

COMPARISON OF STATISTICAL AND OPTIMIZATION-BASED METHODS FOR DATA-DRIVEN NETWORK RECONSTRUCTION OF BIOLOGICAL SYSTEMS

Behrang Asadi

Mano Maurya

Daniel Tartakovsky

Shankar Subramaniam

University of California, San Diego



OUTLINE

- Introduction
- Methods
 1. LS: least squares
 2. PCA/PCR
 3. LASSO
 4. LMI
- Metrics
- Result of comparison of methods
 1. Experimental data
 2. Synthetic data
- Conclusion



INTRODUCTION

- Data-driven network reconstruction
 - Deriving relationships between input/output data
 - Represent the relationships as a network
- Applied in many areas
 - Chemometrics
 - Biology
- Examples:

Collection of different types of data related to signaling, gene regulatory, and metabolic pathways

* [Il-Gyo Chong, Chi-Hyuck Jun, *Chemometrics and Intelligent Laboratory Systems*, 2005,78,103-112]



INTRODUCTION

- Various methods
 - Optimization-based approaches (leas-squares)
 - Dimensionality reduction methods (PCR and PLS)
 - Partial-correlation-related
 - Bayesian networks analysis
 - Hybrid methods (LMI and LASSO)
- Static, vs. dynamic networks
- Some efforts to compare the performance of various methods
 - More systematic comparison needed with respect to properties of the data such as noise, size, missing data.



METHODS

- Let \mathbf{X} be an input data set (each column normalized to zero-mean and unit standard deviation) and \mathbf{y} (mean-centered) be the corresponding observed response (output) in all methods.
- **Standard least-squares***:
 - Suppose that $\hat{\mathbf{b}}$ is the candidate estimate for the parameter \mathbf{b} in a linear (affine) system. Then the linear regression model of the system becomes:

$$\hat{\mathbf{b}} = \arg \min \{e^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})\}$$

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$RMSE_{LS} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - y_{i,p})^2} = std(\mathbf{y} - \mathbf{y}_p) \times \sqrt{(m-1)/m}$$

* [Å. Björck, SIAM, 1996]



METHODS

- **Principal Component Regression (PCR)***

- Based on principal component analysis.
- Used when $\mathbf{X}^T\mathbf{X}$ is (nearly) singular.
- PCs corresponding to only the first several eigenvalues are used.

$$\mathbf{b}_k = \mathbf{V}_k \times \mathbf{\Gamma}_k^{-1} \times \mathbf{T}^T \times \mathbf{y}$$

$$\mathbf{\Gamma}_k = \{\gamma_j, j = 1, \dots, k\}, \gamma_j : j\text{th eigen value}$$

$$\mathbf{T}_K = \mathbf{X} \times \mathbf{V}_K$$

- \mathbf{V} is the set of corresponding k eigenvectors, and \mathbf{T} is the matrix of latent variables.
- The number of latent variables on the basis of fraction of cumulative variance (say $0.8 < r < 0.95$) captured.
- Fit-error: root mean squared error (RMSE)

$$RMSE_{PCR} = std(\mathbf{y} - \mathbf{y}_p) \times ((m-1) / m)^{1/2}$$

* [Ian T. Jolliffe , *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 1989, Vol. 31, No. 3) **31** (3): 300–303]



METHODS

- **Least Absolute Shrinkage and Selection Operator (LASSO)***

- The problem of reconstruction is cast into a quadratic optimization problem with an additional nonlinear constraint.

$$\hat{\mathbf{b}} = \arg \min \{e^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})\} \quad s / t \quad \sum_j |\hat{b}_j| \leq t$$

- t controls the amount of shrinkage in the estimation of parameters \mathbf{b} .
- For certain values of t ($t \leq t_{LS}$, $t_{LS} = \sum_j |\hat{b}_j|$ obtained from LS) the algorithm shrinks some of the larger parameter-values and sets some of the parameters to zero

* [Tibshirani, R., *Journal of the Royal Statistical Society Series B-Methodological*, 1996. **58**(1): p. 267-288]



METHODS

- **Linear Matrix Inequalities (LMI)***

- Converts a nonlinear optimization problem into a linear optimization problem.

$$\min_{B \in \mathbb{R}^{n \times p}} (e) \quad s.t. \quad (Y - XB)(Y - XB)^T < eI_{m \times m}$$

- Congruence transformation:

$$\begin{pmatrix} -eI_{m \times m} & Y - X\hat{b} \\ (Y - X\hat{b})^T & -I_{p \times p} \end{pmatrix} < 0$$

- Pre-existing knowledge of the system (e.g. $a_{13} > 0$, $a_{21} < 0$) can be added in the form of LMI constraints:

$$v_i^T B u_j + u_j^T B^T v_i = (><)0 \quad v_i = \begin{cases} v_r = 0, r \neq i \\ v_r = 1, r = i \end{cases} \quad u_i = \begin{cases} u_r = 0, r \neq i \\ u_r = 1, r = i \end{cases}$$

- Threshold the coefficients:

$$\bar{\hat{b}}_{ij} = |\hat{b}_{ij}| / \left(\|\hat{b}_{i..}\|_2 \|\hat{b}_{.:j}\|_2 \right)$$

* [Cosentino, C., et al., *IET Systems Biology*, 2007. **1**(3): p. 164-173]



METRICS

- **Metrics for comparing the methods**

- Reconstruction from 80% of datasets and 20% for validation
- RMSE on the test set, and the number and the identity of the significant predictors as the basic metric to evaluate the performance of each method

1. Fractional error in the estimating the parameters

$$\Delta b_{frac,j} = \text{mean} \left(\left| \frac{b_{method,j}}{b_{true,j}} - 1 \right| \right)$$

parameters smaller than 10% of the standard deviation of all parameter values were set to 0 when generating the synthetic data

2. Sensitivity, specificity, G, accuracy

$$\text{Accuracy} : \frac{TN + TP}{TN + TP + FN + FP}$$

$$\text{Sensitivity} : \frac{TP}{TP + FN}$$

$$\text{Specificity} : \frac{TN}{TN + FP}$$

TP : True Positive

FP : False Positive

TN : True Negative

FN : False Negative



COMPARISON OF PCR, LASSO, AND LMI

- Why these methods?
- PCR and PLS from the same family: dimensionality reduction (linear feature extraction)
- LASSO: Least squares with L_1 norm constraint on the coefficients
- LMI: from control theory
 - Based on L_∞ norm
 - Ability to add linear constraints



RESULTS: DATA SETS

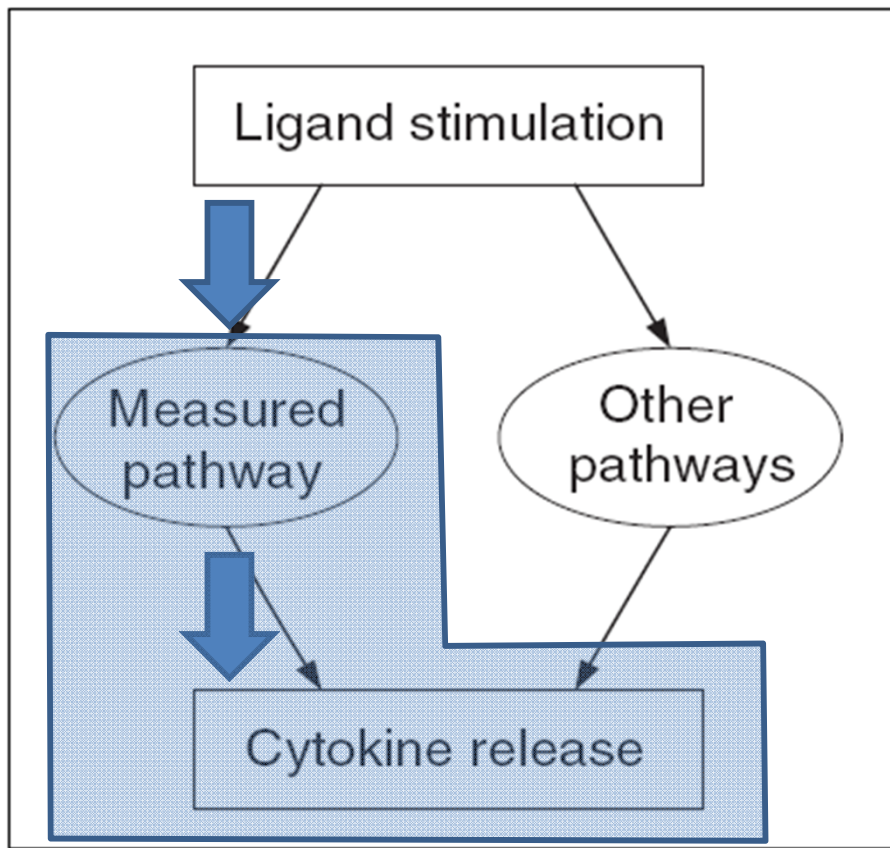
- Data sets for benchmarking: Two data sets
 1. First set: experimental data measured on macrophage cells (Phosphoprotein (PP) vs Cytokine)*
 2. Second sets consist of synthetic data generated in Matlab. We build the model using 80% of the data-set (called training set) and use 20% of data-set to validate the model (called test set).

* [Pradervand, S., M.R. Maurya, and S. Subramaniam, *Genome Biology*, 2006. 7(2): p. R11].



Phosphoprotein-Cytokine Data Set

- Schematic representation of Phosphoprotein (PP) vs Cytokine



- Signals were transmitted through 22 recorded signaling proteins and other pathways (unmeasured pathways).
- Only measured pathways contributed to the analysis

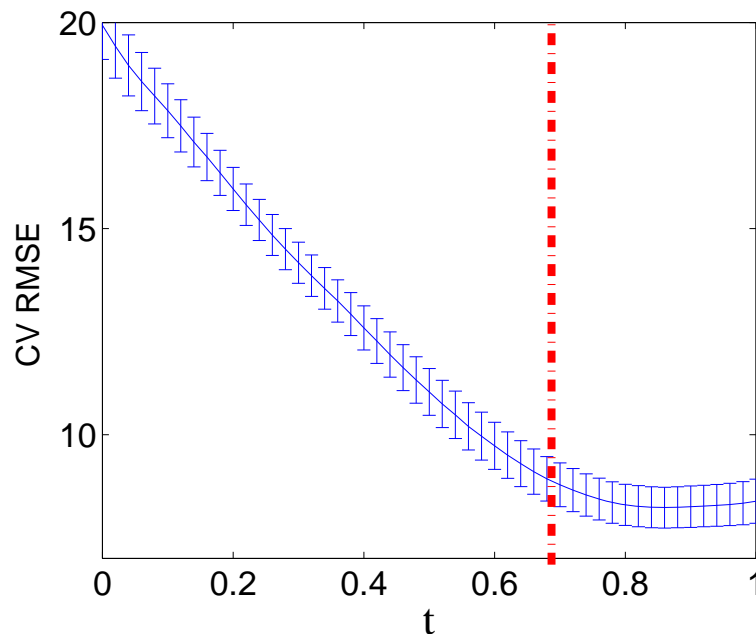
Schematic graphs from:

[Pradervand, S., M.R. Maurya, and S. Subramaniam, *Genome Biology*, 2006. 7(2): p. R11].

RESULTS

- Tuning parameter for LMI and LASSO

The tuning parameters in LASSO and LMI (the threshold parameters t and r_{LMI} , respectively) were identified through k -fold (with $k = 10$) cross-validation on associated dataset.



Optimal values of the tuning parameters in LASSO: 0.66

Optimal values of the tuning parameters in LMI: 0.33

Validation error versus selection threshold t for LASSO on synthetic data set



RESULTS

- Comparison on PP/Cytokine Data
 - The PP/Cytokine data set has 22 inputs and 6 outputs.
 - RMSE of the resulting model by each method was calculated for all the outputs.

RMSE on training set for different methods (PP/cytokine data)

Output	1	2	3	4	5	6
LS	0.73	0.41	0.61	0.61	1.30	0.99
LS_sig	1.16	0.56	0.80	0.80	2.30	1.27
PCR	0.79	0.44	0.73	0.76	1.45	1.06
LASSO	0.92	0.56	0.79	0.72	1.84	1.27
LMI	0.76	0.42	0.67	0.67	1.34	1.09

Output	LS_sig	PCR	LASSO	LMI
G-CSF	3	11	6	12
IL-1a	4	12	6	15
IL-6	1	5	2	8
IL-10	3	7	7	8
MIP-1a	2	11	6	16
RANTES	1	9	4	9
TNFa	4	12	6	13

- LS with significant inputs tends to retain lesser number of inputs.
- LMI tends to retain more inputs.



RESULTS

- **Comparison on synthetic noisy data**

- The methods are applied on synthetic data with **22** inputs and **1** output. The true coefficients for the inputs (about 1/3rd) are made zero to test the methods if they identify them as insignificant.

- **Effect of noise level**

Four outputs with 5, 10, 20 and 40% noise levels, respectively, are generated from the noise-free (true) output.

- **Effect of noise type**

Three outputs with White, t-distributed, and uniform noise types, respectively are generated from the noise-free (true) output

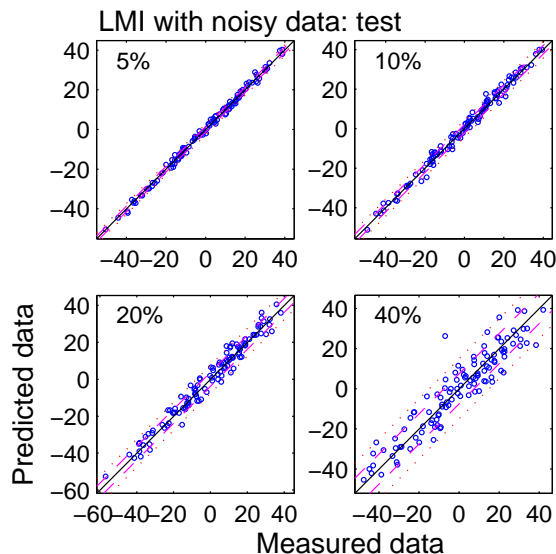


RESULTS

RMSE on all data: methods vs. noise level (synthetic data)

Noise %	5	10	20	40
On training set				
LS_sig	0.94	1.5	2.47	4.22
PCR	2.28	2.79	3.99	6.86
LASSO	3.46	3.52	3.91	5.12
LMI	0.57	1.31	2.39	4.23
On validation set				
LS_sig	0.95	1.64	2.66	4.82
PCR	2.3	2.8	4.05	9.08
LASSO	3.64	3.71	3.89	5.77
LMI	0.97	1.57	2.65	4.9

➤ LS and LMI perform better than PCR and LASSO



Accuracy, Sensitivity, and Specificity of methods for white noise

Noise	5	10	20	40
PCR				
ACC.	0.71	0.71	0.70	0.69
Sense.	0.72	0.70	0.66	0.57
Spec.	0.71	0.73	0.77	0.87
G	0.71	0.71	0.71	0.70
LASSO				
ACC.	0.88	0.89	0.88	0.83
Sense.	0.81	0.83	0.83	0.84
Spec.	0.99	0.99	0.96	0.81
G	0.89	0.90	0.89	0.82
LMI				
ACC.	0.97	0.97	0.93	0.79
Sense.	0.95	0.94	0.93	0.90
Spec.	1.00	1.00	0.92	0.63
G	0.97	0.97	0.92	0.75

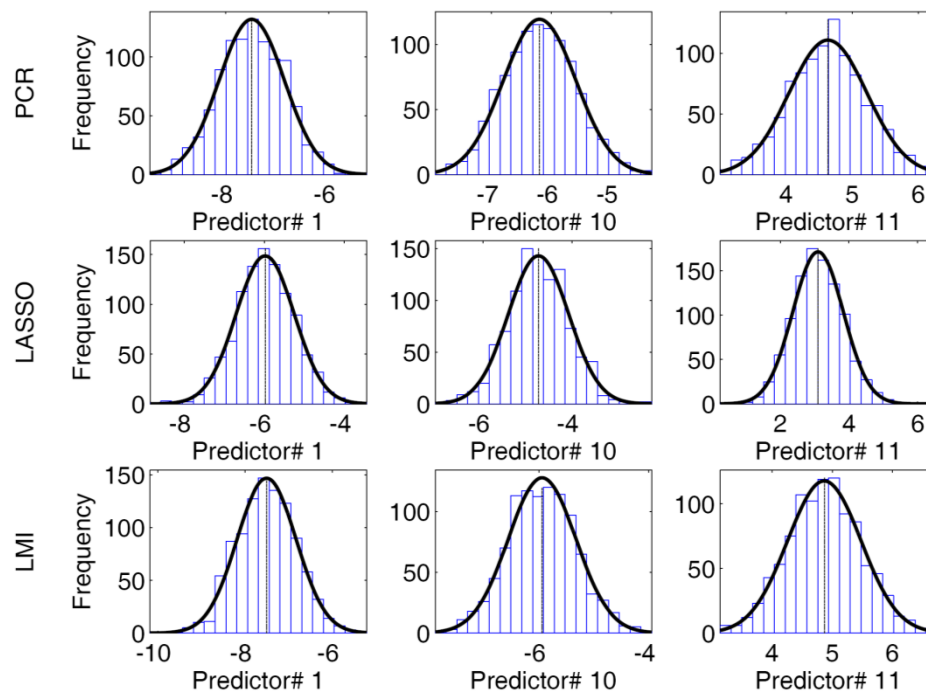
➤ LASSO and LMI perform better than PCR

RESULTS

- **Variability between realizations of data with white noise**

PCR, LASSO, and LMI—are used to identify significant predictors for 1000 input-output pairs.

Histograms of the coefficients in the three significant predictors common to the three methods:



Mean and standard deviation in the histograms of the coefficients computed with PCR, LASSO, and LMI.

Method	Predictor #	1	10	11
	True value	-3.40	5.82	-6.95
PCR	Mean	-3.81	4.73	-6.06
	Std.	0.33	0.32	0.32
	Frac. Err. in mean	0.12	0.19	0.13
LASSO	Mean	-2.82	4.48	-5.62
	Std.	0.34	0.32	0.33
	Frac. Err. in mean	0.17	0.23	0.19
LMI	Mean	-3.70	4.74	-6.34
	Std.	0.34	0.32	0.34
	Frac. Err. in mean	0.09	0.18	0.09

RESULTS

- **Effect of noise distributions of (type of noise in) the output data**
white noise (noise type 1), t -distributed noise (noise type 2), and shifted uniform noise (noise type 3)

Accuracy, Sensitivity, and Specificity of methods for noise with t -distribution (noise-type = 2) and uniform distribution (noise-type = 3) ($m = 100$, $n = 22$).

	Noise type = 2		Noise type = 3	
Noise %	5	20	5	20
PCR				
ACC.	0.74	0.73	0.72	0.72
Sense.	0.73	0.69	0.74	0.68
Spec.	0.75	0.81	0.70	0.78
G	0.73	0.74	0.71	0.73
LASSO				
ACC.	0.89	0.89	0.89	0.89
Sense.	0.82	0.84	0.82	0.85
Spec.	0.99	0.95	1.00	0.95
G	0.90	0.89	0.90	0.90
LMI				
ACC.	0.97	0.92	0.98	0.93
Sense.	0.94	0.94	0.97	0.95
Spec.	1.00	0.89	1.00	0.91
G	0.97	0.91	0.98	0.93



Accuracy of LASSO does not change with the noise level, but its counterparts for LMI and LASSO decrease

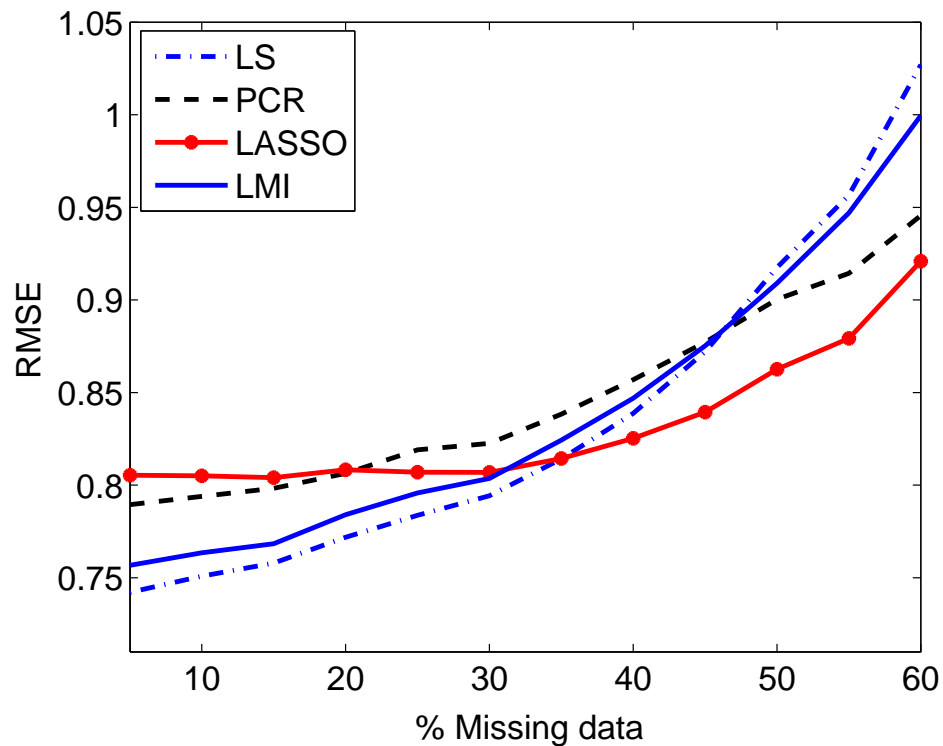
RESULTS

- Effect of missing data:
 - The outputs from the both real and synthetic data (with 20% noise contamination) is chosen. 0-60% data, in increments of 5%, was assumed to be missing. The remaining data was used for training and RMSE was computed on the test (missing) data.
 - This was repeated 100 times by choosing the selected fraction of data randomly, and average RMSE was computed.



RESULTS

- Effect of missing data:

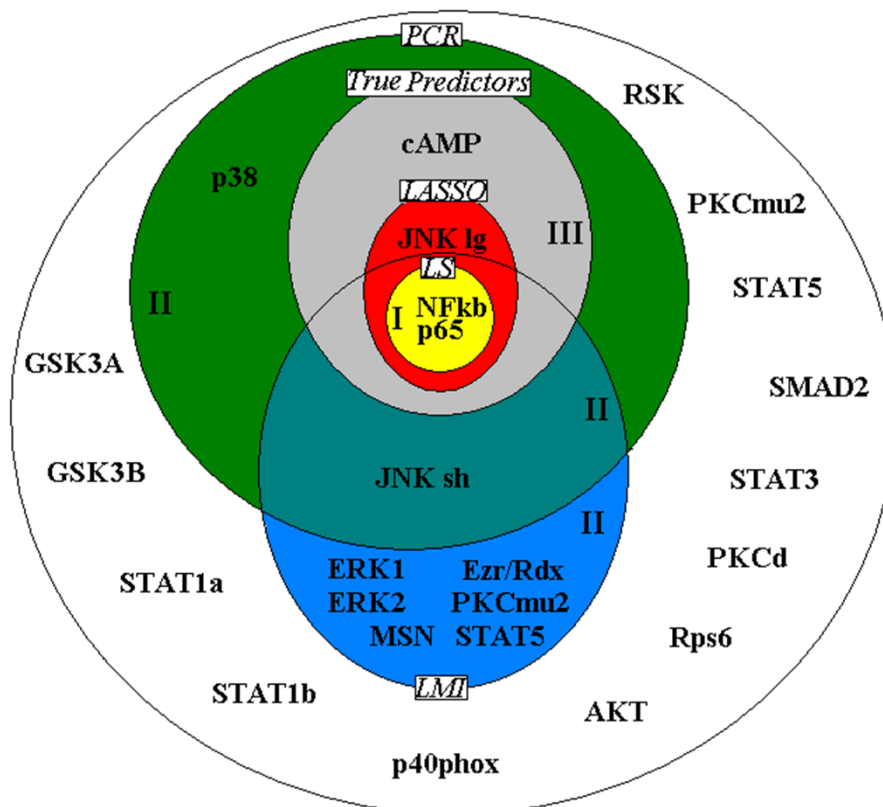


Method	Fractional standard deviation	Fractional max deviation
LS	0.12	0.40
PCR	0.07	0.20
LASSO	0.05	0.14
LMI	0.10	0.32

RMSE versus percentage of missing data for different methods on PP/Cytokine data

SUMMARY

- Comparison of outcome of different methods on the real data
- Different methods identified unique sets of common and distinct predictors for each output



- Only the PCR method detects the true input cAMP
- zone I provides validation and it highlights the common output of all the methods

Graphical illustration of methods PCR, LASSO, and LMI in detection of significant predictors for output IL-6 in PP/cytokine experimental dataset

SUMMARY

- Accuracy, G , and RMSE for different noise levels:
 - With increasing noise level, accuracy and G of LMI and LASSO decreases more rapidly with increasing noise level than accuracy and G of PCR does
 - With respect to RMSE, LMI gets the best score (0.94) and LASSO gets the worst score (0.56).
(Scores are computed as $\text{RMSE}_{\text{method}} / \text{RMSE}_{\text{LS}}$)
- Comparison with respect to the distribution of estimated coefficients:
 - LMI method performs best.
 - The numeric values of the coefficients obtained from LMI method are least affected due to the presence of noise as compared to other methods.

CONCLUSION

- Compared four methods for reconstruction of networks (LS, PCR, LASSO, and LMI) on two different data-sets.
- The least-squares method has lowest RMSE.
- Other three methods better in capturing most of the true inputs.
- PCR better for the synthetic data with increasing noise.
- PCR most robust for both the real data and the synthetic data with medium level of noise.
- LMI method performs best according to comparison with respect to the distribution of estimated coefficients
- LASSO is the most robust method in terms of $RMSE_{val}$ when a portion of the dataset is missing/unavailable.



ACKNOWLEDGEMENTS

- National Heart, Lung and Blood Institute (NHLBI) grant 5 R33 HL087375-02 (SS)
- National Science Foundation (NSF) grant DBI-0641037 (SS)
- NSF collaborative grant DBI-0835541 (SS)

Thank you for your attention

Questions?

