

# LassoOrderSearch:

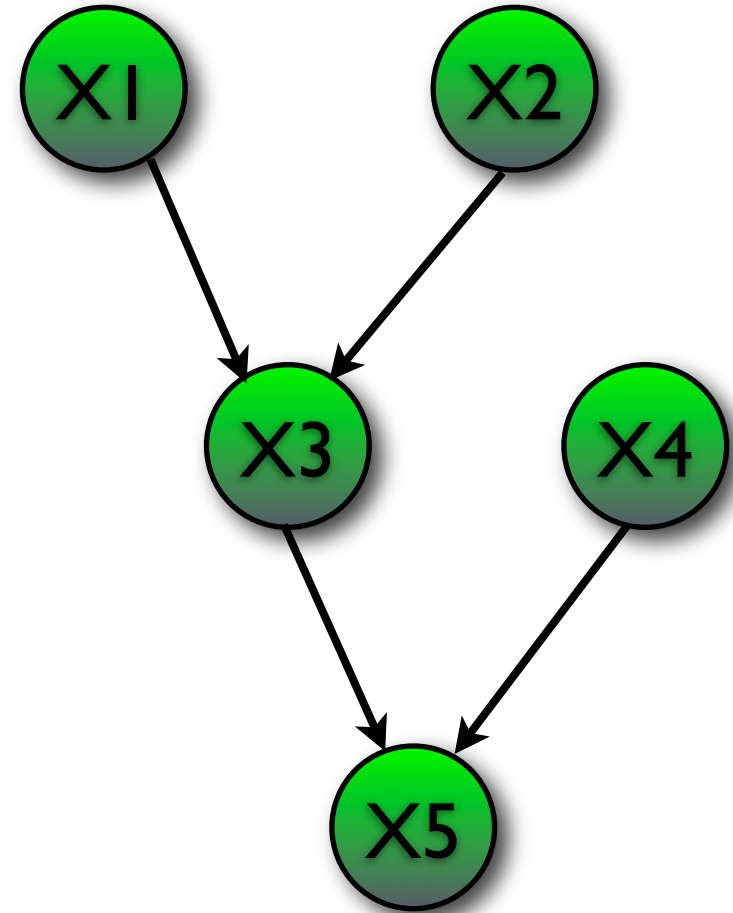
Learning Directed Graphical Model Structure  
using L1-Penalized Regression and Order Search

Mark Schmidt and Kevin Murphy  
University of British Columbia



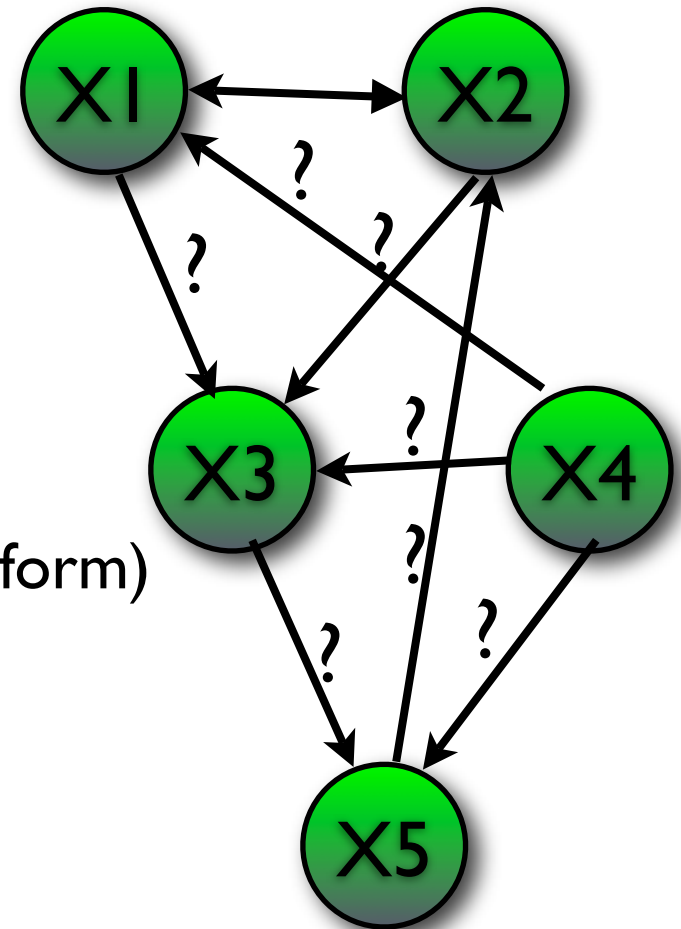
# Outline

- Bayes Net Structure Learning
- Order-Search
- LassoOrderSearch
- Experimental Results
- Interventional Data
- Parent Pruning



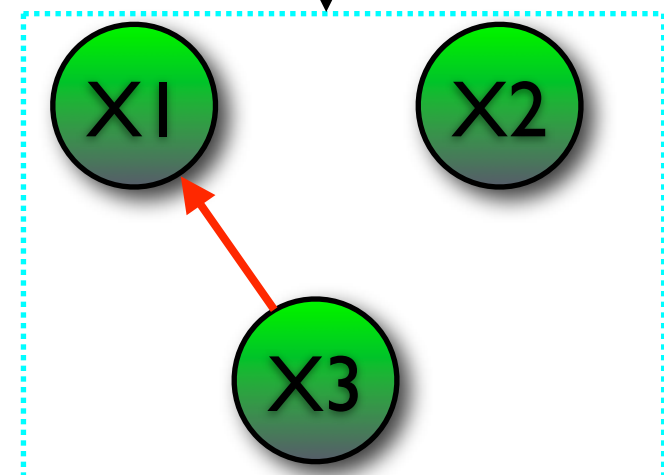
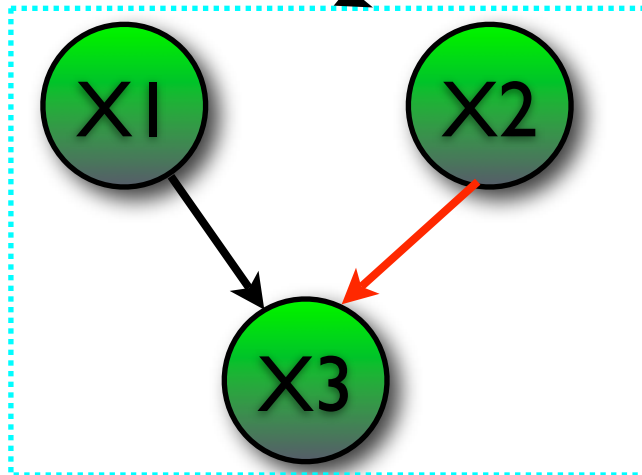
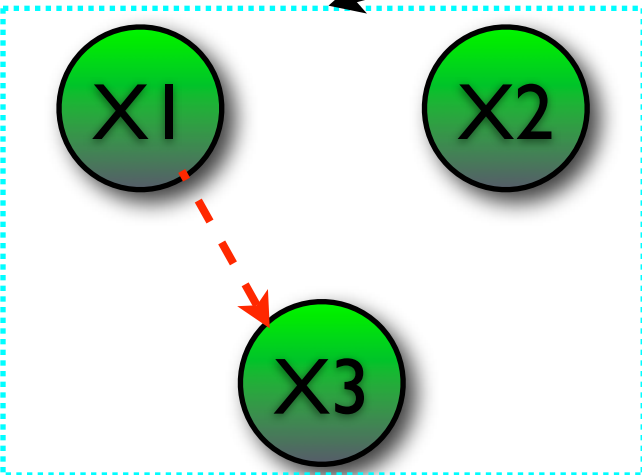
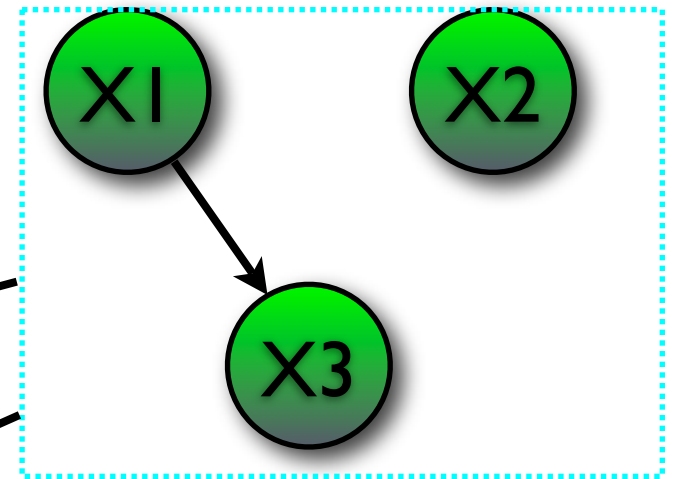
# Bayes Net Structure Learning

- Given Data:
  - Find 'best' Bayes Net structure
  - NP-Hard
- Scoring Metrics:
  - Bayesian Score
  - BIC (integral can not be done in closed form)
- Strategies:
  - Greedy Search
  - Constraint-Based
  - Convex Relaxation



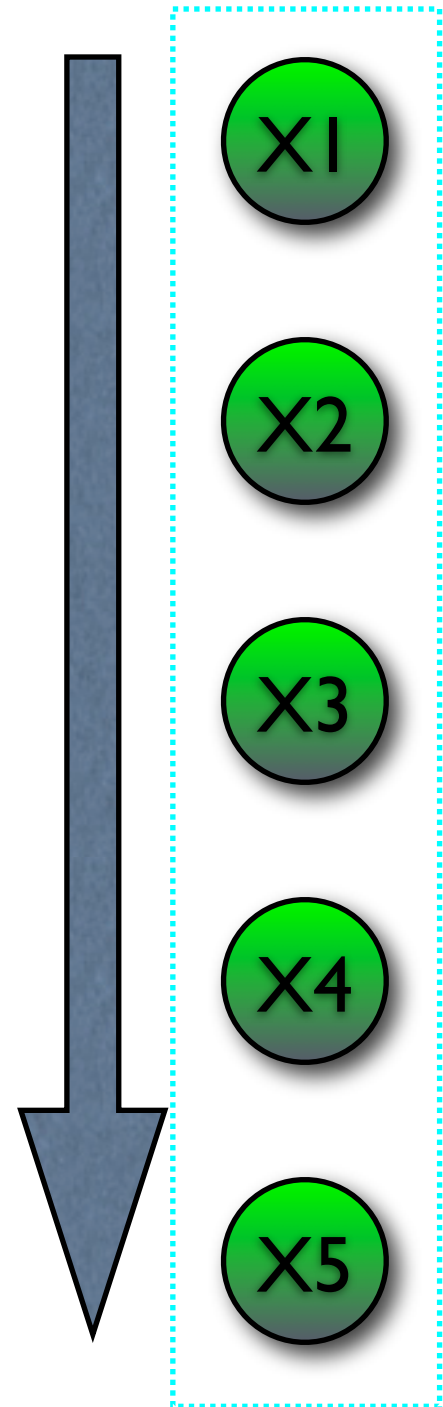
# DAG-Search

- Takes advantage of 'decomposable' network score
- Greedily add/delete/reverse 1 edge, repeat. (naive, but hard to beat in practice)
- State of the Art:
  - Max-Min Hill-Climbing (MMHC):
  - Prune potential edges, then run DAG-Search



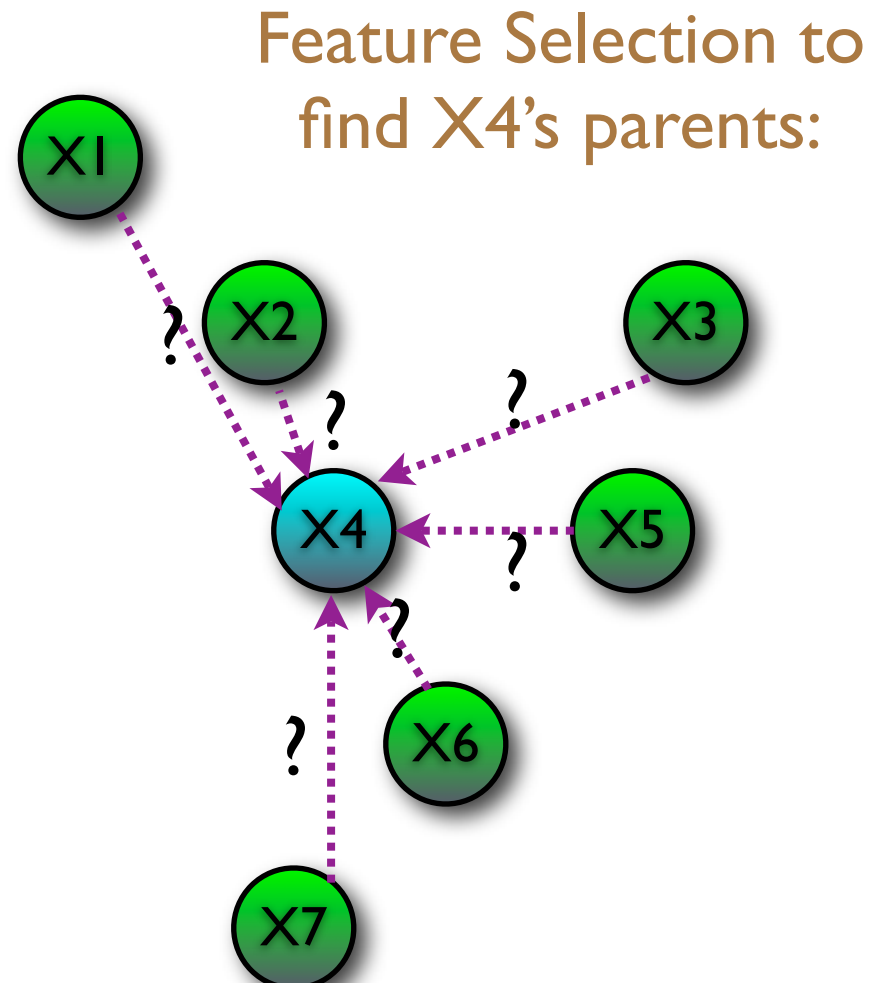
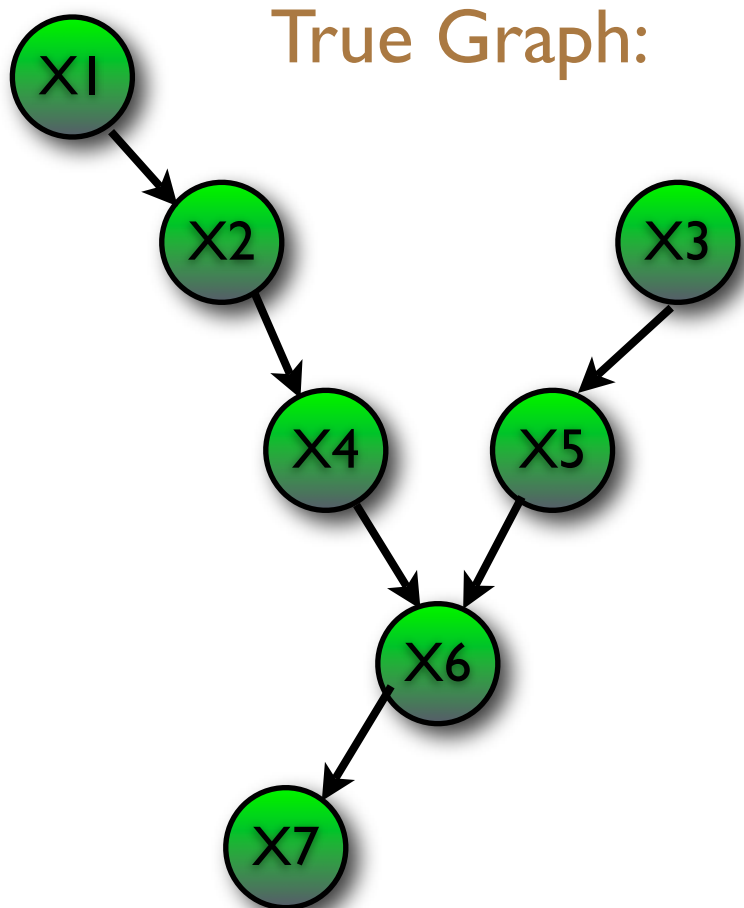
# Outline

- Bayes Net Structure Learning
- Order-Search
- LassoOrderSearch
- Experimental Results
- Interventional Data
- Parent Pruning

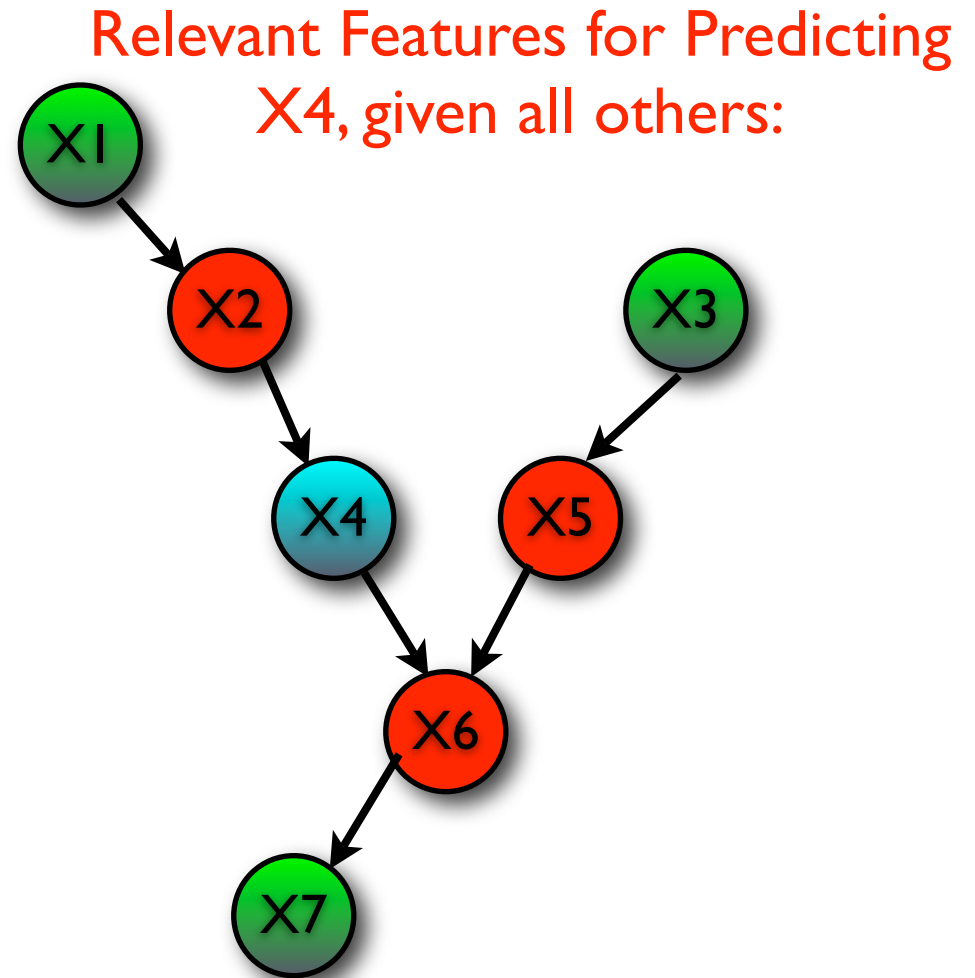
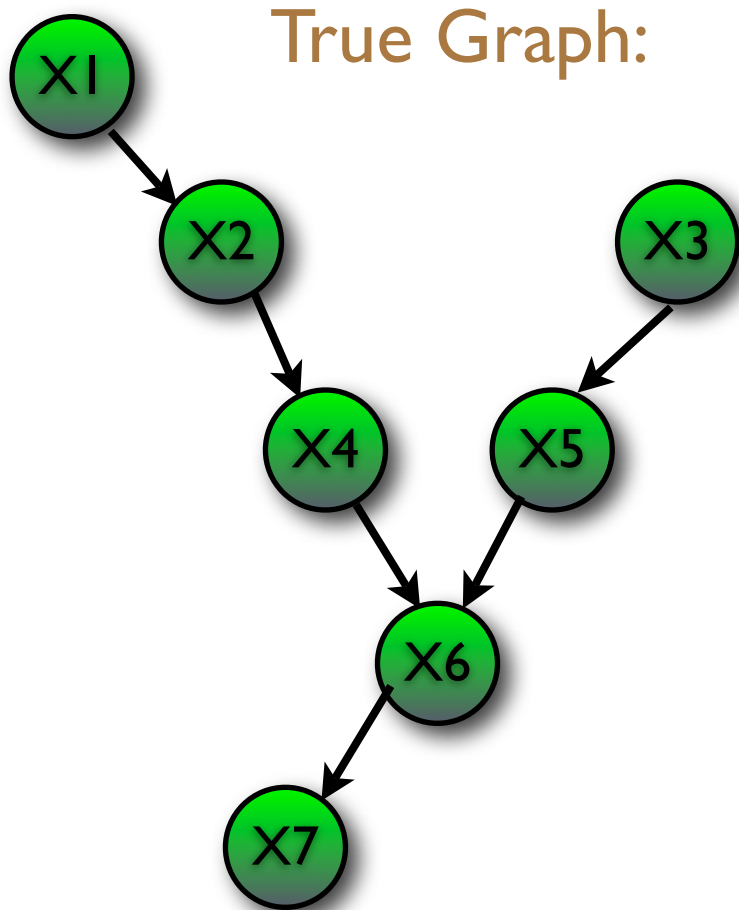


# Feature Selection?

- Is finding a node's parents just feature selection?



# Feature Selection?

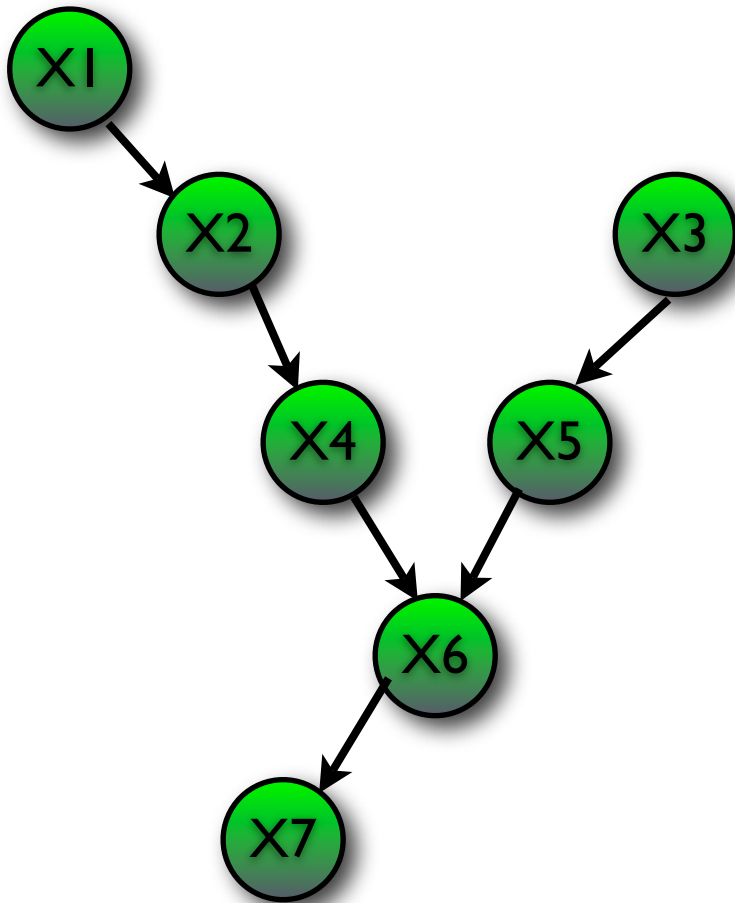


- can't distinguish between parents and children
- explaining away makes coparents relevant

But, this will work if we use a 'Topological Ordering'

# Topological Ordering

- Order on nodes, such that:  
Parents are before Children



Topological Orderings:

$\{1,2,3,4,5,6,7\}$

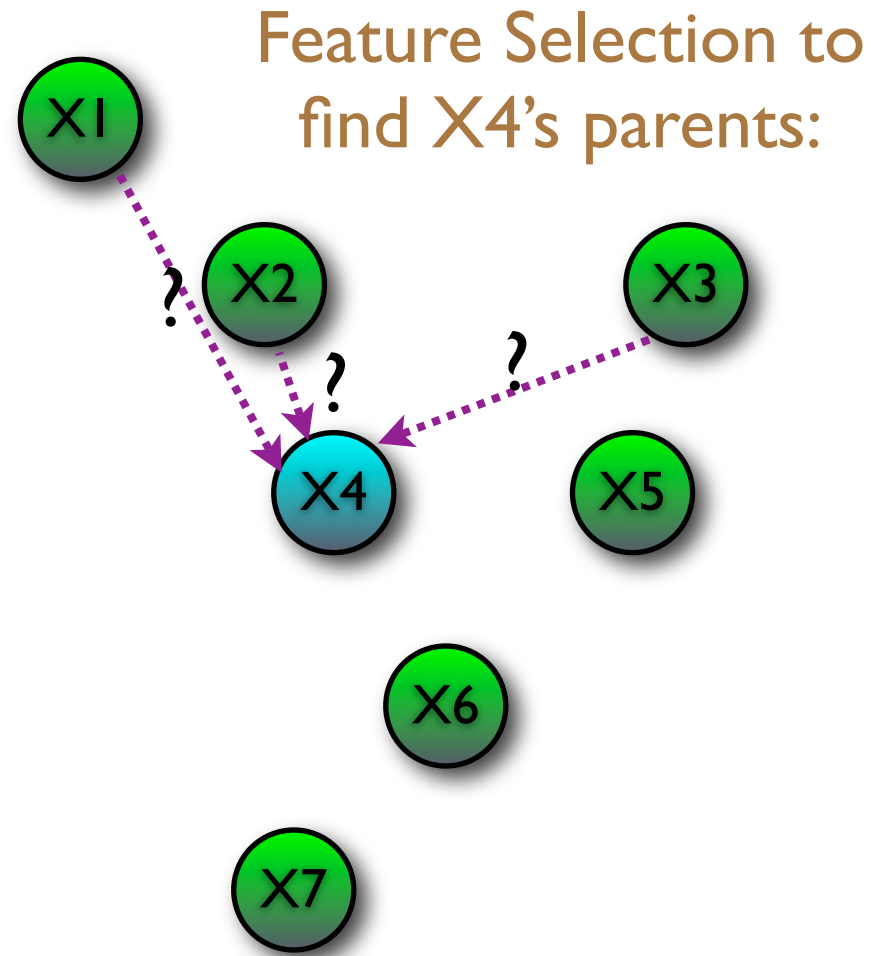
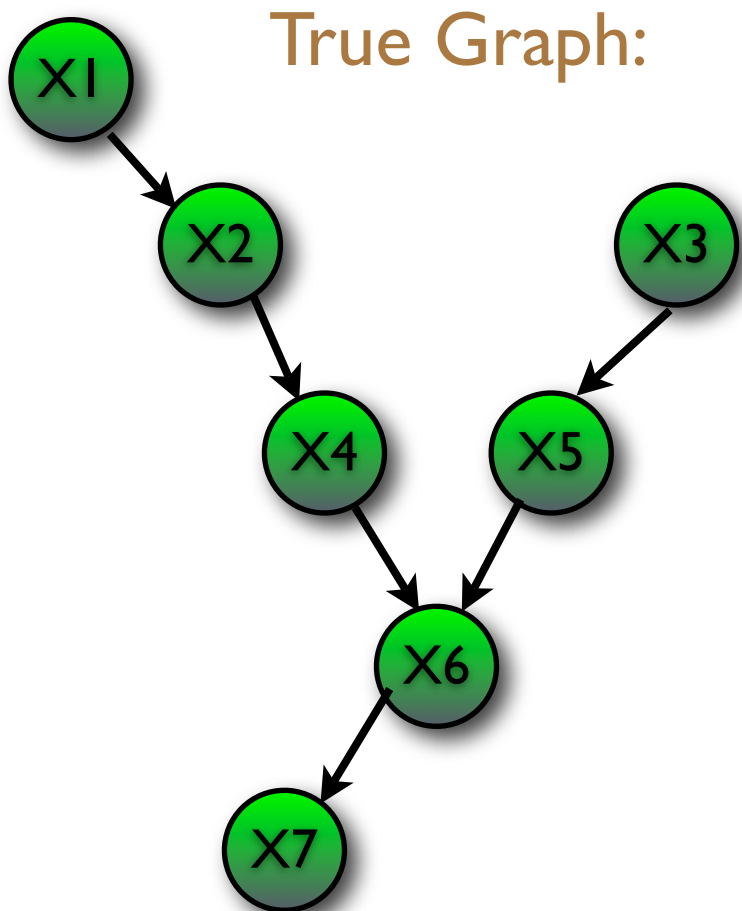
$\{3,5,1,2,4,6,7\}$

$\{1,3,2,4,5,6,7\}$

etc.

# Feature Selection with Topological Ordering

- Given Topological Ordering:
  - children removed as potential parents
  - coparents no longer explained away by parents
  - resulting graph is acyclic
- Independent feature selection problems



# Order-Based Search

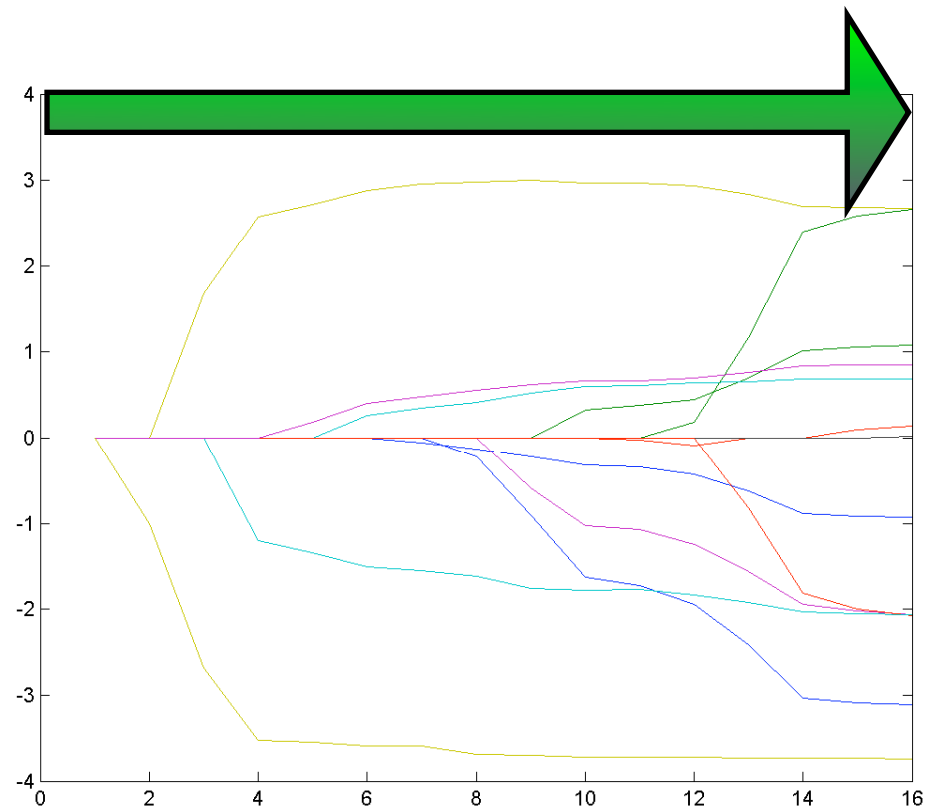
- [Teyssier and Koller, 2005]:
  - Search through space of Topological Orderings (smaller than space of DAGs)
  - Move between orderings by swapping adjacent elements:
    - $p-1$  candidate moves at each step, but only 4 new problems need to be solved after each move
  - Search over all  $2^k$  potential parents of node  $k$  in ordering
- Avoiding Exponential Work:
  - Restriction on Number of Parents
  - Restriction on Candidate Set of Parents

# Order-Based Search

- Problems:
  - Tabular Potentials have an exponential number of parameters
  - Can't model graphs with high fan-in
  - Sparse Candidate may remove true parents

# Outline

- Bayes Net Structure Learning
- Order-Search
- LassoOrderSearch
- Experimental Results
- Interventional Data
- Parent Pruning



# Parsimonious CPDs

- We use linearly parameterized CPDs rather than Tabular CPDs, requiring only a linear number of parameters:

- Gaussian:  
Continuous Data (defines joint Gaussian)

$$P(y|x) = N(y|\mu + w^T x, \sigma)$$

- Sigmoid:  
Binary Data (integral not analytic)

$$P(y = 1|x) = \frac{1}{1 + \exp(-w^T x)}$$

- Softmax:  
Multinomial Data (integral not analytic)

$$P(y = k|x) = \frac{w^T x}{\sum_{k^*} \exp(-w_{k^*}^T x)}$$

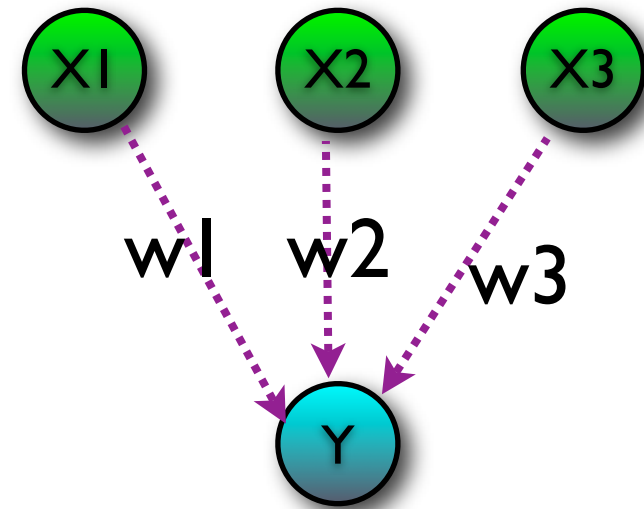
- Non-linear interactions can be modeled by replace  $x$  with a non-linear basis  $f(x)$

# L1-Penalization for Structure Learning

- Learn one or more directed learning model, then convert to undirected GGM:
  - *[Huang et al., Biometrika 2006]: Covariance selection and estimation via penalized normal likelihood.*
  - *[Li & Yang, AAAI 2005]: Using modified lasso regression to learn large undirected graphical models.*
  - *[Meinshausen & Buhlman, Annals of Statistics 2006]: High dimensional graphs and variable selection with the lasso.*
- Sample orderings, then convert to undirected GGM:
  - *[Dobra et al., J. Multivariate Analysis 2004]: Sparse graphical models for exploring gene expression data.*
- L1-Penalty directly on GGM covariance matrix:
  - *[Dahl et al., UCLA TR 2005]: Maximum likelihood estimation of gaussian graphical models*
  - *[Yuan and Lin, GT TR 2005]: Model selection and estimation in the gaussian graphical model.*
  - *[Banerjee et al., ICML 2006]: Convex optimization techniques for fitting sparse gaussian graphical models*
- Discrete version of [Meinshausen & Buhlman] undirected GGM consistency proof:
  - *[Wainwright et al., NIPS 2006]: Inferring graphical model structure using L1-regularized pseudo-likelihood*
- L1-Penalty on Boltzmann Machine parameters (loopy BP used to approximate Z)
  - *[Lee et al., NIPS 2006]: Efficient structure learning of Markov networks using L1-regularization*

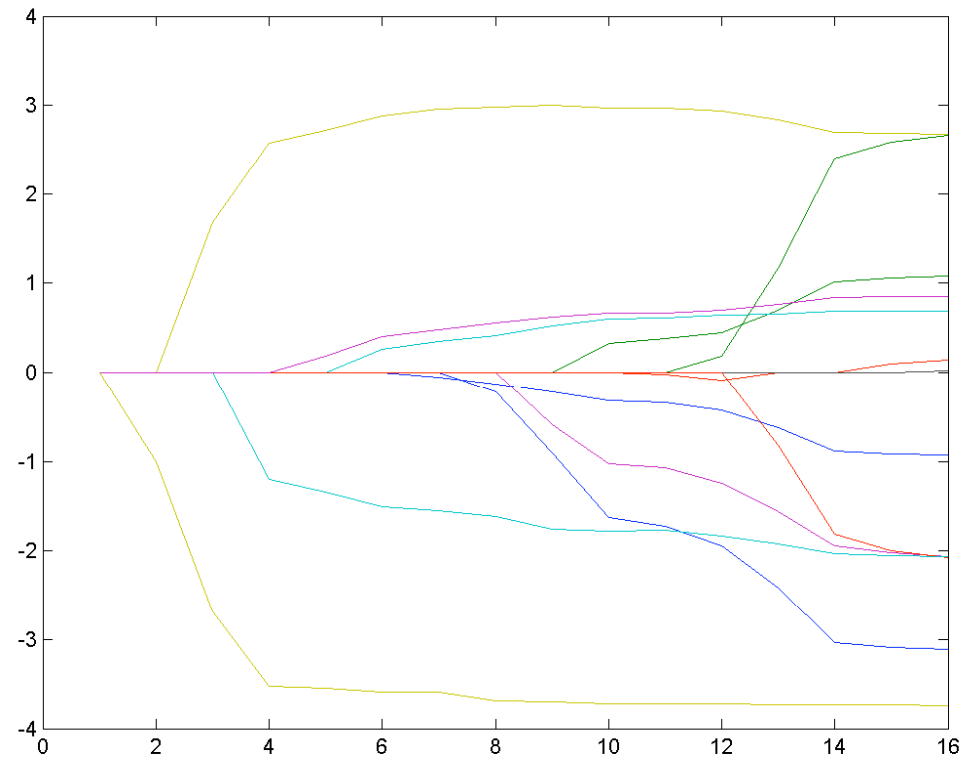
# L1-Penalization

- Exhaustively searching over all possible subsets leads to an exponential time dependency
- We can alternately regularize the parent weights with an L1-penalty, to induce a sparse parent set
- Leads to a convex optimization of the form:
$$\min_w \log[p(y|x, w)]$$
$$s.t. |w|_1 \leq t$$
- 't' controls the degree of parameter 'shrinkage'



# L1-Penalized Optimization

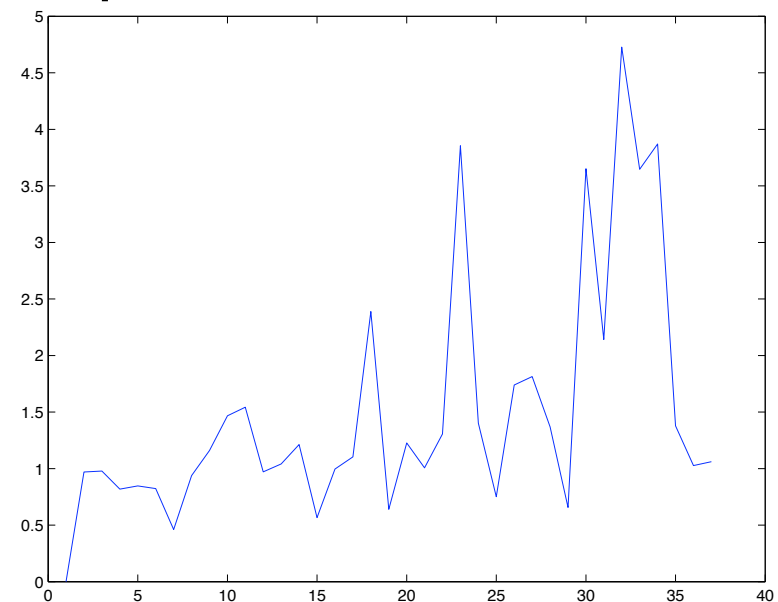
- Gaussian CPDs:
  - Least Angle Regression (LARS) computes the optimal weights for all values of lambda: the 'regularization path'
  - The runtime of LARS is  $O(p^3)$ , the same asymptotic cost as solving a Least Squares problem
- Other CPDs:
  - LARS can be used to solve the sub-problems arising from an Iteratively-Reweighted Least Squares (IRLS) formulation  
[Lee et al., 2006]



# Hyper-parameter Estimation

- A constant value of 't' introduces an ordering-dependent bias
- Therefore, we need a different hyper-parameter at each node
- This is needed for consistency proof of L1-penalized neighborhood selection [Meinshausen & Bühlman, 2006]
- But how should you pick t?

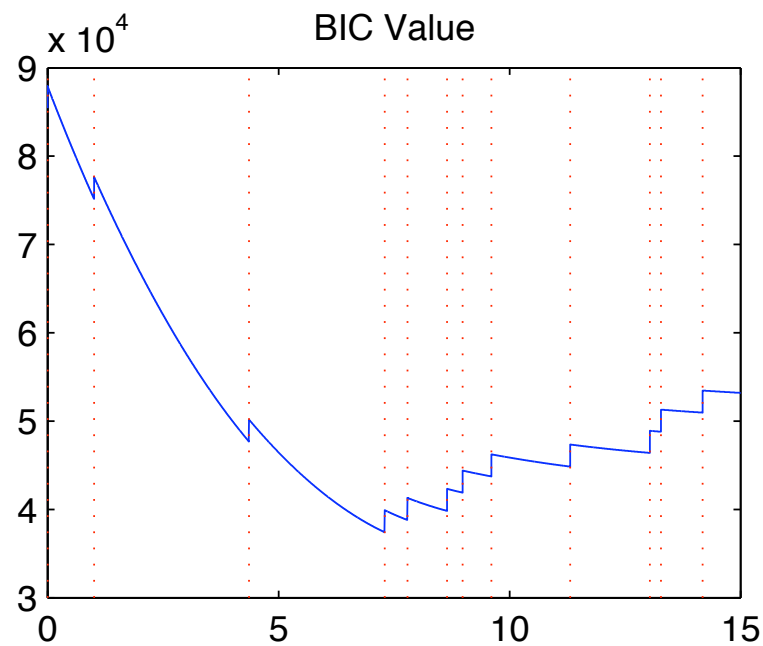
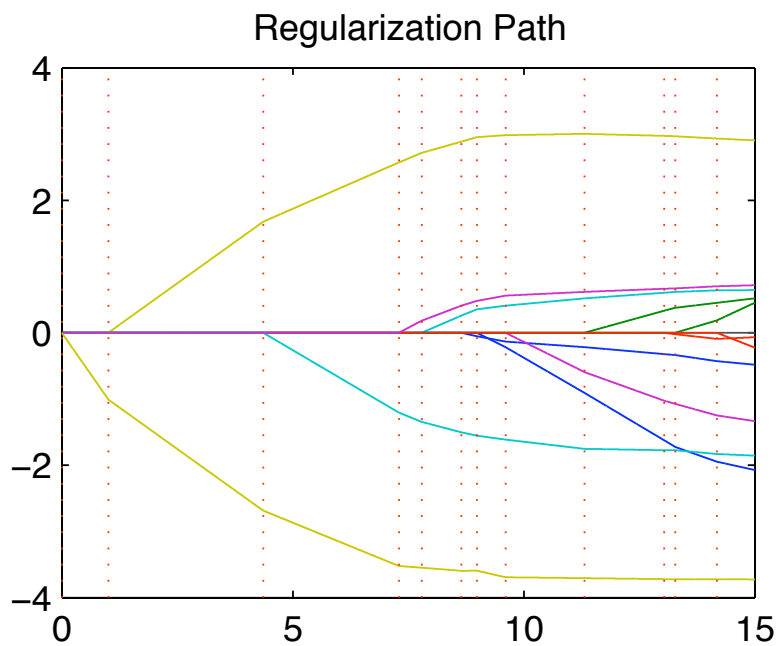
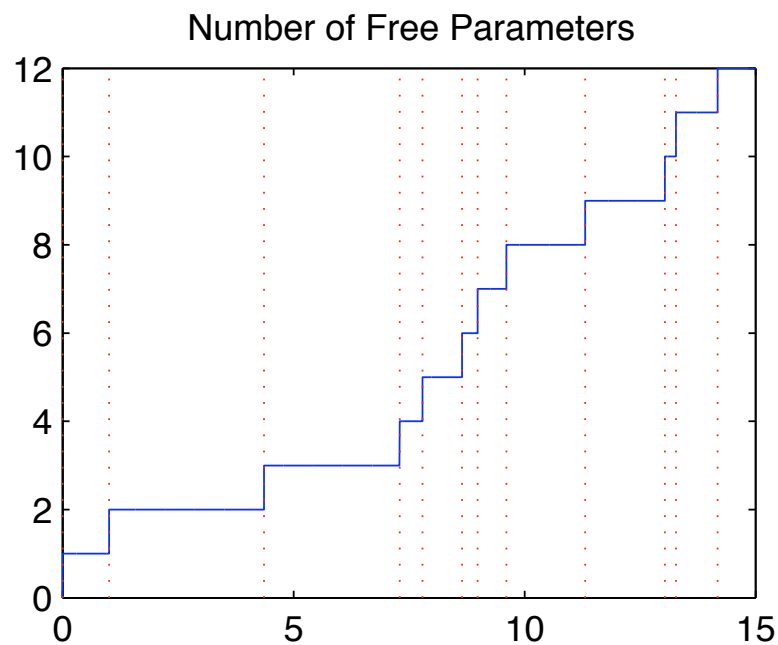
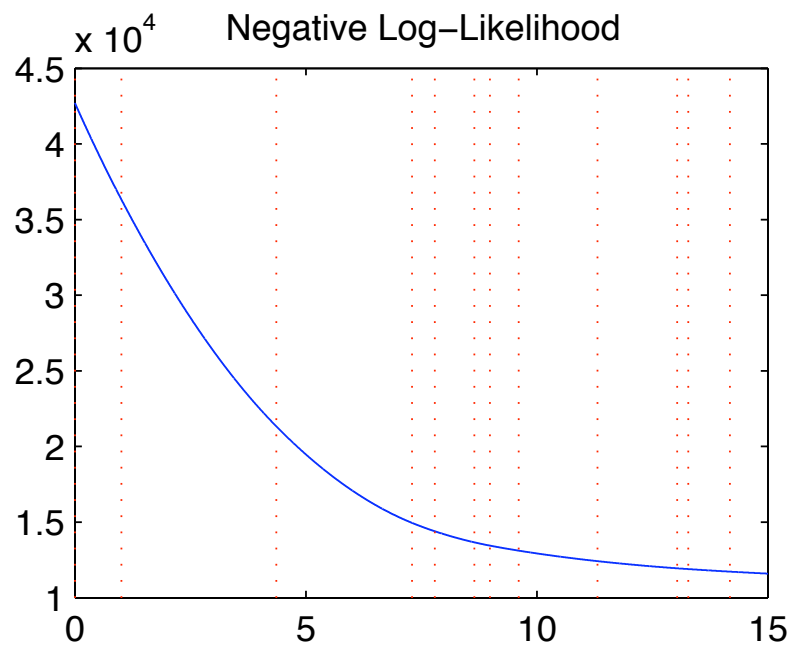
Optimal t vs. Node Order:



# BIC-Optimal Hyperparameter

- Choosing 't' to optimize BIC:
  - The data log-likelihood monotonically increases with 't'
  - The number of free-parameters is piecewise constant as 't' varies
- Therefore, the BIC-optimal 't' must lie at a discontinuity in the regularization path

# BIC-Optimal Hyperparameter



# LARS-MLE

- The LARS algorithm computes ALL the MAP 'w' values at the discontinuities in the regularization path in  $O(p^3)$
- To compare orderings, we typically use the MLE parameters
- Key to the efficiency of LARS is the  $O(p^2)$  updating/downdating of a Cholesky factorization of the Hessian  $X'X$  of the Active Set
- We modify the LARS algorithm, such that this factorization serves the dual purpose of computing the MLE of all subsets encountered along the regularization path

# LassoOrder

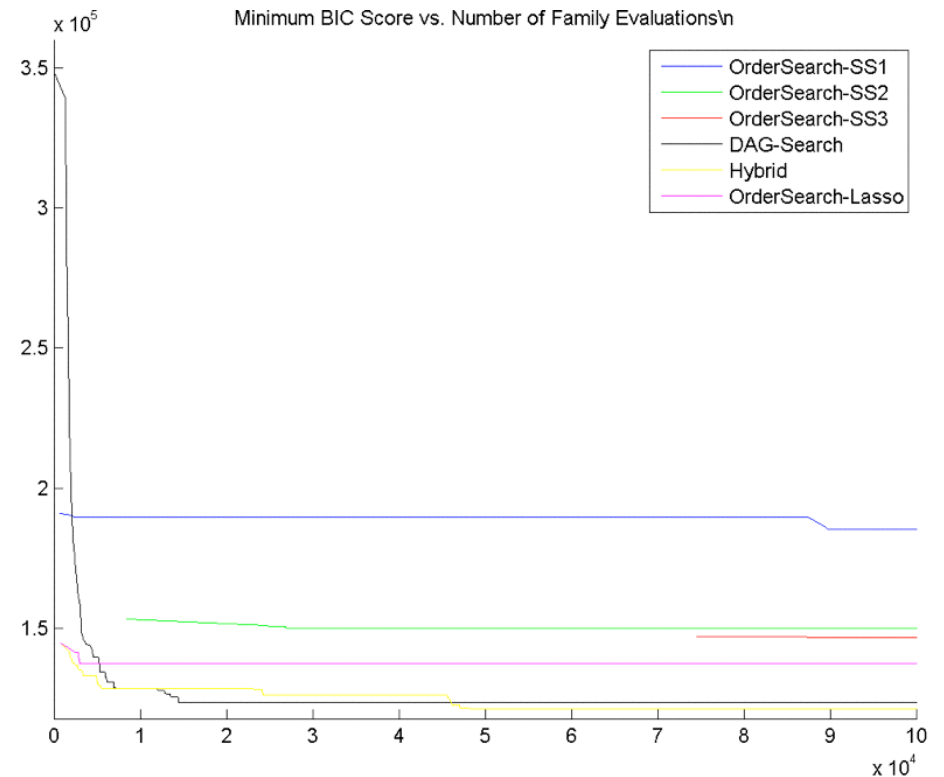
- LassoOrder for the Gaussian Case with  $p$  variables:
  - for node  $k$  in the ordering:
    - run LARS-MLE on variables  $\{1:k-1\}$
    - select the parents with the highest BIC among the  $O(k)$  subsets along the regularization path
- Asymptotic runtime:  $O(p^4)$
- Non-Gaussian (under some assumptions):  $O(p^5)$
- In comparison, exhaustive subset enumeration:  $O(p^4 2^p)$

# LassoOrderSearch

- LassoOrderSearch:
  - Similar as OrderSearch [Teyssier & Koller, 2005], but use parsimonious CPDs and BIC-LI to avoid exponential space and time
- For reasonable sized problems:
  - No need for a restriction on a fan-in
  - No need to use sparse candidate

# Outline

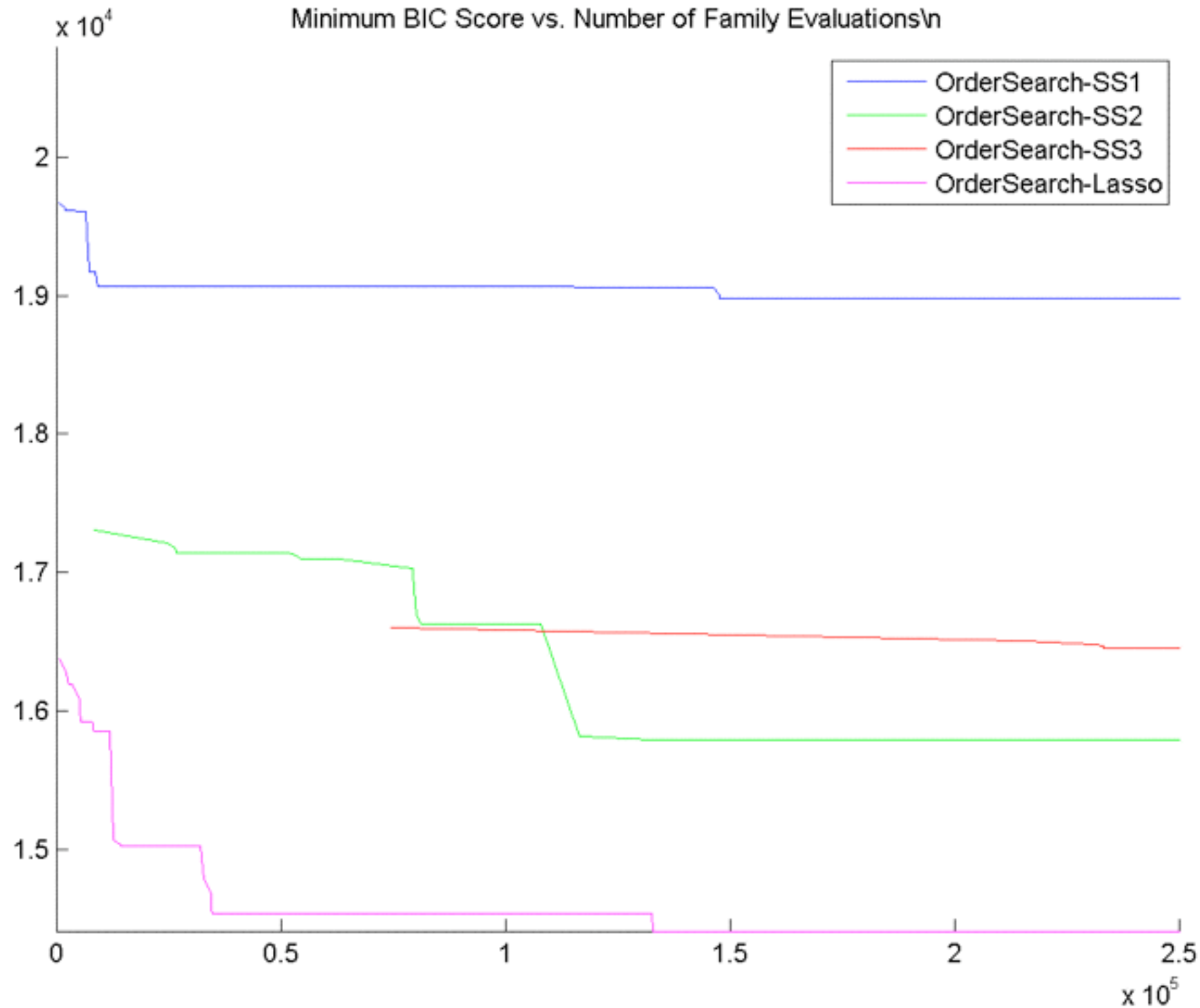
- Bayes Net Structure Learning
- Order-Search
- LassoOrderSearch
- Experimental Results
- Interventional Data
- Parent Pruning



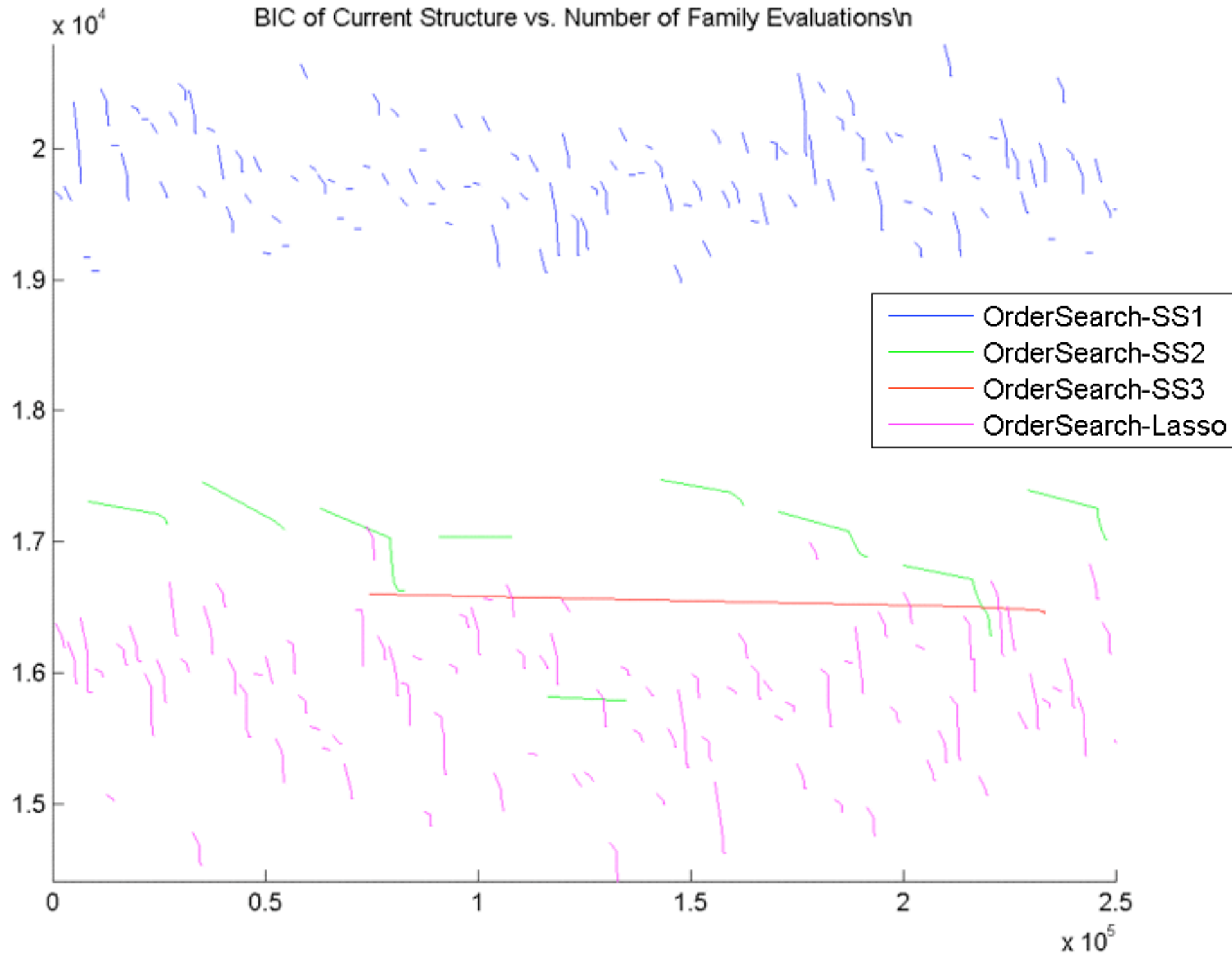
# Known Structure Results

- Generated 10000 samples from Discrete/Continuous versions of the Alarm network with random parameters (bounded away from zero)
- Compared to implementations of DAG-Search, and Order-Search with a fan-in restriction of 1, 2, 3, or 4 (true network has an in-degree of 4).
- All models use multiple restart hill-climbing, the order-based methods receive the same random orderings
- Measured BIC-score compared to number of family evaluations performed by the algorithm

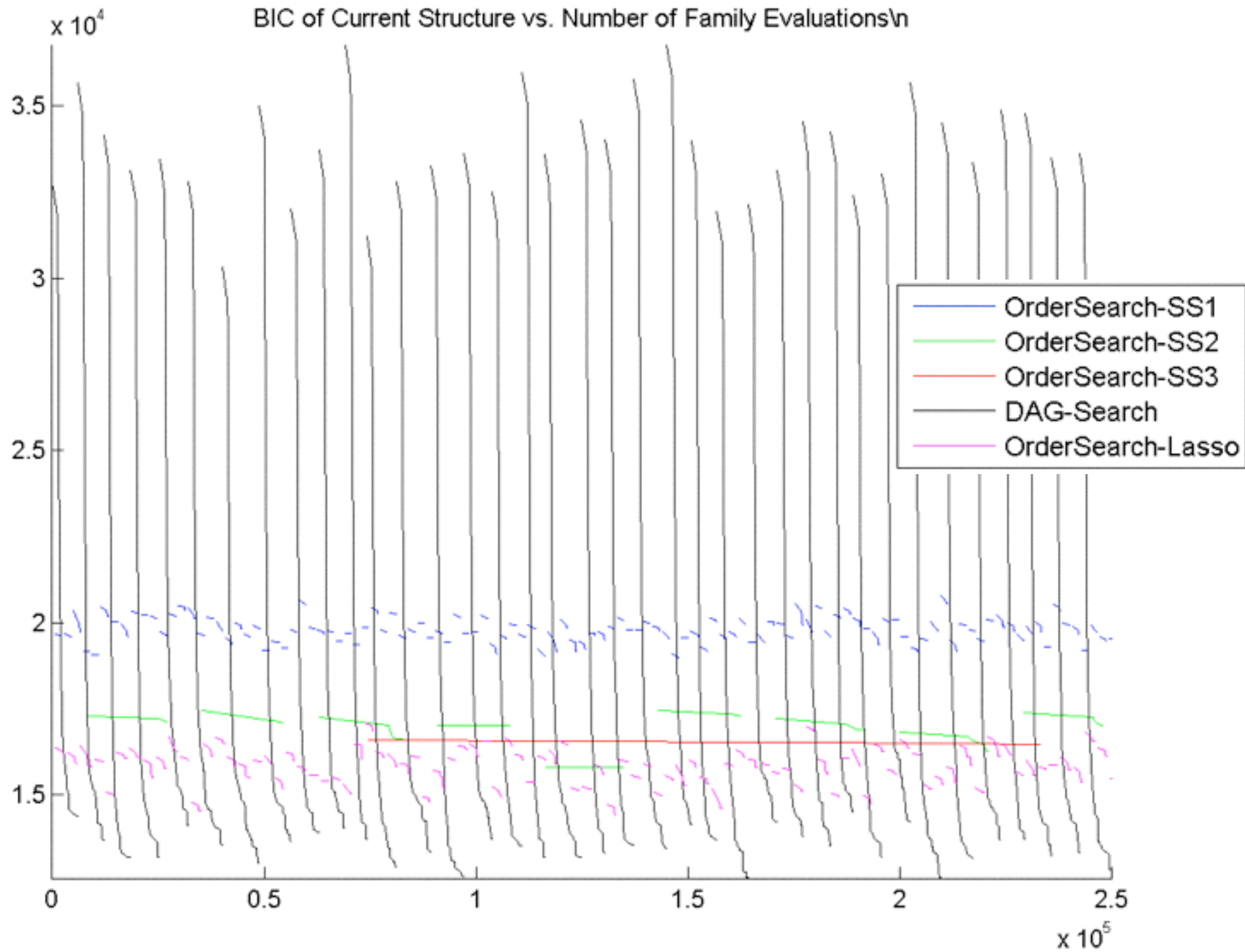
Minimum BIC Score vs. Number of Family Evaluations



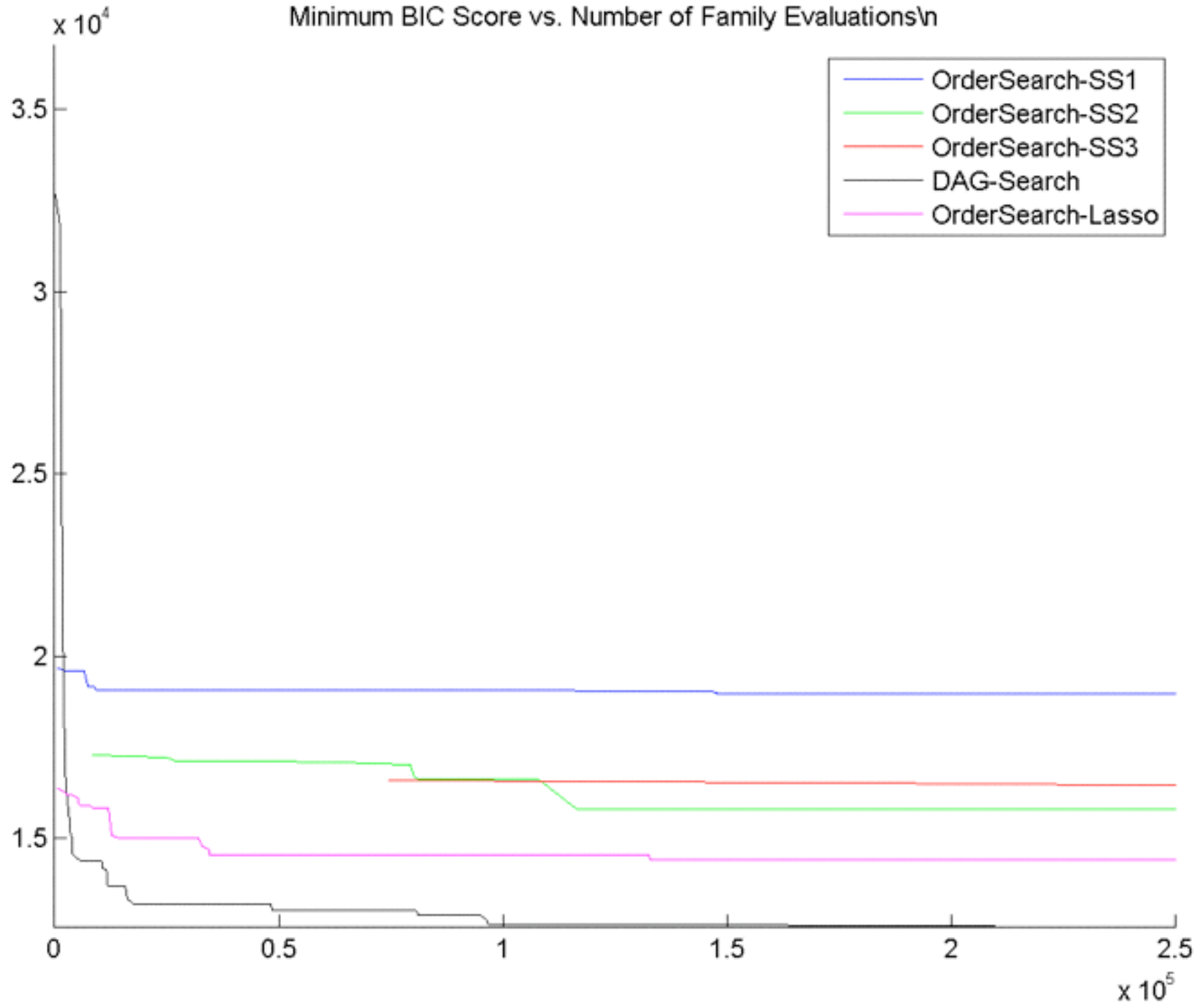
BIC of Current Structure vs. Number of Family Evaluations



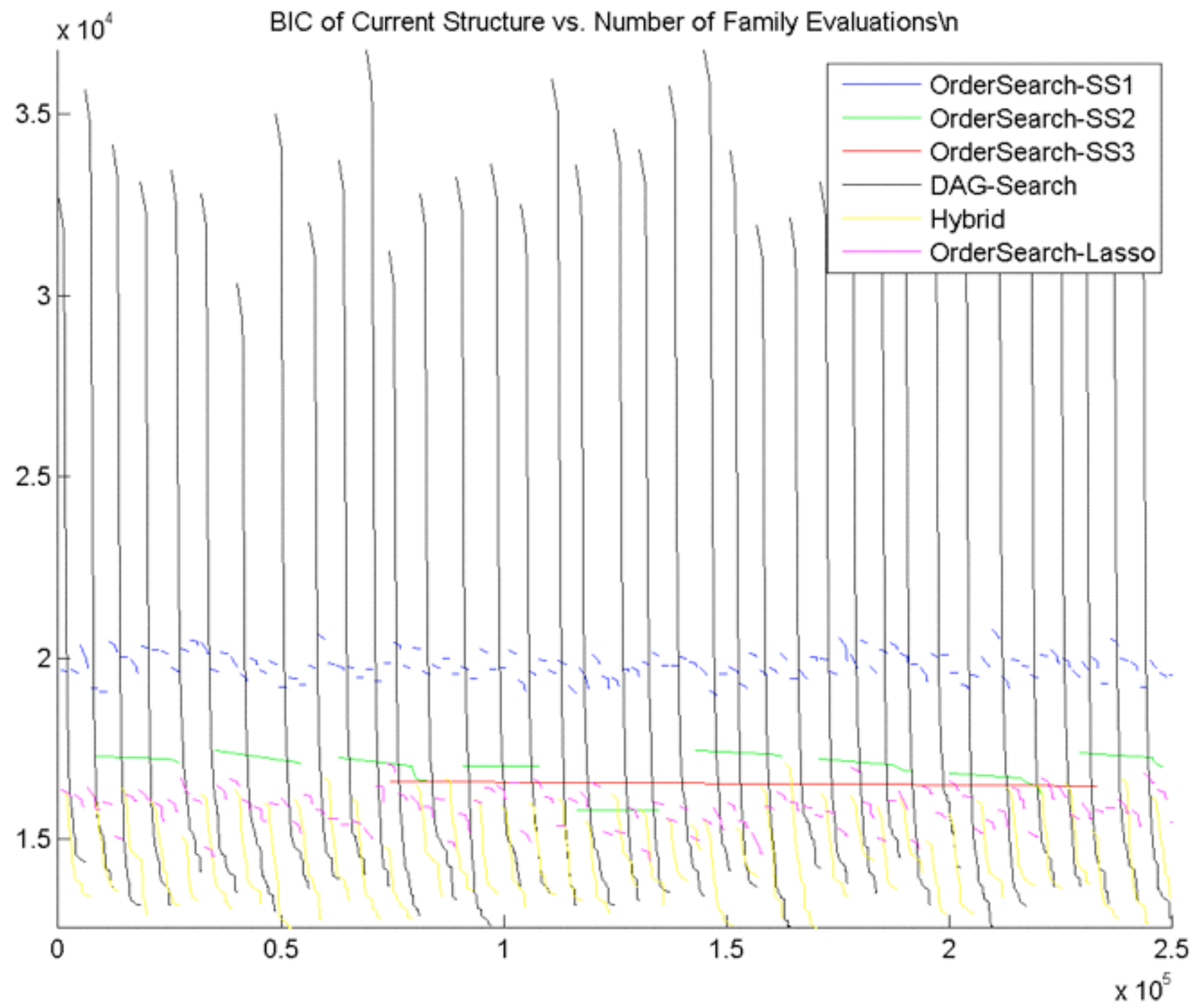
BIC of Current Structure vs. Number of Family Evaluations



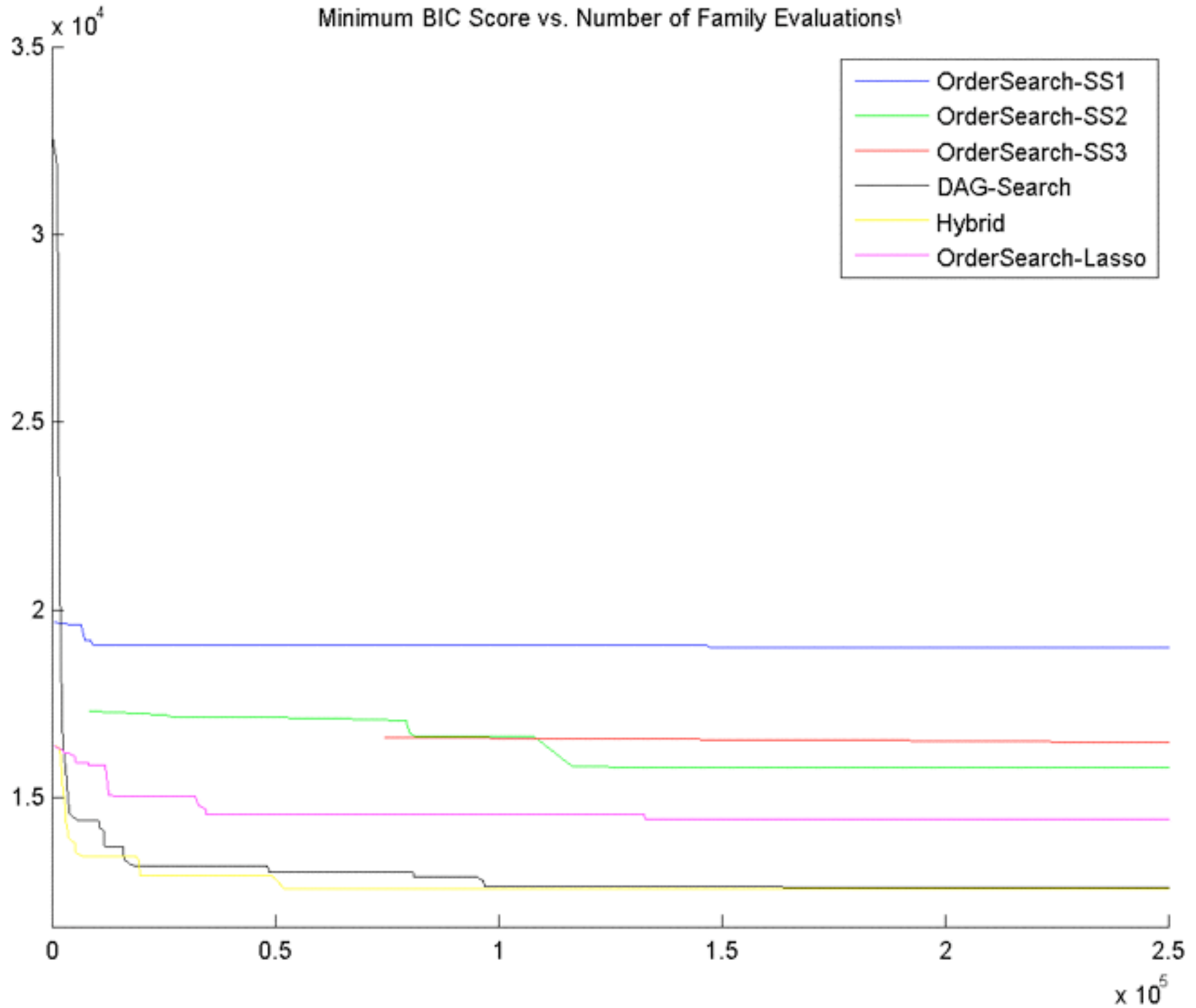
Minimum BIC Score vs. Number of Family Evaluations



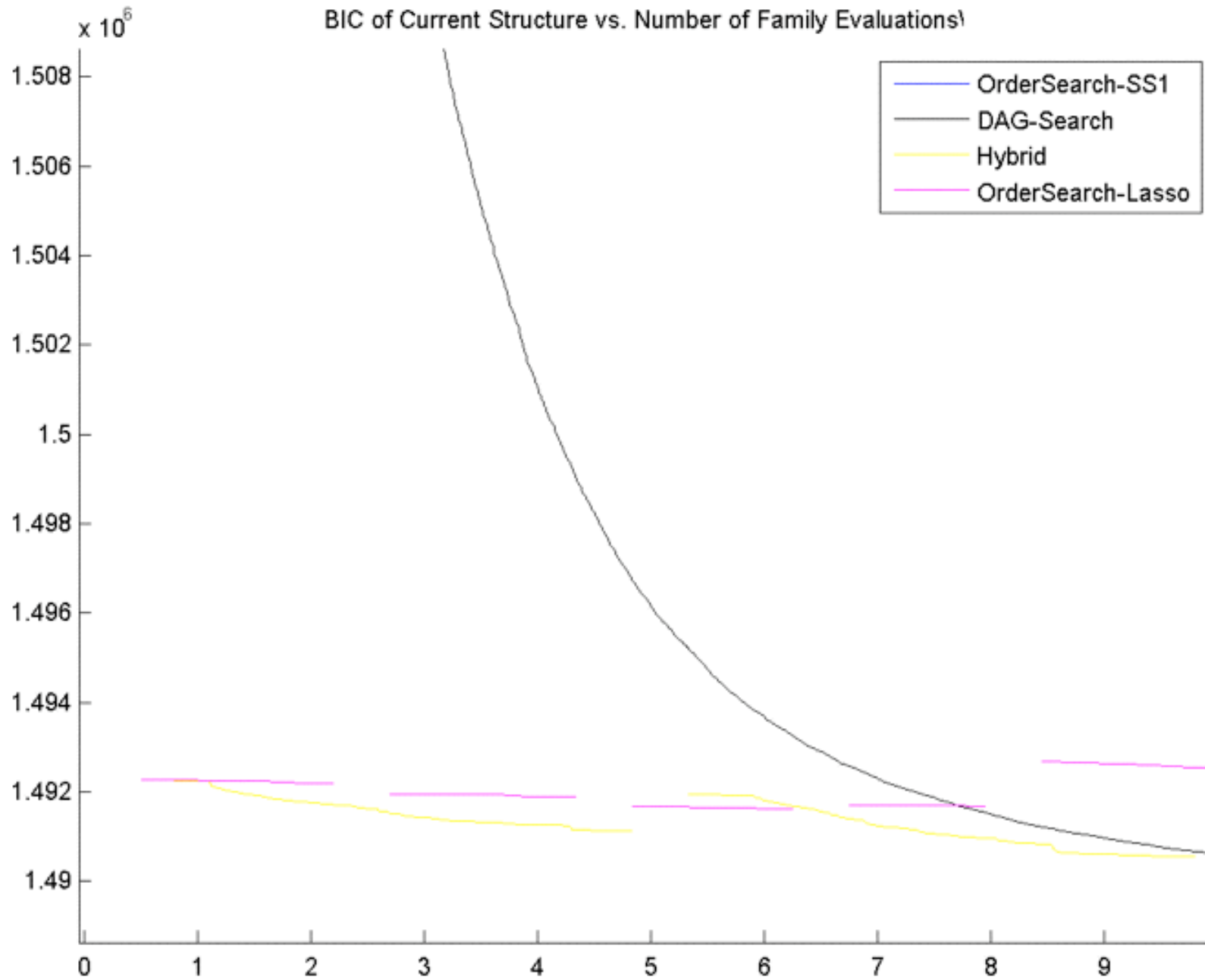
BIC of Current Structure vs. Number of Family Evaluations



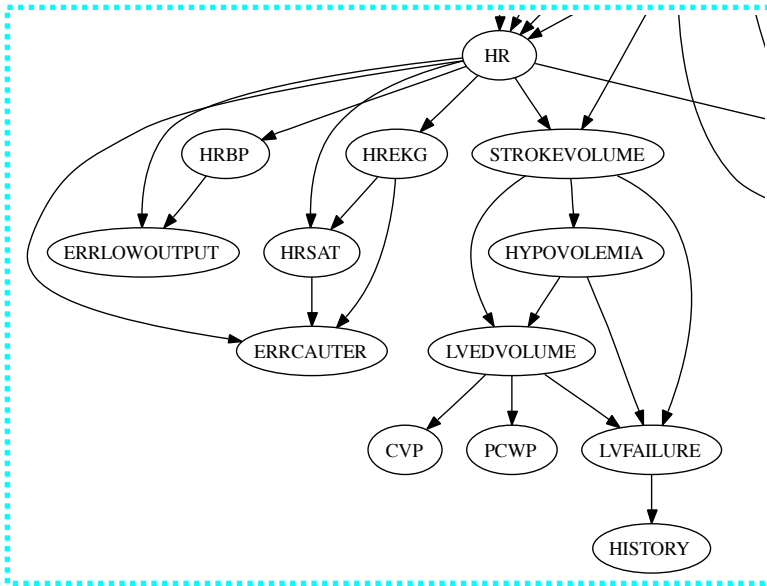
Minimum BIC Score vs. Number of Family Evaluations<sup>1)</sup>



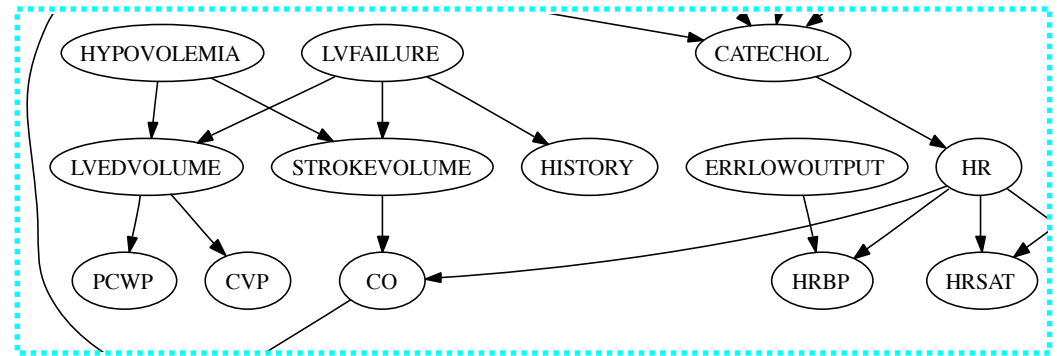
# Algorithms applied to 100-node Non-Causal 'Newsgroups' data



## Part of Learned Graph:



## Similar area in True Graph:



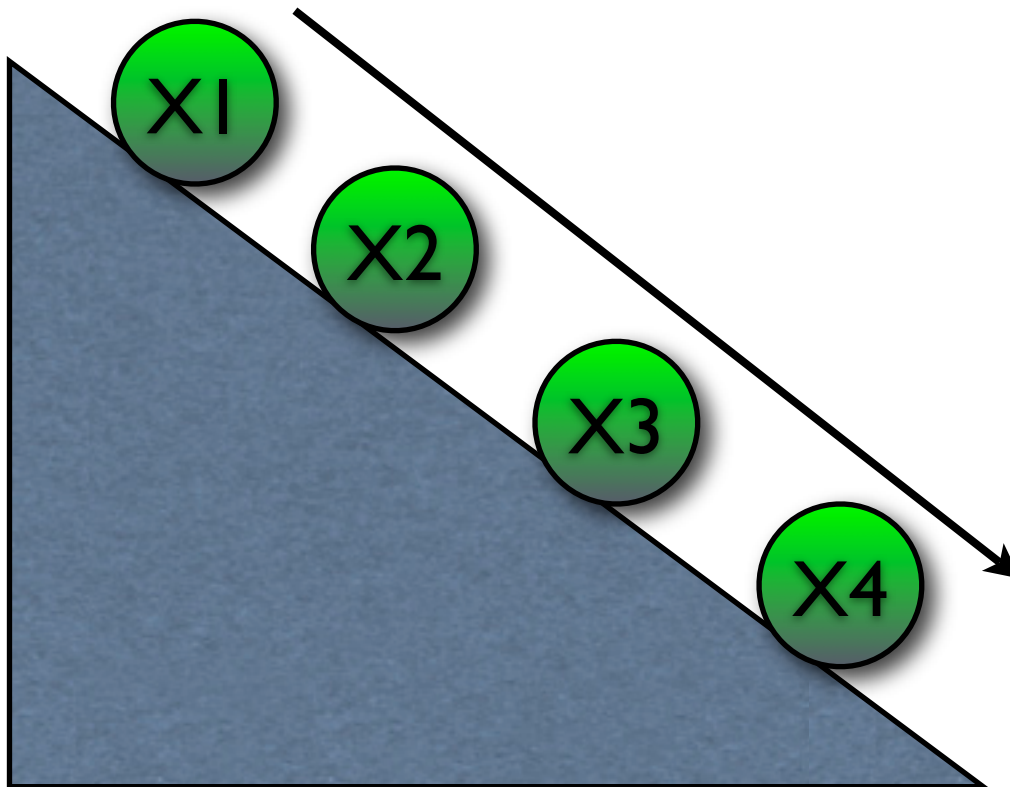
Learned Graph has not uncovered the causal structure

# Outline

- Bayes Net Structure Learning
- Order-Search
- LassoOrderSearch
- Experimental Results
- Interventional Data
- Parent Pruning

# Causal Topological Ordering

- If our topological ordering constrains the relationships between nodes to be causal, then we will learn a causal structure



# Interventional Data

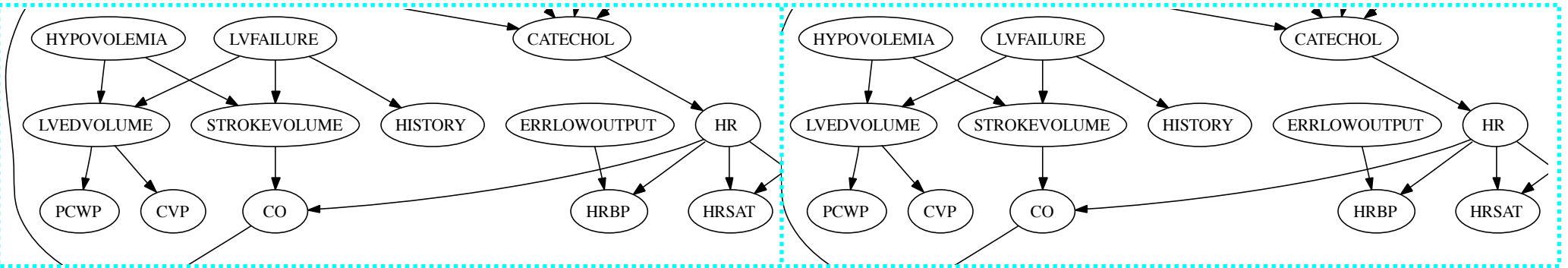
- Given Observational Data
  - It is possible to identify v-structures
  - We can not distinguish  $X \Rightarrow Y$  from  $Y \Rightarrow X$
- However, if we intervene and set either  $X$  or  $Y$ , we can break symmetry and direct the edge

# DAG/Order-Search with Interventions

- Structure learning with perfect interventions:
  - When computing score/parameters of a node, do not use cases where the node was the target of an intervention
- Same experimental set-up as before, but for each sample we do a perfect intervention on a random node

Part of Learned Graph:

Similar area in True Graph:



# Outline

- Bayes Net Structure Learning
- Order-Search
- LassoOrderSearch
- Experimental Results
- Interventional Data
- Parent Pruning

# Scaling Up

- Although polynomial, it may be expensive to evaluate orderings in graphs that have a large number of potential parent nodes
- A logical way to scale to larger problems is to precompute a candidate set of parents for each node
- ‘Sparse Candidate’ restricts the search to the  $k$  best parents based on pairwise associate scores, but the true parents are not guaranteed to be in this set

# Constraint-Based Approaches

- As opposed to finding the k-best, ‘Constraint-based’ approaches to structure learning remove potential parents that fail a conditional independency test
- Key observation: failure of a conditional independency test implies that nodes can’t share a direct edge
- Max-Min Hill-Climbing: Heuristic that prune a substantial amount of edges, then DAG-Search
- LassoOrderSearch could be used on the MMPC-restricted space to initialize the DAG-Search
- For interventional data, conditional independency tests may be non-commutative

# Summary

- Polynomial-time algorithm to compute DAG given ordering, independent of fan-in
- BIC-Optimal Hyperparameter selection and LARS-MLE
- Leads to a hybrid DAG/Order-Search algorithm, that overcomes some of the limitations of both approaches
- Extended to interventional data scenario for causal learning
- This technique can naturally be incorporated in state of the art hybrid pruning/search strategies