Interpretable Sparse High-Order Boltzmann Machines for Transcription Factor Interaction Identification

Martin Renqiang Min
NEC Labs America

Xia Ning
NEC Labs America

Chao Cheng
Dartmouth College

Mark Gerstein
Yale University

1 Introduction

Identifying interpretable high-order feature interactions is important in both machine learning and data visualization, specially for biomedical applications. In this paper, we propose a new model called Sparse High-order Boltzmann Machine (SHBM) to identify interpretable high-order feature interactions in an unsupervised setting. The learning for `SHBM` is decoupled into two steps: interaction neighborhood estimation and interaction weight learning. In order to identify high-order multiplicative interaction neighborhood for each feature, we propose a scalable sparse high-order logistic regression, named `shooter`, based on ℓ_1-norm regularization. We also propose different sampling methods for learning the interaction weights in `SHBM`. We apply `SHBM` to a challenging bioinformatics problem of discovering complex Transcription Factor interactions from ChIP-Seq measurements in the ENCODE project [1]. Compared to conventional Boltzmann Machine and directed Bayesian Network, `SHBM` can identify much more biologically meaningful interactions that are supported by literature and recent biological studies. To the best of our knowledge, `SHBM` is the first working Boltzmann Machine with explicit high-order feature interactions applied to a real-world problem.

2 Sparse High-Order Boltzmann Machines

In practice, it is typically infeasible for High-order Boltzmann Machines (HBMs) to include all possible energy functions of different orders. Thus, we need to perform structure learning, which is a challenging task for high-dimensional discrete graphical models. Following [7], the structure learning of HBMs could be conducted by minimizing the following ℓ_1-regularized negative log-likelihood

$$\min_{\mathbf{W}} E(\mathbf{v}) + \log Z + \lambda \|\mathbf{W}\|_1.$$

[1] http://www.genome.gov/10005107

That is, we constrain the HBM to have only a sparse set of all possible high-order interactions. However, calculating the above negative log-likelihood and its gradient is intractable. To address this, we convert the problem of minimizing the negative log-likelihood of observed data into that of minimizing the negative pseudo log-likelihood as proposed in [5]. Specifically, we solve the following optimization function

$$\min_{W} \sum_i \log p(v_i|\mathbf{v}_{-i}, W) + \lambda \|W\|_1,$$

where $\mathbf{v}_{-i}$ is the set of visible units except v_i. Essentially, the above optimization takes the form of a set of ℓ_1-regularized logistic regression problems that are not independent due to the shared parameters W.

Due to the extremely large space of the parameters for the high-order interactions, we approximate the above pseudo log-likelihood further by utilizing a strategy proposed by Wainwright *et al* [10] and propose the following decoupled 2-step method for learning an `SHBM`.

Step 1: high-order interaction neighborhood estimation: we first estimate the high-order interaction neighborhood structure of each visible unit, i.e., the Markov blanket of each unit. We formulate this problem as a high-order feature selection problem and propose a learning algorithm, denoted as `shooter`, as described in Section 3. In particular, for each visible unit (i.e., each feature), we consider a regression problem from all the other visible units and their high-order interactions.

Step 2: `SHBM` weight learning: once the high-order interaction neighborhood structure of each visible unit is identified, we add the corresponding energy functions with respect to the high-order interaction of that unit into the energy function of HBM as in Equation 1. Then we use Maximum-Likelihood Estimation updates to learn the weights associated with the identified high-order energy functions, which requires drawing samples from the model distribution. In Section 4, we present Gibbs Sampling and Mean-Field updates for obtaining samples. Instead of drawing samples exactly from the equilibrium model distribution, we only

perform sampling a few steps and use Contrastive Divergence (CD) [4] to update the weights.

$$-E(\mathbf{v}) = \sum_{j=1}^{m} \sum_{i_1 i_2 \cdots i_j} W_{i_1 i_2 \cdots i_j} v_{i_1} v_{i_2} \cdots v_{i_j}, \quad (1)$$

3 Sparse High-Order ℓ_1-Regularized Logistic Regression

We extend the conventional ℓ_1-`LR` to have both single features and multiplicative feature interactions of orders up to m as predictors with ℓ_1 regularization, and this method is denoted as **s**parse **h**igh-**o**rder l**o**gis**t**ic r**e**g**r**ession (`shooter`). The optimization problem of `shooter` with feature interactions of maximum order m is as follows:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \sum_{i=1}^{n} \log\{1+ \\ & \exp[-y_i(\sum_{k=1}^{m} \sum_{j_1<j_2<\cdots<j_k} w_{j_1 j_2 \cdots j_k} x_i^{j_1} x_i^{j_2} \cdots x_i^{j_k} + b)]\} \\ & + \sum_{k=1}^{m} \lambda_k \sum_{j_1<j_2<\cdots<j_k} |w_{j_1 j_2 \cdots j_k}|, \end{aligned} \quad (2)$$

where x_i^j denotes the j-th feature of $\mathbf{x}_i$. Solving the problem in Equation 2 directly is intractable even for fair feature set size p and small interaction order m (e.g. $p = 500$, $m = 6$). Thus, we propose a greedy block-wise optimization method to solve Equation 2.

We decompose the above problem into several sub-problems and solve the sub-problems greedily from the lowest order 1 up to the maximum order m as follows.

Step 1: first, we denote the set of all the single features as $F_0^{(1)}$, that is,

$$F_0^{(1)} = \{x_\cdot^j | \forall j\}$$

We use PSSG to solve the optimization problem as in Equation 3.

$$\begin{aligned} \min_{\mathbf{w}^{(1)},b^{(1)}} \quad & \sum_{i=1}^{n} \log\{1 + \exp[-y_i(\sum_{x_i^j \in F_0^{(1)}} w_j^{(1)} x_i^j + b^{(1)})]\} \\ & + \lambda_1 \sum_{x_i^j \in F_0^{(1)}} |w_j^{(1)}|. \end{aligned} \quad (3)$$

The discriminative single features are identified as the ones which have non-zero weights $w_j^{(1)}$ across all the data points. We denote this set of identified single features by $F^{(1)}$, that is,

$$F^{(1)} = \{x_\cdot^j | x_\cdot^j \in F_0^{(1)}, w_j^{(1)} \neq 0\},$$

where $j = 1, ..., p_1$, $p_1 = |F^{(1)}|$.

Step 2: then we multiply each discriminative feature in $F^{(1)}$ with all the rest $p-1$ single features in $F_0^{(1)}$ to construct the set of all possible second-order feature interactions $F_0^{(2)}$, that is

$$F_0^{(2)} = \{x_\cdot^{j_1} x_\cdot^{j_2} | x_\cdot^{j_1} \in F^{(1)}, x_\cdot^{j_2} \in F_0^{(1)}, j_1 \neq j_2\}$$

We solve the optimization problem as in Equation 4

$$\begin{aligned} \min_{\mathbf{w}^{(2)},b^{(2)}} \quad & \sum_{i=1}^{n} \log\{1+ \\ & \exp[-y_i(\sum_{x_i^{j_1} \in F^{(1)}} w_{j_1}^{(2)} x_i^{j_1} \\ & \quad + \sum_{x_i^{j_1} x_i^{j_2} \in F_0^{(2)}} w_{j_1 j_2}^{(2)} x_i^{j_1} x_i^{j_2} + b^{(2)})]\} \\ & + \lambda_1 \sum_{x_i^{j_1} \in F^{(1)}} |w_{j_1}^{(2)}| + \lambda_2 \sum_{x_i^{j_1} x_i^{j_2} \in F_0^{(2)}} |w_{j_1 j_2}^{(2)}|. \end{aligned} \quad (4)$$

so as to identify discriminative second-order feature interaction set $F^{(2)}$, that is,

$$F^{(2)} = \{x_\cdot^{j_1} x_\cdot^{j_2} | x_\cdot^{j_1} x_\cdot^{j_2} \in F_0^{(2)}, w_{j_1 j_2}^{(2)} \neq 0\}.$$

Step 3: similarly, we multiply each discriminative $(k-1)$-th order feature interaction in set $F^{(k-1)}$ with $p-k+1$ other single features in $F_0^{(1)}$ to construct the set of all possible k-th order interactions $F_0^{(k)}$, that is,

$$\begin{aligned} F_0^{(k)} \quad = \{ & x_\cdot^{j_1} x_\cdot^{j_2} \cdots x_\cdot^{j_k} | x_\cdot^{j_1} x_\cdot^{j_2} \cdots x_\cdot^{j_{k-1}} \in F^{(k-1)}, \\ & x_\cdot^{j_k} \in F_0^{(1)}, \\ & j_k \neq j_{k-q}, \forall q = 1, \cdots, k-1\} \end{aligned}$$

Then from $F_0^{(k)}$ we identify discriminative feature interaction set $F^{(k)}$ by solving the optimization problem as in Equation 5.

$$\begin{aligned} \min_{\mathbf{w}^{(k)},b^{(k)}} \quad & \sum_{i=1}^{n} \log\{1+ \\ & \exp[-y_i(\sum_{q=1}^{k-1} \sum_{x_i^{j_1} x_i^{j_2} \cdots x_i^{j_q} \in F^{(q)}} w_{j_1 j_2 \cdots j_q}^{(k)} x_i^{j_1} x_i^{j_2} \cdots x_i^{j_q} \\ & \quad + \sum_{x_i^{j_1} x_i^{j_2} \cdots x_i^{j_k} \in F_0^{(k)}} w_{j_1 j_2 \cdots j_q}^{(k)} x_i^{j_1} x_i^{j_2} \cdots x_i^{j_k} \\ & + b^{(k)})]\} \\ & + \sum_{q=1}^{k-1} \lambda_q \sum_{x_i^{j_1} x_i^{j_2} \cdots x_i^{j_q} \in F^{(q)}} |w_{j_1 j_2 \cdots j_q}^{(k)}| \\ & + \lambda_k \sum_{x_i^{j_1} x_i^{j_2} \cdots x_i^{j_k} \in F_0^{(k)}} |w_{j_1 j_2 \cdots j_k}^{(k)}|. \end{aligned} \quad (5)$$

and the order-k discriminative feature interaction set $F^{(k)}$ is identified as

$$F^{(k)} = \{x_{\cdot}^{j_1} x_{\cdot}^{j_2} \cdots x_{\cdot}^{j_k} | x_{\cdot}^{j_1} x_{\cdot}^{j_2} \cdots x_{\cdot}^{j_k} \in F_0^{(k)}, w_{j_1 j_2 \cdots j_k}^{(k)} \neq 0\}.$$

Note that in Equation 5 we include discriminative single features and discriminative lower-order interactions $F^{(1)}, \cdots, F^{(k-1)}$ into the ℓ_1-regularized optimization problem for order k so as to optimally remove less important lower-order interactions when high-order interactions present. To speed up the optimization, we divide each identified discriminative feature interaction set F into equal-sized blocks, and we expand each block and solve the ℓ_1-regularized optimization problem for the particular block.

The above greedy optimization approach sequentially identifies discriminative feature interactions of different orders that essentially form a tree structure, because each k-th order discriminative feature interactions must have at least one of its $(k-1)$-th order constituents belonging to $F^{(k-1)}$, where $k > 1$. Although this greedy approach can only identify a sub-optimal solution to the original intractable optimization problem in Equation 2, it performs very well in practice as demonstrated by our experimental results.

4 Sampling Methods for SHBM

In this section, we present Contrastive Divergence (CD) learning [4] based on Gibbs Sampling (GS) and damped Mean-Field updates (MF). The weight updates in SHBM based on CD are as follows,

$$\Delta W_{i_1 i_2 \cdots i_j} = \epsilon(\langle v_{i_1} v_{i_2} \cdots v_{i_j} \rangle_{\text{data}} - \langle v_{i_1} v_{i_2} ... v_{i_j} \rangle_T), \tag{6}$$

where $\langle v_{i_1} v_{i_2} ... v_{i_j} \rangle_T$ is calculated using the samples obtained from different sampling methods after T steps. Although CD updates do not exactly follow the gradient of data log-likelihood, it works well in practice.

Gibbs sampling (GS) can be used within CD for drawing samples. To perform Gibbs Sampling, we initialize $\mathbf{r}^{(0)}$ to be a random data vector, and we sample each visible unit v_j sequentially using the conditional probability

$$p^{(t)}(v_j | r_1^{(t)}, \cdots, r_{j-1}^{(t)}, r_{j+1}^{(t-1)}, \cdots, r_p^{(t-1)})$$

to get the sample for unit v_j in step t, where $j = 1, \ldots, p, t = 1, \ldots, T$, and p is the total number of visible units. Then we use the statistics in the T-step samples to calculate the second term in Equation 6 for weight updates.

However, GS is very slow due to the sequential sampling procedure over all the visible units. Instead, we use mean-field approximations (MF) [11] to calculate the sampled values for all the visible units in each step in parallel given the sample values in the previous step. In specific, we use the damped version of mean-field updates [8] to draw samples to increase sampling stability. Starting from a random data vector $\mathbf{r}^{(0)}$, we calculate the t-step sample for each visible unit v_j as follows,

$$r_j^{(t)} = \lambda r_j^{(t-1)} + (1 - \lambda) p(v_j = 1 | \mathbf{v}_{-j}, \mathbf{W}),$$

where $t = 1, \ldots, T$, and $p(v_i = 1 | \mathbf{v}_{-i}, \mathbf{W})$ is the conditional probability of $v_i = 1$ given its neighborhood interactions. Please note that, unlike in GS, we can calculate $\mathbf{r}^{(t)}$ for all the visible units in parallel to speed up our computation because the calculation for $\mathbf{r}^{(t)}$ is only dependent on $\mathbf{r}^{(t-1)}$.

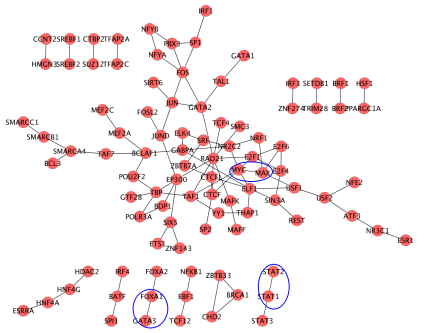

Figure 1: Interaction network from SHBM

5 Experiments

5.1 Datasets

We evaluate SHBM and shooter for interaction identification and feature reconstruction on the TF dataset. The dataset TF is downloaded from Gerstein *et al* [3]. Against a set of regulatory targets which have promoter-proximal binding sites, 116 human TFs were tested through ChIP-seq experiments. On those confident gene-TF interactions shown by the experiments, interaction scores were calculated based on a probabilistic model and weighed by the characteristic profile of the corresponding TF [2]. Then the most confident interactions were selected based on the refined interaction scores so as to construct the TF dataset. In TF, each gene is considered as a data point, each TF is considered as a feature dimension, and each data point is represented by the interaction profile (i.e., 1 for interaction and 0 for non-interaction) of the corresponding gene with respect to the TFs. The dataset TF has in total 9,322 data points and 116 features with density 2.58%.

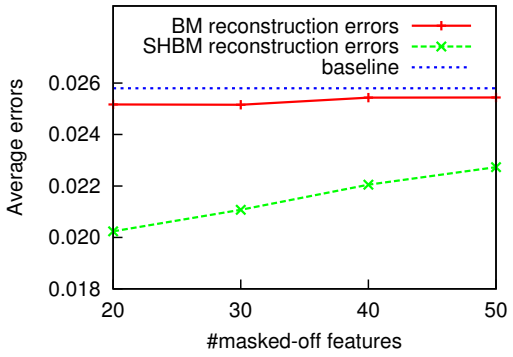

Figure 2: Reconstruction on testing set

5.2 Interaction Identification

Figure 1 shows the interactions identified by SHBM on dataset TF, where only the interactions with weights higher than 2.5 are presented.

As an example, three important interactions: MYC vs MAX, STAT1 vs STAT2 and FOXA1 vs GATA3, are successfully identified (and also highlighted) in the interaction network. MYC and MAX form a MAX/MYC heterodimer, which has been discovered and studied recently in literature [1]. The interaction between MYC and MAX is ranked second by shooter among all the identified interactions. STAT1 and STAT2 also form an heterodimer and the interaction has been studied in literature [6]. The interaction between FOXA1 and GATA3 is also highly ranked by shooter, which also has support from several recent studies [9].

5.3 Data Reconstruction

We compare SHBM and BM on how good the interaction networks that they learn are and how well they can accordingly generate new interactions that are true with high probabilities. We do the comparison by looking at how they can recover missing data. This set of experiments is conducted on TF dataset, which has no labels but its interaction network has important biological significance. 80% of the entire TF dataset is used for SHBM and BM training, whereas the rest 20% is held out for testing.

Figure 2 shows the performance of SHBM and BM on recovering/reconstructing missing data for training set and testing set, respectively. The performance is measured by the average sum of squared errors on each feature per data point, i.e., the total sum of squared errors divided by the product of the number of data points and the number of features. Note that given the density of TF dataset 2.58%, a guess of all 0 values will give a reconstruction sum of squared errors about 0.0258 (the blue baseline in Figure 2).

For the data reconstruction, first a random set of features is selected and masked off from the data, that is, all the corresponding binding between TFs and proteins are reset as none, and it is to utilize the information of the rest features in the data and the interaction relations among features to recover the masked-off part of the original interactions. 20, 30, 40 and 50 features out of 116 are randomly selected and masked off, and then reconstructed by BM and SHBM. Such procedure is repeated 50 times and the average sum of squared errors over the 50 times from BM and SHBM are presented in Figure 2 for testing data. On both training set and testing set, SHBM is constantly superior to BM in terms of its average sum of squared errors on the reconstruction, both of which are better than baseline. In particular, on the testing set, even when 50 features are masked off, which is 43% of the all the features, SHBM reaches a sum of squared errors 0.0227, which is 11.9% better than baseline and 10.7% better than BM.

References

[1] A. Cascón and M. Robledo. Max and myc: a heritable breakup. *Cancer research*, 72(13):3119–3124, 2012.

[2] C. Cheng, R. Min, and M. Gerstein. A probabilistic method for identifying transcription factor target genes from chip-seq binding profiles. *Bioinformatics*, 2011.

[3] M. B. Gerstein and *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, Sept. 2012.

[4] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[5] H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research*, 10:883–906, 2009.

[6] M. G. Katze, Y. He, and M. Gale. Viruses and interferon: a fight for supremacy. *Nature Reviews Immunology*, 2(9):675–687, 2002.

[7] S. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of markov networks using l1 regularization. In *In NIPS*. Citeseer, 2006.

[8] S. Osindero and G. E. Hinton. Modeling image patches with a directed hierarchy of markov random fields. *Advances in neural information processing systems*, 20:1121–1128, 2008.

[9] V. Theodorou, R. Stark, S. Menon, and J. S. Carroll. Gata3 acts upstream of foxa1 in mediating esr1 binding by shaping enhancer accessibility. *Genome research*, 23(1):12–22, 2013.

[10] M. Wainwright, P. Ravikumar, and J. Lafferty. High-dimensional graphical model selection using l̃1-regularized logistic regression. *Advances in neural information processing systems*, 19:1465, 2007.

[11] M. Welling and G. E. Hinton. A new learning algorithm for mean field boltzmann machines. In *Artificial Neural NetworksICANN 2002*, pages 351–357. Springer, 2002.