Objective model selection for sparse Gaussian DAG models

Davide Altomare
Università di Pavia
Guido Consonni
Università Cattolica del Sacro Cuore
Luca La Rocca
Università di Modena e Reggio Emilia

February 27-28, 2014 - University of Warwick

Outline

Publishing Productivity among Academics

Spirtes, Glymour and Scheines (2000)

Publishing Productivity among Academics

Spirtes, Glymour and Scheines (2000)

Publishing Productivity among Academics

Spirtes, Glymour and Scheines (2000)

DAG: Directed Acyclic Graph

1. subject's sex (Sex)
2. score of the subject's ability (Ability)
3. measure of the quality of the graduate program attended (GPQ)
4. preliminary measure of productivity (PreProd)
5. quality of the first job (QFJ)
6. publication rate (Pubs)
7. citation rate (Cites)

Publishing Productivity among Academics

Spirtes, Glymour and Scheines (2000)

DAG: Directed Acyclic Graph

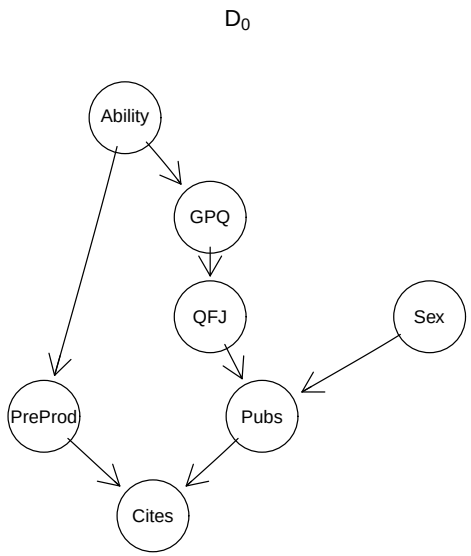
1. subject's sex (Sex)
2. score of the subject's ability (Ability)
3. measure of the quality of the graduate program attended (GPQ)
4. preliminary measure of productivity (PreProd)
5. quality of the first job (QFJ)
6. publication rate (Pubs)
7. citation rate (Cites)

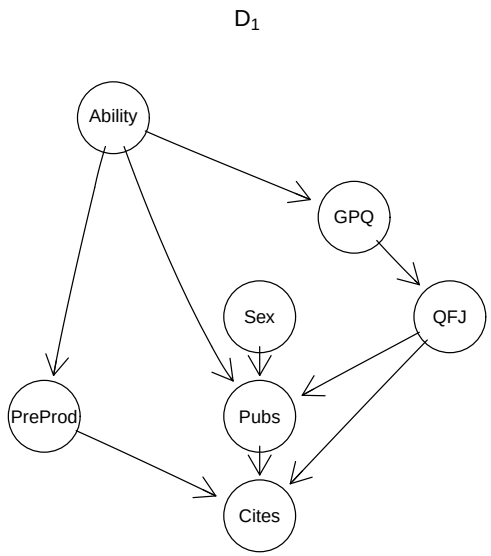
Publishing Productivity among Academics

Spirtes, Glymour and Scheines (2000)

DAG: Directed Acyclic Graph

1. subject's sex (Sex)
2. score of the subject's ability (Ability)
3. measure of the quality of the graduate program attended (GPQ)
4. preliminary measure of productivity (PreProd)
5. quality of the first job (QFJ)
6. publication rate (Pubs)
7. citation rate (Cites)





DAG: Directed Acyclic Graph

DAG: Directed Acyclic Graph

DAG: Directed Acyclic Graph

$\mathcal{D} = (V, E)$ DAG
$V = \{1, \ldots, q\}$ set of its vertices
$E \subseteq V \times V$ set of directed edges.
Total ordering of the vertices.
Vertices of $\mathcal{D}$ are well-numbered: i.e.
if $\exists$ directed path from vertex i to
vertex j, then $i < j$.

DAG: Directed Acyclic Graph

$\mathcal{D} = (V, E)$ DAG
$V = \{1, \ldots, q\}$ set of its vertices
$E \subseteq V \times V$ set of directed edges.
Total ordering of the vertices.
Vertices of $\mathcal{D}$ are well-numbered: i.e.
if $\exists$ directed path from vertex i to
vertex j, then $i < j$.

DAG: Directed Acyclic Graph

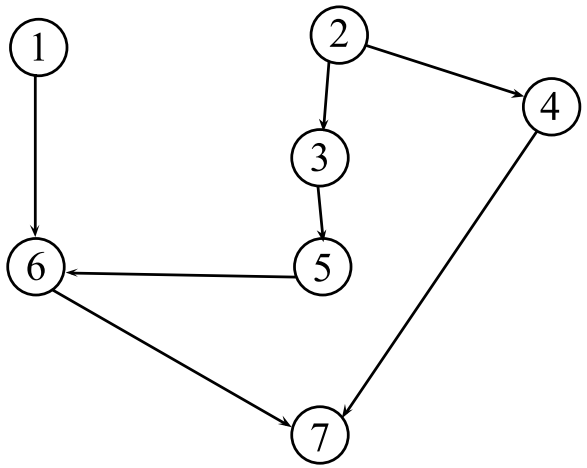
$\mathcal{D} = (V, E)$ DAG
$V = \{1, \ldots, q\}$ set of its vertices
$E \subseteq V \times V$ set of directed edges.
Total ordering of the vertices.
Vertices of $\mathcal{D}$ are well-numbered: i.e.
if $\exists$ directed path from vertex i to
vertex j, then $i < j$.

DAG: Directed Acyclic Graph

$\mathcal{D} = (V, E)$ DAG
$V = \{1, \ldots, q\}$ set of its vertices
$E \subseteq V \times V$ set of directed edges.
Total ordering of the vertices.
Vertices of $\mathcal{D}$ are well-numbered: i.e.
if $\exists$ directed path from vertex i to
vertex j, then $i < j$.

DAG: Directed Acyclic Graph

$\mathcal{D} = (V, E)$ DAG
$V = \{1, \ldots, q\}$ set of its vertices
$E \subseteq V \times V$ set of directed edges.
Total ordering of the vertices.
Vertices of $\mathcal{D}$ are well-numbered: i.e. if $\exists$ directed path from vertex i to vertex j, then $i < j$.



DAG $\mathcal{D}_0$

Probabilistic DAG

Each vertex j corresponds to a random variable u_j.
$W \subseteq V$: u_W is the set of all variables u_j with $j \in W$.

A special subset
$W = \text{pa}(j)$: parents of j.
Factorization of the joint density

$$f(u_1, \ldots, u_q|\theta) = \prod_{j=1}^{q} f(u_j|u_{\text{pa}(j)}; \theta_j)$$

$$u_j \perp\!\!\!\perp u_{\{1,\ldots,j-1\}\setminus \text{pa}(j)} \mid u_{\text{pa}(j)}, \theta_j$$

Cites $\perp\!\!\!\perp$ { *Sex*, *Ability*, *Grad Progr*, *Quality First Job* | *Prelim Meas Product*, *Pub Rate* }

Gaussian DAG

Gaussian DAG $\mathcal{D}$ model
Family of all q-variate normal distributions
satisfying conditional independence implied by $\mathcal{D}$

$$f(u_1, \ldots, u_q | \beta, \gamma) = \prod_{j=1}^{q} f(u_j | u_{\mathsf{pa}(j)}; \beta_j, \gamma_j).$$

Each conditional distribution is a univariate normal
β_j: regression coefficients; γ_j: conditional precision

Bayes factor

Usually DAG $\mathcal{D}$ is unknown

Need to select one among a list of candidates DAG-models

Bayes factor

Usually DAG $\mathcal{D}$ is unknown
Need to select one among a list of candidates DAG-models

Two models $\mathcal{M}_k$, $k = 0, 1$,
Sampling density $f(y|\theta_k)$, $\theta_k \in \Theta_k$, and prior $p(\theta_k)$.
Bayes Factor (BF)

$$BF_{10}(y) = m_1(y)/m_0(y)$$

$m_k(y)$ is the marginal likelihood of $\mathcal{M}_k$,

$$m_k(y) = \int f(y|\theta_k)p(\theta_k)d\theta_k$$

Bayes factor

Usually DAG $\mathcal{D}$ is unknown
Need to select one among a list of candidates DAG-models

Two models $\mathcal{M}_k$, $k = 0, 1$,
Sampling density $f(y|\theta_k)$, $\theta_k \in \Theta_k$, and prior $p(\theta_k)$.
Bayes Factor (BF)

$$BF_{10}(y) = m_1(y)/m_0(y)$$

$m_k(y)$ is the marginal likelihood of $\mathcal{M}_k$,

$$m_k(y) = \int f(y|\theta_k)p(\theta_k)d\theta_k$$

Posterior model probability

$$\mathbb{P}\{\mathcal{M}_0 \,|\, y\} = \frac{\mathbb{P}\{\mathcal{M}_0\}}{\mathbb{P}\{\mathcal{M}_0\} + BF_{10}\mathbb{P}\{\mathcal{M}_1\}}$$

Improper priors

Objective priors typically improper
(defined up to a multiplicative constant)

Improper priors

Objective priors typically improper
(defined up to a multiplicative constant)
Cannot be used to compute BFs
(even when the marginal likelihoods exist)

Improper priors

Objective priors typically improper
(defined up to a multiplicative constant)
Cannot be used to compute BFs
(even when the marginal likelihoods exist)
A few solutions

- intrinsic Bayes factors (Berger and Pericchi, 1996)

Improper priors

Objective priors typically improper
(defined up to a multiplicative constant)
Cannot be used to compute BFs
(even when the marginal likelihoods exist)
A few solutions

- intrinsic Bayes factors (Berger and Pericchi, 1996)
- intrinsic priors (Moreno, 1997)

Improper priors

Objective priors typically improper
(defined up to a multiplicative constant)
Cannot be used to compute BFs
(even when the marginal likelihoods exist)
A few solutions

- intrinsic Bayes factors (Berger and Pericchi, 1996)
- intrinsic priors (Moreno, 1997)
- expected posterior priors (Perez and Berger, 2002)

Improper priors

Objective priors typically improper
(defined up to a multiplicative constant)
Cannot be used to compute BFs
(even when the marginal likelihoods exist)
A few solutions

- intrinsic Bayes factors (Berger and Pericchi, 1996)
- intrinsic priors (Moreno, 1997)
- expected posterior priors (Perez and Berger, 2002)
- fractional Bayes factor (0'Hagan, 1995)
  easy to implement
  marginal likelihoods available in closed-form
  (in exponential family-conjugate prior setup)

Fractional BF

$\mathcal{M}_k$; $f(y|\theta_k)$; $p(\theta_k)$
Fractional marginal likelihood for model $\mathcal{M}_k$

$$w_k(y; g) = \frac{\int f(y|\theta_k)p(\theta_k)d\theta_k}{\int (f(y|\theta_k))^g p(\theta_k)d\theta_k}$$

$0 < g < 1$ (*fraction*)
Fractional BF in favor of $\mathcal{M}_1$

$$FBF_{10}(y; g) = w_1(y; g)/w_0(y; g).$$

Fractional BF

$\mathcal{M}_k$; $f(y|\theta_k)$; $p(\theta_k)$
Fractional marginal likelihood for model $\mathcal{M}_k$

$$w_k(y; g) = \frac{\int f(y|\theta_k)p(\theta_k)d\theta_k}{\int (f(y|\theta_k))^g p(\theta_k)d\theta_k}$$

$0 < g < 1$ (*fraction*)
Fractional BF in favor of $\mathcal{M}_1$

$$FBF_{10}(y; g) = w_1(y; g)/w_0(y; g).$$

Notice

$$w_k(y; g) = \int (f(y|\theta_k))^{(1-g)} p^F(\theta_k|g, y)d\theta_k$$

$p^F(\theta_k|g, y) \propto (f(y|\theta_k))^g p(\theta_k)$ is the implied data-dependent *fractional prior*

Fractional BF

$\mathcal{M}_k$; $f(y|\theta_k)$; $p(\theta_k)$
Fractional marginal likelihood for model $\mathcal{M}_k$

$$w_k(y; g) = \frac{\int f(y|\theta_k)p(\theta_k)d\theta_k}{\int (f(y|\theta_k))^g p(\theta_k)d\theta_k}$$

$0 < g < 1$ (*fraction*)
Fractional BF in favor of $\mathcal{M}_1$

$$FBF_{10}(y; g) = w_1(y; g)/w_0(y; g).$$

Notice

$$w_k(y; g) = \int (f(y|\theta_k))^{(1-g)} p^F(\theta_k|g, y)d\theta_k$$

$p^F(\theta_k|g, y) \propto (f(y|\theta_k))^g p(\theta_k)$ is the implied data-dependent *fractional prior*
Consistency of the Fractional BF holds as long as $g \to 0$ $(n \to \infty)$

Objective priors

Recall the recursive structure of the likelihood

$$f(u_1, \ldots, u_q|\beta, \gamma) = \prod_{j=1}^{q} f(u_j|u_{\text{pa}(j)}; \beta_j, \gamma_j),$$

Objective prior

$$p^D(\beta, \gamma) \propto \prod_{j=1}^{q} \gamma_j^{-1}$$

it satisfies *global parameter independence* (Geiger and Heckerman, 2002)

DAGs $\mathcal{D}_0$ and $\mathcal{D}_1$

same vertex set and vertex ordering

$\mathcal{D}_0$ nested in $\mathcal{D}_1$

Fix vertex j:

L_j: set of vertices which are parents of j under $\mathcal{D}_1$, but not under $\mathcal{D}_0$

$\mathcal{D}_0 \Leftrightarrow \beta_{jl} = 0,\ l \in L_j,\ j = 1, \ldots, q$

DAGs $\mathcal{D}_0$ and $\mathcal{D}_1$

same vertex set and vertex ordering

$\mathcal{D}_0$ nested in $\mathcal{D}_1$

Fix vertex j:

L_j: set of vertices which are parents of j under $\mathcal{D}_1$, but not under $\mathcal{D}_0$

$\mathcal{D}_0 \Leftrightarrow \beta_{jl} = 0,\ l \in L_j,\ j = 1, \ldots, q$

DAGs $\mathcal{D}_0$ and $\mathcal{D}_1$

same vertex set and vertex ordering

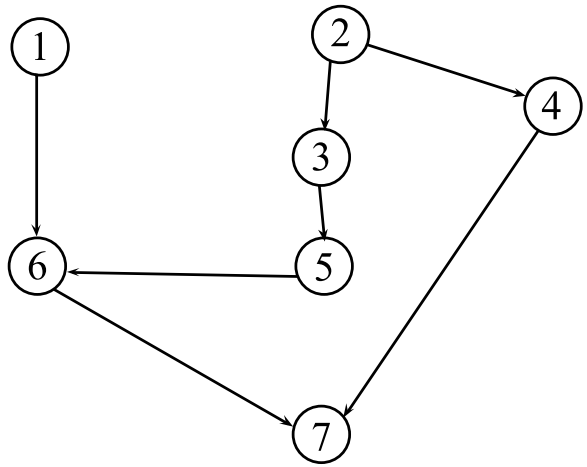
$\mathcal{D}_0$ nested in $\mathcal{D}_1$

Fix vertex j:

L_j: set of vertices which are parents of j under $\mathcal{D}_1$, but not under $\mathcal{D}_0$

$\mathcal{D}_0 \Leftrightarrow \beta_{jl} = 0,\ l \in L_j,\ j = 1, \ldots, q$

$\mathcal{D}_1$



DAGs $\mathcal{D}_0$ and $\mathcal{D}_1$

same vertex set and vertex ordering

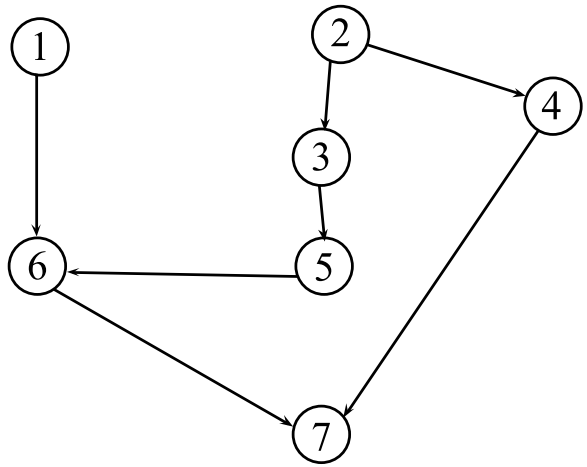
$\mathcal{D}_0$ nested in $\mathcal{D}_1$

Fix vertex j:

L_j: set of vertices which are parents of j under $\mathcal{D}_1$, but not under $\mathcal{D}_0$

$\mathcal{D}_0 \Leftrightarrow \beta_{jl} = 0,\ l \in L_j,\ j = 1, \ldots, q$

$\mathcal{D}_1$



DAGs $\mathcal{D}_0$ and $\mathcal{D}_1$

same vertex set and vertex ordering

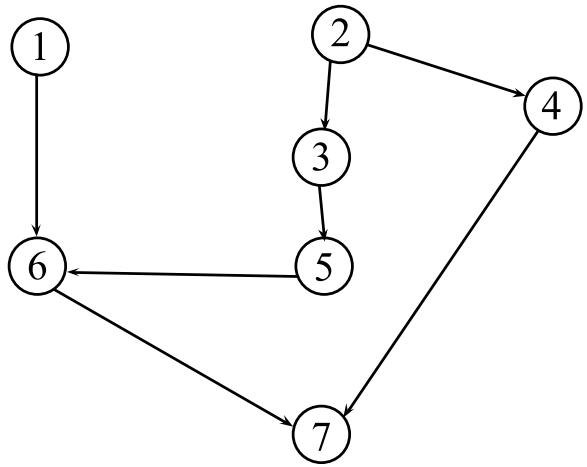
$\mathcal{D}_0$ nested in $\mathcal{D}_1$

Fix vertex j:

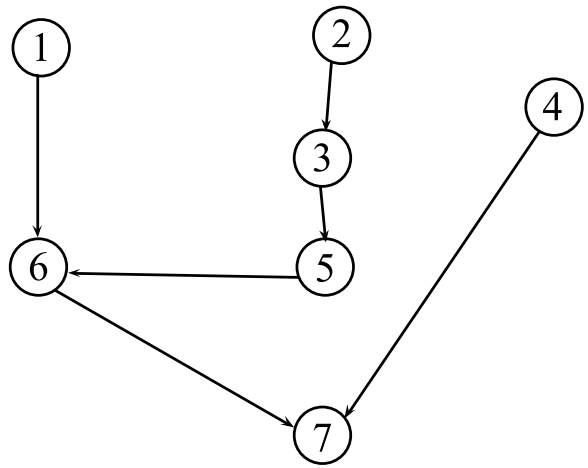
L_j: set of vertices which are parents of j under $\mathcal{D}_1$, but not under $\mathcal{D}_0$

$\mathcal{D}_0 \Leftrightarrow \beta_{jl} = 0,\ l \in L_j,\ j = 1, \ldots, q$

$\mathcal{D}_1$



$\mathcal{D}_0$



$L_4 = \{2\}$

Objective Product Moment Prior

Product moment prior *under* $\mathcal{D}_1$

$$p_1^M(\beta,\gamma) \propto \prod_{j=1}^{q} \left\{ \gamma_j^{-1} \prod_{l \in L_j} \beta_{jl}^{2h} \right\}$$

Fractional marginal likelihood factorizes
Expression for Moment Fractional BF available in closed form
(C and La Rocca, 2011)

Prior on DAG space

A Gaussian DAG model can be viewed as a *sequence* of $(q-1)$ conditional 'regression' models.
$\mathcal{D}_k \Leftrightarrow \mathcal{M}_{k_2}, \ldots, \mathcal{M}_{k_q}$
$\mathfrak{M}_j$: family of all 'regression' models for node j
(there are 2^{j-1} such models)
Prior over the space $\mathfrak{D}$ of all DAG models

$$\mathbb{P}\{\mathcal{D}_k\} = \prod_{j=2}^{q} \mathbb{P}\{\mathcal{M}_{k_j}\} = \prod_{j=2}^{q} \frac{1}{j} \binom{j-1}{|\mathrm{pa}_k(j)|}^{-1}, \quad \mathcal{D}_k \in \mathfrak{D}$$

This is a product of *multiplicity correction priors* (Scott and Berger, 2010)

Finite *collection* of DAGs $\{\mathcal{D}_k\} \in \mathfrak{D}$
$\mathcal{D}_0$ *complete independence* DAG
(DAG with no edges)
nested into every other model $\mathcal{D}_k$
encompassing from below

Finite *collection* of DAGs $\{\mathcal{D}_k\} \in \mathfrak{D}$
$\mathcal{D}_0$ *complete independence* DAG
(DAG with no edges)
nested into every other model $\mathcal{D}_k$
encompassing from below
Compute the (Moment) Fractional BF (FBF) of $\mathcal{D}_k$ against $\mathcal{D}_0$, namely $\{FBF_{k0}(y)\}$
Derive the posterior probability of model $\mathcal{D}_k$

$$\mathbb{P}\{\mathcal{D}_k|y\} = \frac{FBF_{k0}(y;g)\mathbb{P}\{\mathcal{D}_k\}}{\sum_j FBF_{j0}(y;g)\mathbb{P}\{\mathcal{D}_j\}}, \quad \mathcal{D}_k \in \mathfrak{D}$$

Number of DAGs

Grows exponentially with the number of variables

Enumeration is not feasible even for moderately sized vertex sets

Resort to search algorithm to identify the most valuable models.

Number of DAGs

Grows exponentially with the number of variables

Enumeration is not feasible even for moderately sized vertex sets

Resort to search algorithm to identify the most valuable models.

Number of DAGs

Grows exponentially with the number of variables

Enumeration is not feasible even for moderately sized vertex sets

Resort to search algorithm to identify the most valuable models.

q	number of DAGs
10	$3.5 \cdot 10^{13}$
15	$4.1 \cdot 10^{31}$
20	$1.6 \cdot 10^{57}$
30	$8.9 \cdot 10^{130}$
40	$6.4 \cdot 10^{234}$

The Algorithm (based on Berger and Molina, 2005)

1. Start with a base DAG $\mathcal{D}_B$ and obtain deterministically $m \equiv q(q-1)/2$ distinct new DAGs each one differing from $\mathcal{D}_B$ by exactly one edge.

   Compute (the estimated) graph posterior probabilities and edge inclusion probabilities by re-normalization.

The Algorithm (based on Berger and Molina, 2005)

1. Start with a base DAG $\mathcal{D}_B$ and obtain deterministically $m \equiv q(q-1)/2$ distinct new DAGs each one differing from $\mathcal{D}_B$ by exactly one edge.

   Compute (the estimated) graph posterior probabilities and edge inclusion probabilities by re-normalization.

2. *Resampling move*

   Return to one of the previously visited graphs, according to the posterior probabilities.

The Algorithm (based on Berger and Molina, 2005)

1. Start with a base DAG $\mathcal{D}_B$ and obtain deterministically $m \equiv q(q-1)/2$ distinct new DAGs each one differing from $\mathcal{D}_B$ by exactly one edge.

   Compute (the estimated) graph posterior probabilities and edge inclusion probabilities by re-normalization.

2. *Resampling move*

   Return to one of the previously visited graphs, according to the posterior probabilities.

3. *Local move*

   Identify single edges leading to a new DAG.

   Randomly choose one and add/delete according to inclusion probability.

The Algorithm (based on Berger and Molina, 2005)

1. Start with a base DAG $\mathcal{D}_B$ and obtain deterministically $m \equiv q(q-1)/2$ distinct new DAGs each one differing from $\mathcal{D}_B$ by exactly one edge.

   Compute (the estimated) graph posterior probabilities and edge inclusion probabilities by re-normalization.

2. *Resampling move*

   Return to one of the previously visited graphs, according to the posterior probabilities.

3. *Local move*

   Identify single edges leading to a new DAG.

   Randomly choose one and add/delete according to inclusion probability.

4. Usually return directly to step 2

   Periodically make a *global move* to the current Median Probability-DAG

   Return to step 3.

Simulation with high-dimensional sparse DAGs

Three random DAGs of size $q = 50, 100, 200$
generated using R-package `pcalg` (Kalish and Bühlman, 2007)
each DAG has exactly $|E| = 100$ edges

Simulation with high-dimensional sparse DAGs

Three random DAGs of size $q = 50, 100, 200$
generated using R-package `pcalg` (Kalish and Bühlman, 2007)
each DAG has exactly $|E| = 100$ edges
N.B. As q increases, DAG becomes *sparser*.

Simulation with high-dimensional sparse DAGs

Three random DAGs of size $q = 50, 100, 200$
generated using R-package `pcalg` (Kalish and Bühlman, 2007)
each DAG has exactly $|E| = 100$ edges
N.B. As q increases, DAG becomes *sparser*.

For each of the three DAGs, we simulated $n = 100$
observations from the linear structural equation model

$$u_i = \sum_{j \in \mathrm{pa}(i)} \rho_{ij} u_j + \varepsilon_i, \quad i = 1, \ldots, q,$$

with $\epsilon_j \overset{iid}{\sim} N(0, 1)$, $\rho_{ij} = 0.8$ for all i and j, and replicated the simulation 10 times in order to assess sampling variability

Evaluation of search algorithm

Comparison of (Moment) Fractional BF with alternative methods

- Lasso

Evaluation of search algorithm

Comparison of (Moment) Fractional BF with alternative methods

- Lasso
- Adaptive Lasso

Evaluation of search algorithm

Comparison of (Moment) Fractional BF with alternative methods

- Lasso
- Adaptive Lasso
- SIN

Evaluation of search algorithm

Comparison of (Moment) Fractional BF with alternative methods

- Lasso
- Adaptive Lasso
- SIN
- PC-algorithm
  (no ordering of variables is assumed)

Evaluation of search algorithm

Comparison of (Moment) Fractional BF with alternative methods

- Lasso
- Adaptive Lasso
- SIN
- PC-algorithm
  (no ordering of variables is assumed)

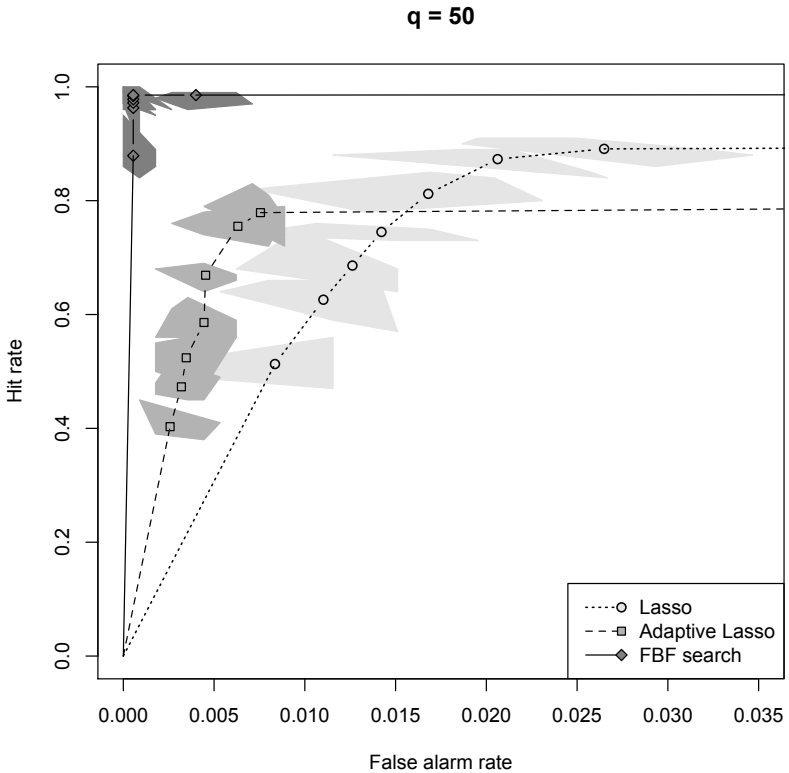
Evaluation of search algorithm

Comparison of (Moment) Fractional BF with alternative methods

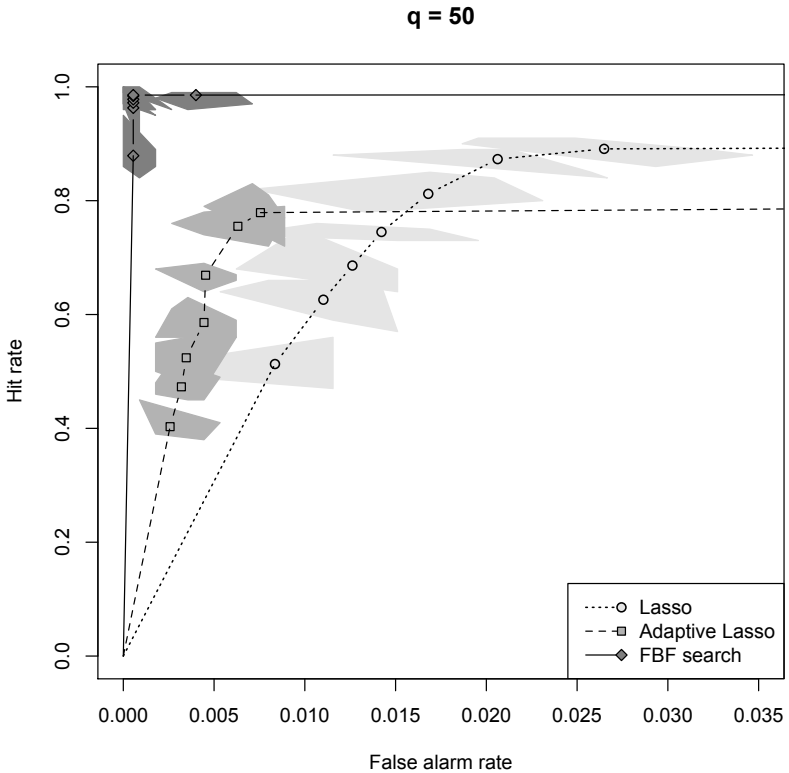
- Lasso
- Adaptive Lasso
- SIN
- PC-algorithm
  (no ordering of variables is assumed)

Receiver Operating Characteristics (ROC) curve

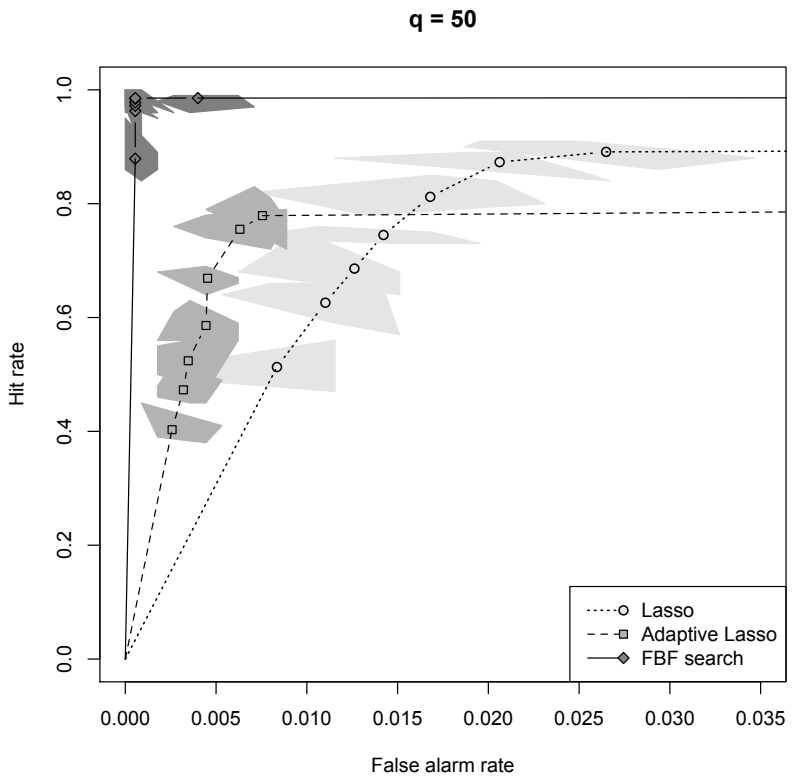
$q = 50$ ROC curve



$q = 50$ ROC curve



$q = 50$ ROC curve



Fractional BF searches
$h = 0, 1, 2, 3, 4, 5, 10$
(from right to left)

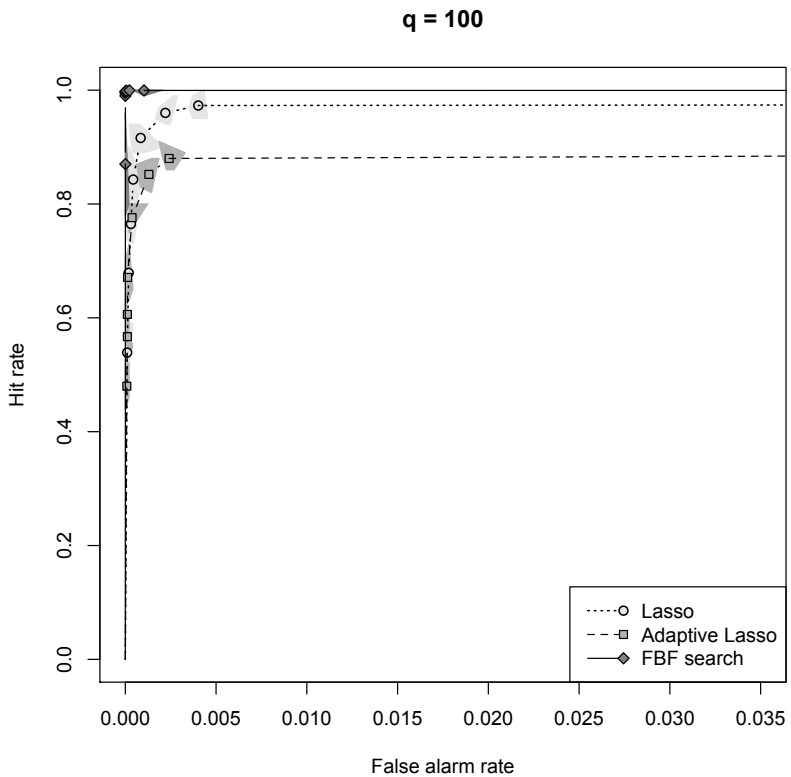
Lasso and Adaptive Lasso with "significance" levels
$\alpha = 0.0001, 0.01, 0.1, 1, 10, 50, 100$
(from left to right).

Fractional BF search outperforms Lasso and Adaptive Lasso

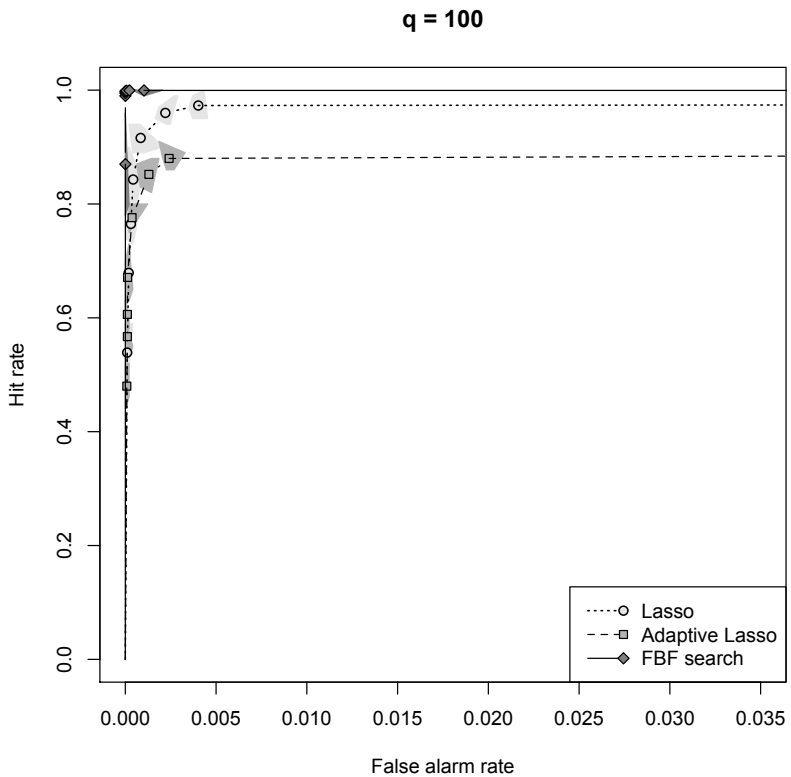
(Adaptive Lasso better than Lasso).

Shaded area represents sampling variability

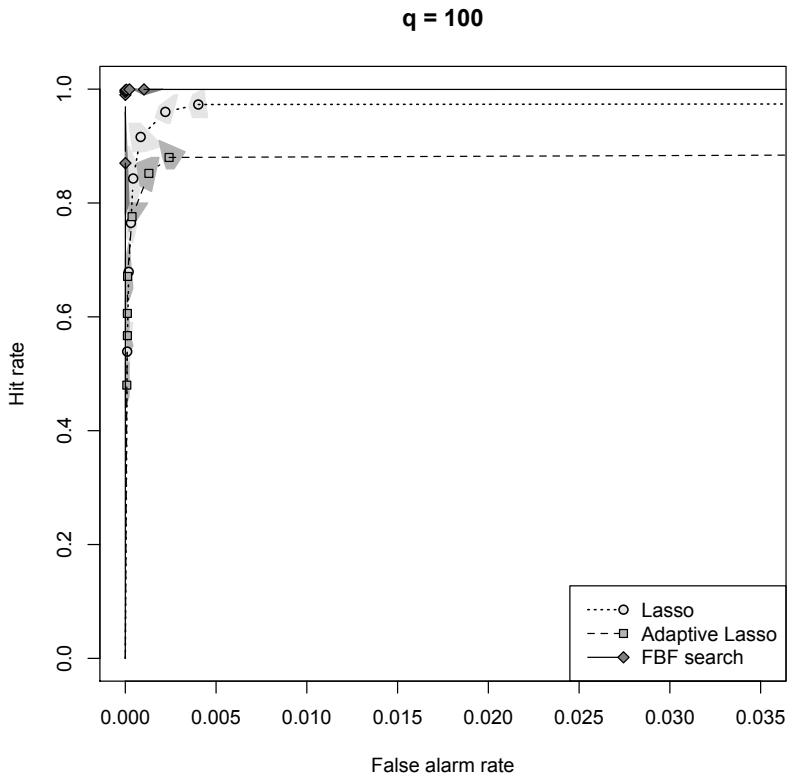
$q = 100$ ROC curve



$q = 100$ ROC curve



$q = 100$ ROC curve



Fractional BF searches
$h = 0, 1, 2, 3, 4, 5, 10$
(from right to left)

Lasso and Adaptive Lasso with "significance" levels
$\alpha = 0.0001, 0.01, 0.1, 1, 10, 50, 100$
(from left to right).

Superiority of Fractional BF still visible but less pronounced.

(Lasso better than Adaptive Lasso).

Human cell signalling pathways

Human cell signalling pathways

Human cell signalling pathways

Flow cytometry experiments. Signalling networks of human cells
(Sachs et al 2003)

Data: $q = 11$ proteins and $n = 7466$

Ordering of the connections assumed known as in Shojaie and Michailidis (2010)

Human cell signalling pathways

Flow cytometry experiments. Signalling networks of human cells
(Sachs et al 2003)

Data: $q = 11$ proteins and $n = 7466$

Ordering of the connections assumed known as in Shojaie and Michailidis (2010)

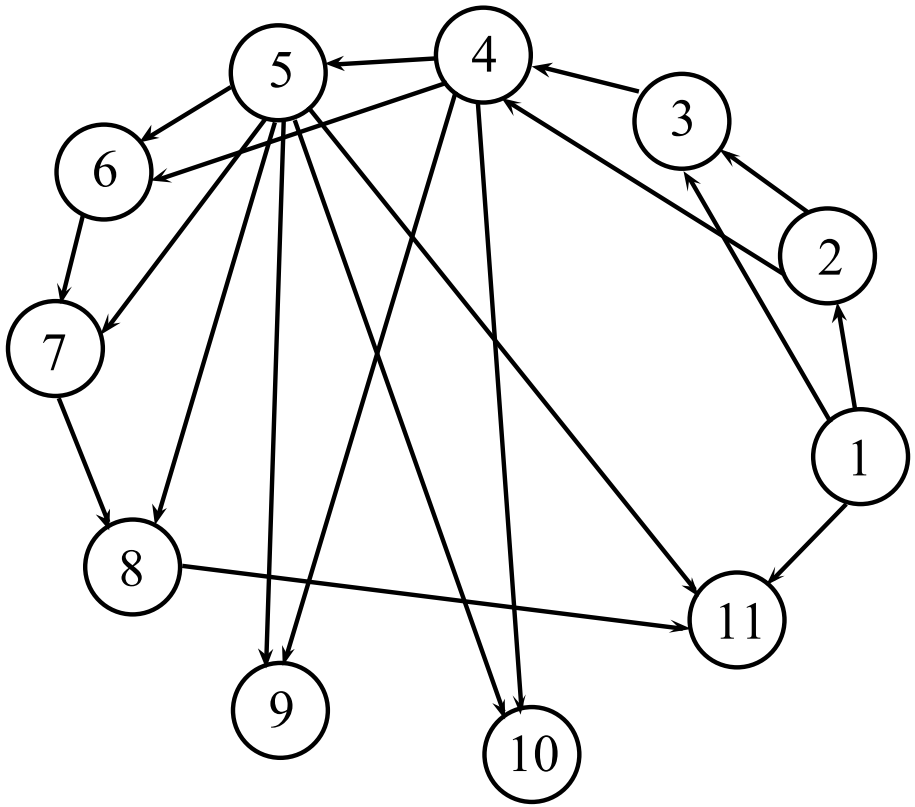
Human cell signalling pathways

Flow cytometry experiments. Signalling networks of human cells
(Sachs et al 2003)

Data: $q = 11$ proteins and $n = 7466$

Ordering of the connections assumed known as in Shojaie and Michailidis (2010)

(Supposedly) known regulatory network

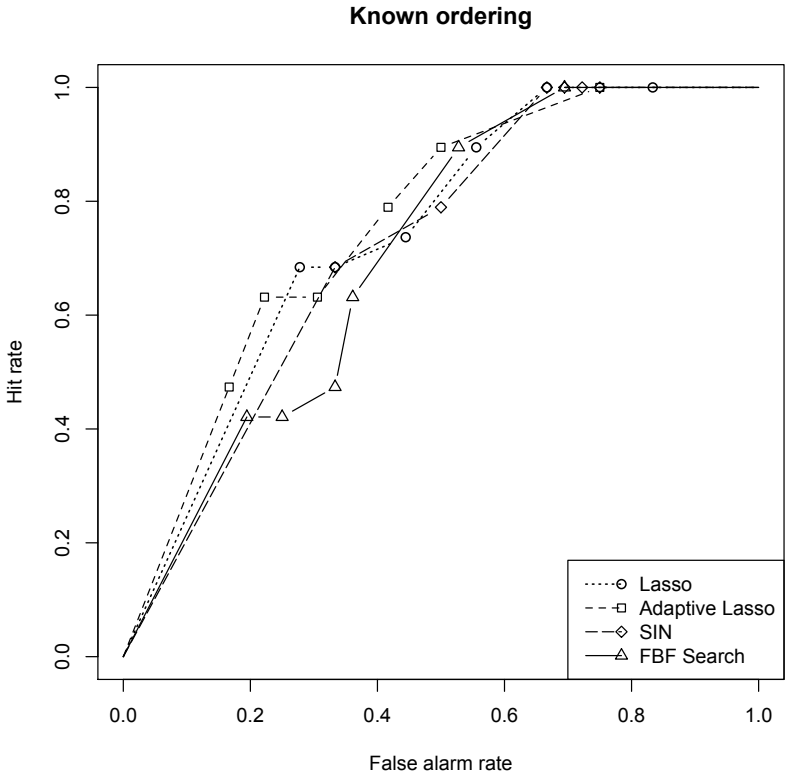


Scenario 1

Scenario 1

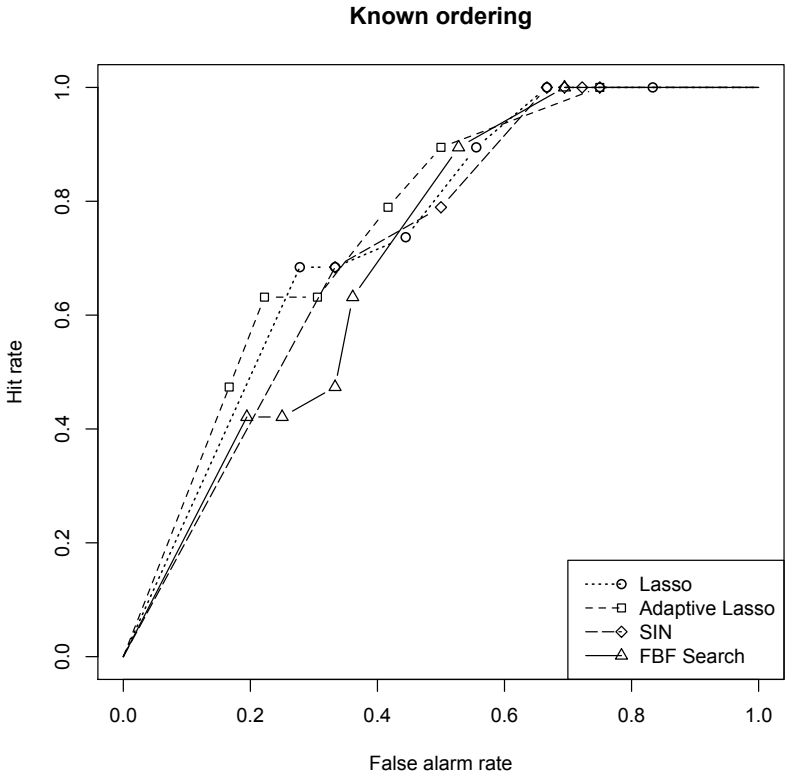
Scenario 1

ROC curve: *real* data



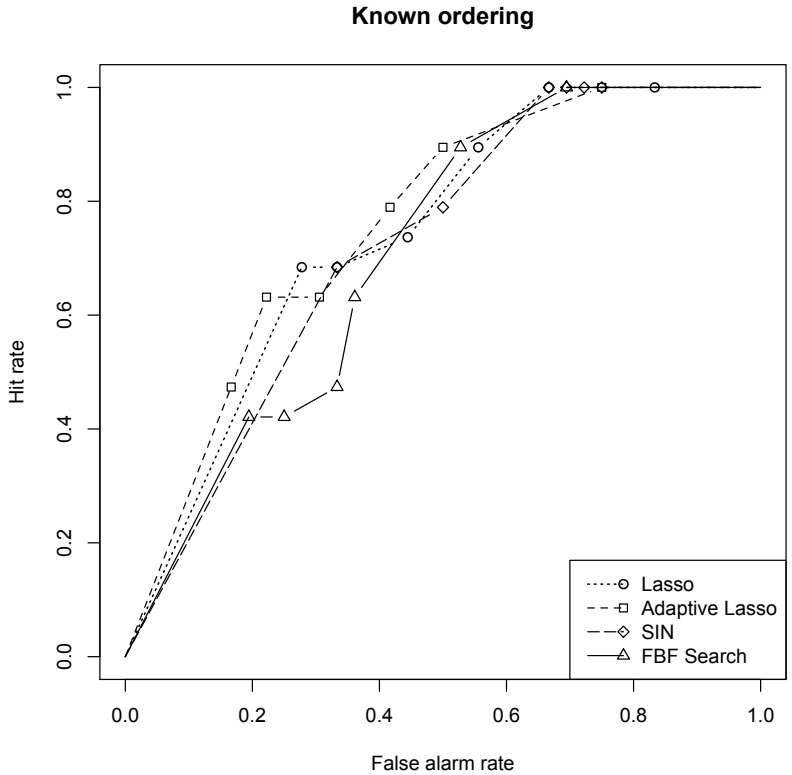
Scenario 1

ROC curve: *real* data



Scenario 1

ROC curve: *real* data



Adaptive Lasso tends to perform better than any of the other methods.

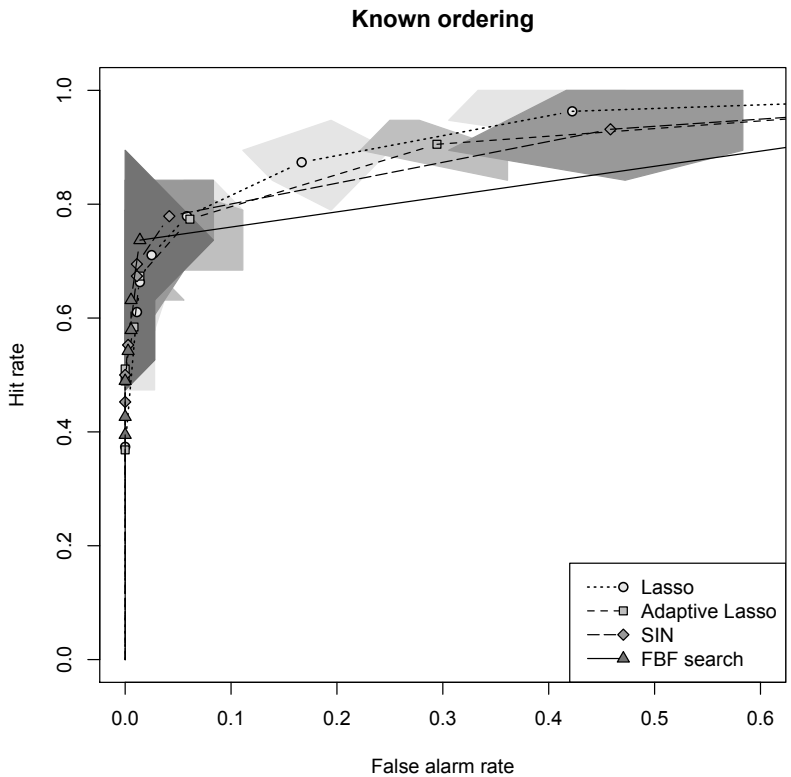
FBF performs rather poorly in the left part of the curve Recall, however, that this experiment uses *real* data while assuming a (supposedly) known underlying network

Scenario 2

Scenario 2

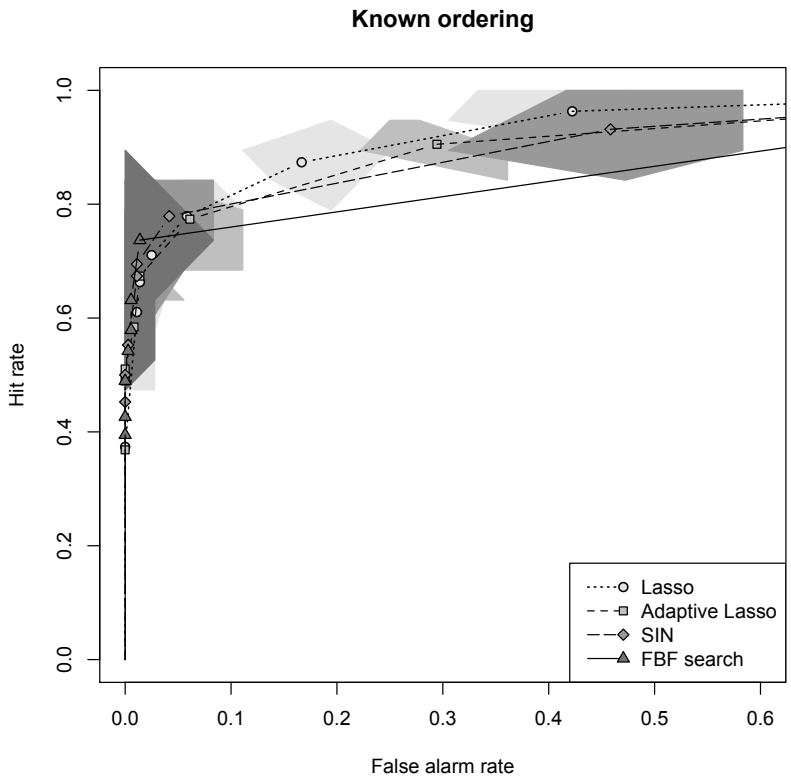
Scenario 2

ROC curve: *simulated* data from estimated known network



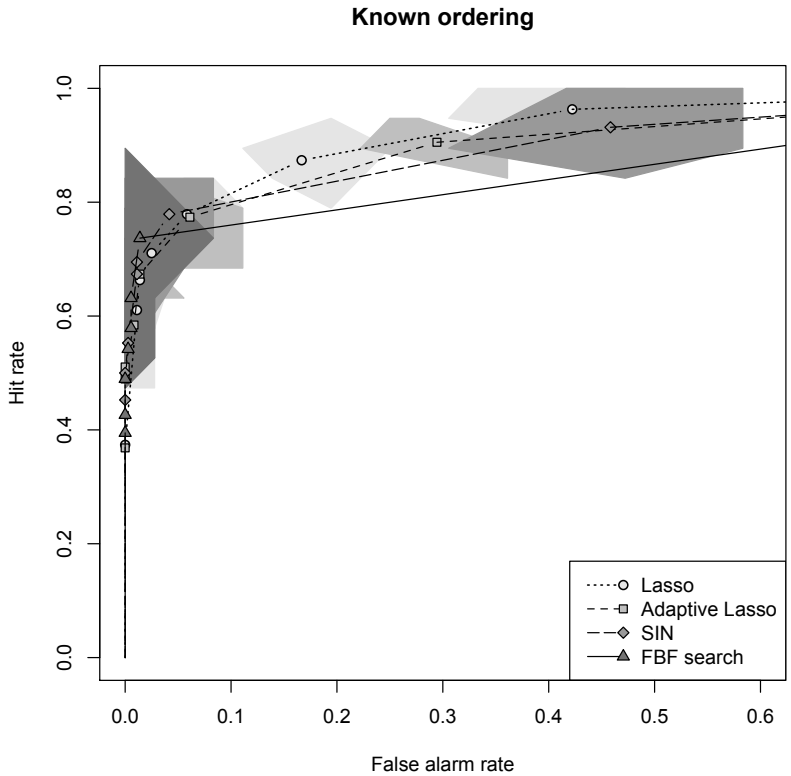
Scenario 2

ROC curve: *simulated* data from estimated known network



Scenario 2

ROC curve: *simulated* data from estimated known network



Another experiment

We used the real data to estimate (OLS) the structural equation model corresponding to the assumed DAG structure.

Then simulated from the estimated model.

The *rationale* behind this experiment is to be faithful both to the actual data *and* to the assumed graphical structure.

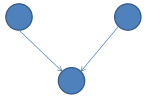
Fractional BF now performs much better.

Discussion

Discussion

Discussion

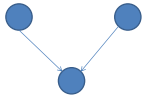
Bad news



Good news

Discussion

Bad news



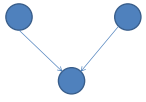
Good news

Discussion

Our method takes as input a fixed ordering of the variables.
What happens if the ordering is mis-specified?
Can the Fractional BF recover the skeleton of a DAG?
How does it compare with methods not requiring the ordering of the variables?
(notably the PC-algorithm)

Bad news

The performance depends crucially on the number of v-structures
As this number increases, the performance of our method deteriorates



Good news

The good news is that sparse graphs have very few v-structures

Main references

Altomare, D., Consonni, G. and La Rocca, L. (2012).

Objective Bayesian search of Gaussian DAG models with non-local priors.
Biometrics **69,** 478–487.

Main references

Altomare, D., Consonni, G. and La Rocca, L. (2012).

Objective Bayesian search of Gaussian DAG models with non-local priors.
Biometrics **69,** 478–487.

Consonni, G. and La Rocca, L. (2011).

Moment priors for Bayesian model choice with applications to directed acyclic graphs.
In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 9 – Proceedings of the Ninth Valencia International Meeting*, pages 119–144. Oxford University Press.

Main references

Altomare, D., Consonni, G. and La Rocca, L. (2012).

Objective Bayesian search of Gaussian DAG models with non-local priors.
Biometrics **69,** 478–487.

Consonni, G. and La Rocca, L. (2011).

Moment priors for Bayesian model choice with applications to directed acyclic graphs.
In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 9 – Proceedings of the Ninth Valencia International Meeting*, pages 119–144. Oxford University Press.

Drton, M. and Perlman, M. D. (2008).

A SINful approach to Gaussian graphical model selection.
J. Statist. Plann. Inference **138,** 1179–1200.

Main references

Altomare, D., Consonni, G. and La Rocca, L. (2012).

Objective Bayesian search of Gaussian DAG models with non-local priors.
Biometrics **69,** 478–487.

Consonni, G. and La Rocca, L. (2011).

Moment priors for Bayesian model choice with applications to directed acyclic graphs.
In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 9 – Proceedings of the Ninth Valencia International Meeting*, pages 119–144. Oxford University Press.

Drton, M. and Perlman, M. D. (2008).

A SINful approach to Gaussian graphical model selection.
J. Statist. Plann. Inference **138,** 1179–1200.

Friedman, N. and Koller, D. (2003).

Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks.
Machine Learning **50,** 95–125.

Main references

Altomare, D., Consonni, G. and La Rocca, L. (2012).

Objective Bayesian search of Gaussian DAG models with non-local priors.
Biometrics **69,** 478–487.

Consonni, G. and La Rocca, L. (2011).

Moment priors for Bayesian model choice with applications to directed acyclic graphs.
In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 9 – Proceedings of the Ninth Valencia International Meeting*, pages 119–144. Oxford University Press.

Drton, M. and Perlman, M. D. (2008).

A SINful approach to Gaussian graphical model selection.
J. Statist. Plann. Inference **138,** 1179–1200.

Friedman, N. and Koller, D. (2003).

Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks.
Machine Learning **50,** 95–125.

Johnson, V. and Rossell, D. (2010).

On the use of non-local prior densities in Bayesian hypotesis tests.
Journal of the Royal Statistical Society, Series B **72,** 143–170.

Main references

Kalisch, M. and Buhlmann, P. (2007).

Estimating high-dimensional directed acyclic graphs with the PC-algorithm.
J. Mach. Learn. Res. **8,** 613–36.

Main references

Kalisch, M. and Buhlmann, P. (2007).

Estimating high-dimensional directed acyclic graphs with the PC-algorithm.
J. Mach. Learn. Res. **8,** 613–36.

O'Hagan, A. (1995).

Fractional Bayes factors for model comparison.
Journal of the Royal Statistical Society. Series B (Methodological) **57,** 99–138.

Main references

Kalisch, M. and Buhlmann, P. (2007).

Estimating high-dimensional directed acyclic graphs with the PC-algorithm.
J. Mach. Learn. Res. **8,** 613–36.

O'Hagan, A. (1995).

Fractional Bayes factors for model comparison.
Journal of the Royal Statistical Society. Series B (Methodological) **57,** 99–138.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2003).

Casual protein-signaling networks derived from multiparameter single-cell data.
Science **308,** 504–6.

Main references

Kalisch, M. and Buhlmann, P. (2007).

Estimating high-dimensional directed acyclic graphs with the PC-algorithm.
J. Mach. Learn. Res. **8,** 613–36.

O'Hagan, A. (1995).

Fractional Bayes factors for model comparison.
Journal of the Royal Statistical Society. Series B (Methodological) **57,** 99–138.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2003).

Casual protein-signaling networks derived from multiparameter single-cell data.
Science **308,** 504–6.

Scott, J. G. and Berger, J. O. (2010).

Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem.
The Annals of Statistics **38,** 2587–2619.

Main references

Kalisch, M. and Buhlmann, P. (2007).

Estimating high-dimensional directed acyclic graphs with the PC-algorithm.
J. Mach. Learn. Res. **8,** 613–36.

O'Hagan, A. (1995).

Fractional Bayes factors for model comparison.
Journal of the Royal Statistical Society. Series B (Methodological) **57,** 99–138.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2003).

Casual protein-signaling networks derived from multiparameter single-cell data.
Science **308,** 504–6.

Scott, J. G. and Berger, J. O. (2010).

Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem.
The Annals of Statistics **38,** 2587–2619.

Shojaie, A. and Michailidis, G. (2010).

Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs.
Biometrika **97,** 519–538.

Local priors

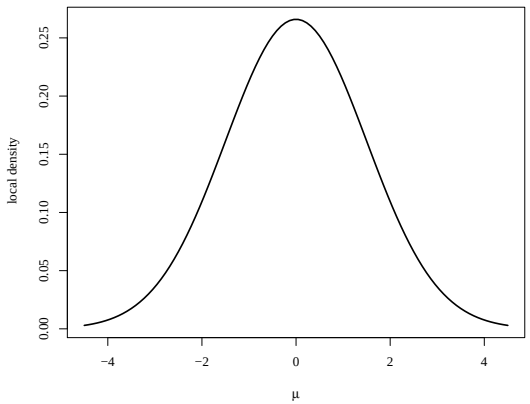
Local priors

Local priors

$\mathcal{M}_0$ *nested* in $\mathcal{M}_1$
$\Theta_0 \subset \Theta_1$
$d_0 = \dim(\Theta_0) < d_1 = \dim(\Theta_1)$

$p(\theta_1)$, $\theta_1 \in \Theta_1$, a *local* prior
continuous, and strictly positive over Θ_0

$\mathcal{M}_0 : N(0,1)$; $\mathcal{M}_1 : N(\mu,1)$, $\mu \neq 0$
$p_1(\mu) = N(\mu \,|\, 0, (1.5)^2)$

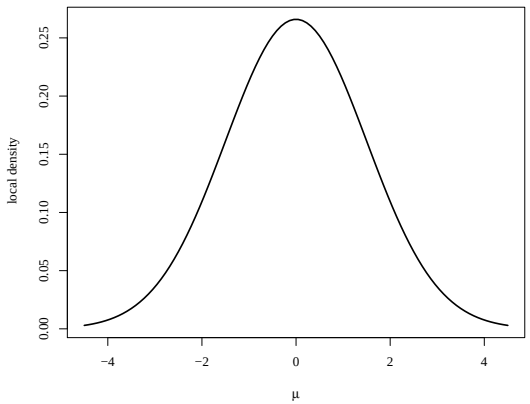


Local priors

$\mathcal{M}_0$ *nested* in $\mathcal{M}_1$
$\Theta_0 \subset \Theta_1$
$d_0 = \dim(\Theta_0) < d_1 = \dim(\Theta_1)$

$p(\theta_1)$, $\theta_1 \in \Theta_1$, a *local* prior
continuous, and strictly positive over Θ_0

$\mathcal{M}_0 : N(0,1)$; $\mathcal{M}_1 : N(\mu,1)$, $\mu \neq 0$
$p_1(\mu) = N(\mu \,|\, 0, (1.5)^2)$

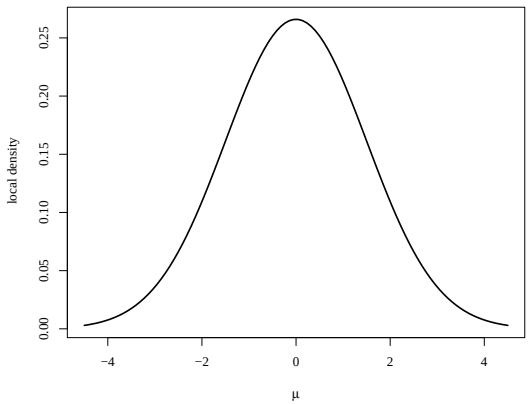


Local priors

$\mathcal{M}_0$ *nested* in $\mathcal{M}_1$
$\Theta_0 \subset \Theta_1$
$d_0 = \dim(\Theta_0) < d_1 = \dim(\Theta_1)$

$p(\theta_1)$, $\theta_1 \in \Theta_1$, a *local* prior
continuous, and strictly positive over Θ_0

$\mathcal{M}_0 : N(0,1)$; $\mathcal{M}_1 : N(\mu,1)$, $\mu \neq 0$
$p_1(\mu) = N(\mu \mid 0, (1.5)^2)$



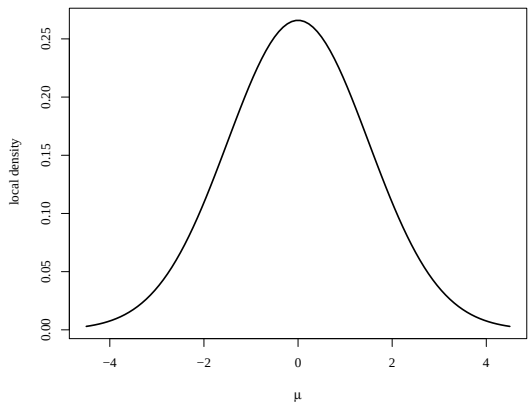
Data $y^{(n)} = (y_1, \ldots, y_n)$
i.i.d. sample from (unknown) distribution

Local priors

$\mathcal{M}_0$ *nested* in $\mathcal{M}_1$
$\Theta_0 \subset \Theta_1$
$d_0 = \dim(\Theta_0) < d_1 = \dim(\Theta_1)$

$p(\theta_1)$, $\theta_1 \in \Theta_1$, a *local* prior
continuous, and strictly positive over Θ_0

$\mathcal{M}_0 : N(0,1)$; $\mathcal{M}_1 : N(\mu,1)$, $\mu \neq 0$
$p_1(\mu) = N(\mu \mid 0, (1.5)^2)$



Data $y^{(n)} = (y_1, \ldots, y_n)$
i.i.d. sample from (unknown) distribution

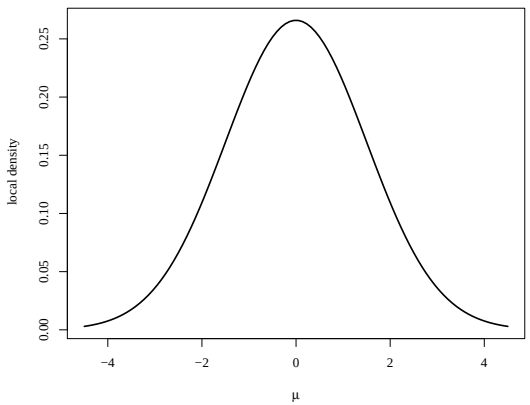
- if $\mathcal{M}_0$ holds
  $BF_{10}(y^{(n)}) = O_p(n^{-(d_1-d_0)/2})$

Local priors

$\mathcal{M}_0$ *nested* in $\mathcal{M}_1$
$\Theta_0 \subset \Theta_1$
$d_0 = \dim(\Theta_0) < d_1 = \dim(\Theta_1)$

$p(\theta_1)$, $\theta_1 \in \Theta_1$, a *local* prior
continuous, and strictly positive over Θ_0

$\mathcal{M}_0 : N(0,1)$; $\mathcal{M}_1 : N(\mu,1)$, $\mu \neq 0$
$p_1(\mu) = N(\mu \,|\, 0, (1.5)^2)$



Data $y^{(n)} = (y_1, \ldots, y_n)$
i.i.d. sample from (unknown) distribution

- if $\mathcal{M}_0$ holds
  $BF_{10}(y^{(n)}) = O_p(n^{-(d_1-d_0)/2})$
- if $\mathcal{M}_1$ holds $BF_{01}(y^{(n)}) = e^{-Kn+O_p(\sqrt{n})}$, for some $K > 0$

  Imbalance in learning rate

Non-local priors

$g(\theta_1)$, $\theta_1 \in \Theta_1$: continuous positive function vanishing on Θ_0.
For given local prior $p(\theta_1)$
define a new non-local prior as $p^M(\theta_1) \propto g(\theta_1)p(\theta_1)$,

Non-local priors

$g(\theta_1)$, $\theta_1 \in \Theta_1$: continuous positive function vanishing on Θ_0.
For given local prior $p(\theta_1)$
define a new non-local prior as $p^M(\theta_1) \propto g(\theta_1)p(\theta_1)$,

Example

θ_1 *a scalar parameter in* $\mathbb{R}$
$\Theta_0 = \{\theta_0\}$, *with* θ_0 *a fixed value*
$g(\theta_1) = (\theta_1 - \theta_0)^{2h}$ *h a positive integer*
moment prior (Johnson and Rossell, 2010)
If $\mathcal{M}_0$ *holds,* $BF_{10}(y^{(n)}) = O_p(n^{-h-1/2})$

For instance if $h = 1$, *the learning rate changes from sub-linear*
$BF_{10}(y^{(n)}) = O_p(n^{-1/2})$
to super-linear
$BF_{10}(y^{(n)}) = O_p(n^{-1-1/2})$

Gaussian model: testing a sharp null hypothesis

Gaussian model: testing a sharp null hypothesis

Gaussian model: testing a sharp null hypothesis

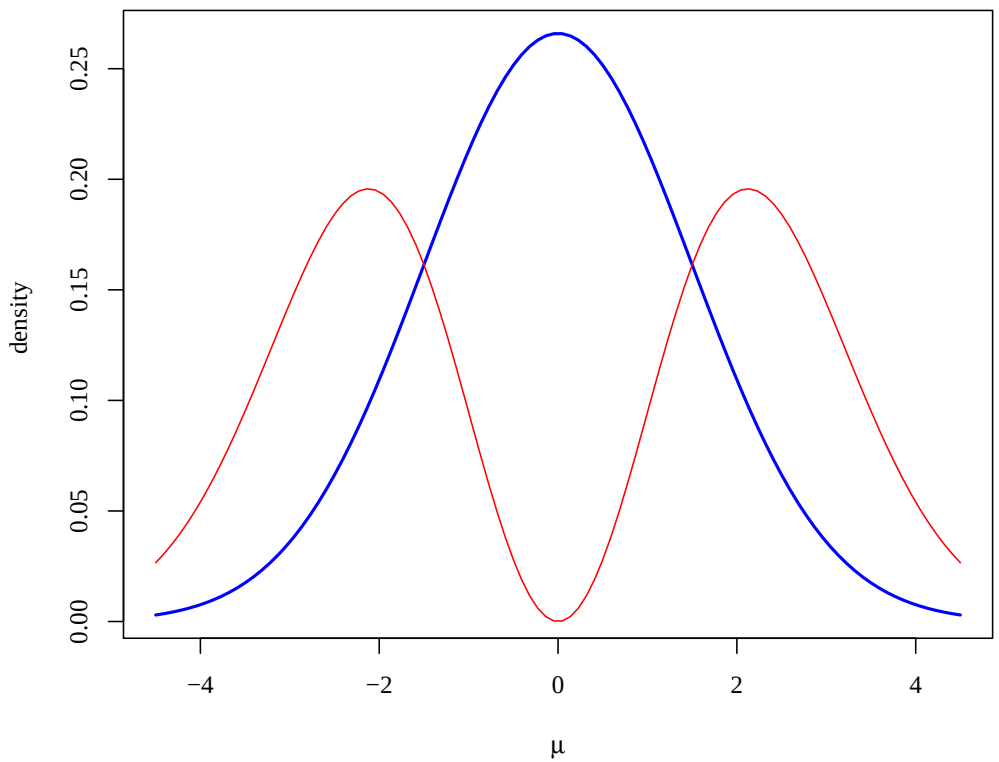
$\mathcal{M}_0 : N(0,1); \mathcal{M}_1 : N(\mu,1), \mu \neq 0$

Local prior: $p_1(\mu) = N(\mu \mid 0, \sigma_\mu^2 = (1.5)^2)$

Nonlocal (moment) prior:

$p_1^M(\mu) \propto \mu^{2h} N(\mu \mid 0, \sigma_\mu^2 = (1.5)^2)$

$h = 1$



Gaussian model: testing a sharp null hypothesis

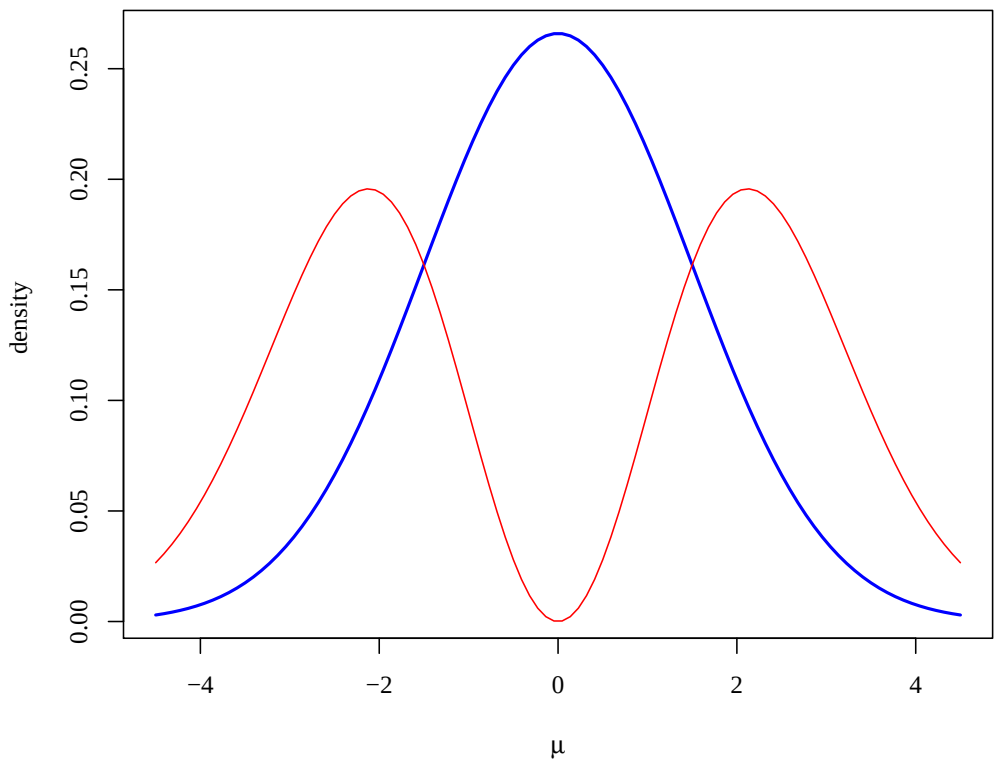
$\mathcal{M}_0 : N(0,1); \mathcal{M}_1 : N(\mu,1), \mu \neq 0$

Local prior: $p_1(\mu) = N(\mu \mid 0, \sigma_\mu^2 = (1.5)^2)$

Nonlocal (moment) prior:

$p_1^M(\mu) \propto \mu^{2h} N(\mu \mid 0, \sigma_\mu^2 = (1.5)^2)$

$h = 1$



Gaussian model: testing a sharp null hypothesis

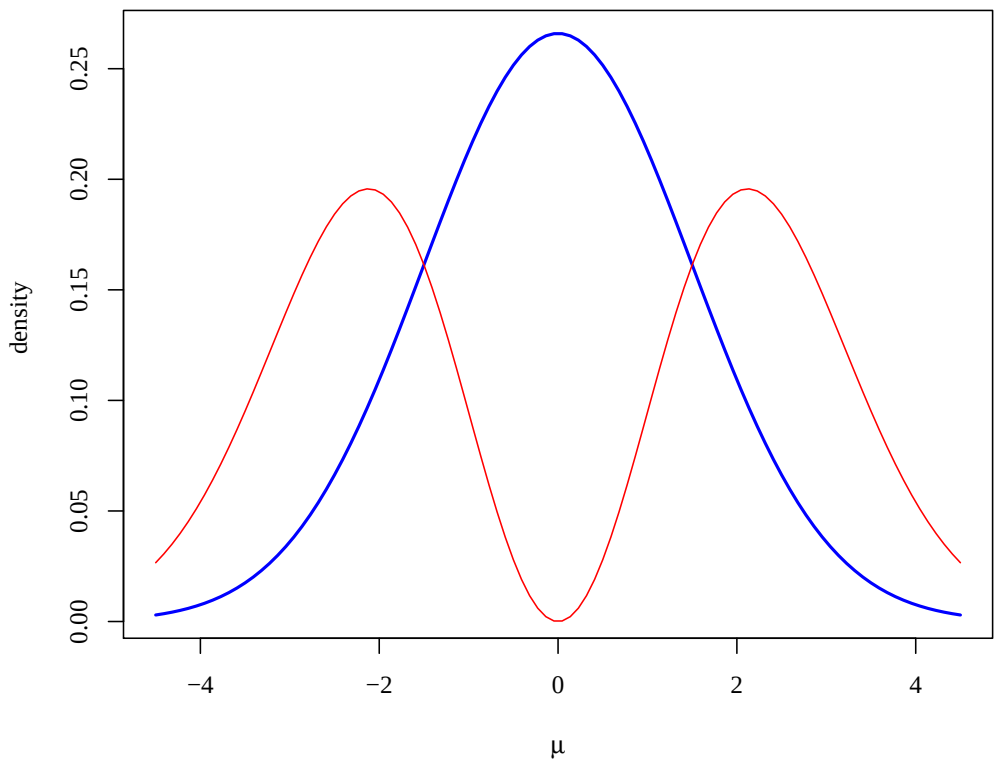
$\mathcal{M}_0 : N(0,1); \mathcal{M}_1 : N(\mu,1), \mu \neq 0$

Local prior: $p_1(\mu) = N(\mu \mid 0, \sigma_\mu^2 = (1.5)^2)$

Nonlocal (moment) prior:

$p_1^M(\mu) \propto \mu^{2h} N(\mu \mid 0, \sigma_\mu^2 = (1.5)^2)$

$h = 1$



Choice of h and σ_μ^2 determines the degree of *separation* between the two models

this can be done subjectively in some ideal situations

but in many situations we must resort to some objective procedure

Sensitivity to order mis-specification

Moment Fractional BF requires an ordering of the variables

Can the Fractional BF recover the skeleton of a DAG?
How does it compare with methods not requiring the ordering of the variables?
(notably the PC-algorithm by Kalish and Bühlman, 2007)

What is the tolerated "distance", based on the number of inversions in a permutation, between the true ordering and the one assumed by our method for a good performance?

Sensitivity to order mis-specification

Moment Fractional BF requires an ordering of the variables

Can the Fractional BF recover the skeleton of a DAG?
How does it compare with methods not requiring the ordering of the variables?
(notably the PC-algorithm by Kalish and Bühlman, 2007)

What is the tolerated "distance", based on the number of inversions in a permutation, between the true ordering and the one assumed by our method for a good performance?

Sensitivity to order mis-specification

Moment Fractional BF requires an ordering of the variables

Can the Fractional BF recover the skeleton of a DAG?
How does it compare with methods not requiring the ordering of the variables?
(notably the PC-algorithm by Kalish and Bühlman, 2007)

What is the tolerated "distance", based on the number of inversions in a permutation, between the true ordering and the one assumed by our method for a good performance?
$0 < d < 1$*: relative distance of permutation from the true one*

Sensitivity to order mis-specification

Moment Fractional BF requires an ordering of the variables

Can the Fractional BF recover the skeleton of a DAG?
How does it compare with methods not requiring the ordering of the variables?
(notably the PC-algorithm by Kalish and Bühlman, 2007)

What is the tolerated "distance", based on the number of inversions in a permutation, between the true ordering and the one assumed by our method for a good performance?
$0 < d < 1$*: relative distance of permutation from the true one*
Fractional BF search outperforms the PC-algorithm when $d = 0$.
It is outperformed when $d = 1$.

Up to a moderate mis-specification ($d = 0.25$*) it is comparable*

A measure of distance between permutations

Ordered sequence $1, 2, \ldots, n$
(identity permutation)

Permutation $\pi(1), \pi(2), \ldots, \pi(n)$

A pair $(\pi(i), \pi(j))$ *is called an inversion in* π
if $i > j$ *and* $\pi(i) < \pi(j)$

The number of (#) inversions assesses how far the permutation is from the naturally ordered sequence

$\pi_{\max}$*: reversed identity sequence*

relative distance $d \in [0, 1]$

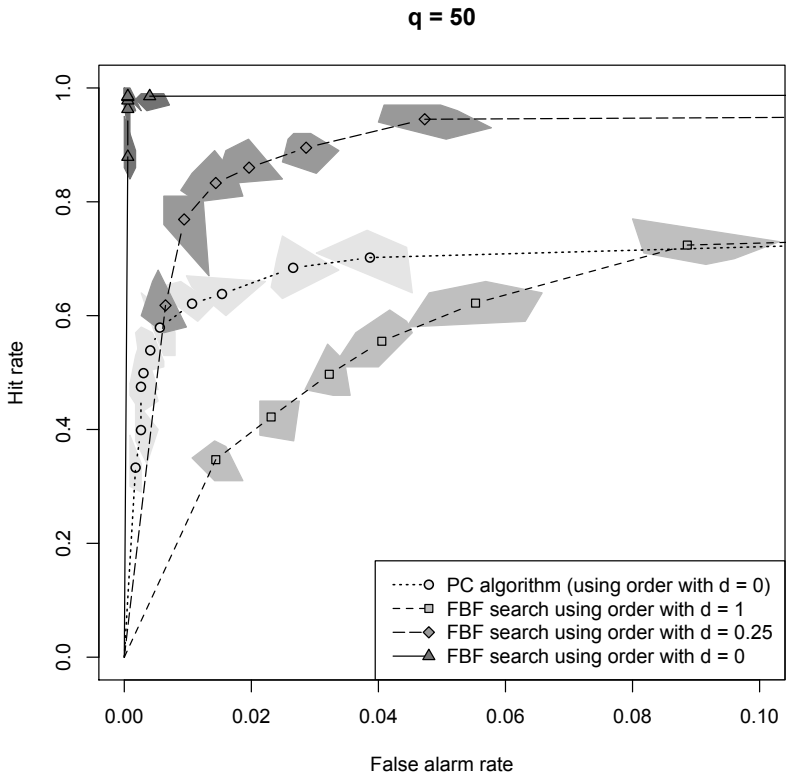
$$d = \#\textit{inversions in } \pi / (\#\textit{inversions in } \pi_{\max})$$

Simulated data from sparse DAGs

Simulated data from sparse DAGs

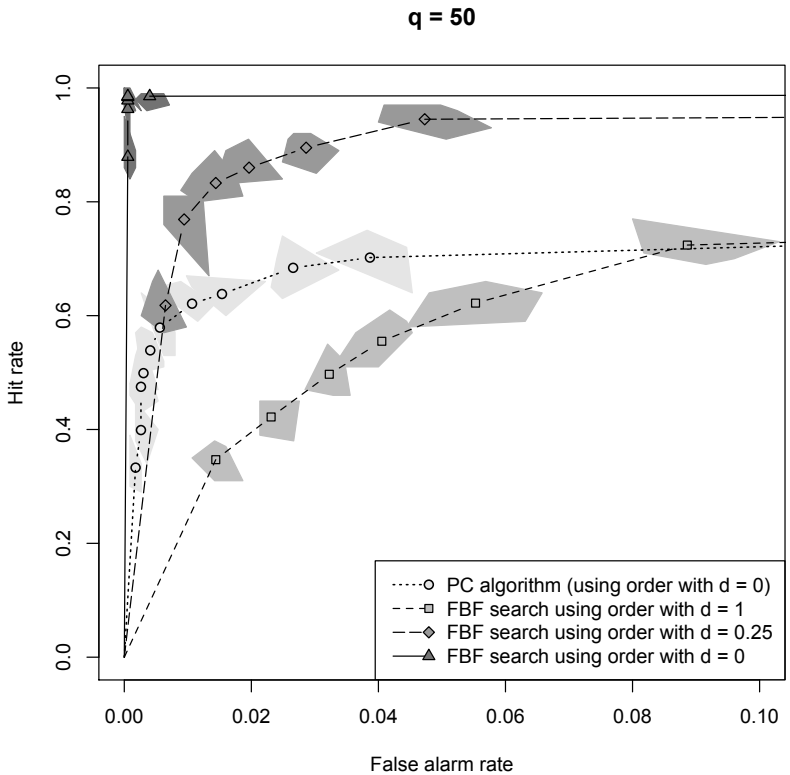
Simulated data from sparse DAGs

ROC curves $q = 50$



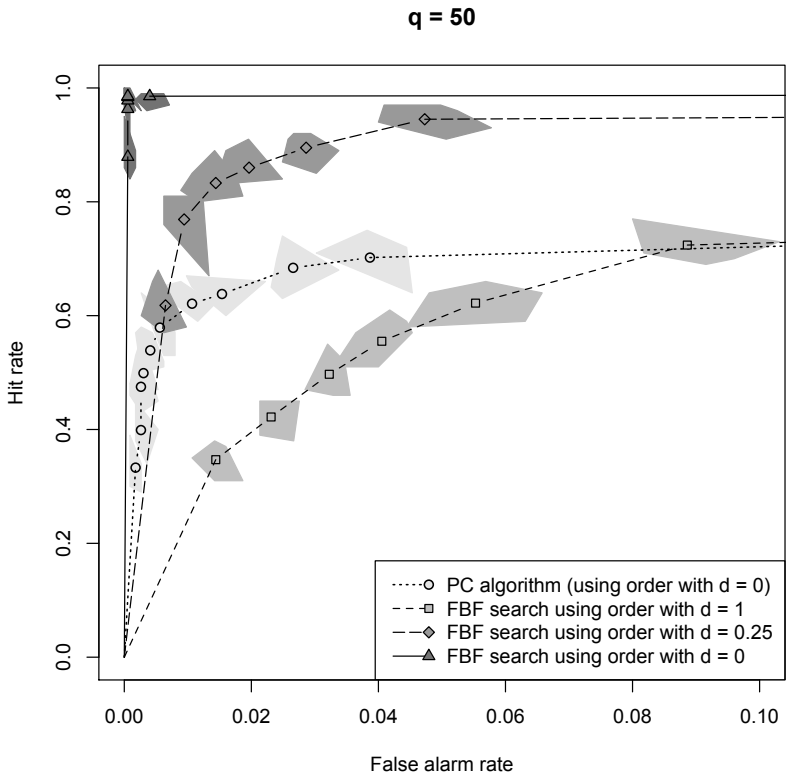
Simulated data from sparse DAGs

ROC curves $q = 50$



Simulated data from sparse DAGs

ROC curves $q = 50$



Simulated data from sparse DAGs

Methods: PC-algorithm and Fractional BF

Order mis-specification
$d = 0$: null
$d = 0.25$: moderate
$d = 1$: max

Fractional BF search outperforms the PC-algorithm when $d = 0$.
It is outperformed when $d = 1$.

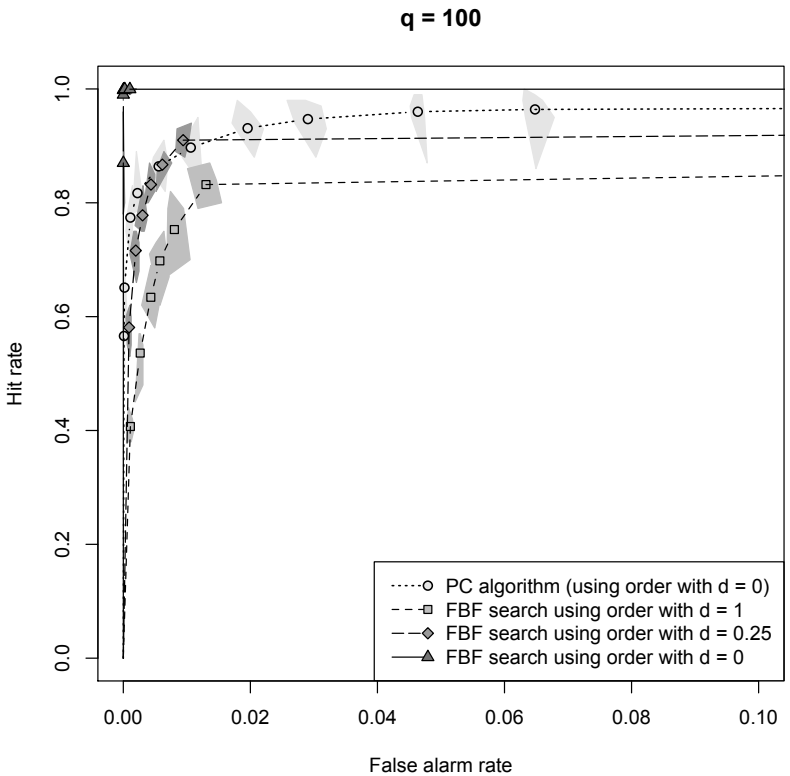
Up to a moderate mis-specification ($d = 0.25$)
FBF search outperforms the PC-algorithm

Simulated data from sparse DAGs

Simulated data from sparse DAGs

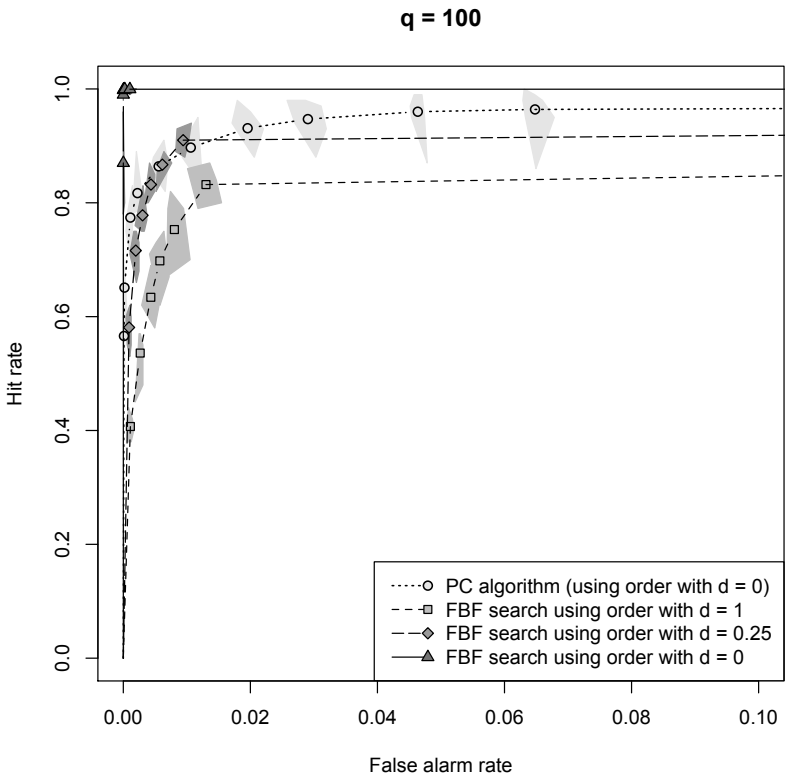
Simulated data from sparse DAGs

ROC curves $q = 100$



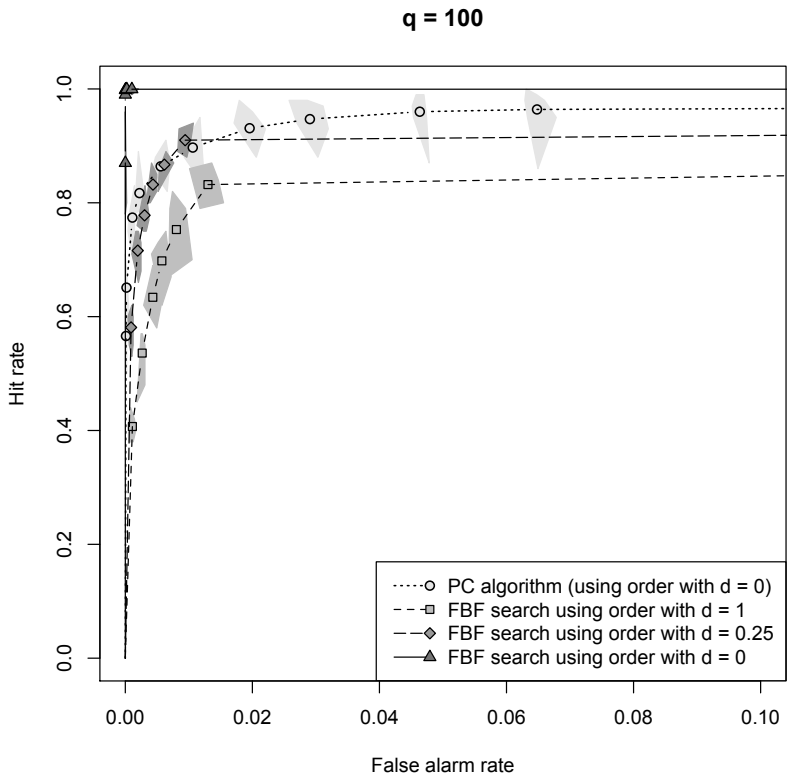
Simulated data from sparse DAGs

ROC curves $q = 100$



Simulated data from sparse DAGs

ROC curves $q = 100$



Methods: PC-algorithm and Fractional BF

Fractional BF search outperforms the PC-algorithm when $d = 0$.
It is outperformed when $d = 1$.

Performance of the two methods when $d = 0.25$ is now comparable.