USC Viterbi School of Engineering

NEC Laboratories America
Relentless passion for innovation

Factorized Sparse Learning Models with Interpretable High Order Feature Interactions

Sanjay Purushotham*, Renqiang (Martin) Min#, C.-C. Jay Kuo*, Rachel Ostroff^

*University of Southern California, Los Angeles
#NEC Labs, Princeton
^SomaLogic Inc.

Introduction

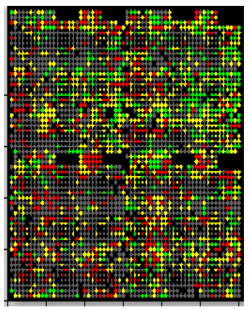
NEC Laboratories America
Relentless passion for innovation

- **High-dimensional** problems
  - Number of observations n << number of variables p
  - Bioinformatics, Vision, Financial Analysis,...
- **Low-dimensional** Structure
  - Sparsity, Low-rank, Block Sparsity,...
- This Talk:
  - Identify interpretable high-order interactions between input features without heredity assumptions

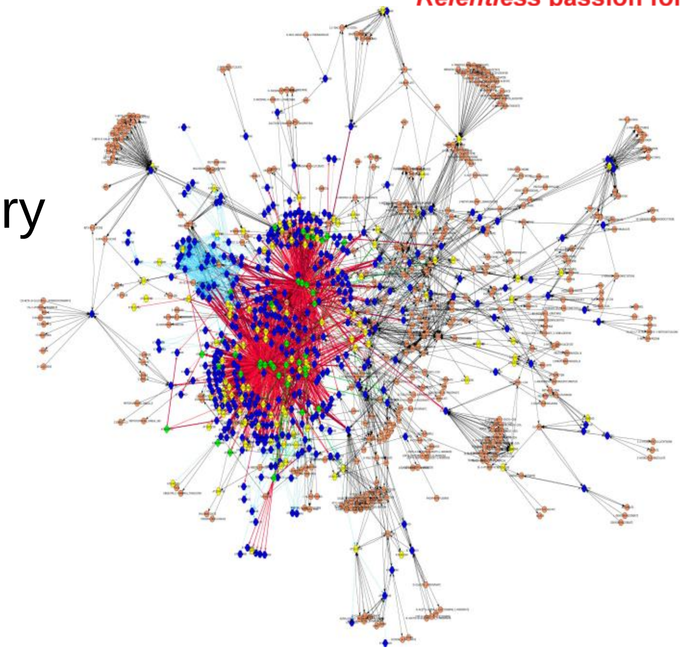
Motivation

Genes

Expression level



Gene regulatory network



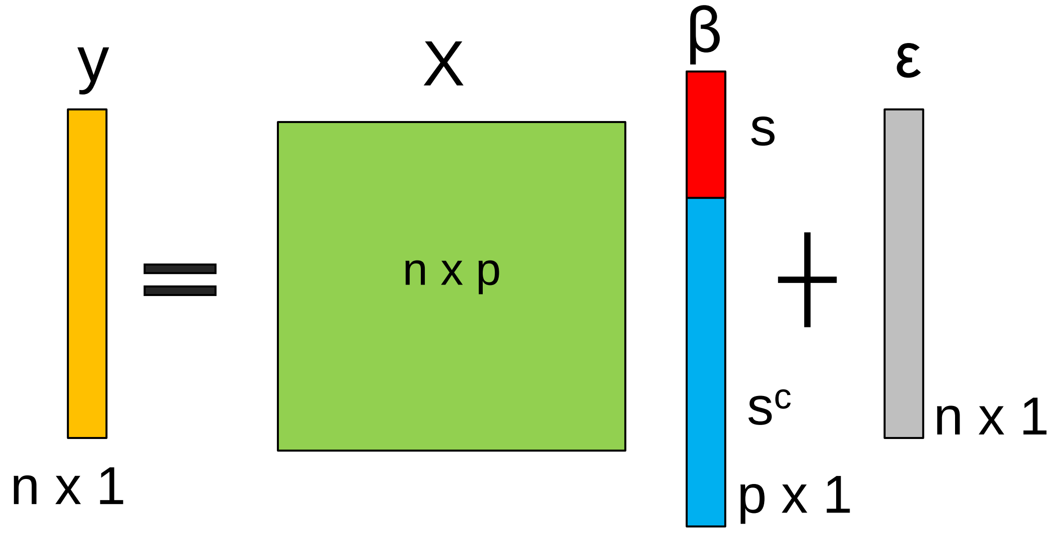
Hey #GOP, no matter how you slam Obama, you own our credit rating downgrade. It is ALL your fault, & we'll remind you in Nov 2012

Image Credit: Diane Oyen

Regression Models

- Linear Regression:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$



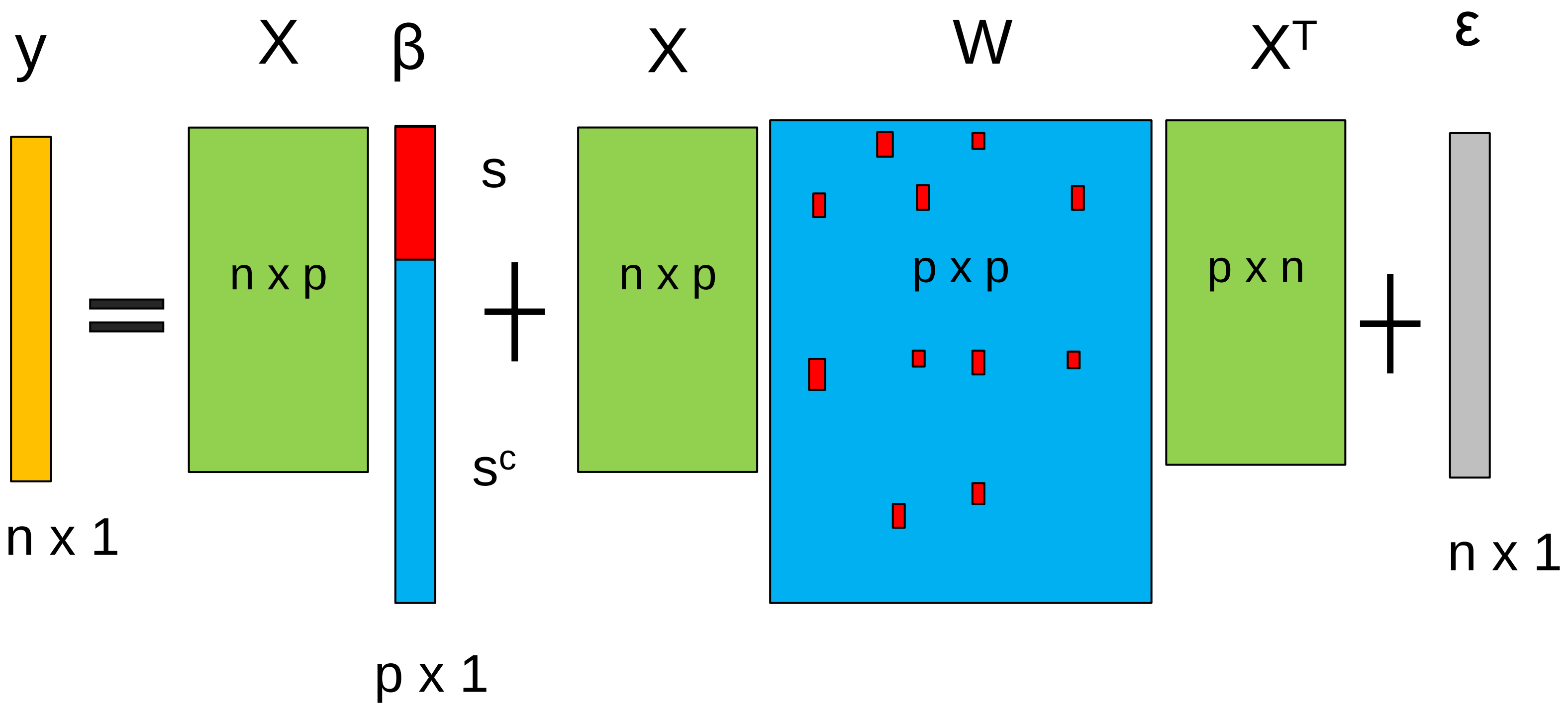
- Logistic Regression:

$$\mathrm{P}\left(y^{(i)} = 1 \middle| \mathbf{X}^{(i)}\right) = \frac{1}{1+exp(-\boldsymbol{X}^{(i)}\boldsymbol{\beta}-\beta_0)}$$

Regression Models

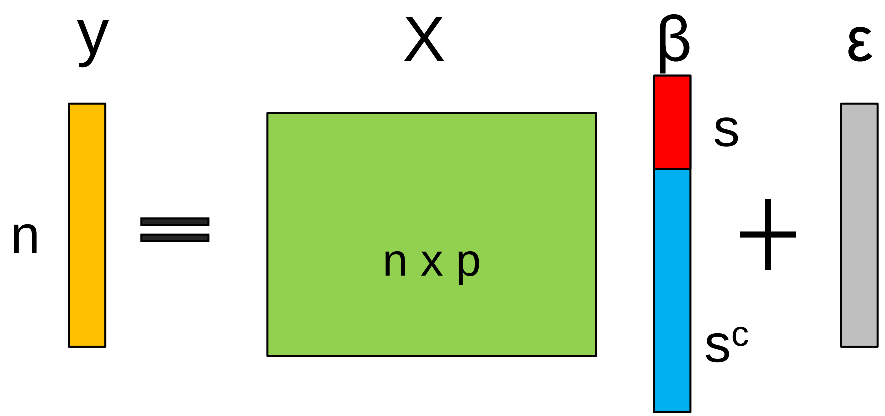
- Linear Regression with high order interactions

$$y = \epsilon + X\beta + \left(X\gamma^T\right)^2 + \ldots$$



Previous Work

NEC Laboratories America
Relentless passion for innovation



- **Lasso** (Tibshirani, 1996)

$$\arg\min_{\beta} \frac{1}{2}\|\boldsymbol{Y}-\mathbf{X}\beta\|_2^2+\lambda\|\beta\|_1$$

- Group Lasso (Yuan et. al, 2006)

$$\min_{\beta} \sum_{k=1}^{r}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|y_i^{(k)}-X_i^{(k)}\beta^{(k)}\right\|_2^2+\lambda\left\|\beta\right\|_{1,\infty}$$

Previous Work

NEC Laboratories America
Relentless passion for innovation

- Variable Selection with **Strong Heredity Constraints** (Choi et. al, 2010)

$$\{\beta^*, \gamma^*_{kk'}\} = \arg\min_{\gamma_{kk'},\beta} \frac{1}{2}\sum_i ||y^{(i)} \boxed{- g(X_i)||_2^2} + \lambda_\beta|\beta|_1 + \boxed{\lambda_{\gamma_{kk'}}|\gamma_{kk'}|_1}$$

- **Hierarchical** Lasso (Tibshirani et. al, 2013)

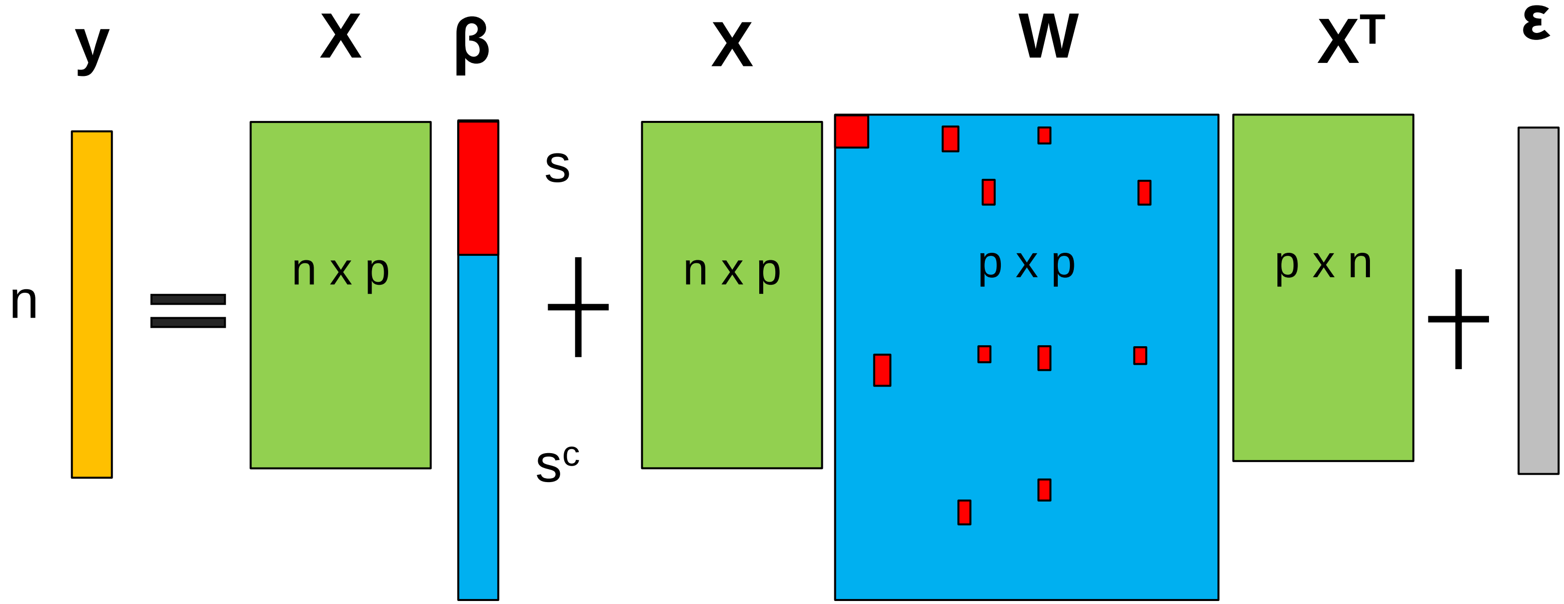
$$\{\beta^*, \Theta\} = \arg\min_{\Theta,\beta} q(\beta,\Theta) + \lambda|\beta|_1 \boxed{+ \frac{\lambda}{2}|\Theta|_1}$$

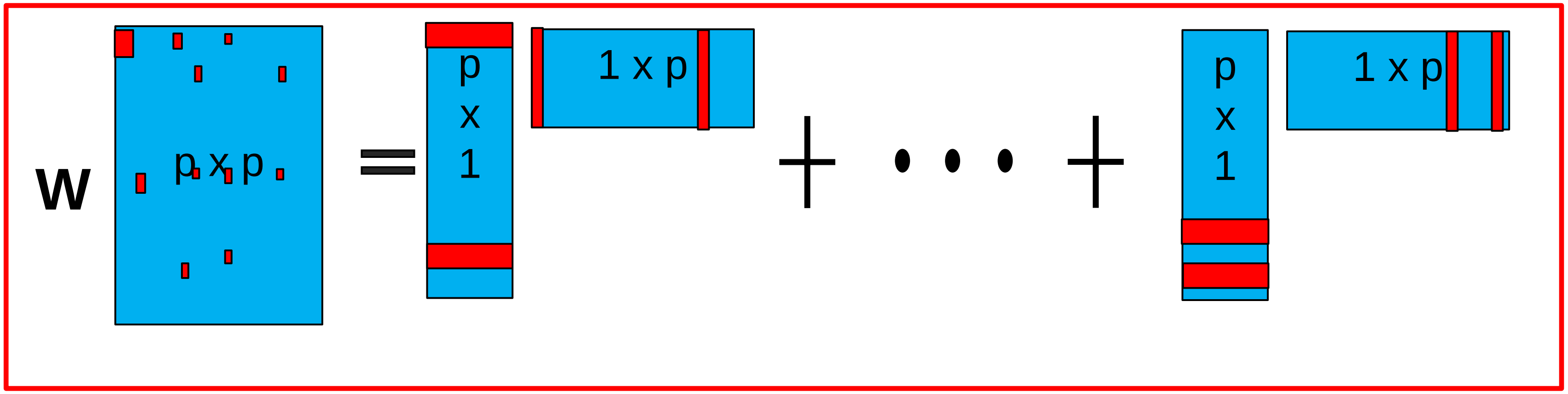
- Our **QUIRE and Shooter** (Martin et. al, 2013, 2014)

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

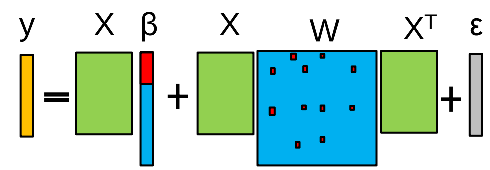
Pairwise interaction coefficients are dependent on main terms

Factorizing Feature Interactions





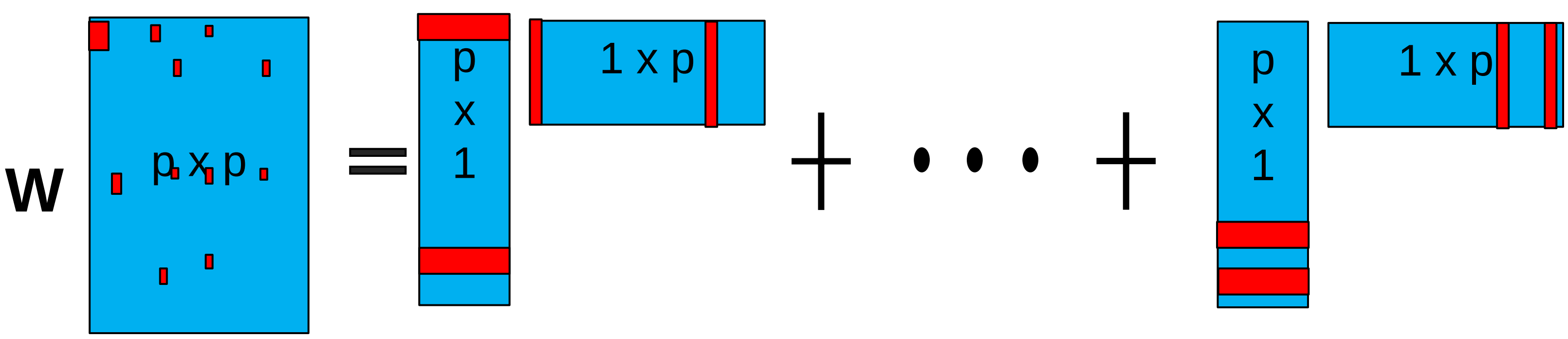
Factorized High order Interactions Model (FHIM)



- Our approach – FHIM
  - Captures pairwise interactions using **tensor product**
  - **Algorithm: Greedy alternating optimization**

$$\{\beta^*, a_k^*\} = \arg\min_{a_k,\beta} \frac{1}{2}\sum_i ||y^{(i)} - \beta X^{(i)} - X^{T(i)} W X^{(i)}||_2^2 + \lambda_1|\beta|_1 + \lambda_2 \sum_k (|a_k|_1)$$

$$\boxed{W = \textstyle\sum_k a_k \odot a_k}$$



Our Approach - FHIM

NEC Laboratories America
Relentless passion for innovation

- Optimization methods
  - **Sub-gradient methods**
    - Orthant-wise Learning (Andrew et. al, 2007)
    - Projected Scaled Subgradient (M. Schmidt, 2010)
  - **Soft-thresholding methods**

$$\tilde{\beta}_j^t(\lambda_\beta) \leftarrow S\Big(\tilde{\beta}_j^{t-1}(\lambda_\beta) + \sum_{i=1}^{n} X_{ij}(y_i - \sum_{k\neq j} X_{jk}\tilde{\beta}_k - \sum_{k} X_{ik}\mathbf{W}X_{ki}), \lambda_\beta\Big)$$

$$\tilde{a}_{kj}^t(\lambda_{a_k}) \leftarrow S\Big(\tilde{a}_{kj}^{t-1}(\lambda_{a_k}) + \sum_{i=1}^{n} X_{ij}(\sum_{r=1}^{p} a_{kr}X_{ir})[y_i - \sum_{k\neq j} X_{jk}\tilde{\beta}_k - \sum_{k} X_{ik}\mathbf{W}_{\sim j}X_{ki}], \lambda_{a_k}\Big)$$

Theoretical Properties

NEC Laboratories America
Relentless passion for innovation

Asymptotic Oracle Properties when $n \to \infty$

Lemma (5.1)

Assume that $a_n = o(1)$ *as* $n \to \infty$. *Then under some regularity conditions (C1)-(C3), there exists a local minimizer* $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ *of* $Q_n(\boldsymbol{\theta})$ *such that*
$\boxed{||\hat{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}^*|| = O_P(n^{-1/2} + a_n)}$

λ's of non-zero coefficients → 0 faster than root-n

Theorem (Sparsity)

Assume that $\sqrt{n} b_n \to \infty$ *and the local minimizer* $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ *given in Lemma 5.1 satisfies* $||\hat{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}^*|| = O_P(n^{-1/2})$. *Then under regularity conditions (C1)-(C3), we have*

$$P(\hat{\beta}_{\mathcal{A}_1^C} = 0) \to 1 \tag{7}$$

$$P(\hat{\alpha}_{\mathcal{A}_2^C} = 0) \to 1 \tag{8}$$

Noise terms are consistently removed with Prob. → 1

Experiments

- Datasets
  - Synthetic Data:
    - Case 1: n>p (n~100-10000, p~50-1000)
    - Case 2: p>n (n~100-500, p~500-2000)
  - Real Datasets
    - RCC- Renal Cell Carcinoma
    - Data collected by SOMAmer technology
    - 212 samples from benign and 4 different stages of cancer

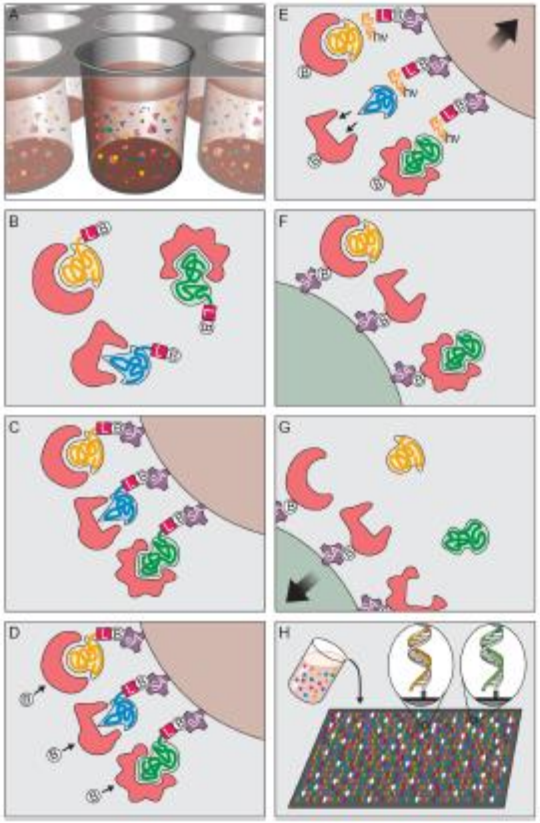


Image Credit: The SOMAmer assay: Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery, Gold et al., 2010

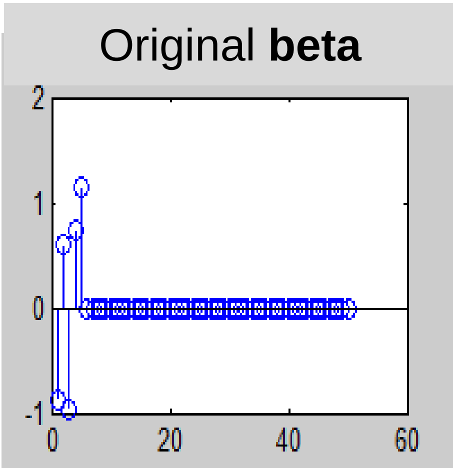
Experiments

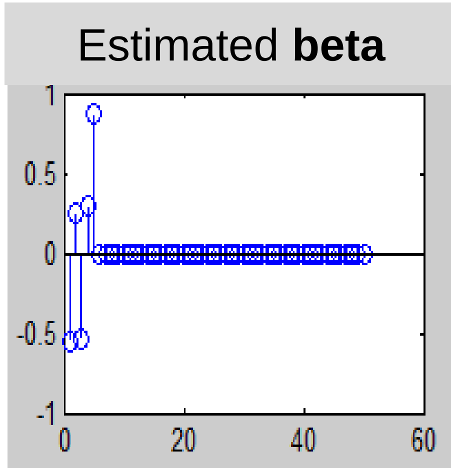
NEC Laboratories America
Relentless passion for innovation

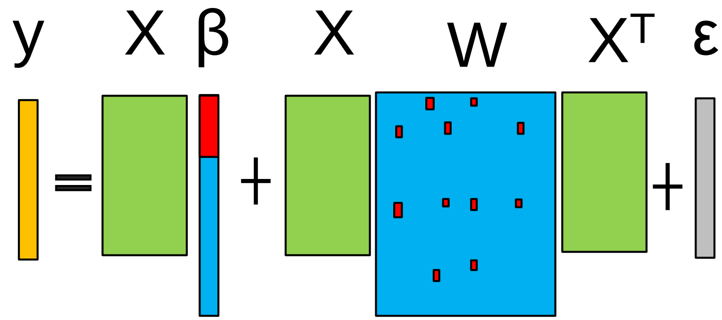
- Experimental Design
  - Prediction error and Support recovery on synthetic data
  - Classification experiments on RCC dataset
    - Case 1: Benign vs. Stage 1-4
    - Case 2: Benign, Stage 1 vs. stage 2-4
    - Case 3. Benign, Stage 1,2 vs. Stage 3,4
  - Compare with state-of-art techniques
  - Interpretability of interactions in real dataset
- Evaluation Metric
  - Prediction error (MSE & std. dev.)
  - Avg. ROC score
  - Avg. F1-score

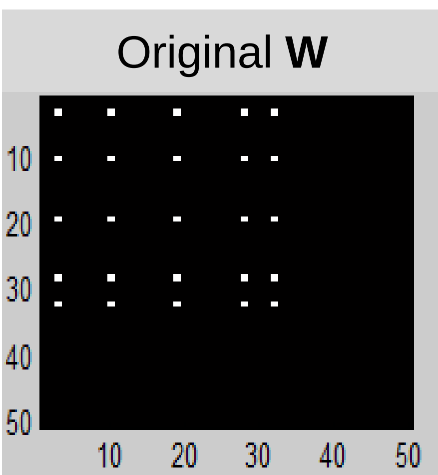
Support Recovery on Synthetic Data (n>p)

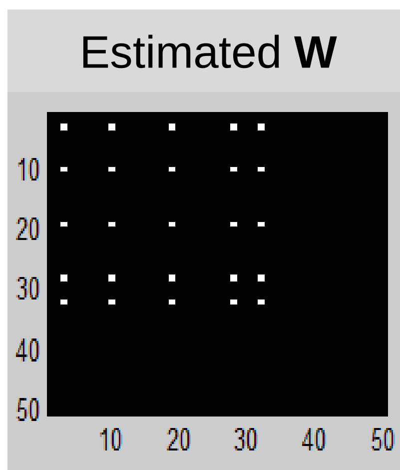
NEC Laboratories America
Relentless passion for innovation

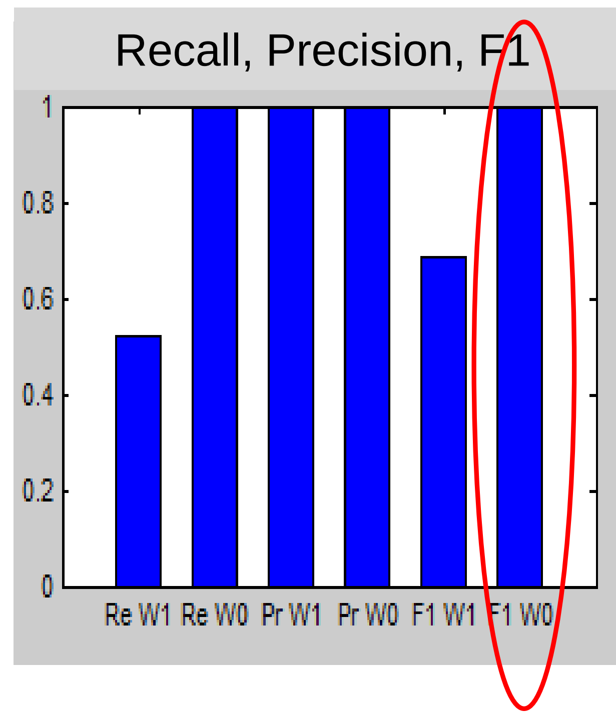






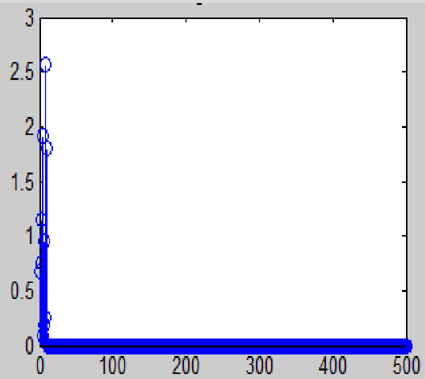




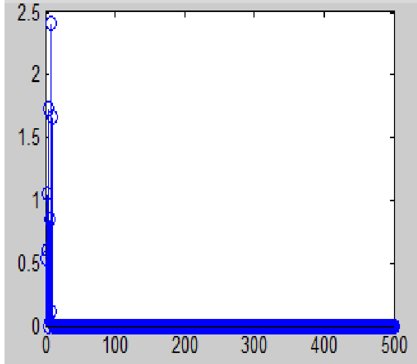


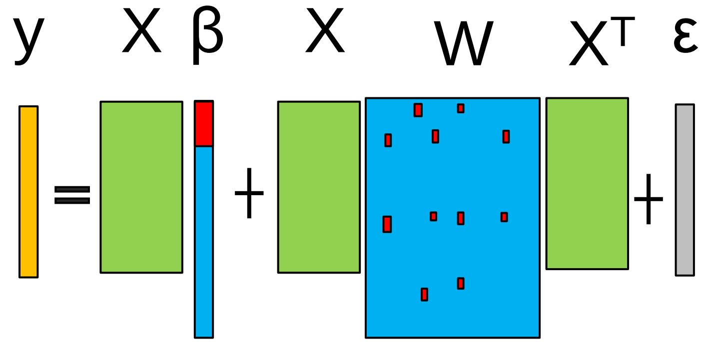
Support Recovery on Synthetic Data (p>n)

Original **beta**

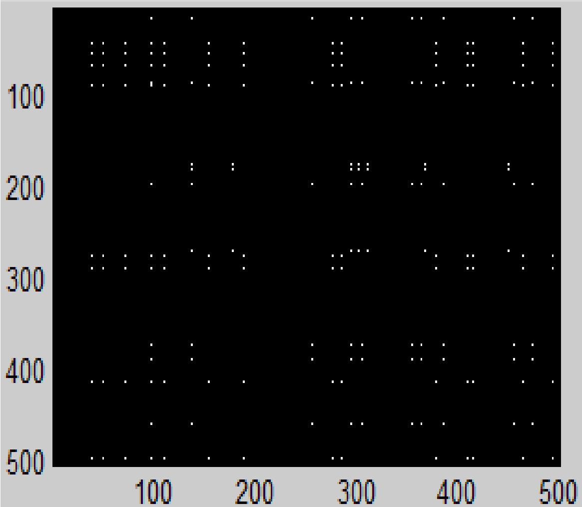


Estimated **beta**

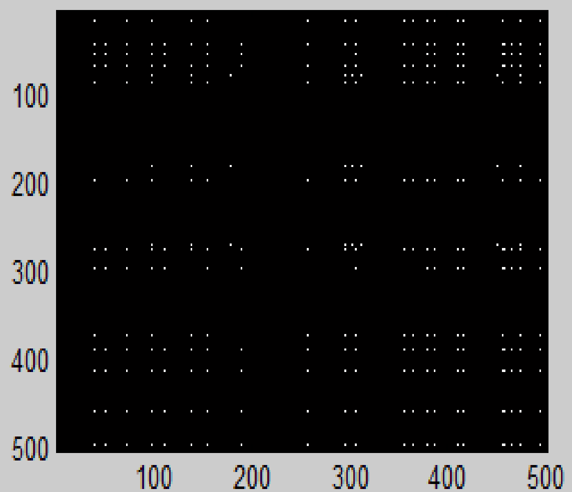




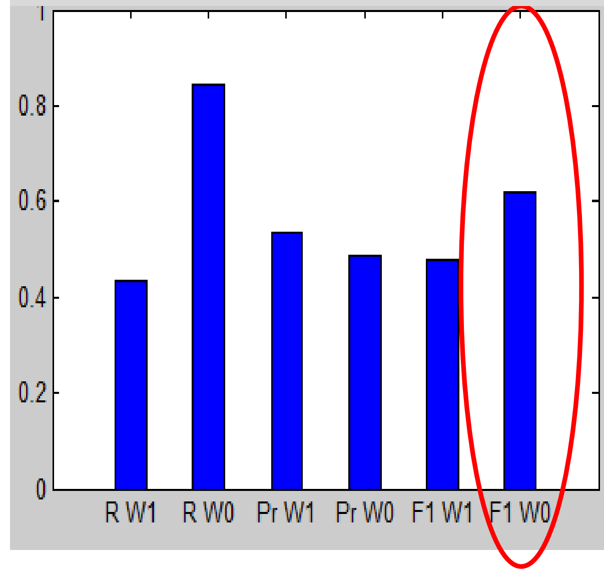
Original **W**



Estimated **W**



Recall, Precision, F1



Support Recovery on Synthetic data

NEC Laboratories America
Relentless passion for innovation

n, p	Sparsity β, a_k	K	Support recovery β, W (F1 score)
1000, 50	5, 5	1	1.0, 1.0
1000, 50	5, 5	3	1.0, 0.95
1000, 50	5, 5	5	1.0, 0.82
10000, 500	10, 20	1	0.95, 0.72
10000, 500	10, 20	3	0.80, 0.64
10000, 500	10, 20	5	0.72, 0.55

Table 4: Support recovery of β, W

n, p	true K	estimated K	W support recovery F1 score
1000, 50	1	1	1.0
1000, 50	3	3	1.0
1000, 50	5	5	0.8
100, 100	1	2	0.75
100, 500	3	2	0.6
100, 1000	5	4	0.5

Table 5: Recovering K using greedy strategy

Prediction Error on Synthetic data

NEC Laboratories America
Relentless passion for innovation

	n, p, K	FHIM	Fused Lasso	Lasso	HLasso	Trace norm	Dirty Model
$q > n$	1000, 50, 1	**338.4(14.5)**	425.9(20.7)	474.7(15.3)	354.32 (24.82)	464.4(36.3)	613.5(0.76)
	1000, 50, 5	**343.7(12.9)**	1888.3(121.1)	1922.9(143.9)	889.1 (112.5)	1822.6(99.8)	2453.8(0.76)
	10000, 500, 1	**1093.1(19.5)**	2739.57(155.1)	3896.3(129.5)	-	3887.9(101.1)	4674.7(0.8)
	10000, 500, 5	**1090.76(12.21)**	22720(597.8)	23279.6(231.3)	-	22916.5(321.4)	29214(0.8)
$p > n$	100, 500, 1	**230.49 (50.3)**	1157.2(355.0)	1335.0(159.2)	-	1160.3(299.7)	1651.9(62.6)
	100, 1000, 1	**340.1 (40.02)**	770.9(127.6)	879.1(180.3)	-	699.9(208.7)	808.1(5.1)
	100, 2000, 1	**907.8 (100.1)**	1022.3(406.2)	919.2(132.1)	-	880.42(471.6)	1916.7(63.4)

Performance comparison for Synthetic data on Linear Regression model with high order interactions

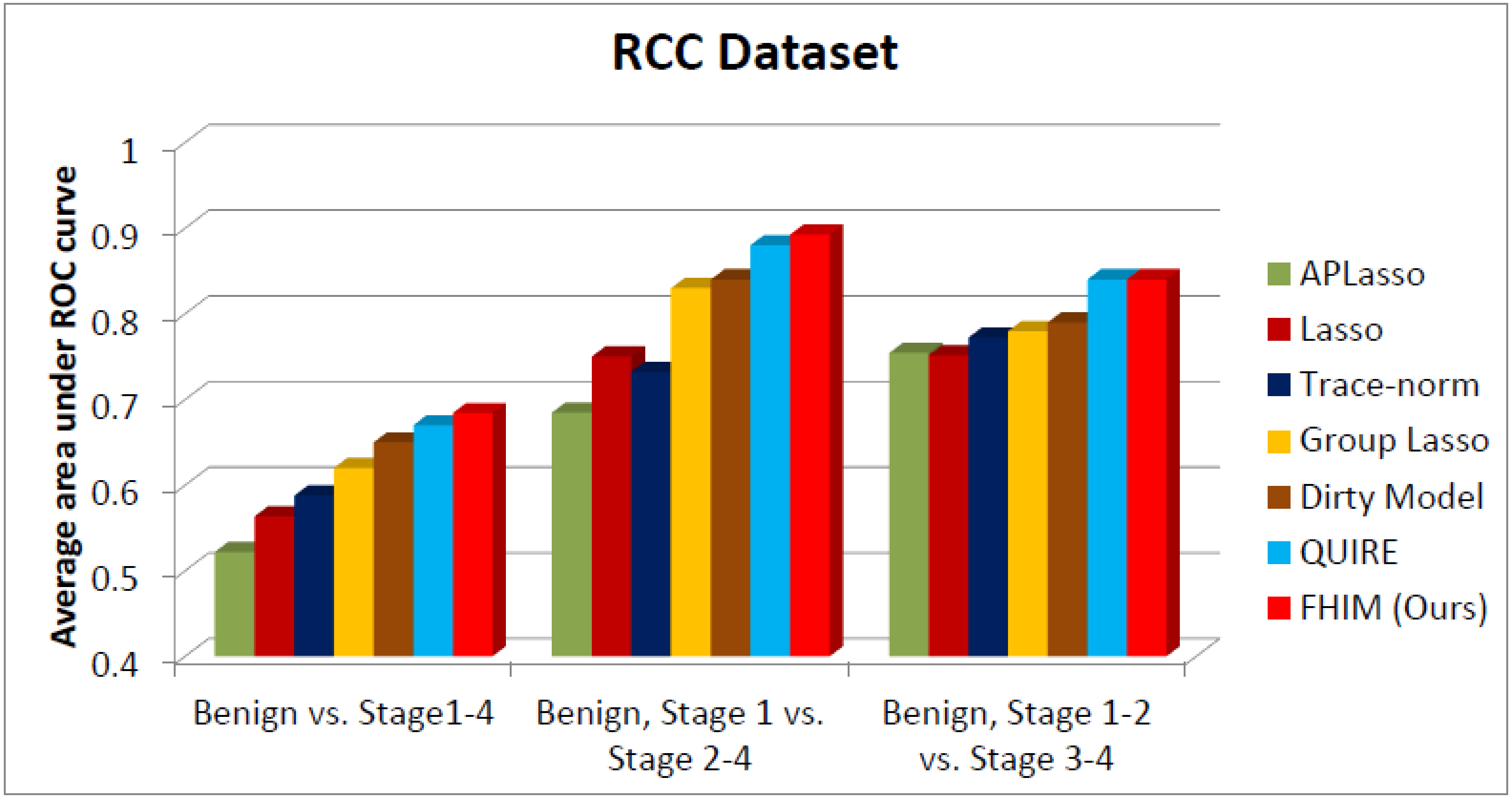
	n, p, K	FHIM	Fused Lasso	Lasso	HLasso	Trace norm
$q > n$	1000, 50, 1	**0.127 (0.009)**	0.128 (0.017)	0.156 (0.017)	0.136 (0.02)	0.128 (0.016)
	1000, 50, 5	**0.189 (0.03)**	0.227 (0.024)	0.292 (0.042)	0.257 (0.022)	0.503 (0.027)
	10000, 500, 1	**0.135 (0.002)**	0.265 (0.007)	0.161 (0.012)	-	0.225 (0.077)
	10000, 500, 5	**0.390 (0.05)**	0.514 (0.006)	0.507(0.108)	-	0.514 (0.006)
$p > n$	100, 500, 1	**0.325 (0.04)**	0.352 (0.086)	0.4323(0.054)	-	0.40(0.079)
	100, 1000, 1	**0.390 (0.056)**	0.409(0.086)	0.458(0.083)	-	0.438(0.011)

Performance comparison for Synthetic data on Logistic Regression model with high order interactions

Classification on RCC Dataset

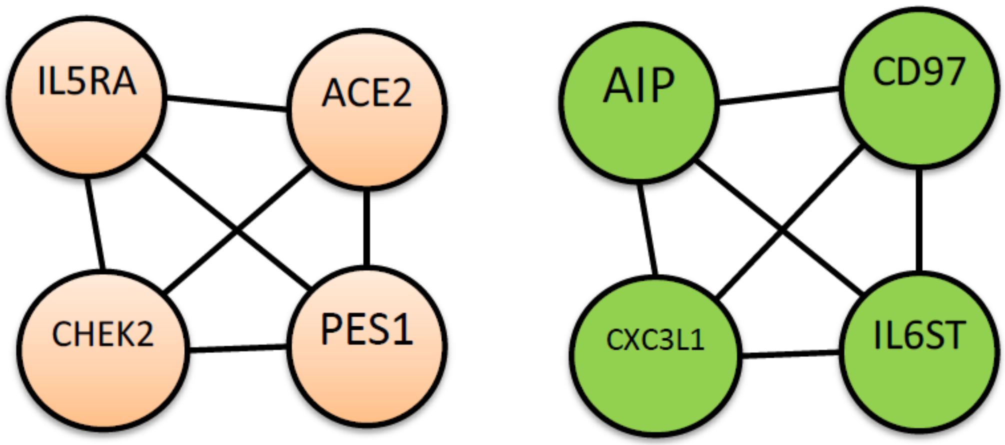
NEC Laboratories America
Relentless passion for innovation

- RCC –212 patients, 1092 proteins measured
- Benign: 40, Stage 1: 101, Stage 2: 17, Stage 3: 24, Stage 4: 31



Interactions in RCC

NEC Laboratories America
Relentless passion for innovation



- CD97 was recently found to promote colorectal cancer[1]
- CHEK2 is known to play a role in several cancers such as lung, kidney, colon, thyroid cancers [2]

[1] M. Wobus, O. Huber, J. Hamann, and G. Aust. Cd97 overexpression in tumor cells at the invasion front in colorectal cancer (cc) is independently regulated of the canonical wnt pathway. **Molecular carcinogenesis**, 45(11):881-886, 2006.
[2] http://ghr.nlm.nih.gov/gene/CHEK2

Summary

- Conclusions
  - Proposed novel sparse learning methods for identify high order feature interactions
  - Promising results on synthetic and real datasets
- Future Work
  - Estimating structure of high-order graphical models
  - Incorporating prior/domain knowledge into the model

Acknowledgements

NEC Laboratories America
Relentless passion for innovation

- Dr. Hans Peter Graf
- Dept. of Machine Learning, NEC Labs

Questions ?

NEC Laboratories America
Relentless passion for innovation

Thank you for listening!