Imputing Missing Data in Sensor Networks via Markov Random Fields

UCIrvine University of California

Nicholas Navaroli, Scott Triglia, Padhraic Smyth

University of California, Irvine

Scalable Methods for the Analysis of Network-Based Data

This research was supported by grant MURI ONR 446667-23055

Main Ideas

- Efficiently impute missing data in network
  - Markov blankets
  - Conditionally independent subsets
- Allow unknown model parameters

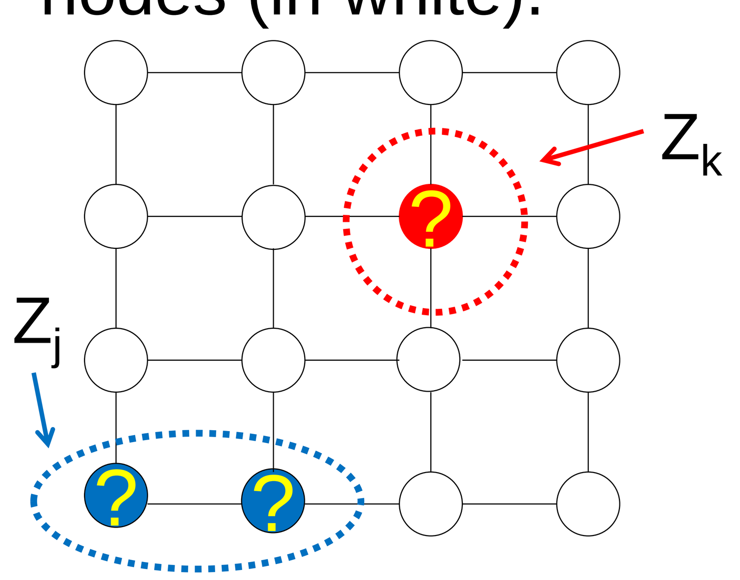
Conditional Independence in Markov Random Fields (MRFs)

Basic MRF:

- Nodes correspond to measurements
- Edges connect directly related nodes

Conditional Independence (C.I.):

- Denote set of unobserved nodes as Z
- Denote set of observed nodes as X (white)
- Missing sets Z_j and Z_k are C.I. if all paths between the two are "cut off" by observed nodes (in white).



Example MRF with two C.I. sets (Z_j and Z_k)

Markov Blankets

- Set of observed nodes "surrounding" missing set Z_j is the *Markov blanket*, denoted B_j.

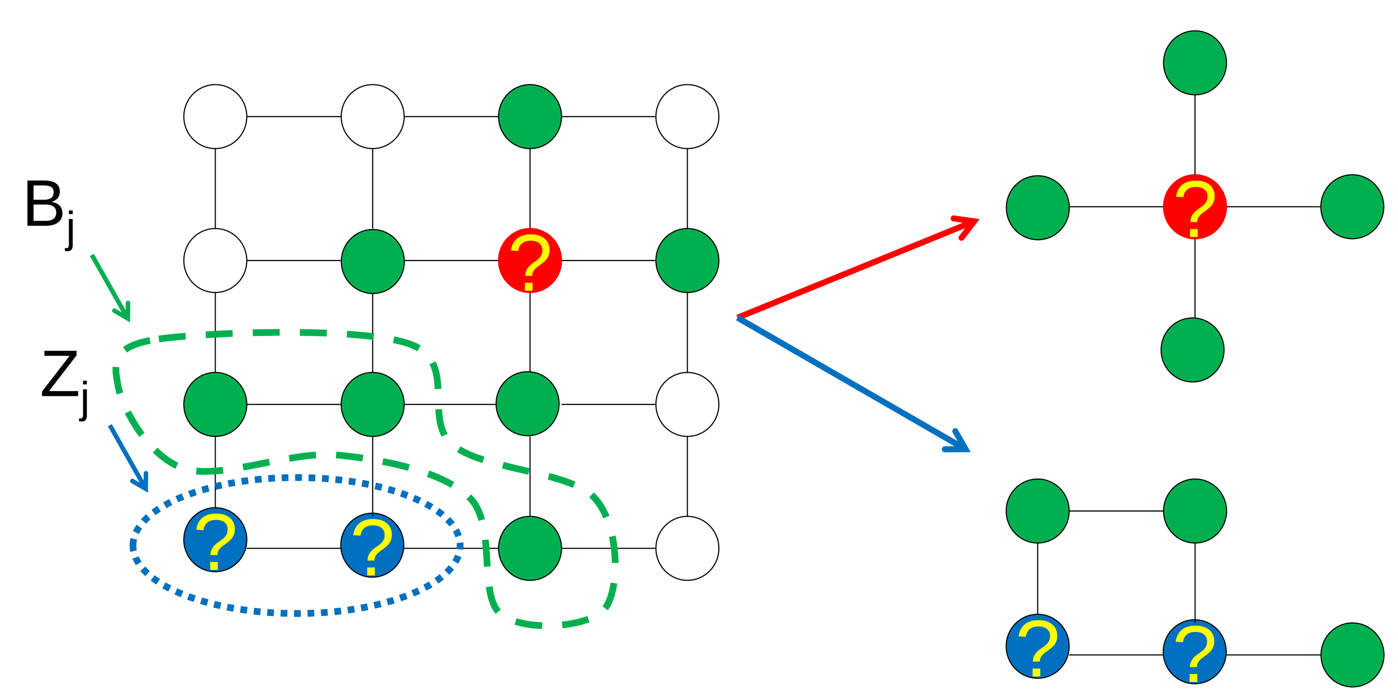
$$P(Z_j|X,\theta) = P(Z_j|B_j,\theta)$$

- Only conditioning on a set's Markov blanket can reduce computational complexity

Conditionally Independent Subsets

- Decomposition into conditionally independent sub-problems
- Can be combined with Markov blankets as shown below

left: *Markov Blanket of Z_j*
right: *Decomposition into C.I. sub-problems*



Computational Complexity

This gives three methods to calculate P(Z|X,ϴ). The bound on each method's computational complexity is below, along with the approximate number of operations required to find P(Z|X,ϴ) in the above graph.

1. Assume Gaussian Data

Method	Complexity	Small Example
Naive	$O(\|X\|^3)$	2197
MB	$O(\|B\|^3)$	343
MB/CI	$O(\max_i \|B_i\|^3)$	64

2. Assume Discrete Data (binary case, k=2)

Method	Complexity	Small Example
Naive	$O(k^{\|X\|})$	8192
MB	$O(k^{\|B\|})$	128
MB/CI	$O(\max_i k^{\|B_i\|})$	16

Parameter Estimation via EM

- Simultaneously estimate parameters and impute missing data using Expectation Maximization (EM) algorithm:

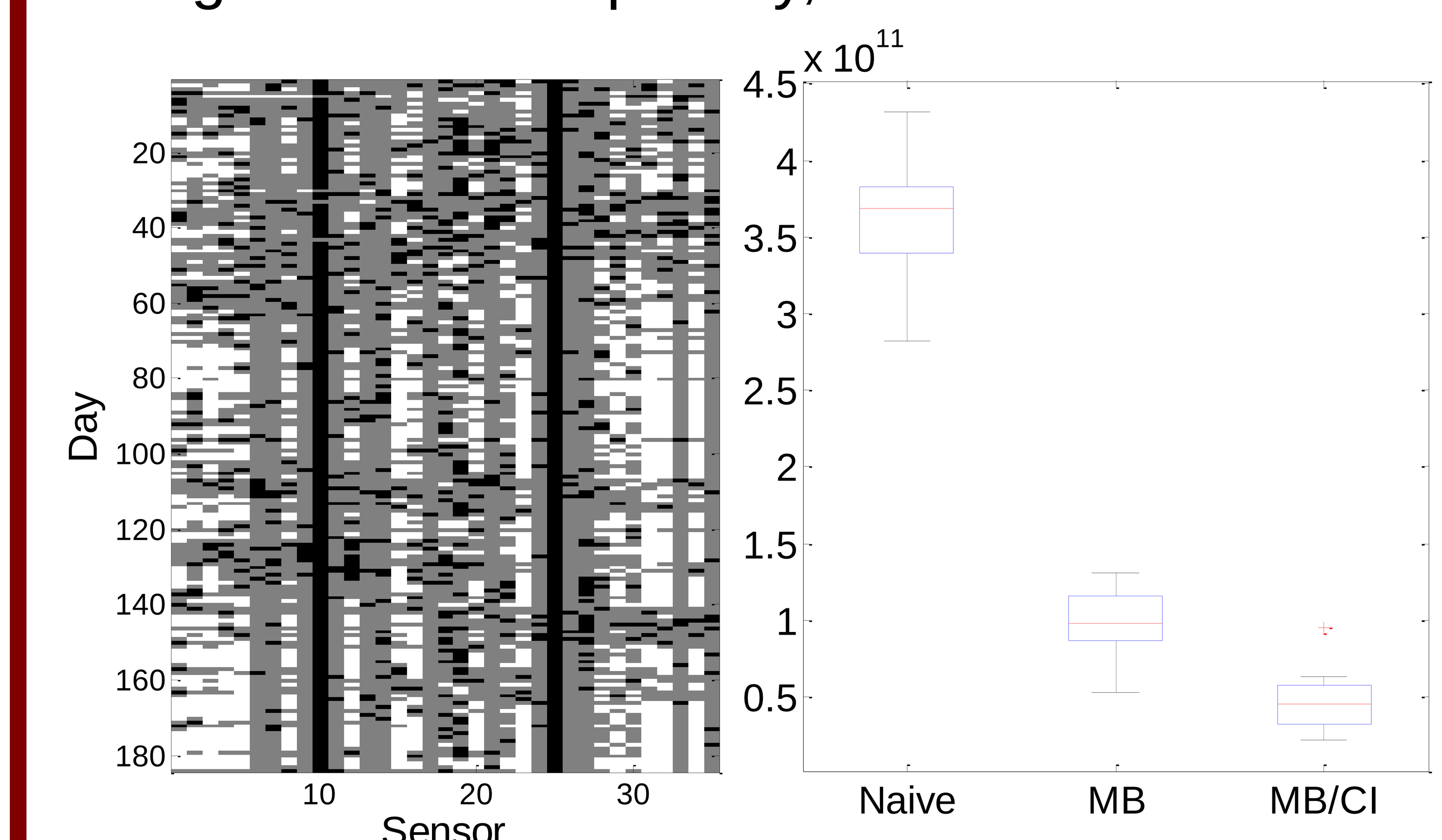
E Step Z^{t+1} = Most likely Z given X and Θ^t

M Step Θ^{t+1} = Most likely ϴ given X and Z^{t+1}

- E-step is equivalent to previously discussed methods for imputing missing data

Example: Rainfall Measurements

- 35 rainfall sensors across northern India
- MRF has a node for each node/day combination in a 122 day season.
- Edges are KNN spatially, Markov in time.



Left: *Missing data pattern (X=white,Z=black,B=gray)*
Right: *Complexity of imputation methods*

Current Work

- Efficiently balance multiple C.I. sub-problems during imputation
- Learning the edge structure of the MRF