Intelligent data analysis
Biomarker discovery II.

Peter Antal antal@mit.bme.hu

Overview

- Biomarkers
  - Endophenotypes
- The Bayesian statistical approach
- Partial multivariate analysis
  - Sub, sup, Ockham
  - Frontlines
  - clustering
- Causal, confounded extension
- Multivariate (multidimensional)extension
- Reporting
- Interpreting
- Genagrid
  - Goals
  - Status
- BayesEye
  - Preprocessing

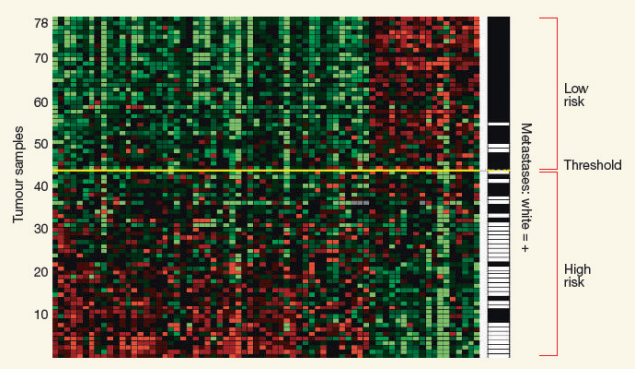
Biomarker challenges in biomedicine

- Better outcome variable
  - „Lost in diagnosis”: phenome
- Better and more complete set of predictor variables
  - „Right under everyone’s noses”: rare variants (RVs)
  - „The great beyond”: Epigenetics, environment
- Better statistical models
  - „In the architecture”: structural variations
  - „Out of sight”: many, small effects
  - „In underground networks”: epistatic interactions

- **Causation (confounding)**
- **Statistical significance („multiple testing problem”)**
- **Complex models: interactions, epistatis**
- **Interpretation**

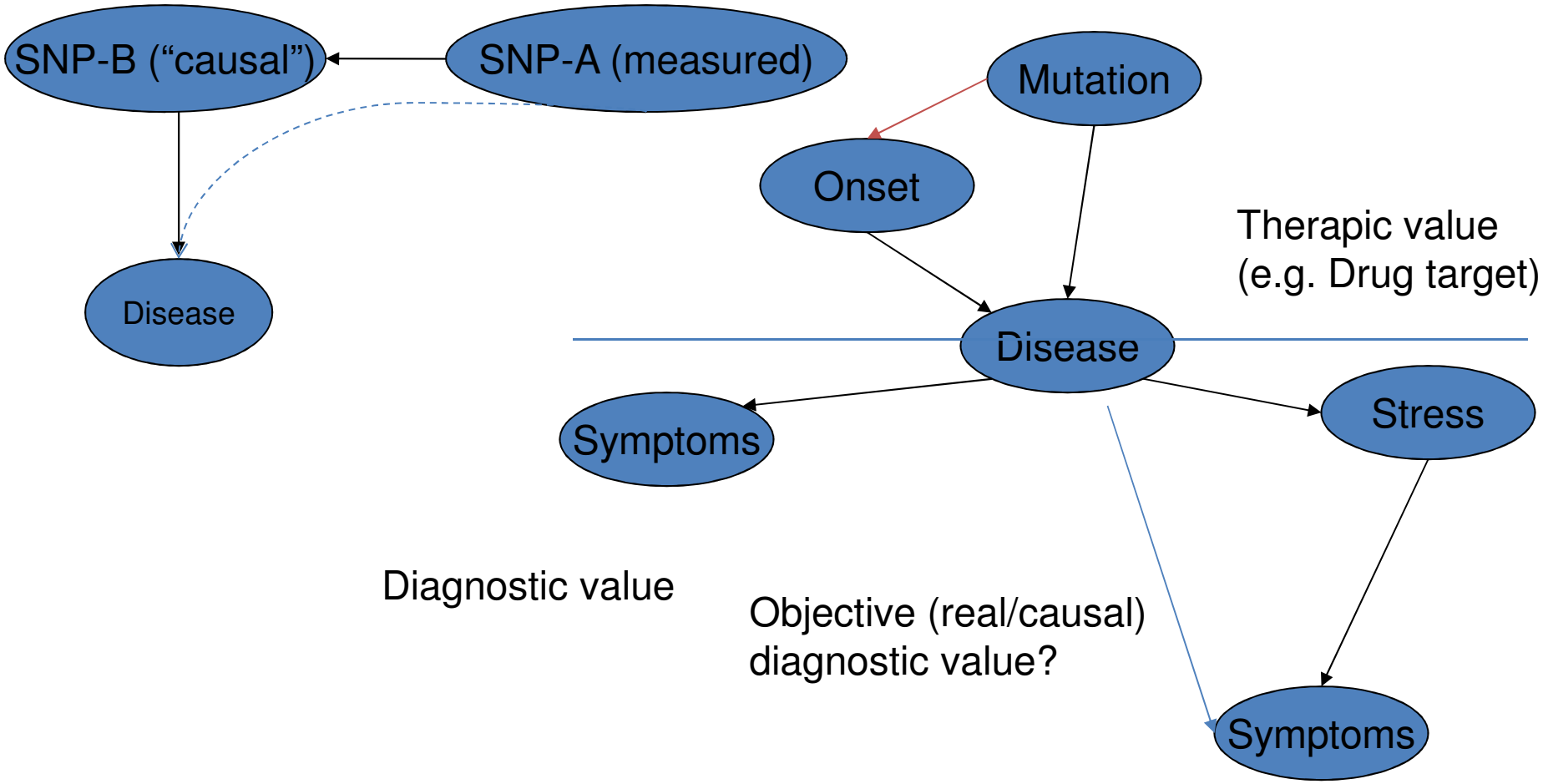
CA125 and its pretenders

- CA125 is very indicative tumor marker in cancer.
- Missing the mark, 2007, Nature

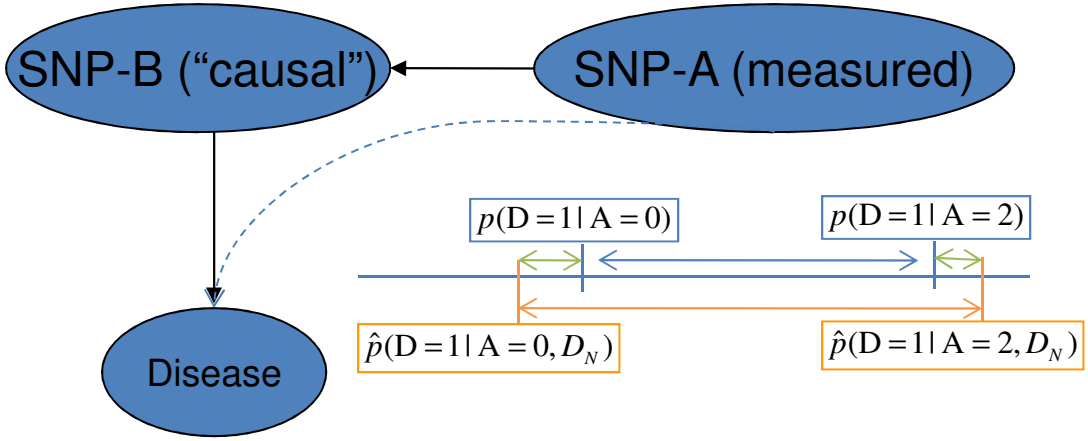


- MISSING THE MARK: *Why is it so hard to find a test to predict cancer?, 2011, March*
  - Series of proteomic kits... with poor performance.
  - The „prolactin” case

Causal vs. diagnostic markers
Direct =/= Causal



Fundamental questions in statistics



Real difference

Estimation errors

Estimated difference

Estimation error because of finite data D_N: $\hat{p}(\mathrm{D}=1|\mathrm{A}=0,D_N) - p(\mathrm{D}=1|\mathrm{A}=0)$

Inequalities for finite(!) data (ε accuracy,δ confidence)

sample complexity: $N_{\varepsilon,\delta}$ $p(D_{N_{\varepsilon,\delta}} : \varepsilon < |\hat{p}(\mathrm{D}=1|\mathrm{A},D_{N\varepsilon,\delta}) - p(\mathrm{D}=1|\mathrm{A})|) < \delta$

The hypothesis testing framework

- Terminology:
  - False/true x positive/negative
  - Null hypothesis: independence

reported	Ref.:0/N	Ref.1/P
0/N	TN	FN
1/P	FP	TP

  - Type I error/error of the first kind/α error/FP: $p(\neg H_0 | \underline{H_0})$
    - Specificity: $p(H_0 | \underline{H_0}) = 1-\alpha$
    - Significance: α
    - p-value: „probability of more extreme observations in repeated experiments”
  - Type II error/error of the second kind/β error/FN: $p(H_0 | \neg \underline{H_0})$:
    - Power or sensitivity: $p(\neg H_0 | \neg \underline{H_0}) = 1-\beta$

reported	Ref. $\underline{H_0}$	Ref.:$\neg \underline{H_0}$
H_0		Type II
$\neg H_0$	Type I („false rejection”)	

Multiple testing problem (MTP)

- If we perform N tests and our goal is
  - $p(\text{FalseRejection}_1 \text{ or } \ldots \text{ or } \text{FalseRejection}_N) < \alpha$
- then we have to ensure, e.g. that
  - for all $p(\text{FalseRejection}_i) < \alpha/N$

➔loss of power!

E.g. in a GWA study N=100,000, so huge amount of data is necessary….(but high-dimensional data is only relatively cheap!)

Solutions for MTP

- Corrections
- Permutation tests
  - Generate perturbed data sets under the null hypothesis: permute predictors and outcome.
- False discovery rate, q-value
- Bayesian approach

False discovery rate (FDR)

Another aspect of multiple hypothesis testing:

- the probability of Type I. error for any tests
- the expected number of Type I. errors at a given significance level (False discovery rate, FDR)
- q-value: the minimum FDR at which the test may be called significant.

Biomarkers and the feature subset selection (FSS) problem

A probabilistic concept of relevance

Definition 1. *A feature X_i is strongly relevant, if there exists some x_i, y and $s_i = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ for which $p(x_i, s_i) > 0$ such that $p(y|x_i, s_i) \neq p(y|s_i)$. A feature X_i is weakly relevant, if it is not strongly relevant, and there exists a subset of features S'_i of S_i for which there exists some x_i, y and s'_i for which $p(x_i, s'_i) > 0$ such that $p(y|x_i, s'_i) \neq p(y|s'_i)$. A feature is relevant, if it is either weakly or strongly relevant; otherwise it is irrelevant [7, 8].*

A graph-theoretic representation of relevance

Theorem 1 ([16]). *If distribution P is stable w.r.t. the DAG G, then the variables corresponding to the nodes in the boundary of Y, $\mathrm{bd}(Y, G)$ (the parents and children of Y and other parents of its children) forms a unique and minimal Markov blanket of Y, $\mathrm{MB}_P(Y)$ (the Markov boundary). Furthermore, $X_i \in \mathrm{MB}_P(Y)$, if X_i is strongly relevant.*

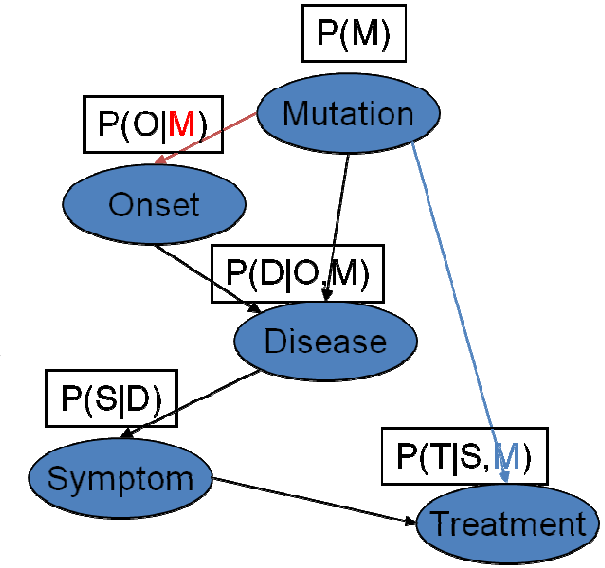
Bayesian networks

Directed acyclic graph (DAG)

- nodes – random variables/domain entities
- edges – direct probabilistic dependencies (edges- causal relations

Local models - $P(X_i | Pa(X_i))$

Three interpretations:



1. Causal model

2. Graphical representation of (in)dependencies

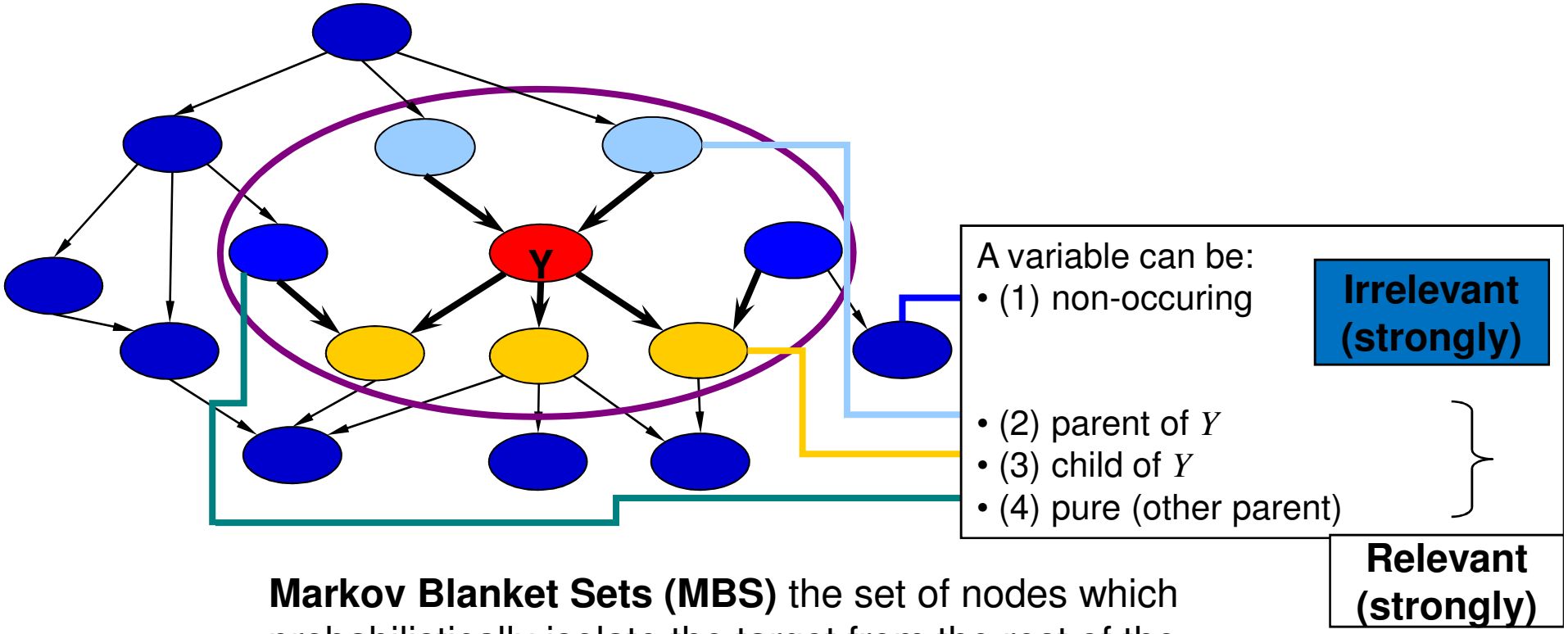
$M_P = \{I_{P,1}(X_1; Y_1 | Z_1), \ldots\}$

3. Concise representation of joint distributions

$$P(M, O, D, S, T) = P(M)P(O \mid M)P(D \mid O, M)P(S \mid D)P(T \mid S, M)$$

The Markov Blanket

A minimal sufficient set for prediction/diagnosis.



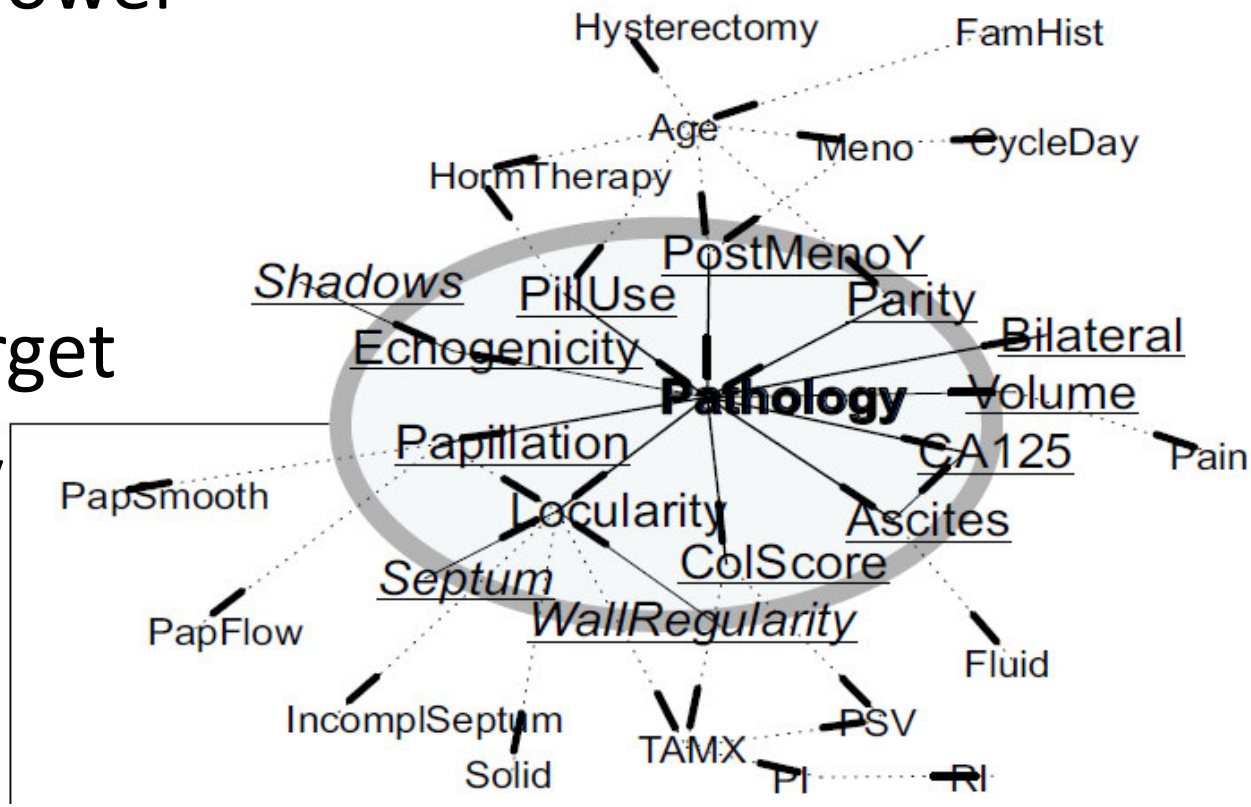
Markov Blanket Sets (MBS) the set of nodes which probabilistically isolate the target from the rest of the model

Markov Blanket Membership (MBM)
(symmetric) pairwise relationship induced by MBS

Aspects of biomarkers

„Maximum predictivity, minimum redundancy”

- Predictive power
- Directness
- Causality
- Multiple target
- Uncertainty



Bayes rule, Bayesianism

„all models are wrong, but some are useful”

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)}$$

A scientific research paradigm

$$p(Model \mid Data) \propto p(Data \mid Model)\, p(Model)$$

A practical method for inverting causal knowledge to diagnostic tool.

$$p(Cause \mid Effect) \propto p(Effect \mid Cause) \times p(Cause)$$

Bayesian prediction

In the frequentist approach: Model identification (selection) is necessary

$$p(prediction \mid data) = p(prediction \mid BestModel(data))$$

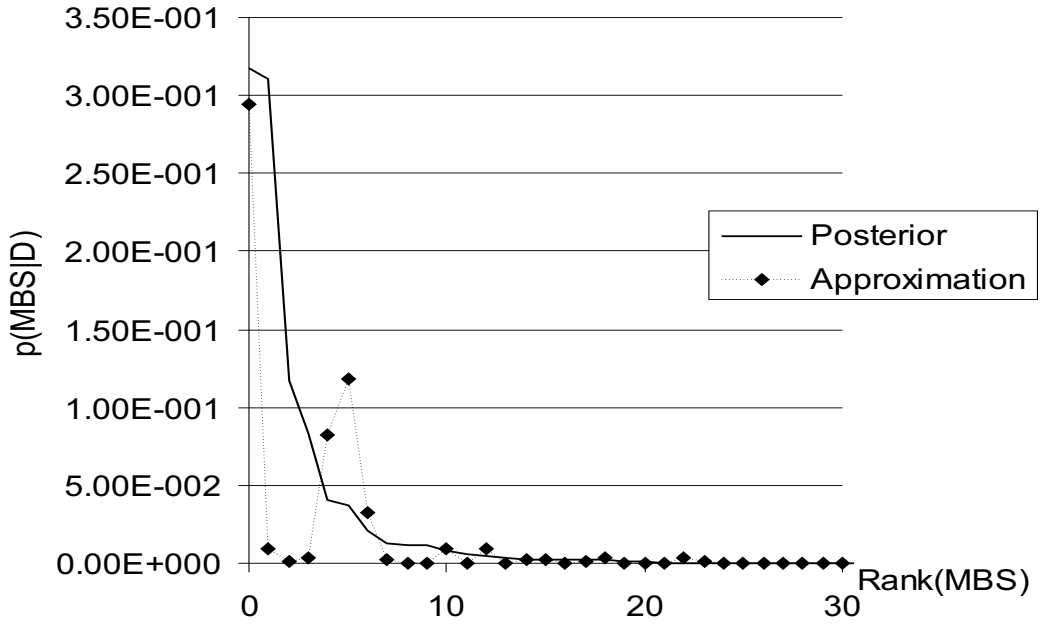
In the Bayesian approach models are weighted

$$p(prediction \mid data) = \sum_i p(pred. \mid Model_i)\, p(Model_i \mid data)$$

Note: in the Bayesian approach there is no need for model selection

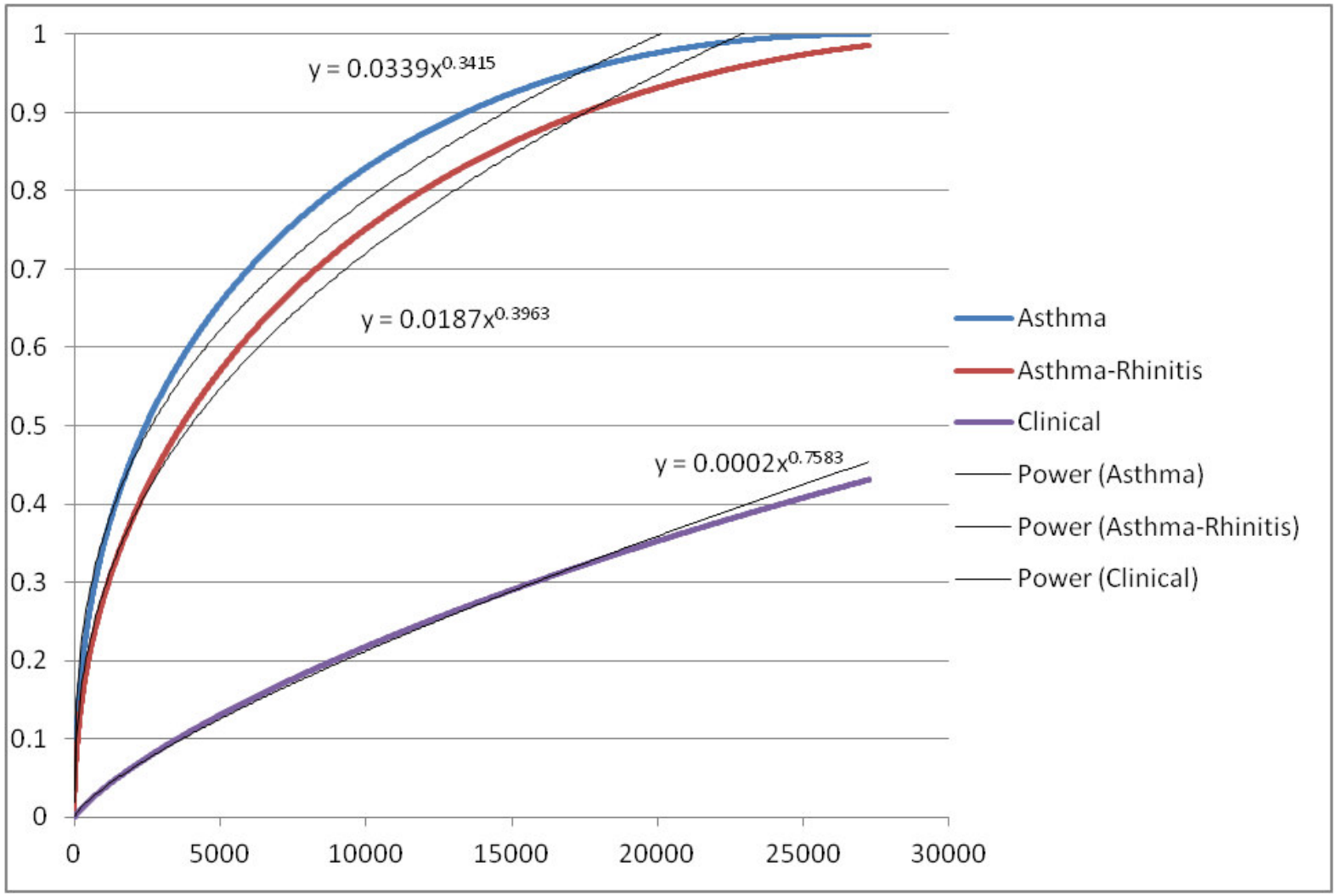
Posterior for complete sets

- High-level of uncertainty in multivariate analysis
- There are stable sub-parts (e.g., subset, subgraphs)
- Results for target variables and for certain SNPs could be aggregated

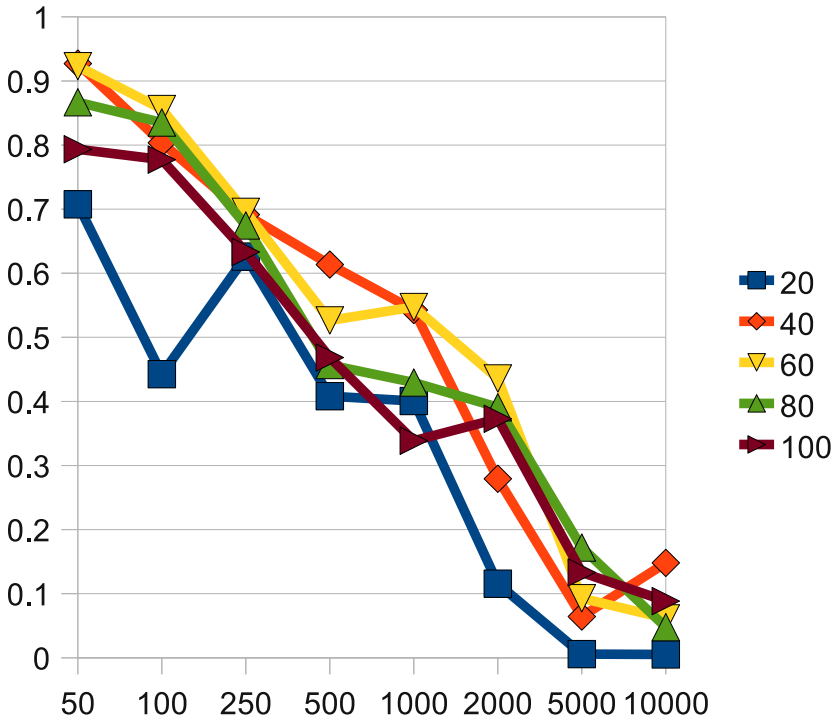


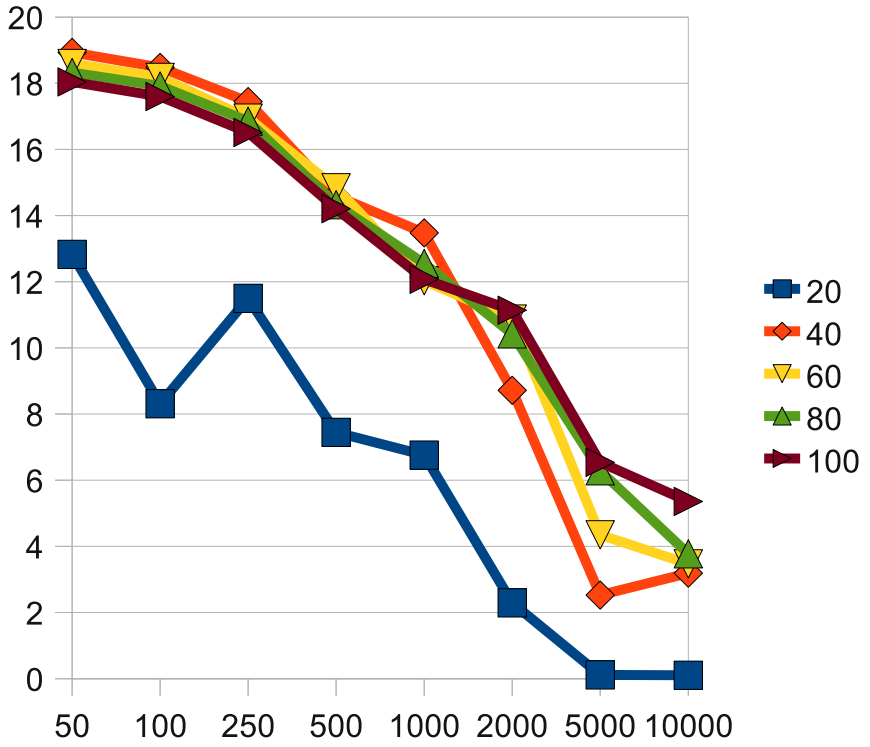
The peakness of the posteriors of the most probable MB sets and their MBM-based approximations.
(46 variables, 1000 samples)

Cumulative posterior of the most probable strongly relevant sets

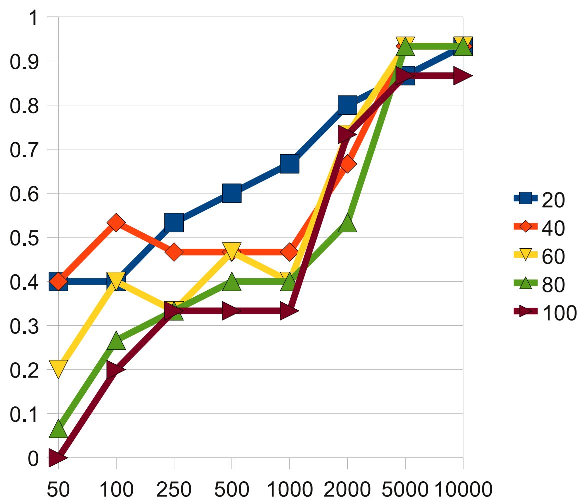


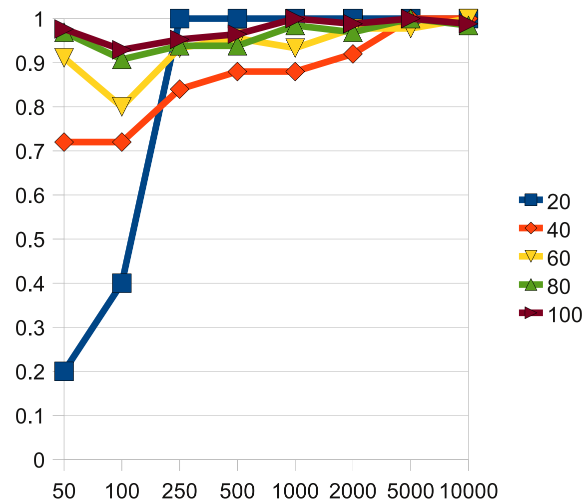
Learning rate of MBM and MBS (entropy)

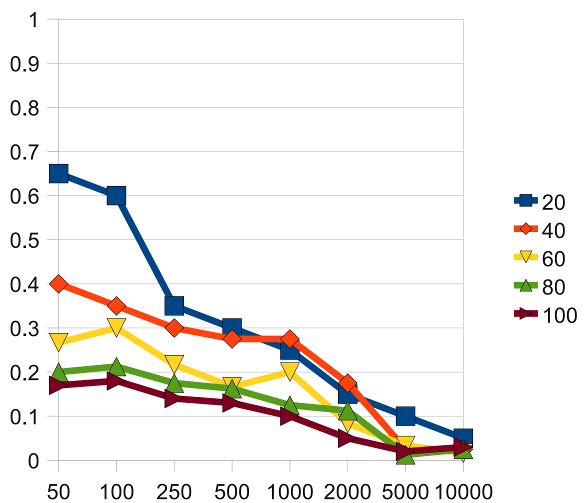


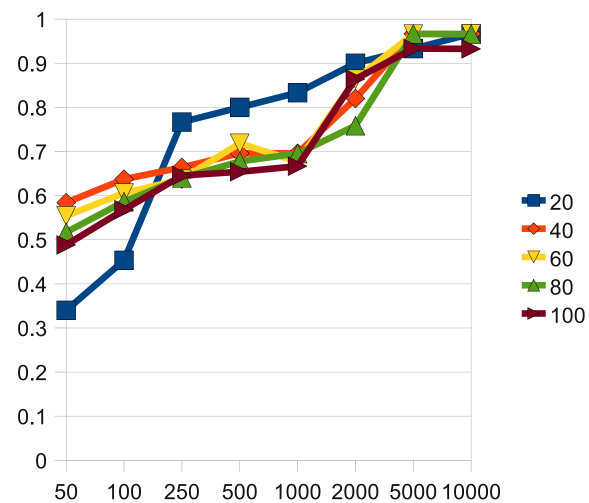


Learning rate of MBM and MBS (sens, spec, MR, AUC)

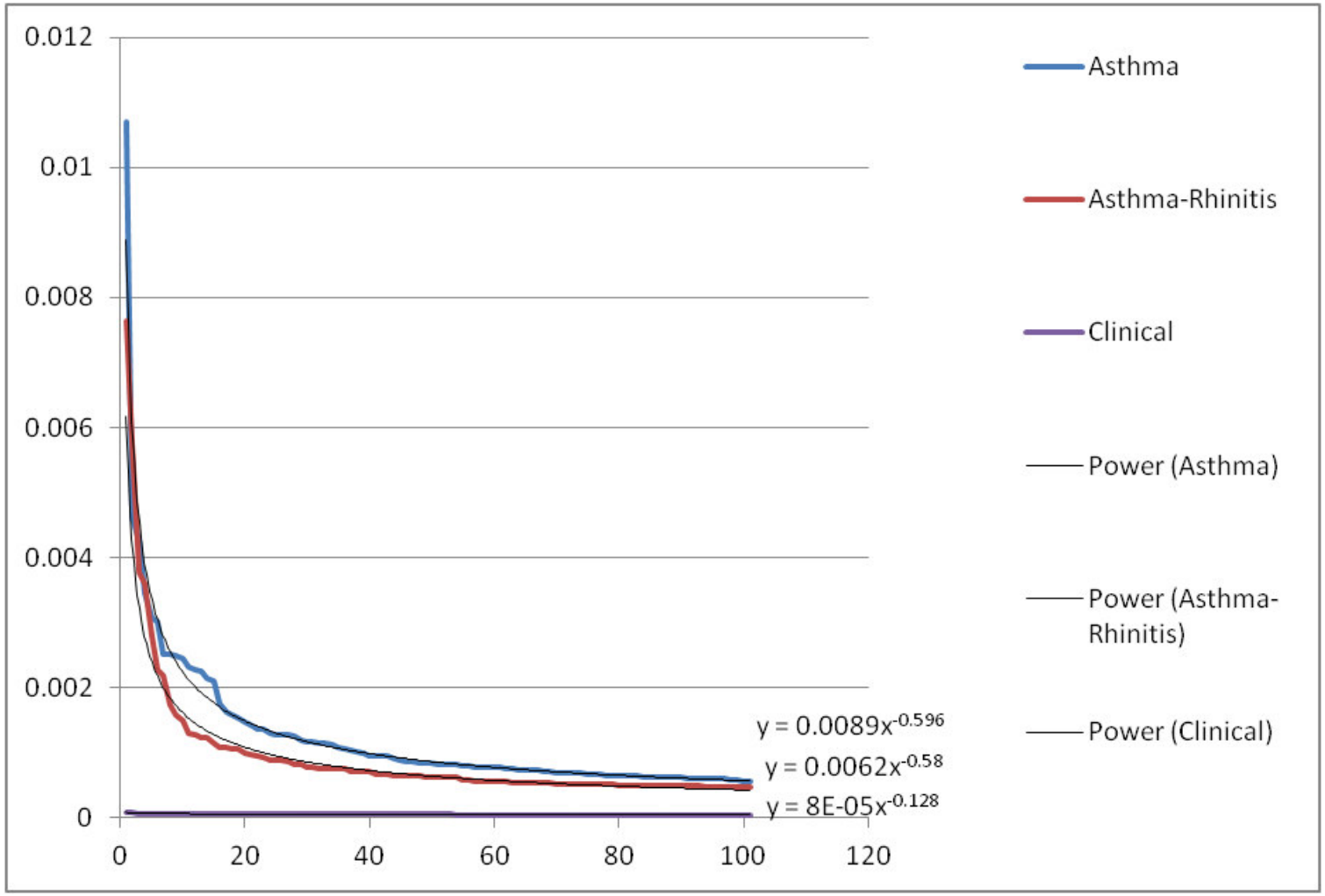




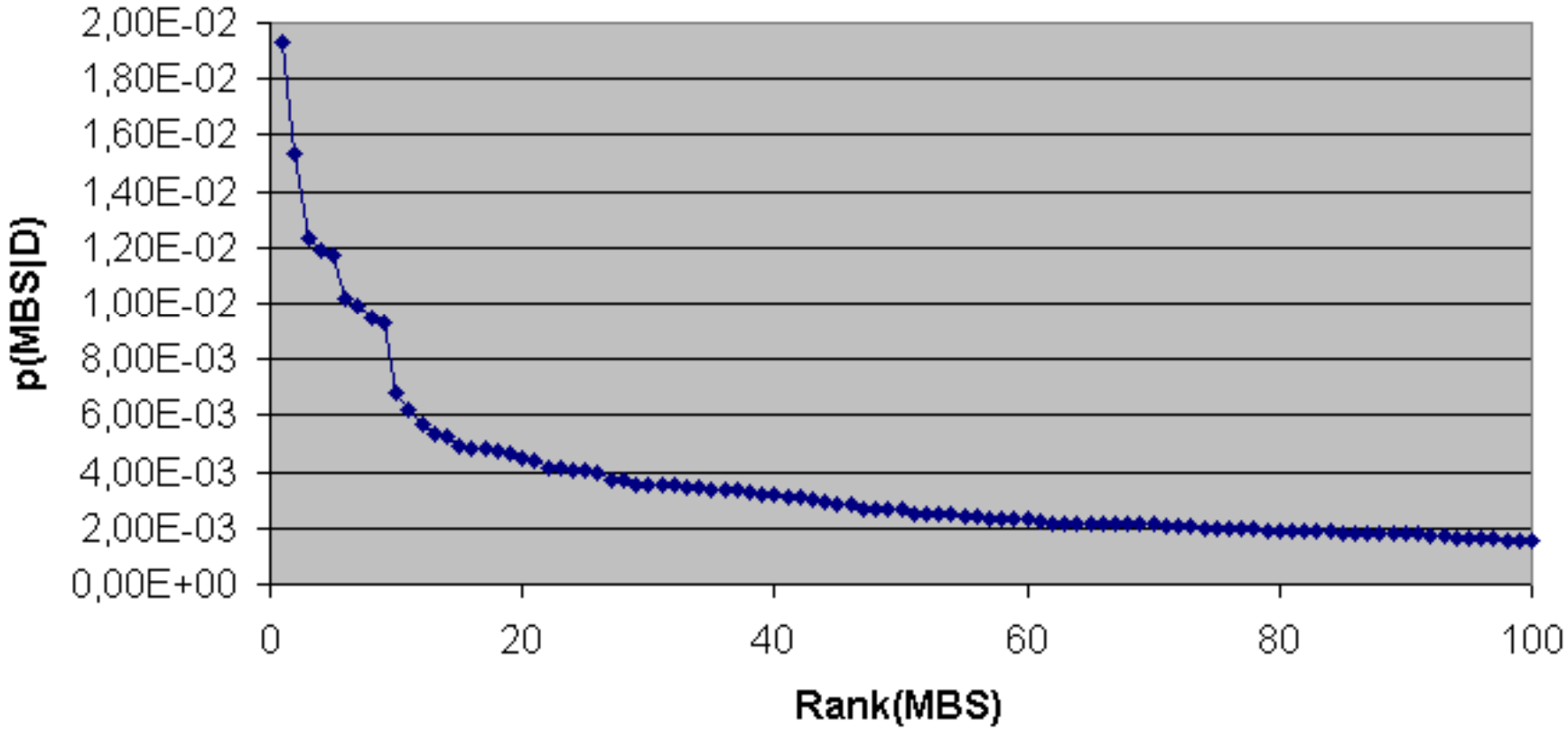




Posterior of the most probable strongly relevant sets

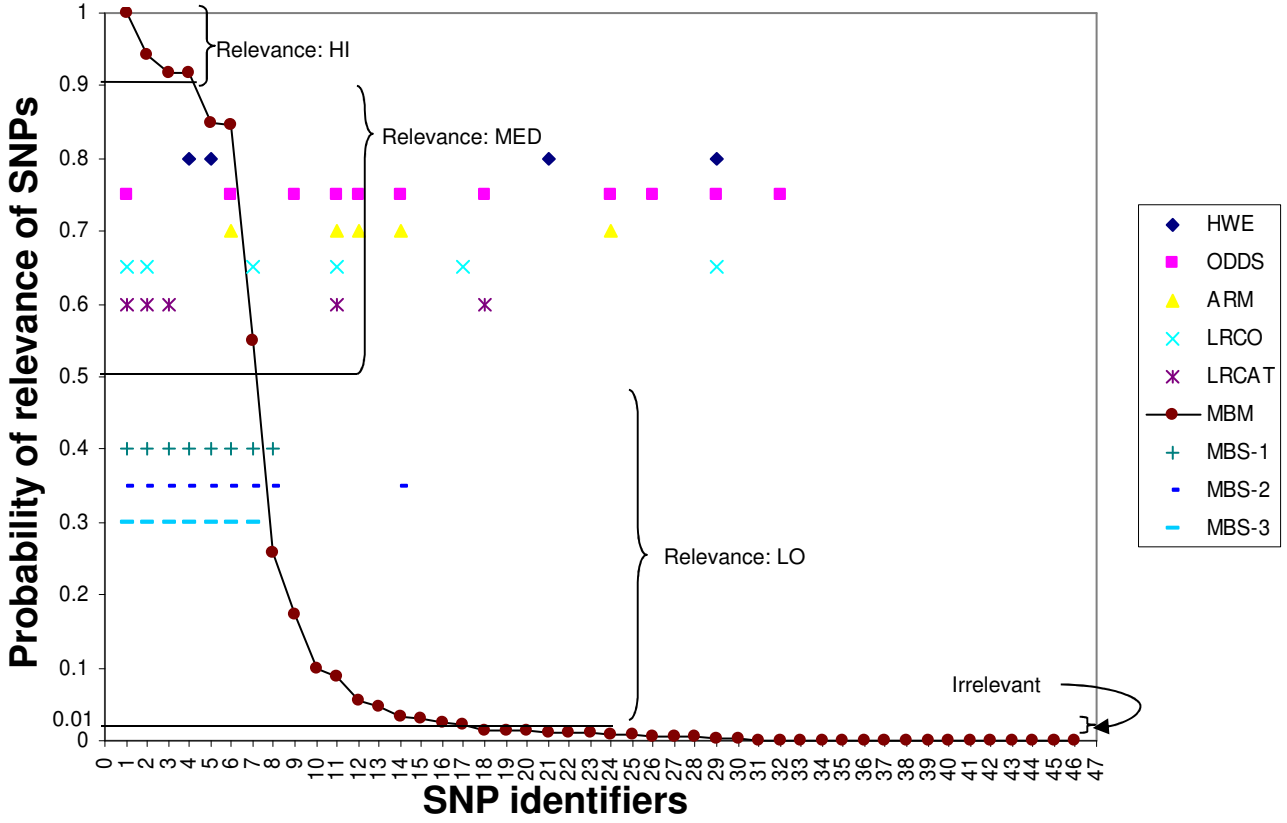


MBS posteriors in Asthma



Posteriors of strong relevance

HWE – Hardy-Weinberg equilibrium test, *ODDS* – odds ratio, *ARM* – Cochran-Armitage trend test, *LRCO* – logistic regression (continuous case), *LRCAT* – logistic regression (categorical case), *MBM* – Bayesian pairwise relevance, *MBS-1–9* relevant sets by Bayesian analysis. (only MBM values are numeric, others are arbitrary values for visualization)

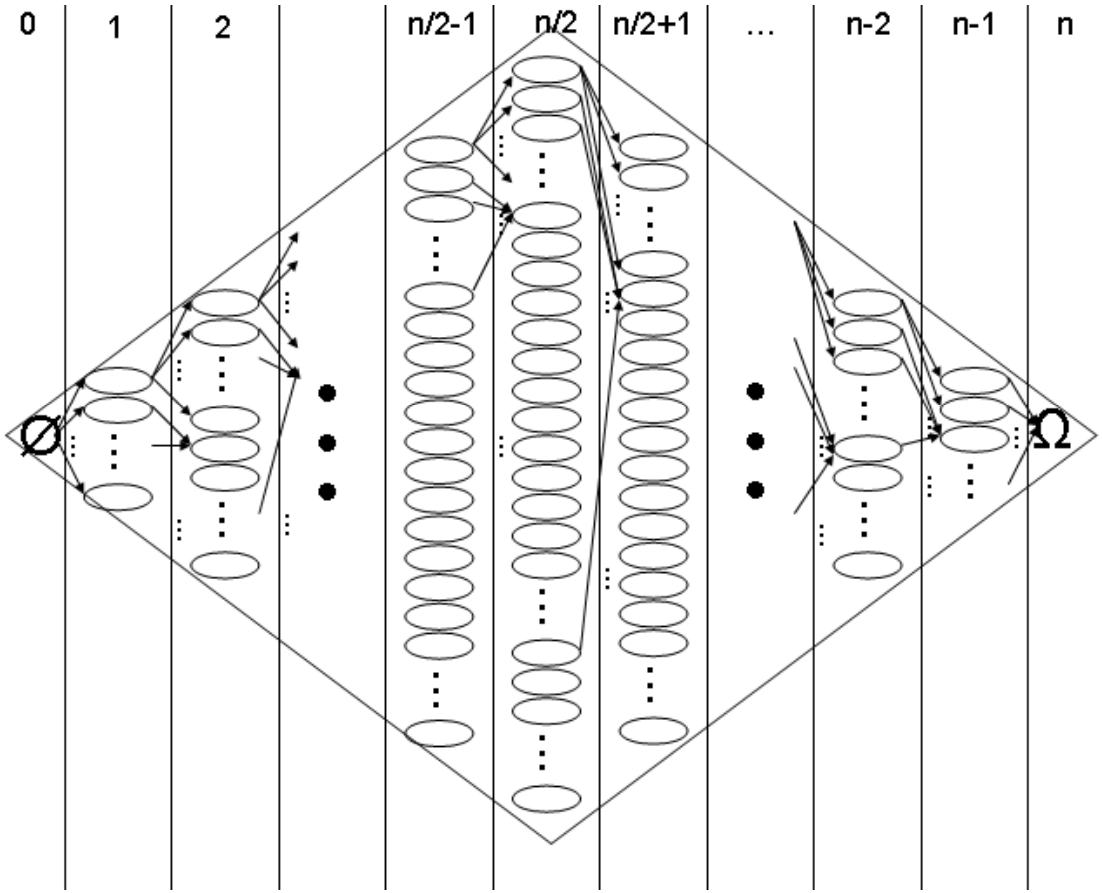


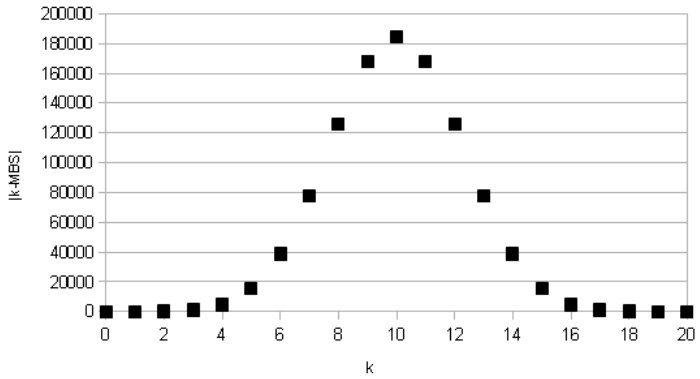
Frequentist vs Bayesian statistics

Frequentist	Bayesian
-	Prior probabilities
Null hypothesis	-
Indirect: proving by refutation	Direct
Model selection	Model averaging
Likelihood ratio test	Bayes factor
p-value	-!
-!	Posterior probabilities
Confidence interval	Credible region
Significance level	Optimal decision based on Exp.Util.
Multiple testing problem	Remains, so → **complex model**
Model complexity dilemma	**Best achievable alternative**

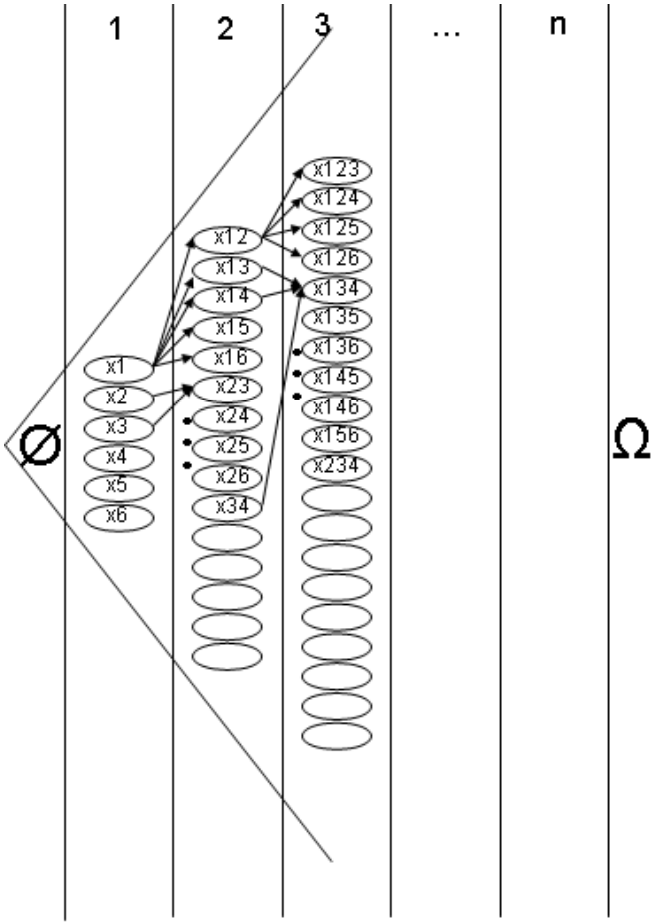
- Note: direct probabilistic statement!

The subset space

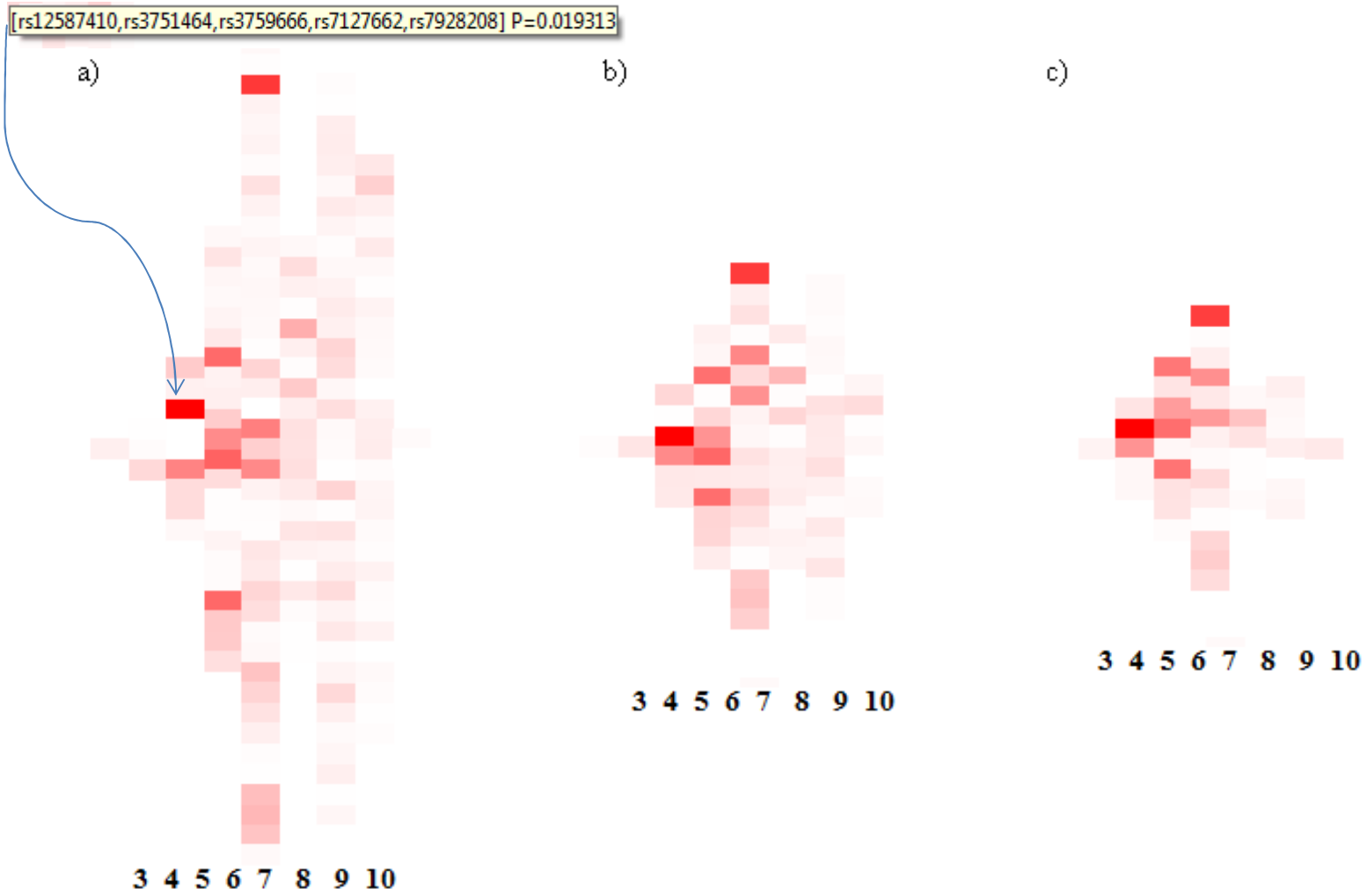




The subset space II.



An MBS heatmap in the subset space



Bayesian-network based Bayesian multilevel analysis (BN-BMLA)

Hierarchic statistical questions about typed relevance can be translated to questions about Bayesian network structural features:

Pairwise association ➔ Markov Blanket Memberhsips (MBM)
Multivariable analysis ➔ Markov Blanket sets (MB)
Multivariable analysis with interactions ➔ Markov Blanket Subgraphs (MBG)
Complete dependency models ➔ Partially Directed Acyclic Graphs (PDAG)
Complete causal models ➔ Bayesian network (BN)

Hierarchy of levels

BN ➔ PDAG➔ MBG ➔ MB ➔ MBM

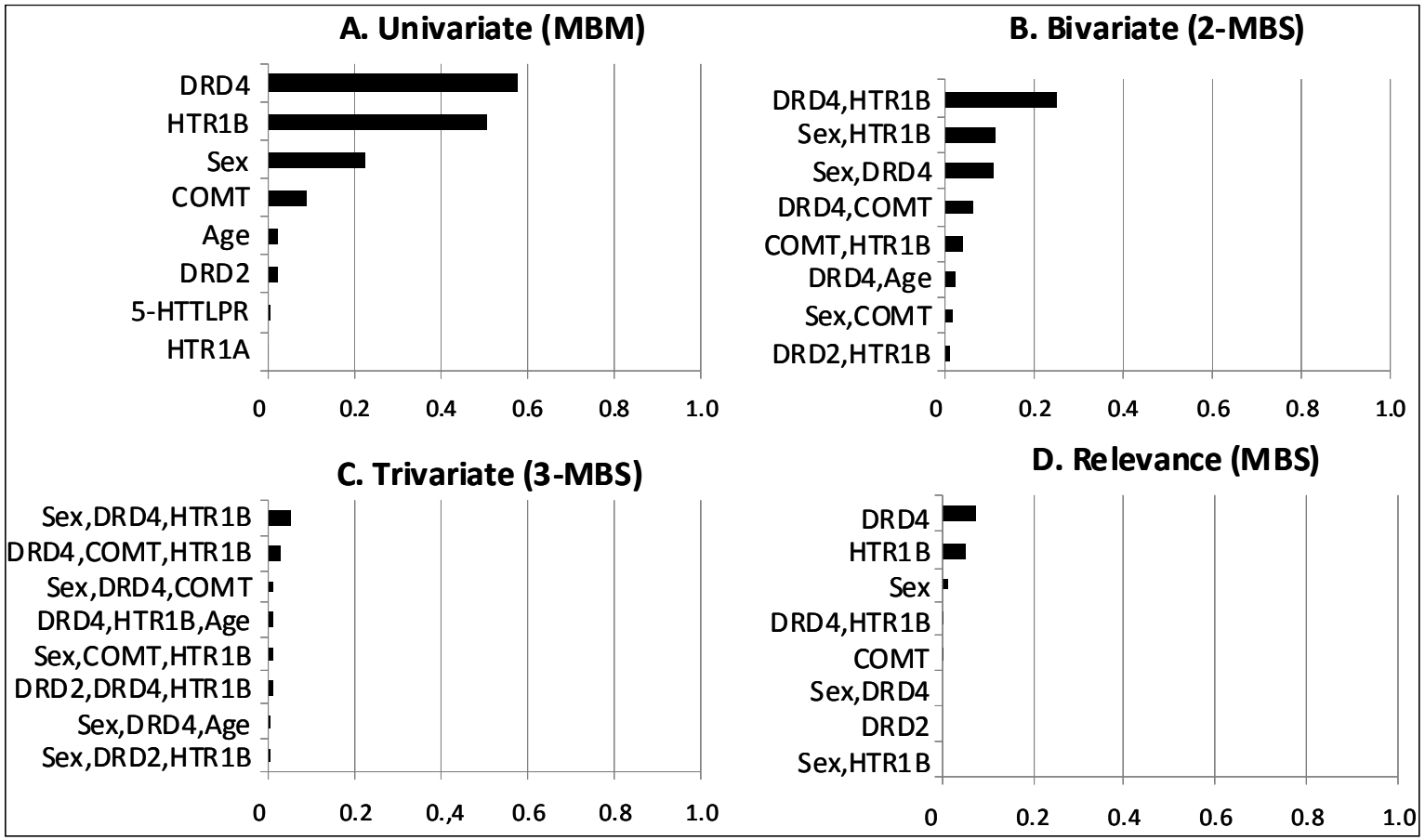
Bayesian inference of Bayesian network features

- Simple features vs. complex features
  - Edges (n^2), MBMs (n^2)
  - MBSs (2^n), MBGs ($2^{O(knlog(n))}$)
  - (Types of pairwise, but model-dependent relations (n^2)?)
- **Simple features**
  - Edges: DAG-based MCMC, Madigan et al., 1995
  - MBMs: ordering-based MCMC, Friedman et al., 2000
  - Modular features: exact averaging, Cooper,2000, Koivisto,2004
- **Complex features**
  - MBSs,MBGs : integrated ordering-based MCMC&search, 2006
  - Bayesian multilevel analysis of relevance (BMLA)
    - Ovarian cancer
    - Rheumatoid arthritis
    - Asthma
    - Allergy

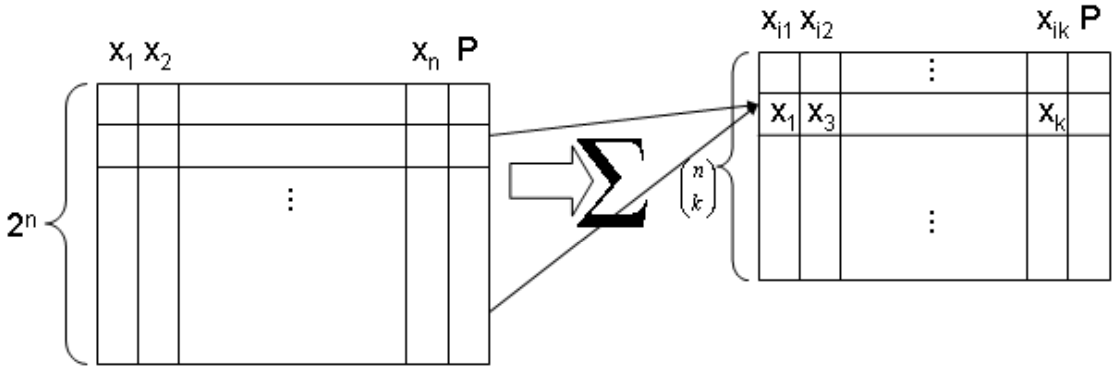
The marginal multivariate analysis

Problem: the "polynomial"gap between simple and complex features
(e.g., MBM (n^2) and MBS (2^n))

Idea: If all X_i in set S with size k are members of a Markov Boundary set, then S is called a k-ary Markov Boundary subset ($O(n^k)$).



Marginal posteriors for multivariate relevance: the definition



Operations:

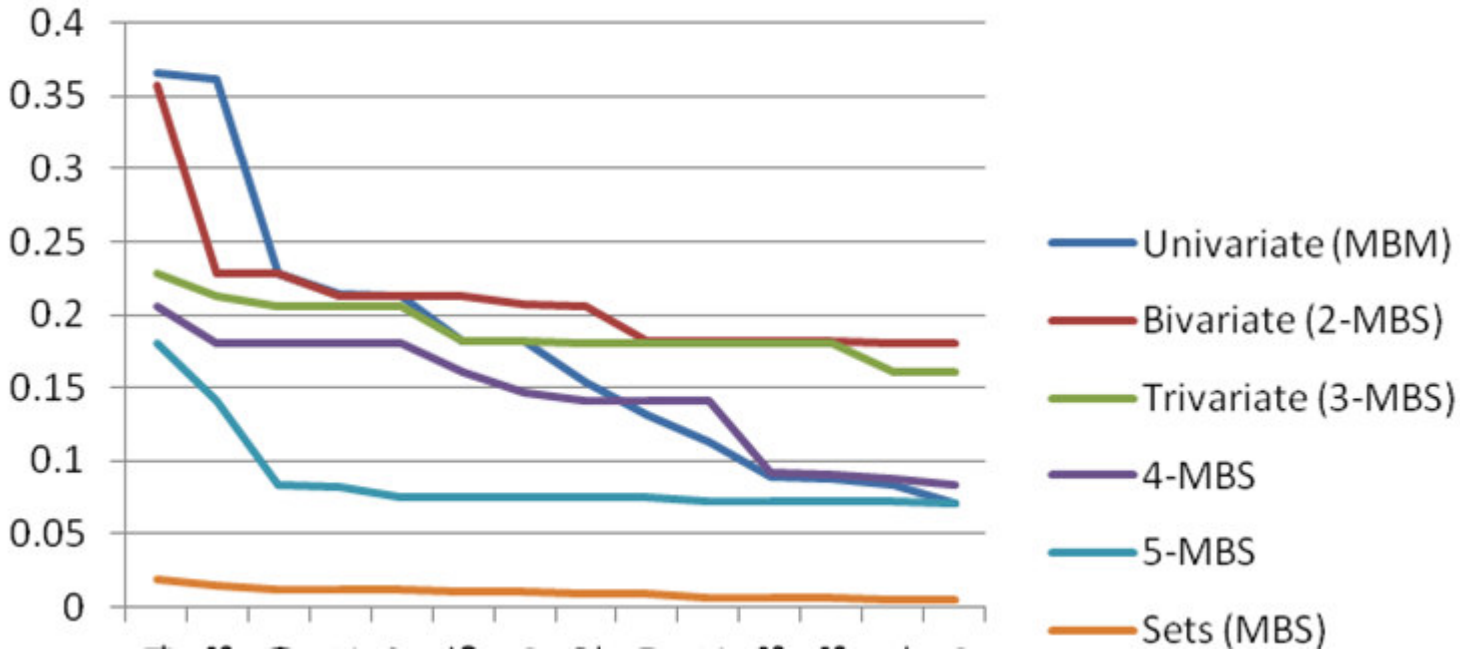
projection/marginalization

truncation

Methods???: heuristics

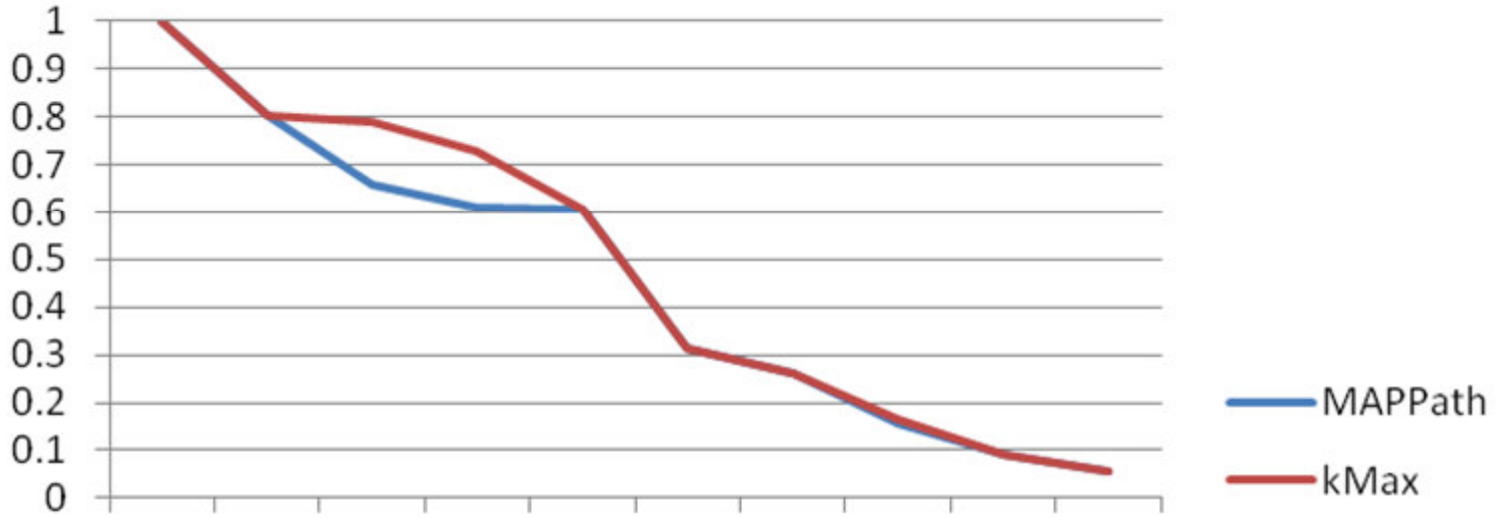
The marginal multivariate analysis in asthma research

$p(G : s \subseteq MBS(G) \mid D_N)$



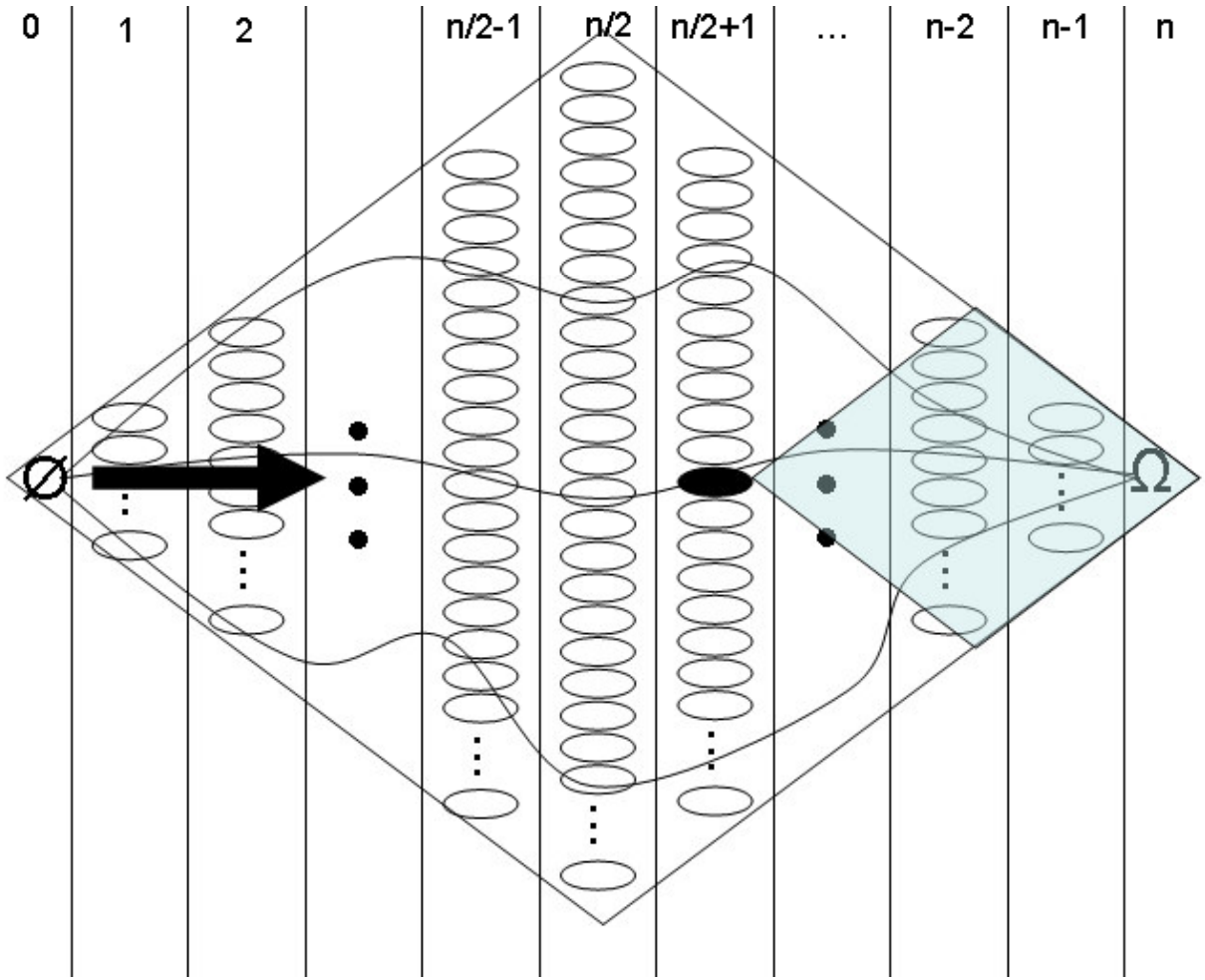
The marginal multivariate analysis in asthma research

$p(G : s \subseteq MBS(G) \mid D_N)$



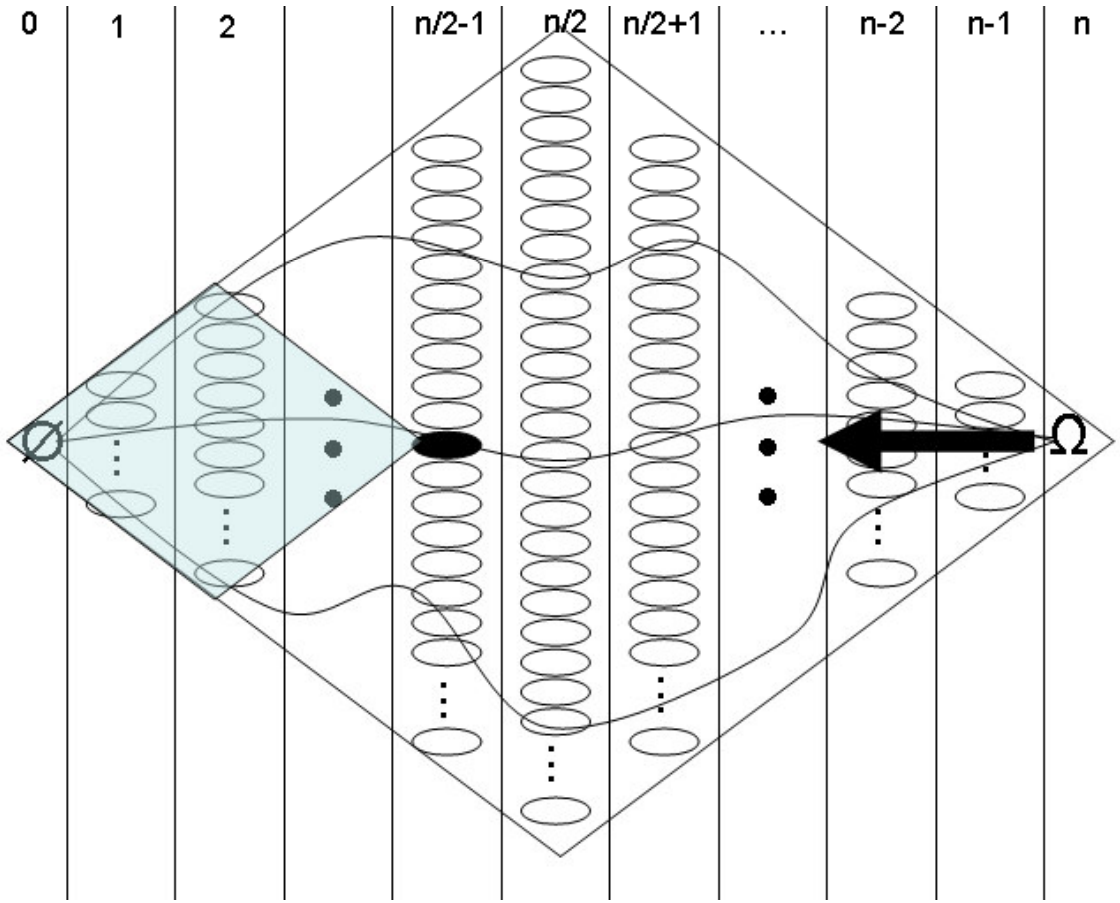
The k-MBS-sub

$p(G : s \subseteq MBS(G) \mid D_N)$



The k-MBS-sup

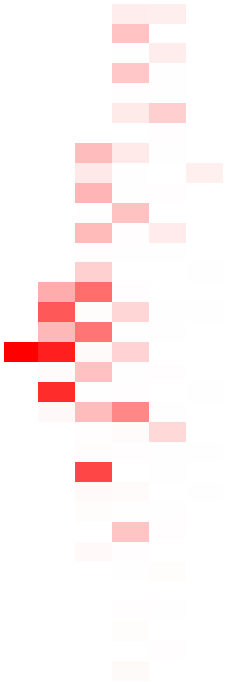
$p(G : s \supseteq MBS(G) \mid D_N)$



Marginal multivariate posteriors in the subset space

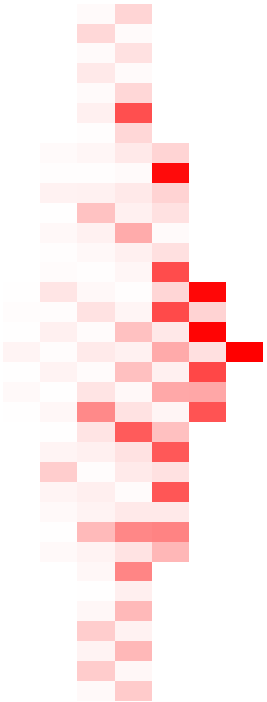
k-MBS-sub

$p(G : s \subseteq MBS(G) \mid D_N)$

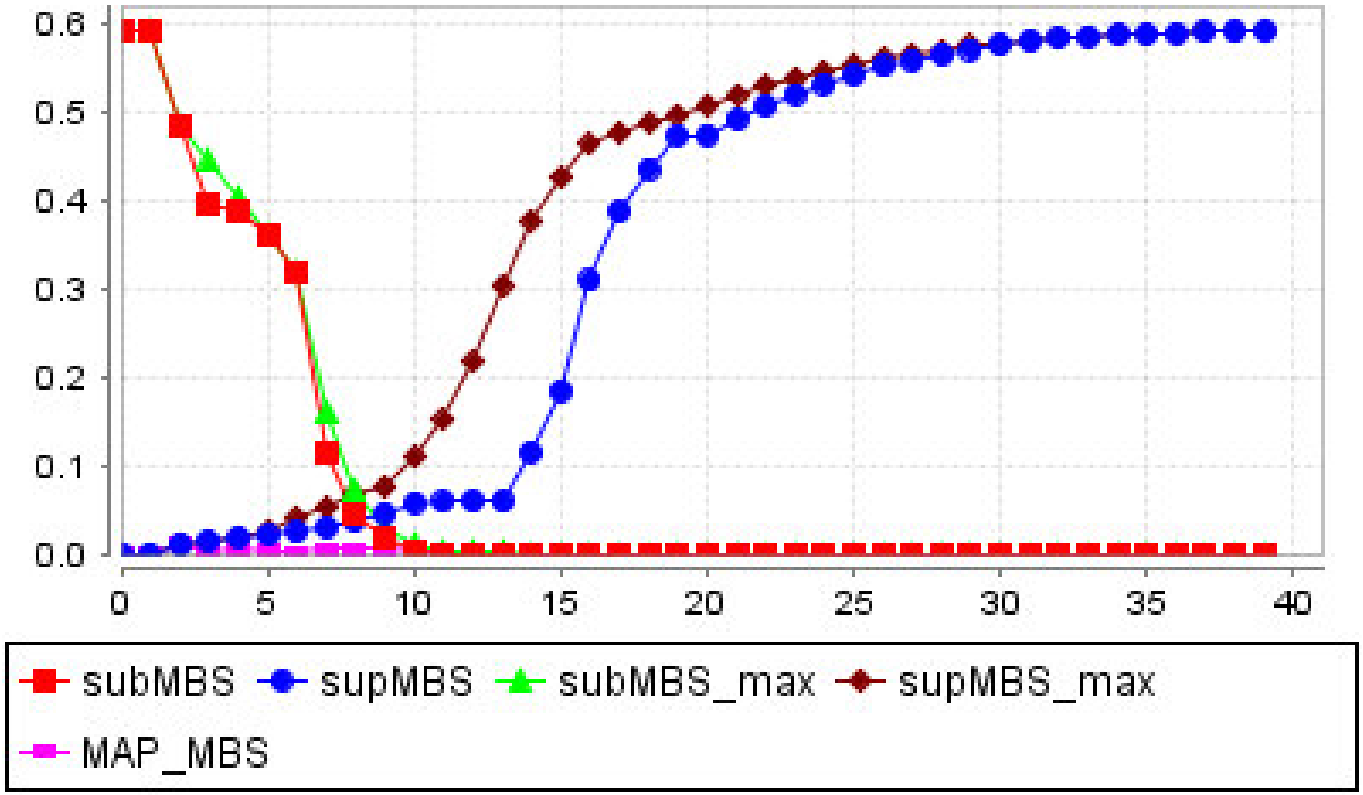


k-MBS-sup

$p(G : s \supseteq MBS(G) \mid D_N)$



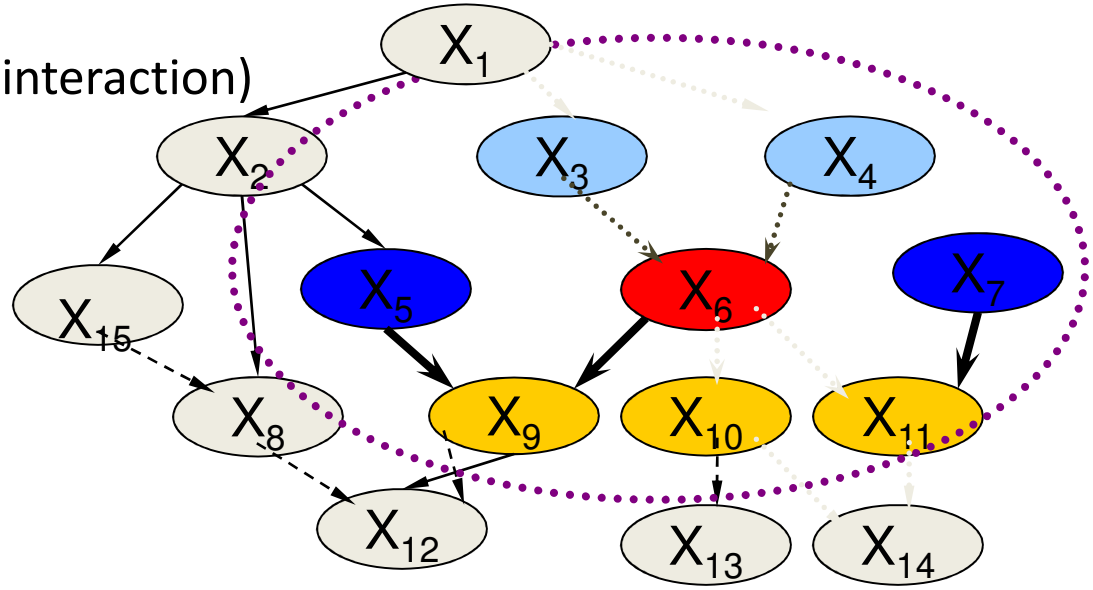
Marginal multivariate posteriors in the subset space



Clusterin in the MBS space

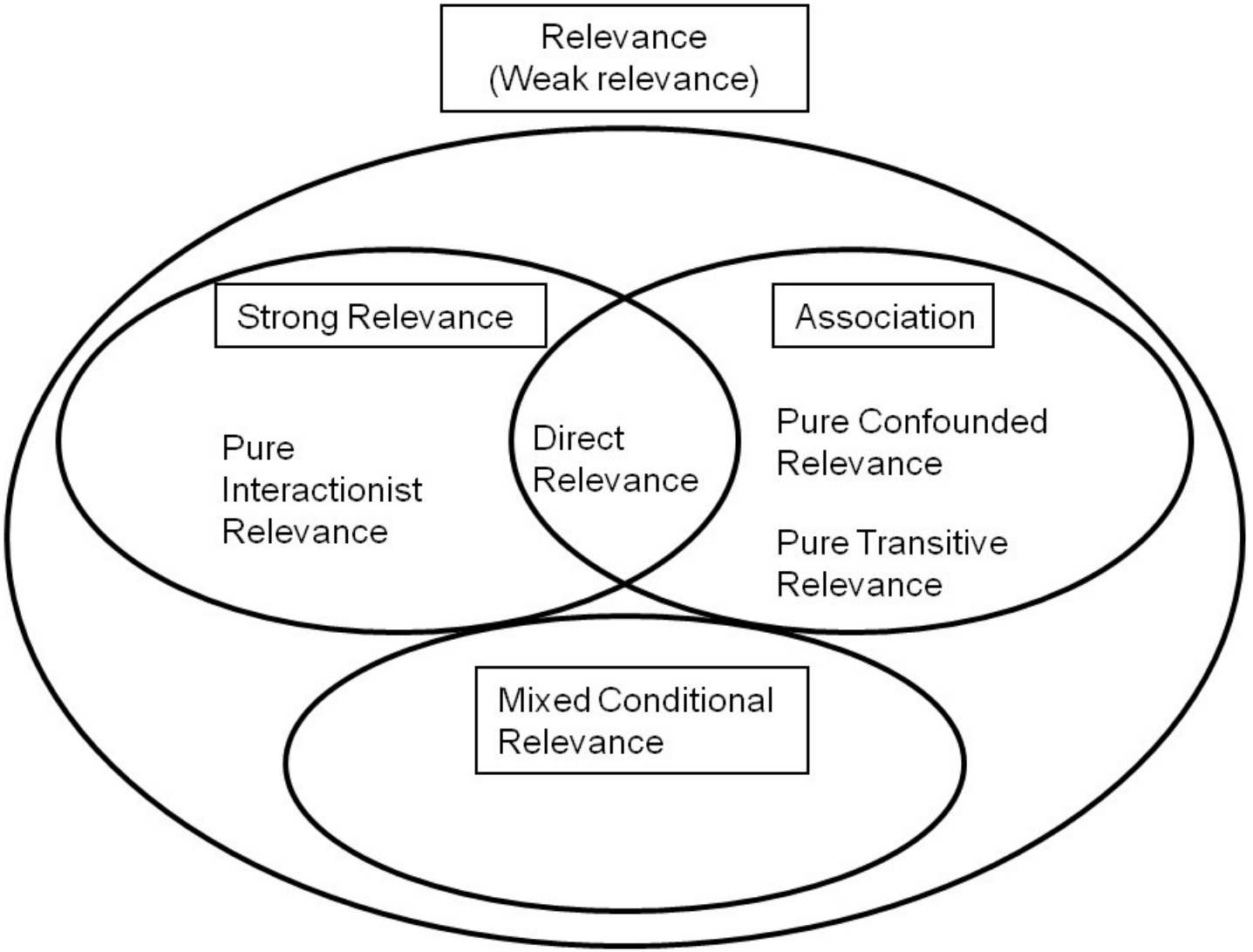
A more detailed language for associations: typed relevance

- Weak relevance
- Strong relevance
- Conditiontional relevance (pure interaction)
- Direct relevancia
  - With hidden variable
  - No hidden variable
- Causal relevancia
- Effect modifier
  - Probabilistic, direct, causal



- Typed relevance
  - Parent, Child
  - Direct=Parent or Child
  - Ascendant=Parent+, Descendant=Child+
  - Markovian=Parent, or Child or Pure interaction
  - Confounded
  - Associated= Ascendant or Descendant or Confounded

A more detailed language for associations: typed relevance



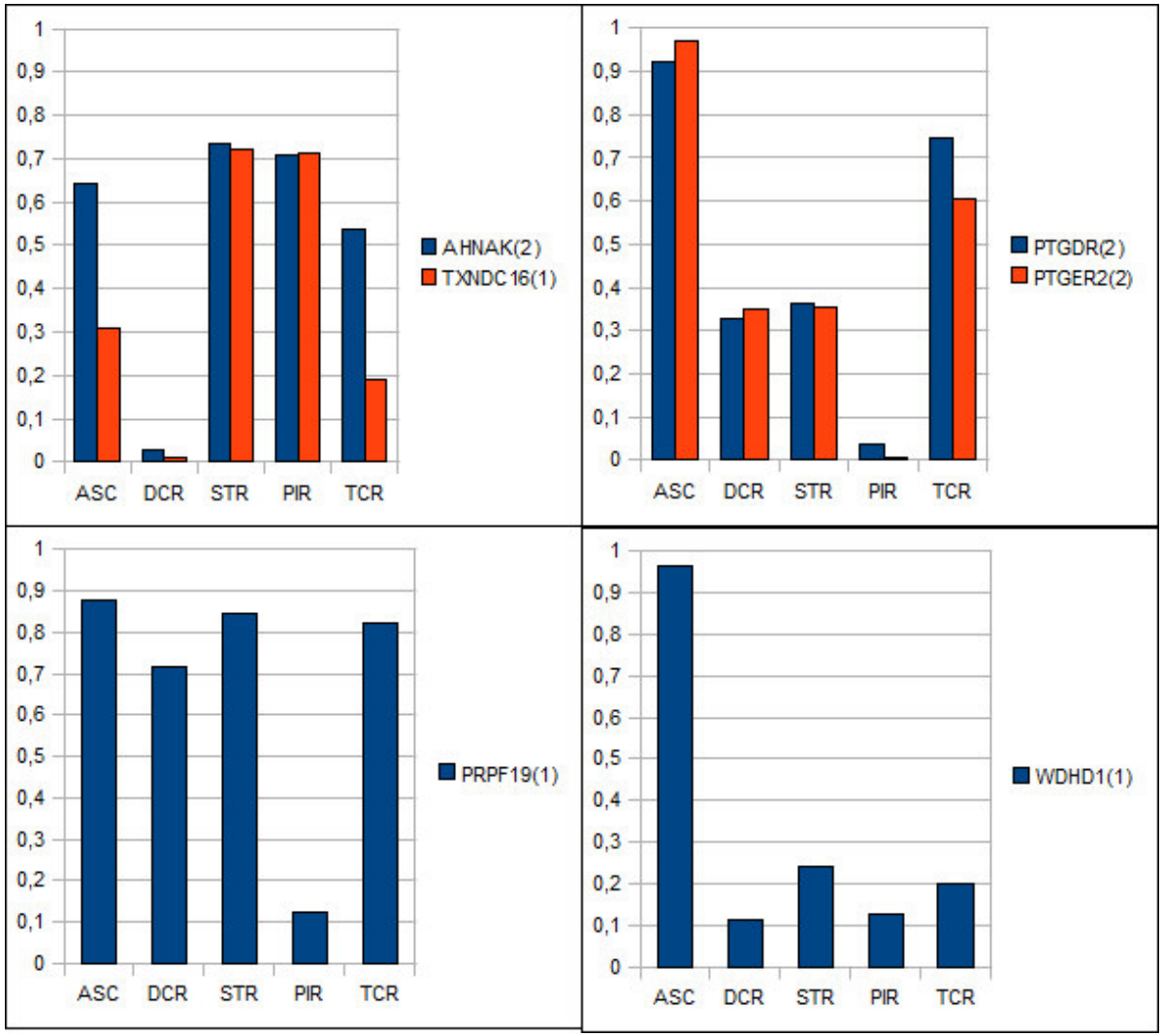
Subtypes of association relations - Causal

Relation	Direct graph definition	Causal interpretation under Causal Markov Assumption
Parent(X,Y)	X is a parent of Y	Cause
Child(X,Y)	X is a child of Y	Effect
PureAscendant(X,Y)	Not parent, but ascendant	IndirectCause
PureDescendant(X,Y)	Not child, but descendant	IndirectEffect

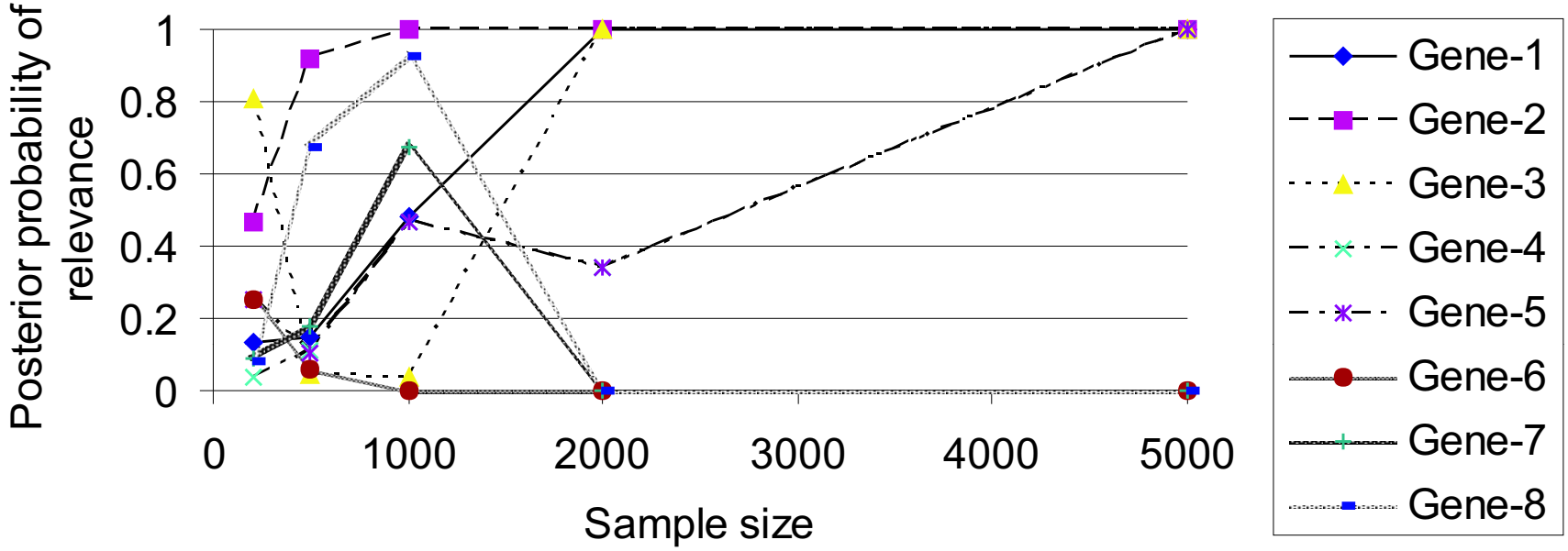
Subtypes of association relations - Acausal

Relation	Direct graph definition	Probabilistic interpretation
PureCommonAncestor(X,Y)	No directed path between X,Y, but there is a common ancestor	*PureConfounded*
PureCommonChild(X,Y)	No directed path between X,Y, but there is a common child	PureInteraction
Independent(X,Y)	No edge, directed path or common ancestor.	Independent
Edge(X,Y)	*Parent or Child*	DirectDependency
Path(X,Y)	*Ascendant or Descendant*	
BoundaryGraphMembership(X,Y)	*Parent or Child or CommonChild*	Strong relevance (Markov Blanket Membership)
Associated(X,Y)	*Ascendant or Descendant or Confounded*	Associated (weak relevance)

A more detailed language for associations: typed relevance



SNP-to-gene aggregation



The sequential posteriors that a given gene contains a SNP relevant for asthma

Abstraction levels: SNP, haplo-block, gene,..., pathway

Note that it is different from aggregated multi-variables.

Aggregating to output

- What can we do in case of multiple output?
- E.g. IgE, Eosinophil,Rhinitis, Asthma,AsthmaStatus
- Compute the posterior of „typed relevance” for
  - A given target,
  - Any of of the targets,
  - Excluding a given a target,
  - Being a multitarget.

Note that typed relevance and typed output can be combined, though not arbitrarely.

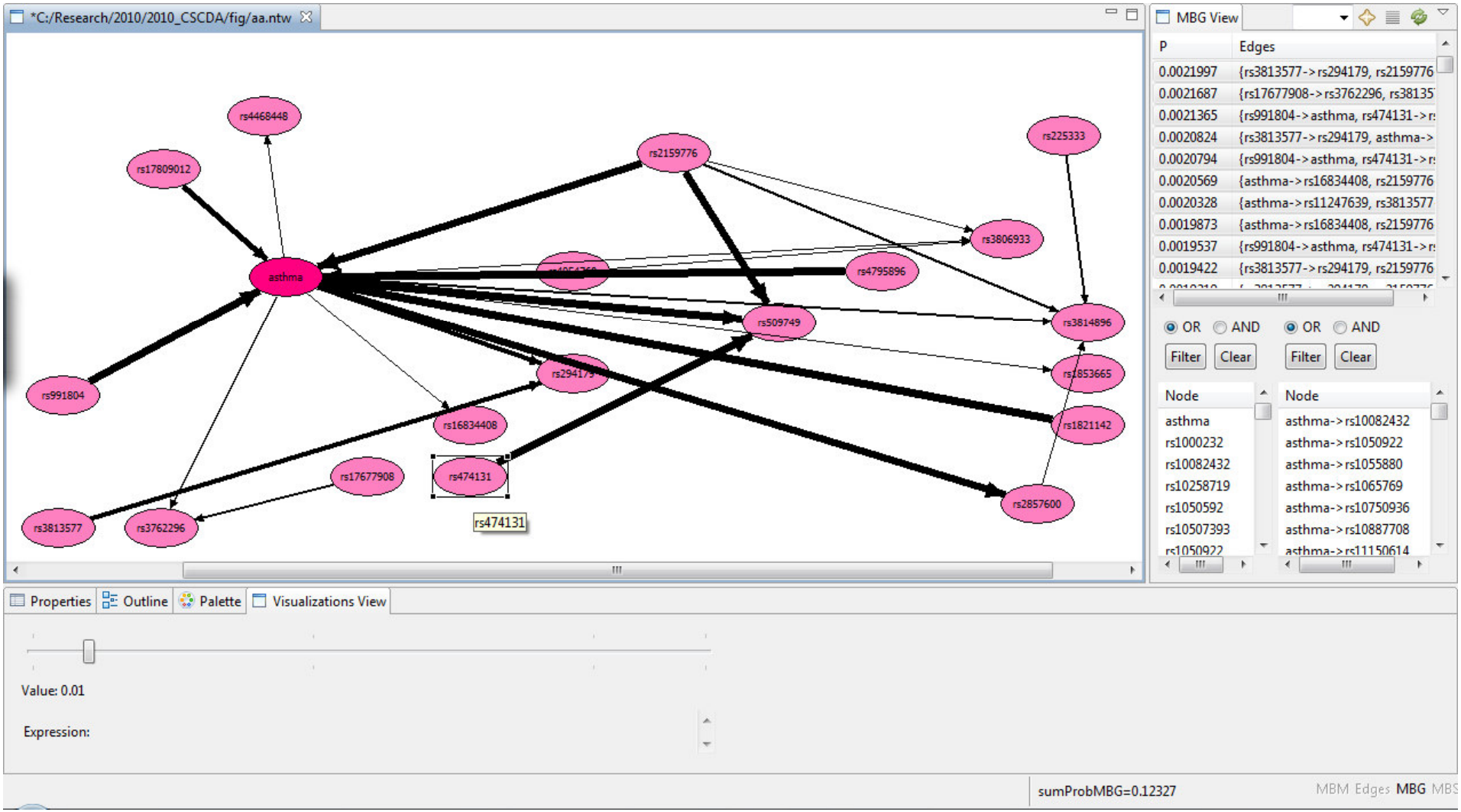
Genagrid

- SGI Altix ICE
  - 5 TFLOPS
  - 1TB memory
  - 64x8 cores
  - FPGAs

High-performance and high-throughput computation (HPC,HTC)

- Moore-law
  - In computer industry
  - In molecular biology(!)
- The grid system vs the SMP architecture
  - Cache-coherence
- Other approaches
  - Field-programmable gate array (FPGA)
  - Graphics processing unit (GPU)

BayesEye


*C:/Research/2010/2010_CSCDA/fig/aa.ntw
rs4468448
rs17809012
rs2159776
rs225333
rs3806933
asthma
rs4795896
rs509749
rs3814896
rs1853665
rs991804
rs16834408
rs1821142
rs17677908
rs474131
rs2857600
rs3813577
rs3762296
rs474131
MBG View
P
Edges
0.0021997 {rs3813577->rs294179, rs2159776
0.0021687 {rs17677908->rs3762296, rs38135
0.0021365 {rs991804->asthma, rs474131->r
0.0020824 {rs3813577->rs294179, asthma->
0.0020794 {rs991804->asthma, rs474131->r
0.0020569 {asthma->rs16834408, rs2159776
0.0020328 {asthma->rs11247639, rs3813577
0.0019873 {asthma->rs16834408, rs2159776
0.0019537 {rs991804->asthma, rs474131->r
0.0019422 {rs3813577->rs294179, rs2159776
OR
AND
OR
AND
Filter
Clear
Filter
Clear
Node
asthma
rs1000232
rs10082432
rs10258719
rs1050592
rs10507393
Node
asthma->rs10082432
asthma->rs1050922
asthma->rs1055880
asthma->rs1065769
asthma->rs10750936
asthma->rs10887708
Properties
Outline
Palette
Visualizations View
Value: 0.01
Expression:
sumProbMBG=0.12327
MBM Edges MBG MBS

Summary

- Challenges in biomarker discovery
  - Robustness (repeatability, transferability)
  - Causation
  - Multiple hypothesis testing
  - Interaction (multivariate approach)
- Feature relevance
- The feature subset selection problem
- Identification of biomarkers
  - Methods
- Challenges
  - Interpretation➔ Bayesian networks
  - Causality➔ Bayesian networks
  - Uncertainty ➔ Bayesian statistics
- A Bayesian network based Bayesian approach to biomarker analysis