mathbb{P}(V_i = 1, V_j = 0)/\mathbb{P}_e(V_i = 1, V_j = 0)$, we can get the same upper bound after similar calculations.

Since $\hat{w}_{ik}\hat{w}_{jk} > 0$, we have $\operatorname{sgn}\{\hat{w}_{ik}\} = \operatorname{sgn}\{\hat{w}_{jk}\}$. According to the assumption, the local minimizer $\widehat{\mathbf{W}}$ satisfies

$$\frac{\partial O(\mathbf{W})}{\partial w_{gh}}\Big|_{\mathbf{W}=\widehat{\mathbf{W}}} = 0, \text{ if } \hat{w}_{gh} \neq 0.$$

Thus we have

$$\frac{\partial L(\mathbf{W})}{\partial w_{ik}}\Big|_{\mathbf{W}=\widehat{\mathbf{W}}} + \lambda_1 \operatorname{sgn}\{\hat{w}_{ik}\} + 2\lambda_2 \hat{w}_{ik} = 0, \tag{5}$$

$$\frac{\partial L(\mathbf{W})}{\partial w_{jk}}\Big|_{\mathbf{W}=\widehat{\mathbf{W}}} + \lambda_1 \operatorname{sgn}\{\hat{w}_{jk}\} + 2\lambda_2 \hat{w}_{jk} = 0. \tag{6}$$

Subtracting Equation (5) from Equation (6) yields

$$\hat{w}_{ik} - \hat{w}_{jk} = \frac{1}{2\lambda_2}\left(\frac{\partial L(\mathbf{W})}{\partial w_{jk}}\Big|_{\mathbf{W}=\widehat{\mathbf{W}}} - \frac{\partial L(\mathbf{W})}{\partial w_{ik}}\Big|_{\mathbf{W}=\widehat{\mathbf{W}}}\right). \tag{7}$$

Furthermore, according to Equation (12) below, the above equation can be written as

$$\begin{aligned}w_{ik} - \hat{w}_{jk} = \frac{1}{2\lambda_2|S|}&((\mathbb{E}_{p(\mathbf{h}|\mathbf{v})q(\mathbf{v})})[V_j H_k] - \mathbb{E}_{p(\mathbf{v},\mathbf{h})}[V_j H_k]) \\ &- (\mathbb{E}_{p(\mathbf{h}|\mathbf{v})q(\mathbf{v})})[V_i H_k] - \mathbb{E}_{p(\mathbf{v},\mathbf{h})}[V_i H_k].\end{aligned}$$

Next follows several computations,

$$\begin{aligned}&D_{\lambda_1,\lambda_2}(i,j,k) = |\hat{w}_{ik} - \hat{w}_{jk}| \\ &= \frac{1}{2\lambda_2|S|} \cdot |\mathbb{E}_{p(\mathbf{h}|\mathbf{v})q(\mathbf{v})}[(V_i - V_j)H_k] - \mathbb{E}_{p(\mathbf{v},\mathbf{h})}[(V_i - V_j)H_k]| \\ &= \frac{1}{2\lambda_2|S|} \cdot |(\mathbb{P}(1|1,0)\mathbb{P}_e(1,0) - \mathbb{P}(1|0,1)\mathbb{P}_e(0,1)) \\ &\quad - (\mathbb{P}(1|1,0)\mathbb{P}(1,0) - \mathbb{P}(1|0,1)\mathbb{P}(0,1))| \\ &\leqslant \frac{1}{2\lambda_2|S|} \cdot (|\mathbb{P}_e(1,0) - \mathbb{P}(1,0)| + |\mathbb{P}_e(0,1) - \mathbb{P}(0,1)|) \\ &\leqslant \frac{1}{2\lambda_2|S|} \cdot (\mathbb{P}_e(1,0) + \mathbb{P}(1,0) + \mathbb{P}_e(0,1) + \mathbb{P}(0,1)) \\ &\leqslant \frac{1}{2\lambda_2|S|} \cdot \left(\mathbb{P}_e(V_i \neq V_j)\left(2 + \frac{\mathbb{P}(1,0)}{\mathbb{P}_e(1,0)} + \frac{\mathbb{P}(0,1)}{\mathbb{P}_e(0,1)}\right)\right) \\ &\leqslant \frac{1 + 2^{m-2}C}{\lambda_2|S|}\mathbb{P}_e(V_i \neq V_j),\end{aligned}$$

where $\mathbb{P}(x_3|x_1,x_2)$, $\mathbb{P}(x_1,x_2)$ and $\mathbb{P}_e(x_1,x_2)$ are short for $\mathbb{P}(H_k = x_3 | V_i = x_1, V_j = x_2)$, $\mathbb{P}(V_i = x_1, V_j = x_2)$ and $\mathbb{P}_e(V_i = x_1, V_j = x_2)$, respectively, $x_1, x_2, x_3 \in \{0,1\}$, and $\mathbb{P}_e(\mathbf{v})$ represents $q(\mathbf{v})$. This completes the proof. □

From Theorem 2, we find that if λ_1 and λ_2 vary, the quantity $D_{\lambda_1,\lambda_2}(i,j,k)$ describes the difference between parameter paths of variables V_i and V_j. Note that Equation (4) bounds D_{λ_1,λ_2} with the empirical probability $\mathbb{P}_e(V_i \neq V_j) = 1 - \mathbb{P}_e(V_i = V_j)$, which represents the positive correlation between binary variables V_i and V_j. If V_i and V_j are highly positively correlated, i.e., $\mathbb{P}_e(V_i \neq V_j)$ is almost 0, their weight difference can be guaranteed to be particularly small. Note that for the negative correlation

where $\mathbb{P}_e(V_i \neq V_j)$ is almost 1, if we replace V_i by $1 - V_i$ and perform the similar discussions, the same conclusion follows. Theorem 2 presents the very nice property that in the general situation, the correlated variables can obtain similar weights in eRBMs, which increases the model flexibility as well as the generalization. We finally note that all the above analysis and conclusions of eRBMs are also valid for the elastic version of ClassRBMs (i.e., considering the elastic regularization in the generative objective of ClassRBMs), since the two base models are the same in nature.

Tests on simulated data

We first evaluated the performance of our eRBMs model on simulated data. Here, we describe the procedure of the data generation. We first set a "true model" $\mathcal{M}_{true}$ that was an RBM with sparse weights, and used this "true model" to generate the simulated data. In $\mathcal{M}_{true}$, we set the numbers of hidden and visible variables to be 250 and 500, respectively. We also set **b** and **c** to be 0, and generated the 250×500 weight matrix **W**, in which each element followed the normal distribution $N(0, 0.1)$. After that, we randomly picked a weight element to be 0 with probability 0.5. Based on this $\mathcal{M}_{true}$, we generated 10 training data $\mathcal{D}_{train}$ and 1,000 test data $\mathcal{D}_{test}$ by running the Gibbs sampling (1,000 iterations for each sample). Here we used a small training set to construct a "$p \gg N$" case ($500 \gg 10$), and generated a large test dataset to evaluate the model performance sufficiently.

Here we tested 25 different combinations of the coefficients of l_1 and l_2 regularization terms, i.e., $(\lambda_1, \lambda_2) \in \{0.1, 0.01, 0.001, 0.0001, 0\} \times \{0.1, 0.01, 0.001, 0.0001, 0\}$. The eRBMs were trained using the eCD algorithm (see Algorithm 1). We evaluated the influence of λ_1 and λ_2 depending on the following procedure: after training an eRBM, another 1,000 samples $\mathcal{D}_{learn}$ were generated based on this learned RBM $\mathcal{M}_{learn}$ (still via the Gibbs sampling). If $\mathcal{M}_{learn}$ captured the statistical characteristics of $\mathcal{M}_{true}$, samples $\mathcal{D}_{learn}$ and $\mathcal{D}_{test}$ should be "close" enough. Since we have derived the eRBMs model according to the correlations between different input dimensions, here we computed the correlation coefficient matrices R_{test} and R_{learn} for $\mathcal{D}_{test}$ and $\mathcal{D}_{learn}$, respectively, in which $R_{ij} = \rho_{V_i,V_j}$ represents the Pearson's correlation coefficient (see Equation (3)) between variables V_i and V_j given corresponding samples. To quantify the difference between R_{test} and R_{learn}, we adopted the two-norm of the matrix, i.e., the maximum singular value of the matrix. Precisely speaking, we computed the *correlation coefficient difference* (CCD) $\Delta_R := \|R_{test} - R_{learn}\|_2$ for each (λ_1, λ_2)-configuration, where the small Δ_R implied similar statistical characteristics of sample sets D_{test} and D_{learn}. Selected results are presented in Table 1 and the complete test results are provided in Appendix A.

Table 1 presents the CCDs of four (λ_1, λ_2)-configurations, including the traditional RBMs without any regularization. We found that the smallest CCD (in all

Algorithm 1 (k-step elastic contrastive divergence)

Input: eRBMs with initialized parameters **W**, **b** and **c**, training set $\mathcal{S}$.

Output: Gradient approximations $\Delta_{w_{ij}}$, Δ_{b_i} and Δ_{c_j} for $i = 1, ..., m, j = 1, ..., n$.

for $\mathbf{v} \in \mathcal{S}$ **do**
  $\mathbf{v}^{(0)} \leftarrow \mathbf{v}$
  for $t = 0, ..., k-1$ **do**
    for $j = 1, ..., n$ **do**
      Sample $h_j^{(t)} \sim p(h_j|\mathbf{v}^{(t)})$
    end for
    for $i = 1, ..., m$, **do**
      Sample $v_i^{(t+1)} \sim p(v_i|\mathbf{h}^{(t)})$
    end for
  end for
  for $i = 1, ..., m, j = 1, ..., n$ **do**
    $\Delta w_{ij} \leftarrow \Delta w_{ij} + p(H_j = 1|\mathbf{v}^{(0)}) v_i^{(0)} - p(H_j = 1|\mathbf{v}^{(k)}) v_i^{(k)} - \lambda_1 \mathrm{sgn}\{w_{ij}\} - \lambda_2 w_{ij}$
    $\Delta b_i \leftarrow \Delta b_i + v_i^{(0)} - v_i^{(k)}$
    $\Delta c_j \leftarrow \Delta c_j + p(H_j = 1|\mathbf{v}^{(0)}) - p(H_j = 1|\mathbf{v}^{(k)})$
  end for
end for

Table 1. Test results on the simulated data[a].

Case	λ_1	λ_2	$\Delta_S(CCD)$
1	0	0	117.58
2	0.0001	0	109.17
3	0	0.01	93.53
4	0.0001	0.01	**67.96**[b]

[a] We only present representative test results here for brevity. The whole results can be found in Appendix A.

[b] The smallest *CCD* in all 25 (λ_1, λ_2)-configurations.

25 (λ_1, λ_2)-configurations) was obtained when $\lambda_1 = 0.0001$ and $\lambda_2 = 0.01$. Though the CCDs of configurations of (0, 0.01) (weight decay case) and (0.0001, 0) (losing the l_2-regularization) were still smaller than that of the traditional RBMs, not appropriately regularized RBMs lacked the ability to fully capture the statistical characteristics of the true distribution described by the "true model". These observations established the necessity of our eRBMs model, namely, *the combination of l_1- and l_2-regularization endows the eRBMs with more power to unravel the objective distributions in the "$p \gg N$" case*. Note that we have confronted the issue of selecting the best (λ_1, λ_2)-configuration (also known as *model selection*) from the candidate set, which yields the most proper regularization. Indeed, certain (λ_1, λ_2)-configurations (e.g., (0.1, 0) and (0.01, 0.01), see Table A1 for details) may lead to worse performance than the unregularized one. In the unsupervised case, such as the test here, we can choose some objective functions (e.g., the CCD) measuring the modeling performance to select the best (λ_1, λ_2). In the supervised case, such as the test in the next section, corresponding classification or regression accuracy can be used based on the cross-validation procedure. More detailed discussions on model selection are given in Section of Model Selection.

Predicting dichotomized survival time

Predicting dichotomized survival time is the first step towards exploiting the clinical utility of the collected molecular profiling of human tumor in TCGA [22]. Though TCGA [4] has yielded large amounts of molecular profiling (e.g., genomic, transcriptomic, epigenomic and proteomic cancer data) of various cancer types, the "$p \gg N$" issue is rather prominent. For example, TCGA maintains hundreds of samples for each tumor type, but every sample contains ~10k features (e.g., DNA methylation and mRNA expression). The state-of-the-art methods on predicting dichotomized survival time using the cancer genomic and proteomic data are not so promising [22]. Here, we adopted the classification version of the eRBMs (see Section of Classification Restricted Boltzmann Machines for technical details) to perform the same task on the datasets of ovarian serous cystadenocarcinoma (OV). More details of the datasets can be found in Section of datasets in Materials and Methods. Our test results showed that the proposed eRBMs models beat other methods, including the traditional RBMs, in the prediction performance, which provides another demonstration of the superiority of the eRBMs.

Model performance under different hyperparameters

To predict the dichotomized survival time of OV from its genomic and proteomic data, here we adopted the more flexible RBM variant — ClassRBM introduced in Section of Classification Restricted Boltzmann Machines. We have shown that ClassRBMs can be trained in both unsupervised and supervised fashions, which are perfect candidate models to evaluate our elastic regularization method by the classification task. For each OV molecular profiling types, we implemented the traditional ClassRBM and its elastic versions on the basis of the MATLAB package [RBM Toolbox], and set the coefficient α (see Equation (14)) adjusting the generative and discriminative objective terms to be the default 0.5. The candidate (λ_1, λ_2)-configuration set was still $\{0.1, 0.01, 0.001, 0.0001, 0\} \times \{0.1, 0.01, 0.001, 0.0001, 0\}$. For the network structure, we set 256 hidden units for the "$p \gg N$" case, i.e., the OV profiling of DNA methylation, mRNA and miRNA expressions. For the SCNA and protein expression, in which the feature dimension is comparable to the sample size, we assigned 100 hidden units to their corresponding RBMs, as previous studies have shown RBMs with equal or less hidden units to/than the visible units will perform better than those with more hidden units [21,23].

To evaluate the prediction performance, here we adopted the area under receiver operating characteristic curve (AUROC), the same as that in [22]. Note that for fairness, we used the same 10-fold data partition as that of [22], which is deposited in [syn1748545]. The result comparison is illustrated in Table 2. Here, we only present the AUROC performance of the best (λ_1, λ_2)-configuration, the traditional ClassRBM, and the state-of-the-art methods for each OV molecular profiling type for brevity. The complete results are provided in Tables B1 to B5 of Appendix B.

We find from Table 2 that, in predicting dichotomized survival time for OV, the elastic version of ClassRBM outperformed all other models, including the traditional ClassRBM, in all OV molecular profiling types. The average AUROC increase of the best eRBM vs. the state-of-the-art method was 7.8%. Interestingly, the traditional ClassRBM, though equipped with more capacity to model objective distributions than other models (e.g., logistic regression and partial least square), only got similar or

Table 2. AUROC scores of the 10-fold cross-validation for predicting dichotomized survival time for OV[a].

Model/Profiling	DDA	KNN	DA	LR	NC	PLS	RF	SVM	RBMs	eRBMs	λ_1	λ_2
Methy	0.597	0.573	0.598	0.599	0.582	0.598	0.579	0.619	0.620	**0.683**	0.001	0.1
mRNA	0.618	0.617	0.640	0.604	0.612	0.640	0.619	0.626	0.608	**0.690**	0	0.1
miRNA	0.586	0.582	0.605	0.594	0.584	0.605	0.589	0.625	0.551	**0.679**	0.1	0.001
SCNA	0.612	0.591	0.558	0.580	0.586	0.558	0.606	0.615	0.563	**0.689**	0.1	0.01
Protein	0.578	0.595	0.575	0.618	0.599	0.589	0.546	0.625	0.605	**0.684**	0.1	0.001

[a] The model acronyms DDA, KNN, DA, LR, NC, PLS, RF, SVM are short for diagonal discriminant analysis, K-nearest neighbor, discriminant analysis, logistic regression, nearest centroid, partial least square, random forest and support vector machine, respectively. The implementation details (e.g., the hyperparameter settings) of those models can be found in [22]. The best AUROC for each OV molecular profiling is shown in bold, and the AUROC of the state-of-the-art method is shown in italic. Only the performance of the best (λ_1,λ_2)-configuration and the traditional ClassRBM is presented here for brevity. The whole test results can be found in Appendix B.

worse prediction performance, probably affected by the "$p \gg N$" issue. For example, for DNA methylation, the RBM model yielded similar AUROC score (0.620) to the state-of-the-art result (0.619), while for mRNA expression, the RBM model acted much worse (AUROC of 0.608) than discriminant analysis and partial least square (AUROC of 0.640). Similar situation happened for other three profiling types. Note that for SCNA and protein expression, in which there was no "$p \gg N$" problem and the feature dimension was less than the sample size (109 vs. 252 and 165 vs. 252 for SCNA and protein expression, respectively), our eRBMs models also obtained the best prediction performance.

We note that to address the "$p \gg N$" problem, Yuan *et al*. [22] adopted the pre-selection strategy, namely, to select important variables before inputing to the corresponding models. Yuan *et al*. [22] used ANOVA and shrinking centroids [24] to select 10 to 50 important input variables via the 10-fold cross-validation and the best AUROC scores were finally reported. In contrast, for our eRBMs models, the feature selection procedure was simultaneously performed during the model training process, i.e., setting the weights of irresponsible visible variables for each hidden variables to be zero. The above observations further demonstrate the superiority of our eRBMs models.

Model selection

The elastic regularization term introduces two hyperparameters, i.e., λ_1 and λ_2, to the model training. In the previous sections, we have demonstrated the effects of different hyperparameters on the final model performance. Here, we take the test of miRNA profiling data of OV as an example to show how we can perform the hyperparameter selection/model selection for eRBMs in a regular manner. Without loss of generality, the candidate (λ_1, λ_2)-configuration was set be {0, 0.1, 0.3, 0.01, 0.03, 0.001, 0.003, 0.0001, 0.0003} × {0, 0.1, 0.3, 0.01, 0.03, 0.001, 0.003, 0.0001, 0.0003}. First, we divided all the samples into 10 folds, in which one fold was randomly selected as the independent test set and the left nine folds were regarded as the training set. Based on the training data, we performed a nine-fold cross-validation procedure for each candidate (λ_1, λ_2)-configuration to select the best hyperparameters. The number of the hidden units of eRBM was set to be 256, the same as that in Section of Model Performance Under Different Hyperparameters. The AUROC scores of the nine-fold cross-validation for different hyperparameters are shown in Table 3. We found that the eRBMs model achieved the best prediction performance when (λ_1, λ_2) = (0.3, 0.001). With this selected (λ_1, λ_2)-configuration, we further trained our

Table 3. AUROC scores for selecting hyperparameters (λ_1, λ_2)[a].

λ_1/λ_2	0	0.1	0.3	0.01	0.03	0.001	0.003	0.0001	0.0003
0	0.517	0.641	0.650	0.564	0.587	0.521	0.534	0.517	0.517
0.1	0.646	0.646	0.641	0.658	0.641	0.642	0.634	0.636	0.658
0.3	0.642	0.644	0.656	0.647	0.653	**0.671**	0.656	0.634	0.661
0.01	0.662	0.648	0.652	0.639	0.644	0.641	0.650	0.649	0.642
0.03	0.656	0.641	0.649	0.665	0.655	0.617	0.634	0.646	0.645
0.001	0.519	0.639	0.652	0.580	0.609	0.531	0.534	0.518	0.535
0.003	0.638	0.623	0.652	0.662	0.646	0.653	0.650	0.658	0.654
0.0001	0.520	0.636	0.639	0.559	0.604	0.552	0.546	0.532	0.538
0.0003	0.549	0.639	0.657	0.541	0.595	0.546	0.543	0.535	0.529

[a] The AUROC score for each (λ_1, λ_2)-configuration is presented here, in which the best performance is shown in bold.

final eRBMs model based on the whole training data and evaluated its performance on the independent test data, in which the AUROC was 0.635 and the area under the precision-recall curve (AUPR) was 0.625.

RELATED WORK

The idea of regularization to solve the "$p \gg N$" problem and increase the model generalization ability is widely used in statistics [7,25]. There are many other regularization/penalty terms with various properties equipping statistical models, see [7] for a nice review. In particular, the traditional ridge regression [5], LASSO [8] and elastic net [9] use similar regularization techniques adopted in our study. However, different from our eRBMs, these base models are linear in nature and have no latent variables, which may lack the ability to model complicated nonlinear features. Moreover, both *supervised* (see the classification RBMs in Section of Classification Restricted Boltzmann Machines) and *unsupervised* learning are practicable in eRBMs, which extends the application scope of our method, like the *semi-supervised scenarios*.

Several code packages have implemented the weight decay technique [21] to train an RBM. We have shown that the weight decay is a special case of our elastic regularization given $\alpha = 1$ in Problem (2), see Section of Theoretical Analysis for details. Our test results in Section 3 also demonstrate that only with l_2-regularization is not enough for high-dimensional data analysis. In particular, our work focuses on addressing the "$p \gg N$" problem from the beginning (by following the bias-variance trade-off principle), and derives the elastic regularization term combining both l_1- and l_2-norms. Both theoretical and empirical analyses verify that our method indeed works as expected.

We have noted that the l_1-norm on weights leads to sparse solutions. For instance, LASSO searches for a few significant and explainable predictors while ignoring other less important factors. In fact, sparsity is always a dominant topic in machine learning, signal processing and statistics [5,13]. The consideration of sparsity first appears in computational neuroscience (the visual system [26]) and is further embodied as sparse coding [27]. Based on this biological observation, researchers in machine learning, especially in neural networks and deep learning, have been exploring various effective distributed sparse representations [28], such as the sparse auto-encoders [29−31] and the sparse RBMs [21,32]. These sparse models are seeking for sparse latent representations of the input data, in which (in the language of RBMs) the hidden variables are activated (computing Equation (10) and getting high probability) in only a few fraction while leaving most others silent. The detailed motivation and advantage of introducing sparsity can be found in an excellent survey [13]. However, we want to emphasize that this representation sparsity (sparse hidden variables) is not our goal (sparse weights) in this study. Indeed, these two kinds of sparsity are not conceptually orthogonal and we can further integrate the representation sparsity into our method.

Recently, Min *et al*. developed a network-regularized sparse logistic regression model for the clinical risk prediction [33], in which the prior knowledge like biological pathways or gene interaction networks were integrated. Test results on both simulated and real cancer data demonstrated the superiority of their regularized method over traditional models. The regularization technique proposed in [33] is complementary to our elastic regularization, and this network-regularized term can also be integrated into our eRBMs model.

DISCUSSION

The "$p \gg N$" issue challenges the extant statistical and computational models for cancer data analysis. In this study, we proposed a novel graphical model, called the elastic restricted Boltzmann machines (eRBMs), to address this problem. In the principle of the bias-variance trade-off, we reformalized the optimization objective of the traditional RBMs by combining the l_1 and l_2 regularization. Both comprehensive theoretical analysis and empirical tests on simulated data verified that our eRBMs models gain nice properties that achieve our primary motivations. Moreover, we developed an efficient training algorithm, referred to as the elastic contrastive divergence (eCD), for eRBMs based on the classic CD algorithm. Our eRBMs models are also consistent with other RBM variants, such as the ClassRBMs, which can be trained in both generative (unsupervised) and discriminative (supervised) settings. This flexibility enlarges the application area of our method. At last, to evaluate the application performance of eRBMs, we performed a task of predicting dichotomized survival time using the molecular profiling of tumors, which is quite challenging as there are many other factors that can influence the survival time, such as the psychological factors, age, gender and tumor stages. Test results revealed that the prediction performance of our method was much superior to that of the state-of-the-art approaches. In this work, we mainly focused on balanced dataset. For the imbalanced high-dimensional dataset, we recommend integrating the regular upsampling or downsampling techniques [34] with our eCD training algorithm.

MATERIALS AND METHODS

Background on restricted Boltzmann machines

Restricted Boltzmann machines (RBMs, see Figure 1)

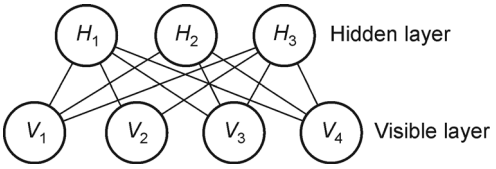

Figure 1. An RBM example where connections only exist between visible and hidden layers.

[10] are undirected graphical models (also referred to as Markov random fields) which describe the probability distributions of binary variables. Various generalizations [21] of RBMs make them effective to model different types of data, e.g., count vectors [18] and real-valued data [21]. Here we introduce the most commonly used RBM (also used in this study) — the Bernoulli-Bernoulli RBM (BBRBM). In the following, we always denote BBRBM by RBM for short.

An RBM can be regarded as a bipartite graph: the visible layer consists of m visible units $\{V_1, ..., V_m\}$ to represent observed data/variables, and the dependency across them is depicted by n hidden units $\{H_1, ..., H_n\}$ which constitute the hidden layer. We note that the both visible and hidden units are binary variables in an BBRBM. In particular, in an RBM, each visible unit is connected to every hidden unit, but there is no edge between each two units in the same layer. This means that the variables in the same layer are conditionally independent given the other layer.

In an RBM, let w_{ij} denote the real-valued weight associated with the edge between the visible variable V_i and the hidden variable H_j, and let b_i and c_i be the bias terms of V_i and H_i, respectively. We further use $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{c}$ to denote the vector representations of corresponding parameters. Then the joint probability mass function of the RBM can be defined by

$$p(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) := \frac{1}{Z(\boldsymbol{\theta})} \exp(-E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})), \tag{8}$$

where $(\mathbf{v}, \mathbf{h}) \in \{0,1\}^{m+n}$, $\boldsymbol{\theta} := \{\mathbf{b}, \mathbf{c}, \mathbf{W}\}$, $Z(\boldsymbol{\theta})$ is the partition function, and the energy function $E : \{0,1\}^{m+n} \to \mathbb{R}$ is defined by

$$E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) := -\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} v_i h_j - \sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j. \tag{9}$$

Note that here m is the dimension of the input features and we have replaced p by m to follow the convention of the machine learning literature. The conditional independence of visible and hidden variables makes it easy to calculate their conditional probabilities:

$$\mathbb{P}(H_j = 1 | \mathbf{v}) = s\left(\sum_{i=1}^{m} w_{ij} v_i + c_j\right) \tag{10}$$

and

$$\mathbb{P}(V_i = 1 | \mathbf{h}) = s\left(\sum_{j=1}^{n} w_{ij} h_j + b_i\right), \tag{11}$$

where $s(x)$ denotes the sigmoid function, i.e., $s(x) := 1/(1 + e^{-x})$.

The principle of training an RBM is based on the maximum likelihood estimation (MLE), i.e., to learn the parameter set $\boldsymbol{\theta}$ that maximizes the log-likelihood $\ln\mathcal{L}(S; \boldsymbol{\theta}) = \Sigma_{\mathbf{v} \in S} \ln p(\mathbf{v}; \boldsymbol{\theta})$ of the training set S. The gradient of the log-likelihood is given by

$$\frac{\partial \ln\mathcal{L}(S; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \propto -\mathbb{E}_{p(\mathbf{h}|\mathbf{v})q(\mathbf{v})}\left[\frac{\partial \mathbb{E}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] + \mathbb{E}_{p(\mathbf{h}, \mathbf{v})}\left[\frac{\partial \mathbb{E}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right], \tag{12}$$

where $q(\mathbf{v})$ denotes the empirical distribution. The above gradient is of exponential complexity and thus intractable. One may approximate the expectations by sampling methods, e.g., Gibbs sampling, but this requires running the Markov chain long enough to achieve its equilibrium, which is also less efficient. Intuitively, during sampling we can run the Markov chain for only a few steps before its convergence, which results in the *Contrastive Divergence* (CD) algorithm [17]. CD has been shown to be sufficient for training products of experts and is now a standard training method for RBMs. With k-step CD (usually $k = 1$), the gradient of the log-likelihood for one training sample $\mathbf{v}^{(0)}$ can be approximated by

$$\mathrm{CD}_k(\mathbf{v}^{(0)}; \boldsymbol{\theta}) = -\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^{(0)}) \frac{\partial E(\mathbf{v}^{(0)}, \mathbf{h}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^{(k)}) \frac{\partial E(\mathbf{v}^{(k)}, \mathbf{h}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{13}$$

where $\mathbf{v}^{(k)}$ is the sample value after running the Markov chain for k steps and $\mathbf{v}^{(0)}$ is the original value of observed data. Note that Equation (13) is in its general form, and the update rules (based on the gradient ascent algorithm) for other RBM-liked models can be derived easily.

Classification restricted Boltzmann machines

The *classification restricted Boltzmann machine* (ClassRBM) [35,36] is a variant of the traditional RBM, which is adapted to both unsupervised and supervised

learning settings. The key idea of ClassRBMs is to treat the target class of each sample as another visible variable and model the joint probability distribution of sample features and target using RBMs. In this principle, the ClassRBMs can be trained based on the *hybrid* objectives, i.e.,

$$H(\mathcal{S};\boldsymbol{\theta}) = -(1+\alpha)\sum_{\mathcal{S}} \ln p(y|\mathbf{x}) - \alpha\sum_{\mathcal{S}} \ln p(\mathbf{x}), \quad (14)$$

where $\mathbf{x}$ is the sample feature, y is the sample target and the coefficient α adjusts the fraction of these two terms. Note that the first term (can be computed directly due to the special structure of the ClassRBM) of the r. h. s. of Equation (14) represents the *discriminative* training objective which estimates the sample target given its input feature. Meanwhile, the second term of the r. h. s. of Equation (14) is the usual log-likelihood function w. r. t. the sample features, which is known as the *generative* training objective.

The word "hybrid" comes from combining the generative objective with the discriminative one, which is the main property of ClassRBMs. In this case, the traditional RBMs are special cases of the ClassRBMs, and the CD algorithm can be integrated into training ClassRBMs efficiently. The major flexibility of the ClassRBMs lies in that we can perform semi-supervised learning and multitask learning in this framework quite naturally. Due to the space limitation, we omit most definitions and derivations here. For details, readers may refer to [35] and [36].

Elastic regularization

We have mentioned that RBMs are universal approximators of discrete distributions [37]. However, in the case of "$p \gg N$", i.e., the dimension of data features is much larger than the sample size, the traditional RBM models and their corresponding training algorithms do fall into a dilemma of model complexity vs. generalization, where the biasvariance trade-off plays a more important role in the model design. Mathematically speaking, limited by the insufficient training samples, the "$p \gg N$" case always incurs an illposed problem [38], whose false solutions discover the false structures in the data that results in the overfitting phenomena. It has been shown that the well-studied regularization theory can solve the ill-posed problem satisfactorily and has been widely used in statistics and optimization fields [7,38].

From complexity to generalization

Based on the discussion given by [21], the number of bits it takes to identify an input data determines the amount of constraint each training example imposes on the model parameters. Thus, it is reasonable to fit $m \times n$ parameters to $m \times N$ training bits if $N \gg n$. This implies that we should determine the number of hidden units to be small enough limited by the sample size, which yields large bias but small variance. On the other hand, since the input feature dimension is severely large, and adding hidden units improves the modeling power at least for BBRBM [37], we need large enough number of hidden units to characterize the complicated probabilistic distribution of high-dimensional input data, which yields small bias but large variance. To address this particular bias-variance trade-off issue, here we consider the following regularized optimization problem,

$$\underset{\mathbf{W}}{\text{minimize}} -\frac{1}{|\mathcal{S}|}\ln\mathcal{L}(\mathcal{S}), \text{subject to } \|\mathbf{W}\|_1 \leqslant t \text{ for some } t, \quad (15)$$

where $|\mathcal{S}|$ is the sample size, $\|\cdot\|_1$ denotes the l_1-norm and $\|\mathbf{W}\|_1 := \Sigma_{ij}|w_{ij}|$. Note that here we only consider the regularization of weights. It is well-known that the nature of l_1 regularization generates the sparsity of the model parameters. For example, in linear regression, LASSO [88] plays the role of variable selection and omits those predictors less responsible for the regression objective. For the RBM model, the l_1 regularization term in Problem (15) will shrink most weights to be 0, which results in that each input bit is only responsible for a few hidden units. This enables us to add more hidden units (i.e., increasing the model complexity) for boosting RBM's modeling power without losing its generalization ability (i.e., preventing overfitting). Therefore, we have tackled the bias-variance trade-off in the "$p \gg N$" case preliminarily. Indeed, since the weights minimizing Problem (15) are sparse, each hidden unit is actually activated by a few visible units (see Equation (10)). Thus optimizing Problem (15) can be treated as automatically selecting prominent variables/features responsible for hidden units similarly to LASSO.

From generalization to complexity

It has been shown that in the "$p \gg N$" problem, there are high correlations across visible variables, which clusters the variables into groups [6,9]. Different from linear models (e.g., linear regression and logistic regression), where variable correlations are embodied by corresponding coefficients, RBMs can model this *grouping effect* with hidden variables, which is the main superiority of latent variable models. However, the l_1 regularization appearing in Problem (15) also restricts the model flexibility, which weakens RBM's ability to fully discover variable correlations. In particular, it has been shown

theoretically that LASSO tends to randomly select only one variable from the correlated group [39]. To illustrate this phenomenon informally, we find that two input variables that share similar weights would be positively correlated, while the l_1 regularization breaks this balance and increases their weight variance, which leads to unpredictable correlation modeling. More theoretical analysis on this topic can be found in Section of Test on Simulated Data.

To make a further compromise between the model generalization and its expression power, the well-known l_2 regularization turns to be our natural choice. Intuitively, let us first consider the following simple optimization problem,

$$\underset{w_{ij}}{\text{minimize}} \sum_{ij} w_{ij}^2, \text{subject to} \sum_{ij} |w_{ij}| = c,$$

where c is a constant. Geometrically speaking, we want to minimize the radius of a ball (expressed as the object function) in a high-dimensional space with the restriction that the ball intersects with a given convex polyhedron (expressed as the constraint). Obviously, the solutions of this problem are given when the ball and the fixed convex polyhedron are tangent, where all coordinates of the solution vectors have equal modulus. Thus, by adding another external l_2 regularization term in Problem (15), the weights in the RBM are averaged to some extent, which imposes similar weights between correlated variables expicitely.

At last, we obtain the following optimization problem,

$$\begin{aligned} &\underset{\mathbf{W}}{\text{minimize}} -\frac{1}{|\mathcal{S}|}\ln\mathcal{L}(\mathcal{S}),\\ &\text{subject to } \|\mathbf{W}\|_1 \leqslant t_1 \text{ and } \|\mathbf{W}\|_2^2 \leqslant t_2 \text{ for some } t_1, t_2, \end{aligned} \tag{16}$$

where $\|\cdot\|_2$ denotes the l_2-norm and $\|\mathbf{W}\|_2^2 := \Sigma_{ij} w_{ij}^2$. In particular, it is equivalent to

$$\underset{\mathbf{W}}{\text{minimize}} -\frac{1}{|\mathcal{S}|}\ln\mathcal{L}(\mathcal{S}) + \lambda_1\|\mathbf{W}\|_1 + \lambda_2\|\mathbf{W}\|_2^2, \tag{17}$$

where $\lambda_1$and λ_2 are two fixed coefficients measuring the contributions of corresponding regularization terms. Let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, then the above problem is further equivalent to the problem

$$\begin{aligned} &\underset{\mathbf{W}}{\text{minimize}} -\frac{1}{|\mathcal{S}|}\ln\mathcal{L}(\mathcal{S}),\\ &\text{subject to } (1-\alpha)\|\mathbf{W}\|_1 + \alpha\|\mathbf{W}\|_2^2 \leqslant t \text{ for some } t. \end{aligned} \tag{18}$$

Model training

With only a few modifications, our eRBMs model can be trained efficiently based on the CD algorithm. In the previous sections, we have noticed that the CD algorithm is used to approximate the gradient of the log-likelihood function, i.e., Equation (12). Meanwhile, we note that the optimization function of eRBMs (i.e., Problem (1)) contains the elastic regularization term besides the likelihood function. Therefore, to learn parameters of eRBMs, we should integrate the derivatives of the l_1- and l_2-norms into the CD algorithm.

It is trivial to compute the derivative of l_2-norm for model parameters, i.e.,

$$\frac{\partial\|\mathbf{W}\|_2^2}{\partial w_{ij}} = 2w_{ij}.$$

As for the l_1-norm, here we adopt the subgradient method [40], i.e.,

$$\frac{\partial\|\mathbf{W}\|_1}{\partial w_{ij}} \approx \text{sgn}\{w_{ij}\},$$

which is widely used in convex optimization.

The adapted CD algorithm, which is called *elastic contrastive divergence* (eCD), is shown Algorithm 1. Though intermediate variables have to be sampled (for k times) during the CD/eCD algorithm, attributed to the conditional independence of visible and hidden variables (see Equations (10) and (11)), the Gibbs sampling can be performed quite efficiently. Also, it has been verified empirically that the 1-step CD algorithm can yield satisfiable training results [21], which was also adopted in our tests. We note that several training parameters, e.g., learning rate, training batch size and momentum [21], are omitted in Algorithm 1 for brevity. We also notice that to hybridly train the elastic version of ClassRBMs, we can simply replace the CD algorithm for its generative part by the eCD described in Algorithm 1.

Datasets

The genomic and proteomic profiling types of ovarian serous cystadenocarcinoma (OV) include: (i) mRNA expression (AgilentG4502A, Agilent 244K Custom Gene Expression G4502A), (ii) DNA methylation (27k, Illumina Infinium Human DNA Methylation 27K), (iii) SCNA (SNP 6, Affymetrix Genome-Wide Human SNP Array 6.0), (iv) microRNA (miRNA) expression (H-miRNA 8×15K, Agilent 8×15K Human miRNA-specific microarray platform) and (v) protein expression (RPPA, MD Anderson reverse phase protein array). The datasets we used in tests came from [22], which can be downloaded from [syn1710282]. In particular, Yuan *et al*. [22] first compiled a sample set (called the core set) in which each sample has the above five kinds of information as well as the survival time. Then Yuan *et al*.

Table 4. Statistics of the OV datasets[a].

Overall survival	Selected samples	Positive	Negative	Methy	mRNA	miRNA	SCNA	Protein
563	252	153	99	24, 980	17, 813	798	109	165

[a] The dichotomizing cutoff time was set to be three years to balance the positive and negative sample sizes [22].

[22] dichotomized the censored continuous survival time by setting a cutoff time, i.e., three years, which balanced the positive and negative sample sizes in the prediction task. Individuals who lived beyond the cutoff time were assigned label one (positive samples), while those died before were labeled as zero (negative samples). The ambiguous samples with the survival time censored before three years were excluded. Note that here predicting dichotomized survival time is a classification task in nature, which is not the same as predicting the exact survival time. Moreover, since the RBM models mainly deal with binary variables, we dichotomized the real-valued molecular profiling data based on the corresponding mean cutoff, i.e., values above the mean were set to be one (highly expressed), while those under the mean were set to be zero (lowly expressed). The overview of the OV datasets is illustrated in Table 4.

APPENDIX A

Table A1. Complete test results on the simulated data[a].

λ_2/λ_1	0	0.0001	0.001	0.01	0.1
0	117.58	80.89	85.07	93.53	129.89
0.0001	109.17	94.30	137.38	**67.96**	122.98
0.001	103.83	98.66	133.40	101.74	127.12
0.01	83.67	84.77	92.11	138.16	94.69
0.1	151.16	131.57	83.37	85.83	100.65

[a] The CCDΔ_R for each (λ_1,λ_2)-configuration is presentedin the table, where the smallest one is shown in bold.

APPENDIX B

Table B1. AUROC scores of the 10-fold cross-validation for predicting dichotomized survival time using DNA methylation[a].

λ_2/λ_1	0	0.0001	0.001	0.01	0.1
0	0.620	0.631	0.629	0.610	0.693
0.0001	0.579	0.623	0.679	0.612	0.693
0.001	0.689	0.684	0.689	0.675	**0.696**
0.01	0.672	0.674	0.679	0.675	0.673
0.1	0.677	0.676	0.678	0.679	0.678

[a] The best AUROC score is shown in bold.

Table B2. AUROC scores of the 10-fold cross-validation for predicting dichotomized survival time using mRNA expression[a].

λ_2/λ_1	0	0.0001	0.001	0.01	0.1
0	0.608	0.566	0.552	0.614	**0.709**
0.0001	0.591	0.597	0.602	0.639	0.687
0.001	0.691	0.691	0.696	0.693	0.695
0.01	0.697	0.694	0.698	0.696	0.697
0.1	0.696	0.697	0.694	0.698	0.694

[a] The best AUROC score is shown in bold.

Table B3. AUROC scores of the 10-fold cross-validation for predicting dichotomized survival time using miRNA expression[a].

λ_2/λ_1	0	0.0001	0.001	0.01	0.1
0	0.551	0.569	0.579	0.634	0.657
0.0001	0.549	0.557	0.569	0.631	0.664
0.001	0.607	0.635	0.624	0.659	0.692
0.01	0.692	0.687	0.687	0.692	0.686
0.1	0.693	0.695	**0.6973**	0.691	0.6969

[a] The best AUROC score is shown in bold.

Table B4. AUROC scores of the 10-fold cross-validation for predicting dichotomized survival time using SCNA[a].

λ_2/λ_1	0	0.0001	0.001	0.01	0.1
0	0.562	0.560	0.600	0.635	0.660
0.0001	0.577	0.559	0.585	0.643	0.670
0.001	0.591	0.585	0.595	0.644	0.700
0.01	0.697	0.696	0.693	0.695	0.695
0.1	0.704	0.700	0.700	**0.709**	0.704

[a] The best AUROC score is shown in bold.

Table B5. AUROC scores of the 10-fold cross-validation for predicting dichotomized survival time using protein expression[a].

λ_2/λ_1	0	0.0001	0.001	0.01	0.1
0	0.605	0.611	0.619	0.687	0.690
0.0001	0.618	0.608	0.615	0.690	0.695
0.001	0.630	0.609	0.625	0.676	0.697
0.01	0.695	0.701	0.694	0.695	0.701
0.1	0.698	0.701	**0.703**	0.699	0.702

[a] The best AUROC score is shown in bold.

ACKNOWLEDGMENTS

This work was supported in part by the National Basic Research Program of China (Nos. 2011CBA00300 and 2011CBA00301), the National Natural Science Foundation of China (Nos. 61033001, 61361136003 and 61472205), and China's Youth 1000-Talent Program, the Beijing Advanced Innovation Center for Structural Biology.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Sai Zhang, Muxuan Liang, Zhongjun Zhou, Chen Zhang, Ning Chen, Ting Chen and Jianyang Zeng declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Ding, L., Wendl, M. C., McMichael, J. F. and Raphael, B. J. (2014) Expanding the computational toolbox for mining cancer genomes. Nat. Rev. Genet., 15, 556–570
2. Jiang, P. and Liu, X. S. (2015) Big data mining yields novel insights on cancer. Nat. Genet., 47, 103–104
3. Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A. and Børresen-Dale, A.-L. (2014) Principles and methods of integrative genomic analyses in cancer. Nat. Rev. Cancer, 14, 299–313
4. The Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013) The cancer genome atlas pan-cancer analysis project. Nat. Genet., 45, 1113–1120
5. Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, 2nd ed., NewYork: Springer
6. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A. Jr, Marks, J. R. and Nevins, J. R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA, 98, 11462–11467
7. Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. Stat Sin, 20, 101–148
8. Tibshirani, R. (1994) Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B, 58, 267–288
9. Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. J. R. Statist. Soc. B, 67, 301–320
10. Fischer, A. and Igel, C. (2012) An Introduction to Restricted Boltzmann Machines. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Alvarez, L., Mejail, M., Gomez, L. and Jacobo, J. eds., Vol. 7441 of Lecture Notes in Computer Science, pp. 14–36, Berlin: Springer
11. Hinton, G. E. and Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks. Science, 313, 504–507
12. Hinton, G. E., Osindero, S. and Teh, Y.-W. (2006) A fast learning algorithm for deep belief nets. Neural Comput., 18, 1527–1554
13. Bengio, Y. (2009) Learning deep architectures for AI. Found. Trends Mach. Learn., 2, 1–127
14. Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C. and Zeng, J. (2016) A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res., 44, e32
15. Salakhutdinov, R. and Hinton, G. E. (2009) Deep boltzmann machines. In International Conference on Artificial Intelligence and Statistics, 448–455
16. Le Roux, N. and Bengio, Y. (2008) Representational power of restricted boltzmann machines and deep belief networks. Neural Comput., 20, 1631–1649
17. Hinton, G. E. (2002) Training products of experts by minimizing contrastive divergence. Neural Comput., 14, 1771–1800
18. Hinton, G. E. and Salakhutdinov, R. R. (2009) Replicated Softmax: an Undirected Topic model. In Advances in Neural Information Processing Systems 22. Bengio,Y., Schuurmans, D., Lafferty, J., Williams, C. and Culotta, A. eds., pp. 1607–1614. New York: Curran Associates, Inc
19. Salakhutdinov, R., Mnih, A. and Hinton, G. (2007) Restricted boltzmann machines for collaborative filtering. In Proceedings of the 24th International Conference on Machine Learning, 791–798
20. Wang, Y. and Zeng, J. (2013) Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics, 29, i126–i134
21. Hinton, G. (2010) A practical guide to training restricted Boltzmann machines. In Neural Networks: Tricks of the Trade, pp. 599–619. Berlin: Springer
22. Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., Byers, L. A., Xu, Y., Hess, K. R., Diao, L., *et al.* (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat. Biotechnol., 32, 644–652
23. Bengio, Y. (2012) Practical recommendations for gradient-based training of deep architectures. arXiv:1206.5533
24. Schervish, M. J. (1995) Theory of Statistics. In Springer series in statistics. New York: Springer. Corrected second printing: 1997
25. Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc., 96, 1348–1360
26. Olshausen, B. A. and Field, D. J. (1997) Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res., 37, 3311–3325
27. Olshausen, B. A. and Field, D. J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381, 607–609
28. Bengio, Y., Courville, A. and Vincent, P. (2013) Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell., 35, 1798–1828
29. Ranzato, M. A., Boureau, Y. L. and Le Cun, Y., (2008) Sparse Feature Learning for Deep Belief Networks. In Advances in Neural Information Processing Systems 20. Platt, J., Koller, D., Singer, Y. and Roweis, S. eds., pp. 1185–1192, New York: Curran Associates, Inc
30. Ranzato, M. A., Poultney, C., Chopra, S. and Le Cun, Y. (2007) Efficient Learning of Sparse Representations with an Energy-based Model. In Advances in Neural Information Processing Systems 19. Schölkopf, B. Platt, J. and Hoffman, T., eds., pp. 1137–1144. Cambridge : MIT Press
31. Ranzato, M., Huang, F., Boureau, Y. and LeCun, Y. (2007) Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. In Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, 1–8
32. Nair, V. and Hinton, G. E. (2009) 3D Object Recognition with Deep Belief Nets. In Advances in Neural Information Processing Systems 22. Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. and Culotta, A.

eds., 1339–1347. New York: Curran Associates, Inc
33. Min, W., Liu, J. and Zhang, S. (2016) Network-regularized sparse logistic regression models for clinical risk prediction and biomarker discovery. arXiv:1609.06480
34. Chawla, N. V. (2005) Data Mining for Imbalanced Datasets: an Overview. In Data Mining and Knowledge Discovery Handbook, pp. 853–867. New York: Springer
35. Larochelle, H. and Bengio, Y. (2008) Classification using discriminative restricted boltzmann machines. In Proceedings of the 25th International Conference on Machine Learning, 536–543
36. Larochelle, H., Mandel, M., Pascanu, R. and Bengio, Y. (2012) Learning algorithms for the classification restricted boltzmann machine. J. Mach. Learn. Res., 13, 643–669
37. Le Roux, N. and Bengio, Y. (2008) Representational power of restricted boltzmann machines and deep belief networks. Neural Comput., 20, 1631–1649
38. Vapnik, V. N. (1998) Statistical Learning Theory. 1 ed, New Jersey: Wiley
39. Efron, B., Hastie, T., Johnstone, L. and Tibshirani, R. (2004) Least angle regression. Ann. Stat., 32, 407–499
40. Boyd, S. and Vandenberghe, L. (2004) Convex Optimization. New York: Cambridge University Press

