Multi-way Interacting Regression via Factorization Machines

Mikhail Yurochkin (moonfolk@umich.edu), XuanLong Nguyen, Nikolaos Vasiloglou
Department of Statistics, University of Michigan, LogicBlox

LSA STATISTICS UNIVERSITY OF MICHIGAN

Regression with interactions

Why go beyond linear models?

- Improve predictive power
- Obtain additional insights about the data
- When linear model is not enough: political science, economics, genetics, retail

Problem and challenges

$$\mathbb{E}(Y|x) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{j=1}^{J} \beta_j \prod_{i \in Z_j} x_i.$$

- $2^D - D - 1$ of possible interactions - infeasible even for $D = 30$
- how to model coefficients β_js for all interactions

Modeling interactions

Definition 1. Let $S = \{e_1, \ldots, e_D\}$ be a set of D objects (e.g. indices of variables) and $Z = \{Z_1, \ldots, Z_J\}$ set of J subsets of S: $Z_j \subset S$, for $j = 1, \ldots, J$. Then we say that $G = (S, Z)$ is a hypergraph with D vertices and J hyperedges.

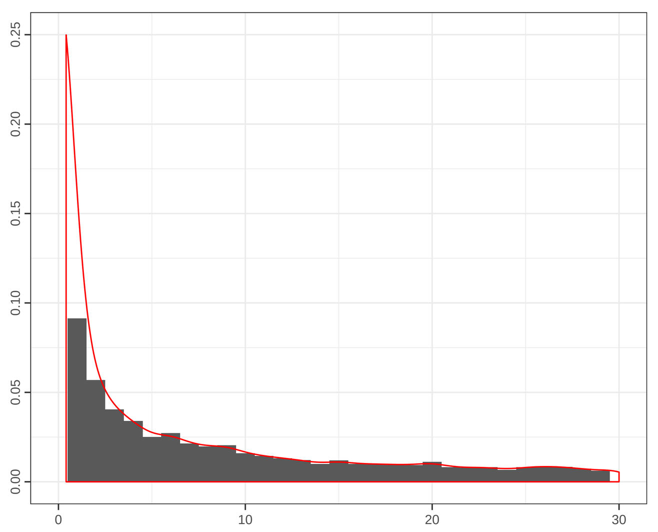
Interactions form a *hypergraph*. Z - incidence matrix of interactions: $Z \in \{0,1\}^{D \times J}$, where $Z_{i_1 j} = Z_{i_2 j} = 1$ iff i_1 and i_2 are part of a hyperedge indexed by column/interaction j.

Prior Challenges

Finite Feature Model (Ghahramani and Griffiths, 2005):

$$\mathbb{P}(Z_i = 1|M_{i-1}) = \frac{M_{i-1} + \gamma_1}{i - 1 + \gamma_1 + \gamma_2},$$

$$\mathbb{E}_{\text{FFM}} M_n = n\frac{\gamma_1}{\gamma_1 + \gamma_2}.$$



mean = 5.0, variance = 60.0

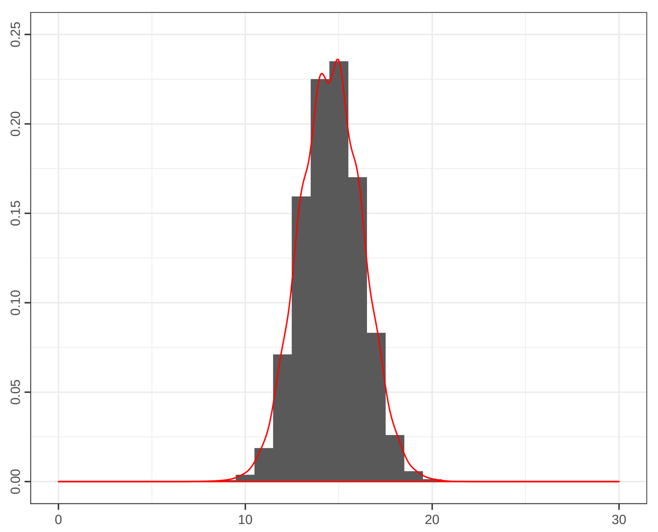
$$M_i := Z_1 + \ldots + Z_i$$

$$\gamma_1 = 0.2; \gamma_2 = 1; n = 30$$

Reversed construction:

$$\mathbb{P}(Z_i = 1|M_{i-1}) = \frac{i - 1 - M_{i-1} + \gamma_1}{i - 1 + \gamma_1 + \gamma_2},$$

$$\mathbb{E}M_n = \frac{n^2 + 2\gamma_1 n - n}{2(n - 1 + \gamma_1 + \gamma_2)}.$$



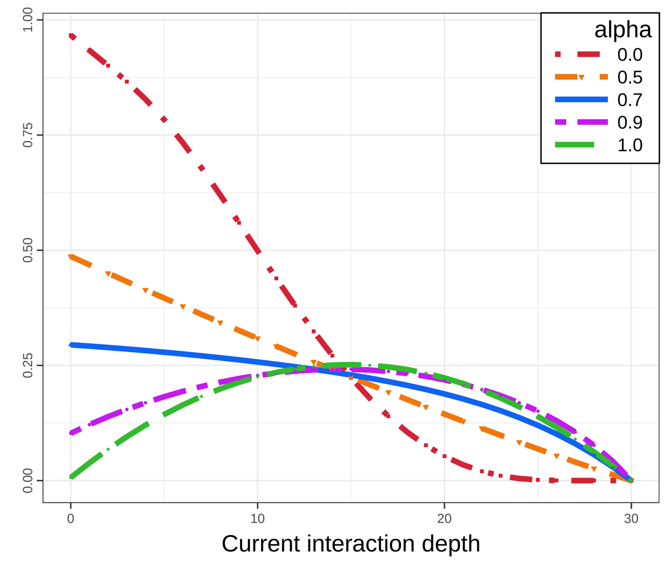
mean = 15.0, variance = 2.6

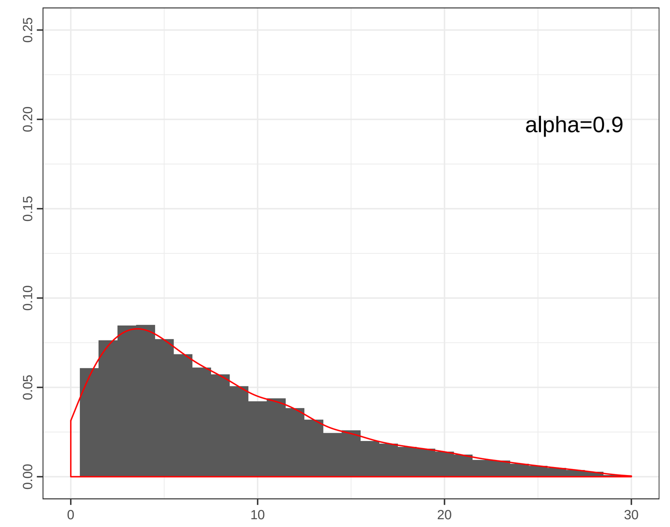
Balancing the two with FFM$_\alpha$

$$\mathbb{P}(Z_{\sigma_i} = 1|Z_{\sigma_1}, \ldots, Z_{\sigma_{i-1}}) = \frac{\alpha M_{i-1} + (1-\alpha)(i - 1 - M_{i-1}) + \gamma_1}{i - 1 + \gamma_1 + \gamma_2}$$

$\sigma(\cdot)$ is a random uniform permutation of $\{1, \ldots, D\}$ and $\sigma_1 = \sigma^{-1}(1), \ldots, \sigma_D = \sigma^{-1}(D)$

$$\mathbb{E}M_n = \\ = \frac{n + 2\alpha + \gamma_1 + \gamma_2 - 2}{n + \gamma_1 + \gamma_2 - 1}\mathbb{E}M_{n-1} + \\ + \frac{n(1-\alpha) + \alpha + \gamma_1 - 1}{n + \gamma_1 + \gamma_2 - 1}.$$





mean = 8.3, variance = 38.7

Modeling coefficients

Factorization Machines (FM) (Rendle, 2010):

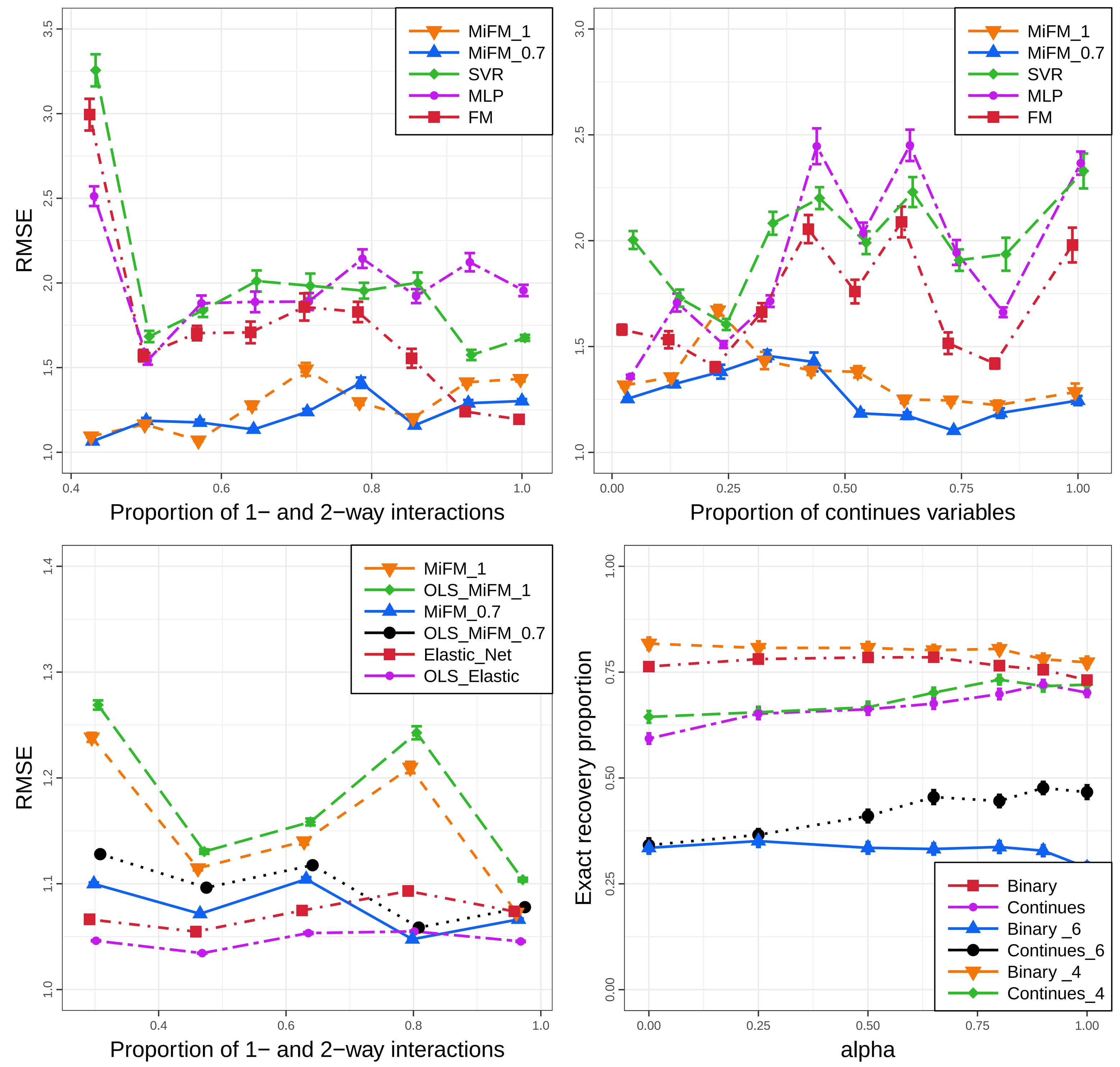
- $\mathbb{E}(Y|x) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{1 \le i < j \le D} \beta_{i,j} x_i x_j$
- factorize interaction weights using PARAFAC (Harshman, 1970): $\beta_{i,j} := \sum_{k=1}^{K} v_{ik} v_{jk}$, where $V \in \mathbb{R}^{D \times K}$ and $K \ll D$
- **MiFM model**: $\mathbb{E}(Y|x) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{j=1}^{J} \sum_{k=1}^{K} \prod_{i \in Z_j} x_i v_{ik}$
- categorical example $z_j = \{\text{color, year, country}\}$; $V_k = \{v_{red,k}, v_{blue,k}; v_{2013,k}, v_{2014,k}, v_{2015,k}; v_{FR,k}, v_{IT,k}, \ldots\}$; and $x = \{\text{blue, 2014, IT}, \ldots\}$, then weight of interaction z_j is
$$\sum_{k=1}^{K} v_{blue,k} \cdot v_{2014,k} \cdot v_{IT,k}.$$

Expressivity of MiFM:

- FFM$_\alpha$ assigns positive probability to any hypergraph of interactions
- system of polynomial equations
$$\sum_{k=1}^{K} \prod_{i \in Z_j} v_{ik} = \beta_j, \; j = 1, \ldots, J$$
should have a solution.

Lemma 1. Given natural number $J \geq 1$, $\beta_j \in \mathbb{R} \setminus \{0\}$ and $Z_j \subset \{1, \ldots, D\}$ for $j = 1, \ldots J$, exists $K_0 < J : \forall K \geq K_0$ system of polynomial equations above has at least one solution in terms of $v_{11}, \ldots, v_{DK}$.

Results



RMSE for experiments: (a) interactions depths; (b) data with different ratio of continues to categorical variables; (c) quality of the MiFM$_1$ and MiFM$_{0.7}$ coefficients; (d) MiFM$_\alpha$ exact recovery of the interactions with different α and data scenarios

Prediction Accuracy on the Held-out Samples for the Gene Data

	MiFM$_1$	MiFM$_0$	LMiFM$_1$	LMiFM$_0$	MLP	RF
3-, 4-, 5-way	0.775	0.771	0.883	0.860	0.870	0.887
only 5-way	0.649	0.645	0.628	0.623	0.625	0.628

Data comes from Himmelstein et al. (2011). In all scenarios true 5-way interaction was present in at least 95% MiFM posterior samples.

Forthcoming

- Extend MiFM to handle power terms via Beta Negative Binomial construction
- Analyze nonparametric version of FFM$_\alpha$