Hierarchical Bayesian Modelling Identifies Shared Gene Function

Peter Sykacek

Bioinformatics Research Group, Dept. of Biotechnology,
BOKU University Vienna
peter.sykacek@boku.ac.at

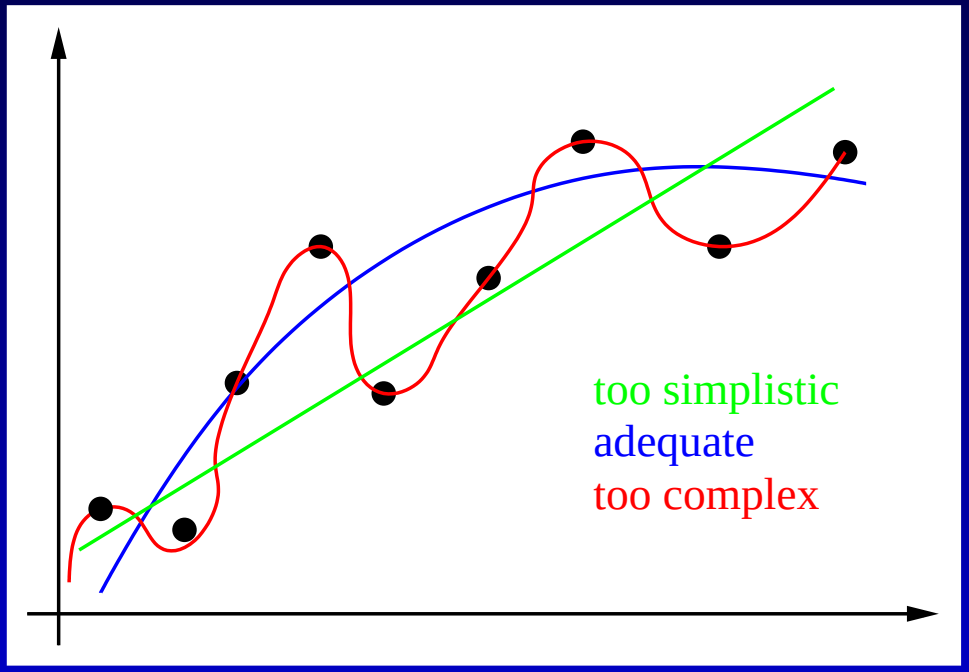
Working Philosophy

- Systems biology aims at providing quantitative models from biological data sources. This is thus mainly an empirical discipline.
- Information gain is data driven. This suggests the main task is inverse modelling or “inference”, i.e. finding suitable model classes and parameters.
- A key problem in inference is the concept of “noise”, which is caused by
  - measurement noise
  - intentional or accidental simplifications (like ignoring certain influence factors)
  - and last but not least by erroneous reports that contribute to background knowledge.

– > no certain background information
– > no point estimates as this implies certainty

Adequate models

Capture underlying structure and avoid overfitting. Fiddle parameters affecting model complexity can have adverse effects.



Idea: overfitting is a result of tuning the model towards the training data. Over or under-complex models that do not capture the underlying data generating mechanism will perform worse on novel data obtained from the generating model than an appropriate model.

Getting the model class right is thus imperative for success!

A Simple Regression Model

Suppose a life science experiment provided some noisy data $\mathcal{Z} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N)\}$ with $\boldsymbol{x}_n$ possibly multivariate i.e. vectors.

Based on $\mathcal{Z}$, we have an inference problem of finding an "optimal" relation between $\boldsymbol{x}$ and y:

$$p(y|\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) + \epsilon(\lambda)$$

A Simple Regression Model

Suppose a life science experiment provided some noisy data $\mathcal{Z} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N)\}$ with $\boldsymbol{x}_n$ possibly multivariate i.e. vectors.

Based on $\mathcal{Z}$, we have an inference problem of finding an "optimal" relation between $\boldsymbol{x}$ and y:

$$p(y|\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) + \epsilon(\lambda)$$

Noise requires a deterministic and a random component.

$-> $ Inherent uncertainty, y is a random variable!

Assessing Model Parameters

Idea: subtract the deterministic part from y_n:

$$\epsilon_n = y_n - f(\boldsymbol{x}_n; \boldsymbol{\theta})$$

For convenience introduce $\mathcal{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ and $\mathcal{D} = \{y_1, ..., y_N\}$. Assuming that ϵ_n are i.i.d samples, we get the likelihood function:

$$p(\mathcal{D}|\boldsymbol{\theta}, \lambda, \mathcal{X}) = \prod_n p(y_n|\boldsymbol{\theta}, \lambda, \boldsymbol{x}_n)$$

which is a suitable objective function for comparing various options for $\boldsymbol{\theta}$ and λ.

MLH's Major Weakness

True model - linear regression, Gaussian noise:

$$p(y|\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta}) + \epsilon(\lambda)$$

$f(\boldsymbol{x};\boldsymbol{\theta}) = [1, \boldsymbol{x}^T]\boldsymbol{\theta}$ and $\epsilon(\lambda) = \mathcal{N}(\epsilon; 0, \lambda)$, with λ denoting "precision" (i. e. inverse variance).
Finite sample size and different model classes: What is the maximum of the likelihood?

Think "phone book": Perfect memorising of all y_n, modelling error 0, $\lambda -> \infty$, $p(\mathcal{D}|\boldsymbol{\theta}, \lambda, \mathcal{X}) -> \infty$.

$->$ likelihood unsuitable objective for inferring model classes!

Note: An additional problem may arise from unidentified models, like $y = abx$, where even an infinite amount of data is insufficient for uniquely defining model coefficients.

Occam's Razor

Human reasoning implicitly applies Occam's Razor

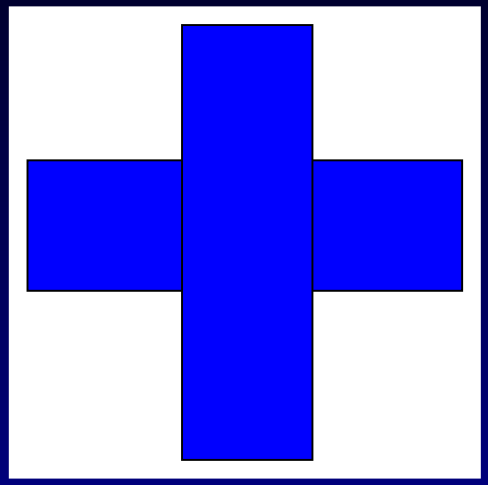
William of Occam (or Ockham) (1288 - 1348)

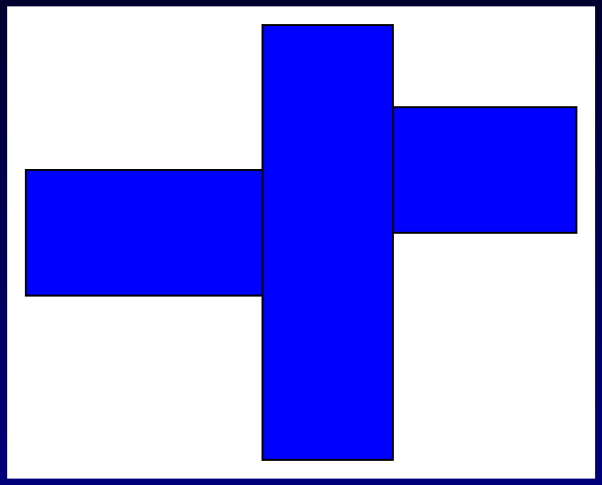
Entia non sunt multiplicanda sine necessitate: Entities are not to be multiplied without necessity.

Interpretation: One should always opt for an explanation in terms of the fewest possible number of causes, factors, or variables.

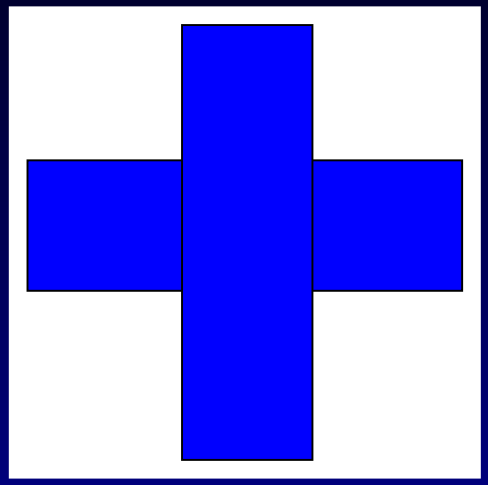
Material from `http://en.wikipedia.org/wiki/William_of_Ockham`.

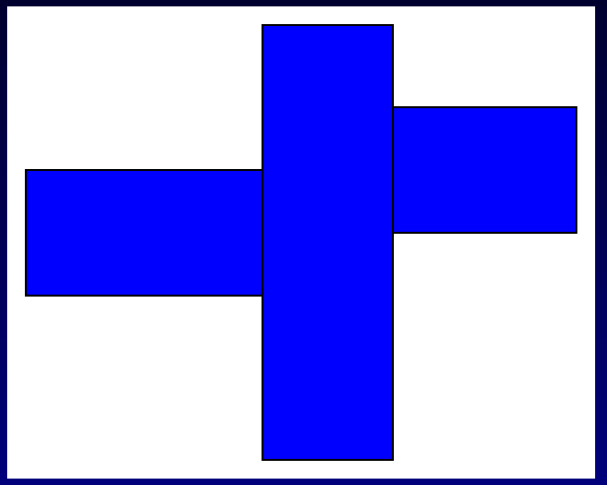
Guess the Correct “Model”

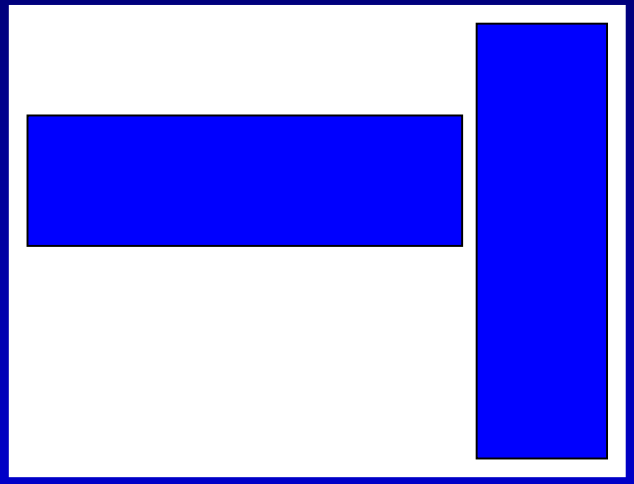


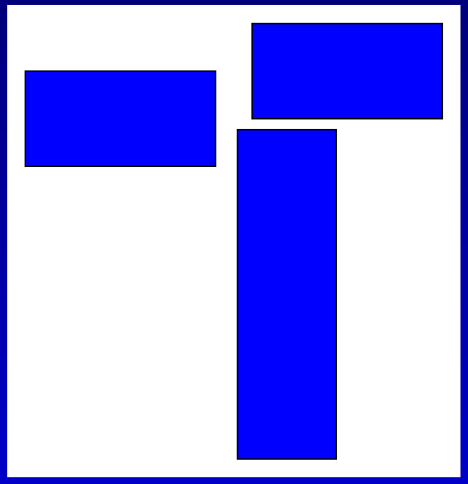


Guess the Correct “Model”



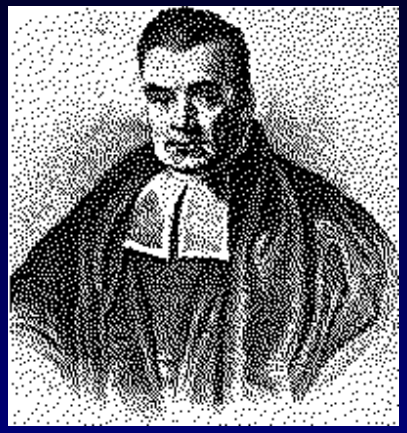






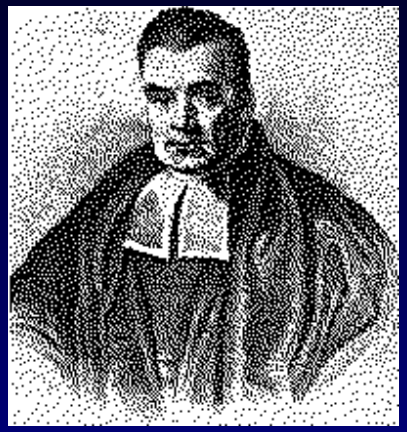
Model comparison requires external penalty on top of likelihood! (AIC, BIC, etc.)

Probabilistic Approaches



Thomas Bayes (1701 - 1763)
Learning from data based on a decision theoretic framework

Probabilistic Approaches

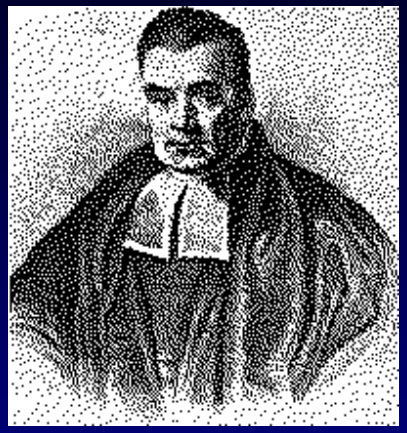


Thomas Bayes (1701 - 1763)
Learning from data based on a decision theoretic framework

$$p(I|\mathcal{D}) = \frac{p(\mathcal{D}|I)p(I)}{p(\mathcal{D})}$$

First consequence: we must revise beliefs according to Bayes theorem

Probabilistic Approaches



Thomas Bayes (1701 - 1763)
Learning from data based on a decision theoretic framework

$$p(I|\mathcal{D}) = \frac{p(\mathcal{D}|I)p(I)}{p(\mathcal{D})}$$

First consequence: we must revise beliefs according to Bayes theorem

$$\alpha_{opt} = \text{argmax}_\alpha < u(\alpha) > \text{, where}$$

$$< u(\alpha) >= \int_I u(\alpha, I)p(I|\mathcal{D})dI.$$

Second consequence: Decisions by maximising expected utilities

Integration replaces maximisation!

Probabilistic Model

Probabilistic model, Bayesian Network or DAG (M. I. Jordan, 1998):

A set of vertexes $V = \{X_1, ..., X_N\}$ and a set of directed edges E define a graph $M = \{V, E\}$ of parent - child relations
$\mathrm{pa}[X_i] = \{X_n | (X_n \rightarrow X_i) \in E \; \forall n\}$.
Conditional probability statements complete the model:

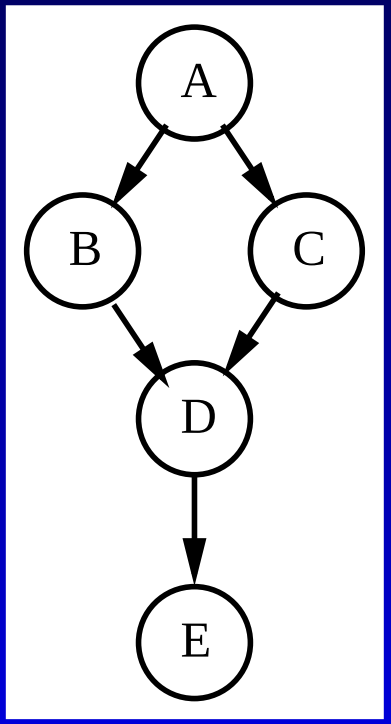
$$P(V) = \prod_{n=1}^{N} P(X_n | \mathrm{pa}[X_n])$$

Example

Rules of probability calculus like $P(A, B) = P(A)P(B|A)$ or $P(A, B, C) = P(A, B)P(C|A, B)$ are simplified by probabilistic independence statements.

Example

Rules of probability calculus like $P(A, B) = P(A)P(B|A)$ or $P(A, B, C) = P(A, B)P(C|A, B)$ are simplified by probabilistic independence statements.



Instead of standard probability calculus where

$$P(A, B, C, D, E) = P(A)P(B|A)P(C|A, B) \\ P(D|A, B, C)P(E|A, B, C, D)$$

we get

$$P(A, B, C, D, E) = P(A)P(B|A)P(C|A) \\ P(D|B, C)P(E|D)$$

The latter requires much fewer parameters.

Bayesian Modelling Applied

- Assumption: Several microarray experiments are obtained such that slides can be mapped to a biological state of interest.
- Shared genetic function: Interesting genes are across experiments informative about these biological states.
- Task: find those genes! Actually two problems:
  - Cross annotation of genes (potentially different species)
  - Calculate a measure across experiments

This talk shows how we may obtain such a measure using a probabilistic approach.

Biological States of Experiments

Mammary Gland tc. (lact. day & hours of involution)

biol. state	L_0	L_5	L_{10}	I_{12}	I_{24}	I_{48}	I_{72}	I_{96}
Type 1 Apoptosis	-	-	-	+	+	?	-	-
Type 2 Apoptosis	-	-	-	-	-	?	+	+
Apoptosis	-	-	-	+	+	+	+	+
Differentiation	+	+	+	?	-	-	-	-
Inflammation	?	-	-	+	+	?	-	-
Remodelling	-(?)	-	-	-	-	?	+	+
Acute Phase	+	-	-	-	+	+	+	+

Serum Deprived Apoptosis (duration in hours)

biol. state	t_0	t_{28}	t_{48}
Type 2 Apoptosis	-	+	+
Apoptosis	-	+	+
Differentiation	+	-	-

Potential Solutions

- Statistical meta analysis (originally proposed by Fisher).
- “Bioinformatics” meta analysis.
- Probabilistic Inference.

Toy Data

Means of Gaussians to generate synthetic data

Experiment	Gene Group	Mean Assay 1	Mean Assay 2
Ranking	1	± 2	± 2
	2	± 0.5	± 0.5
	3	± 0.05	± 0.05
Censoring	1	± 2	± 2
	2	± 4	± 0.5
	3	± 0.1	± 0.1

Data: 4 synthetic "genes" per group generated from Gaussians with unit std. dev. and means as shown.

Bioinformatics Meta Analysis

Simple Approach:

- Take an individual experiment
- Calculate some gene ranking (e.g. using fold change, t-test, LIMMA, etc.)
- Decide upon some threshold
- Search for “genes” found in all lists.

Meta Analysis - Ranking

Assay 1	Assay 2	Combined
gene 1,3	gene 1,1	gene 1,1
gene 1,1	gene 1,4	gene 1,2
gene 1,4	gene 1,2	gene 1,3
gene 1,2	gene 1,3	gene 1,4
gene 2,4		

Rank information gets lost!

Meta Analysis - Censoring

Assay 1	Assay 2	Combined
gene 2,3	gene 1,3	gene 1,1
gene 2,4	gene 1,2	gene 1,2
gene 2,1	gene 1,4	gene 1,3
gene 2,2	gene 1,1	gene 1,4
gene 1,2	gene 2,2	gene 2,2
gene 1,1		
gene 1,3		
gene 1,4		

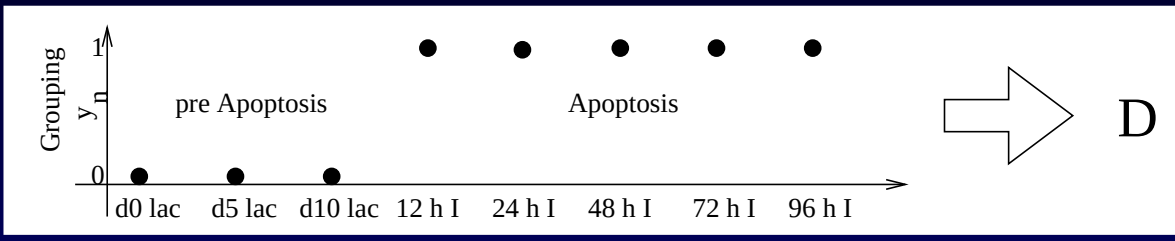
Genes from group 2 get censored at random!

Potential Solutions II

- Statistical meta analysis (originally proposed by Fisher).
- “Bioinformatics” meta analysis $->$ sucks!
- Bayesian Inference.

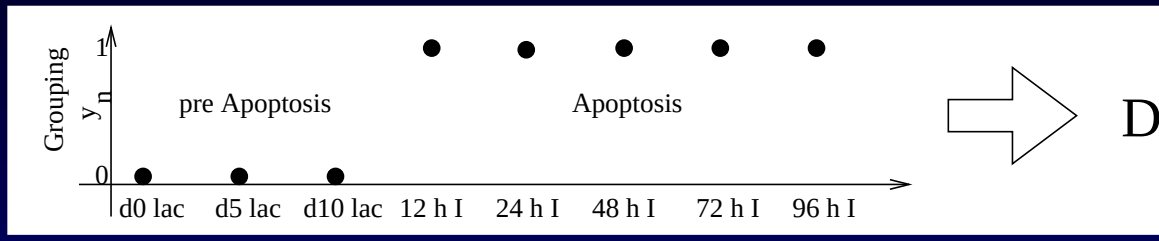
Probabilistic Gene Ranking

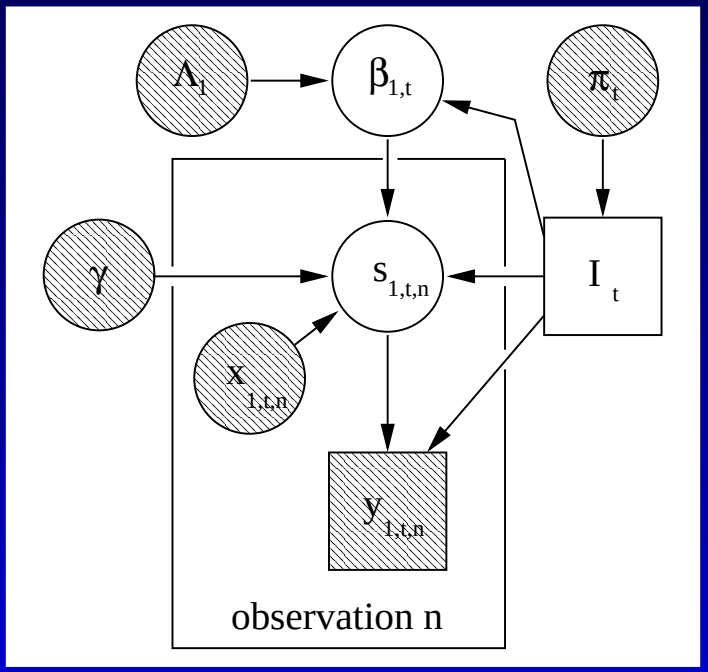
Apoptosis (lac. vs. inv.!) in the Mouse Mammary Gland



Probabilistic Gene Ranking

Apoptosis (lac. vs. inv.!) in the Mouse Mammary Gland





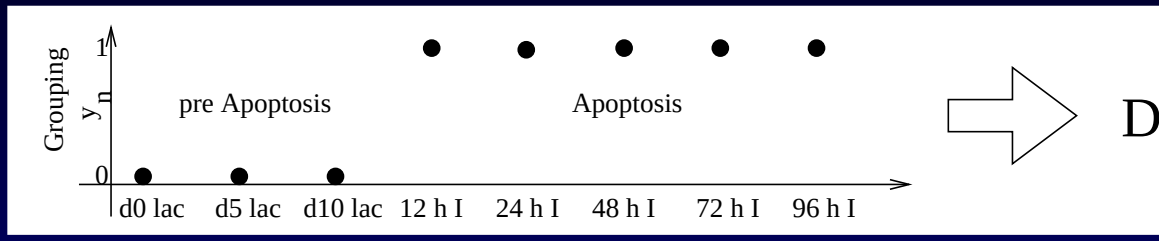
Latent variable probit GLM.

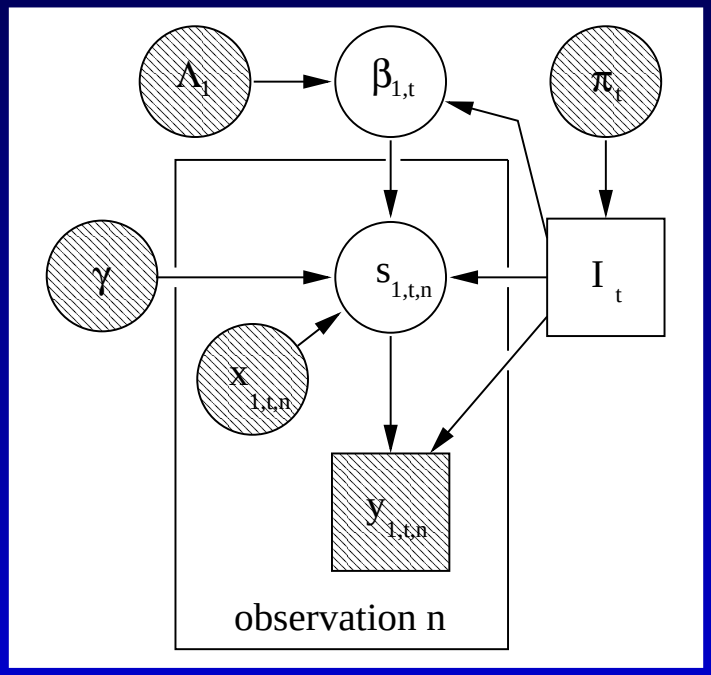
$$\text{if } I_t = \begin{cases} 1 : s_{1,t,n} \sim 1 + x_{t,n} \\ 0 : s_{1,t,n} \sim 1 \end{cases}$$

$s_{1,t,n}$ is a one dimensional Gaussian random variable with mean $\boldsymbol{\beta}_{t,1}^T \boldsymbol{x}_{t,n}$ and precision γ.

Probabilistic Gene Ranking

Apoptosis (lac. vs. inv.!) in the Mouse Mammary Gland





Latent variable probit GLM.

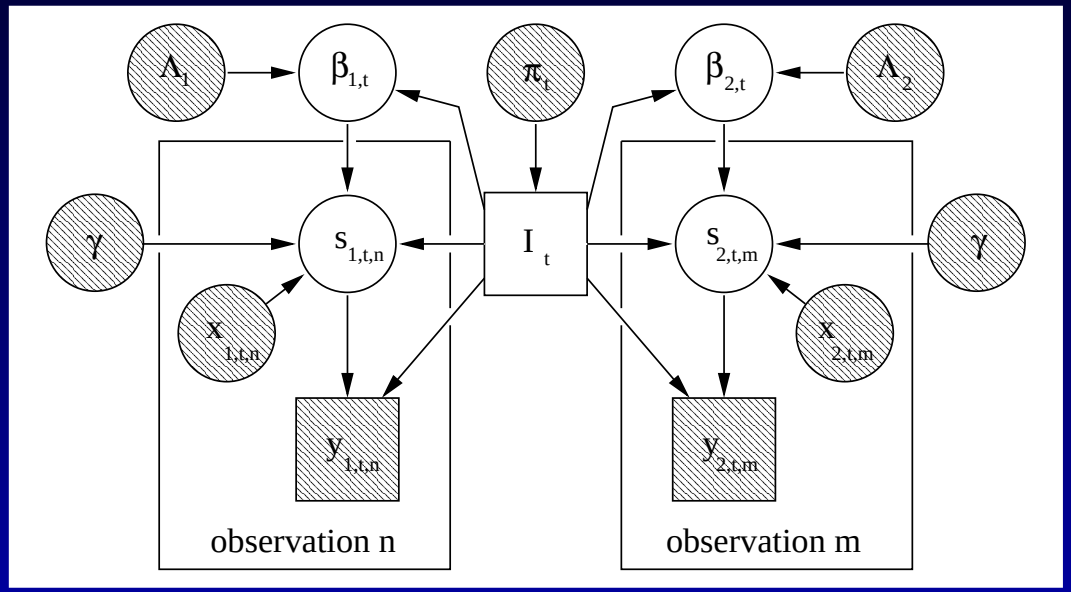
$$\text{if } I_t = \begin{cases} 1 : s_{1,t,n} \sim 1 + x_{t,n} \\ 0 : s_{1,t,n} \sim 1 \end{cases}$$

$s_{1,t,n}$ is a one dimensional Gaussian random variable with mean $\boldsymbol{\beta}_{t,1}^T \boldsymbol{x}_{t,n}$ and precision γ.

As an alternative to p-values, $P(I_t|\mathcal{D}_1)$, serves as a probabilistic rank measure. Gene selection according to $P(I_t|\mathcal{D}_1)$ implies a zero-one utility function and an M-closed model space. (VB-eqns.)

Shared Gene Function

Include Information about Endothelial Cell Death

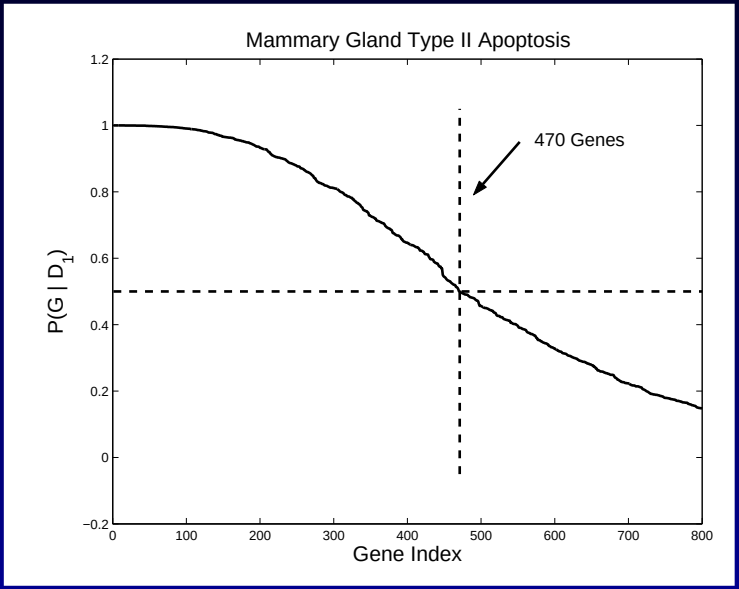


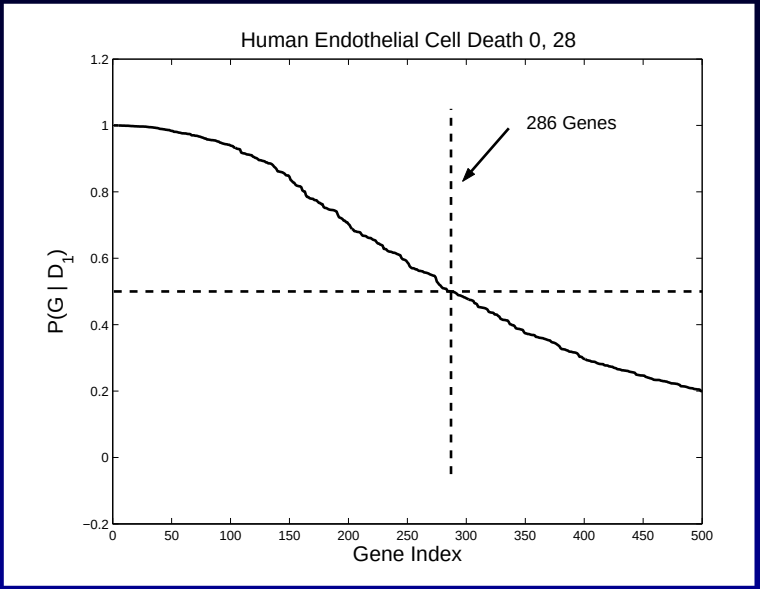
Model 0 hrs. vs. 28 hrs. as latent variable probit GLM. Calculate $P(\mathcal{D}_2|I_t)$, the marginal likelihood.

Bayes theorem gives a *principled* measure for ranking

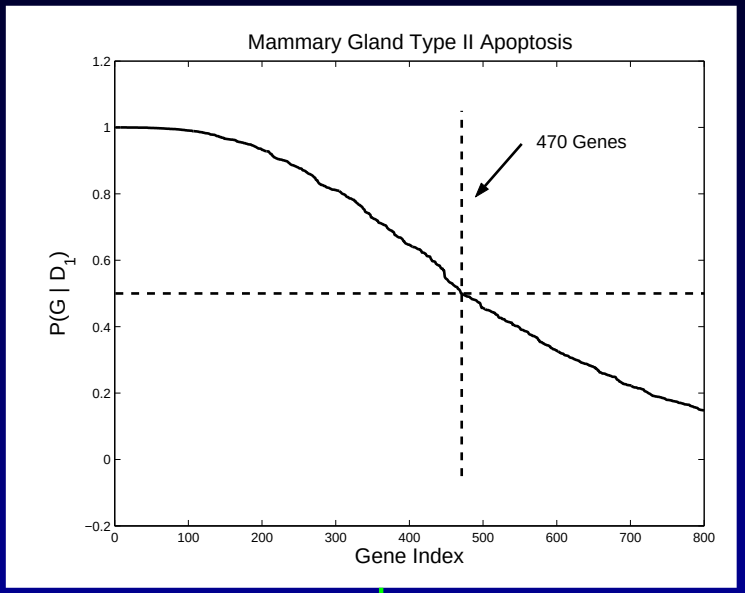
$$P(I_t|\mathcal{D}_1,\mathcal{D}_2) = \frac{P(I_t|\mathcal{D}_1)p(\mathcal{D}_2|I_t)}{p(\mathcal{D}_2|\mathcal{D}_1)}$$

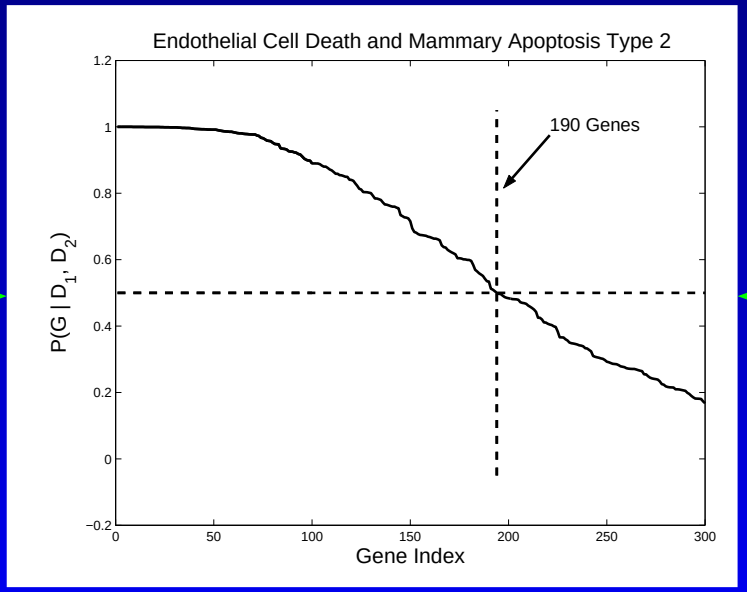
It’s simple and quick





It’s simple and quick





Ranking without Censoring

	Evaluation of Ranking		Evaluation of Censoring	
	gene nr.	$Q(I_t)$	gene nr.	$Q(I_t)$
group 1	gene 1,4	0.999	gene 1,3	0.999
	gene 1,1	0.999	gene 1,1	0.999
	gene 1,3	0.999	gene 1,2	0.999
	gene 1,2	0.999	gene 1,4	0.999
group 2	gene 2,3	0.554	gene 2,2	0.998
	gene 2,4	0.499	gene 2,4	0.995
	gene 2,2	0.400	gene 2,3	0.989
	gene 2,1	0.194	gene 2,1	0.969
group 3	gene 3,4	0.049	gene 3,2	0.147
	gene 3,2	0.040	gene 3,3	0.088
	gene 3,3	0.039	gene 3,4	0.042
	gene 3,1	0.034	gene 3,1	0.033

However Rather Sensitive

The same data is used to calculate rank probabilities in dependency of $\Lambda = \lambda I$.

$P(I|\mathcal{D})$

$1/\lambda$	100	10	3.2	1	0.79	0.5	0.4	0.32	0.25	0.2
$P(I=1\|\mathcal{D})$	0.1	0.35	0.44	0.3	0.26	0.21	0.19	0.19	0.2	0.2
$P(I=2\|\mathcal{D})$	0.29	0.22	0.15	0.22	0.26	0.34	0.36	0.37	0.36	0.35
$P(I=3\|\mathcal{D})$	0.6	0.4	0.34	0.33	0.31	0.28	0.26	0.25	0.24	0.24
$P(I=4\|\mathcal{D})$	0.008	0.031	0.066	0.15	0.16	0.18	0.18	0.19	0.2	0.21

and Rankings

$1/\lambda$	100	10	3.2	1	0.79	0.50	0.4	0.32	0.25	0.2
	3	2	1	2	3	3	3	3	3	4
	2	3	3	3	2	1	1	1	1	1
	1	1	2	1	1	2	2	2	2	2
	4	4	4	4	4	4	4	4	4	3

It should not be surprising that modifying regularisation has an effect of modelling!

Dilemma:

Bayes theorem, which actually allows us to calculate what we want in the first place:

$$P(I_t|\mathcal{D}_1,\mathcal{D}_2) = \frac{P(I_t|\mathcal{D}_1)p(\mathcal{D}_2|I_t)}{p(\mathcal{D}_2|\mathcal{D}_1)}$$

also requires for the above $\boldsymbol{\beta}$ to specify a prior $p(\boldsymbol{\beta}|I_t)$

and that guy introduces nasty side effects when calculating the model probabilities.

What can we do?

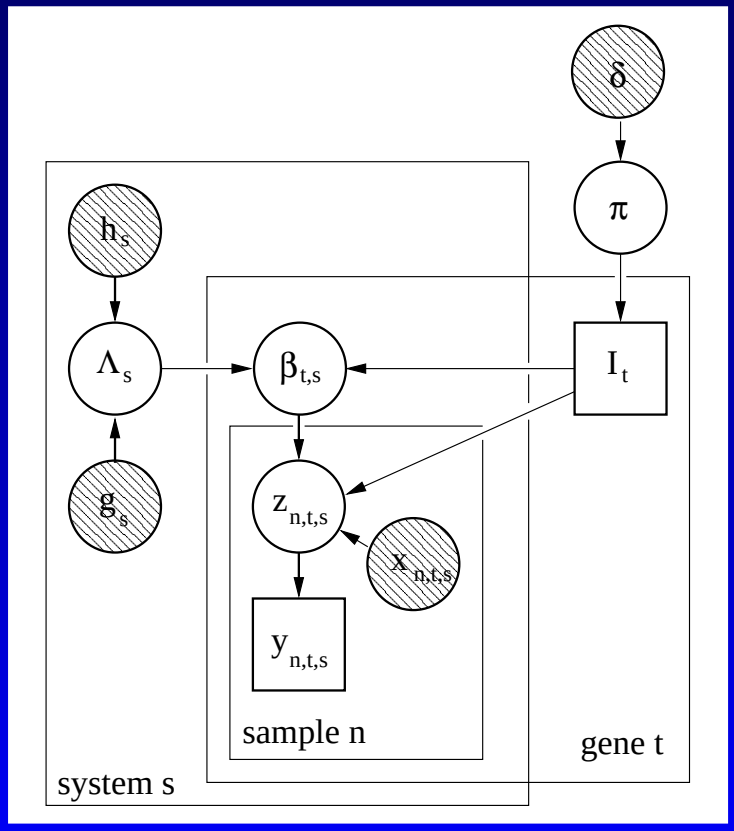
Improving on Previous Model

- Hyper parameters (π_t, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$) influence probability measure $P(I_t|\mathcal{D}_1, \mathcal{D}_2)$.
- Less critical for $P(I_t = 1|\pi_t)$ (e.g. 0.5 for ignorance). However even a pragmatic approach for adjusting $\mathbf{\Lambda}$ like $\min_t p(\hat{\boldsymbol{\beta}}_t|\mathbf{\Lambda}) = 0.95\, p(\mathbf{0}|\mathbf{\Lambda})$ is not convincing. (Why 0.95 ?)

Improving on Previous Model

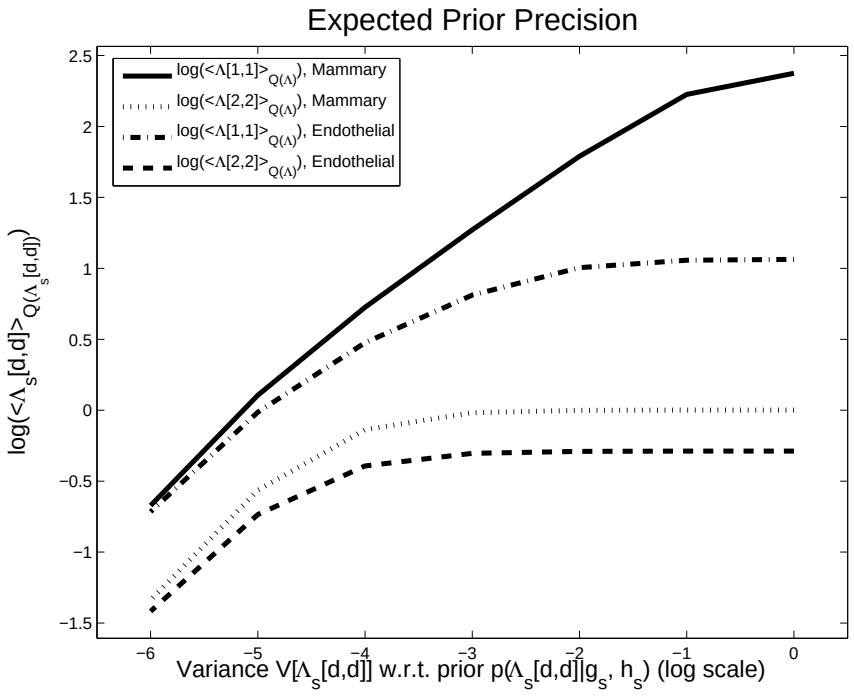
- Hyper parameters (π_t, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$) influence probability measure $P(I_t|\mathcal{D}_1, \mathcal{D}_2)$.
- Less critical for $P(I_t = 1|\pi_t)$ (e.g. 0.5 for ignorance). However even a pragmatic approach for adjusting $\mathbf{\Lambda}$ like $\min_t p(\hat{\boldsymbol{\beta}}_t|\mathbf{\Lambda}) = 0.95\, p(\mathbf{0}|\mathbf{\Lambda})$ is not convincing. (Why 0.95 ?)

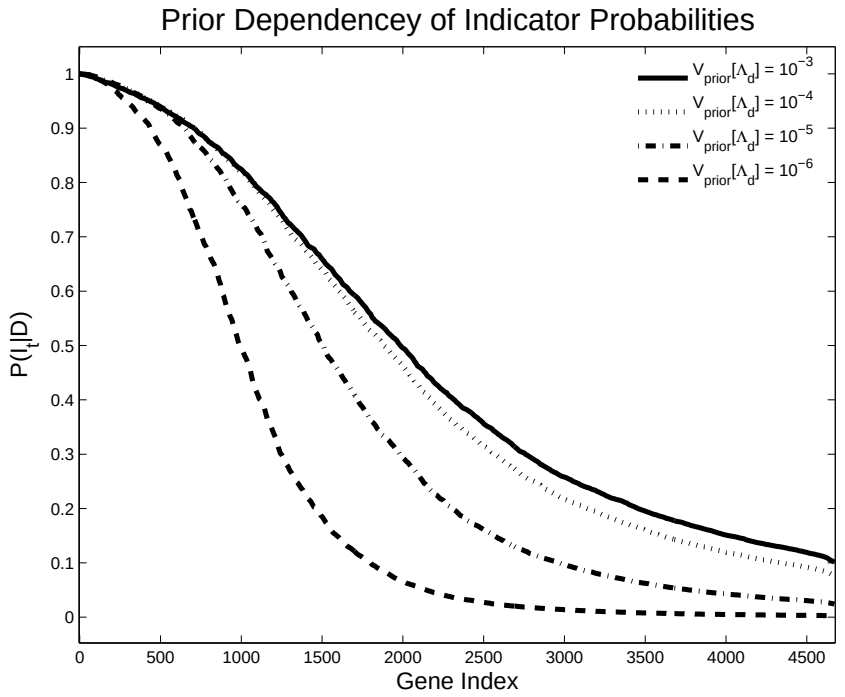
Better solution uses hierarchical priors



- all genes contribute to inference of $\mathbf{\Lambda}_s$
- hierarchical priors for sensitivity analysis
- $Q(I_t)$ approximates gene measure
- using *one* model gets all marginals right

Sensitivity Analysis

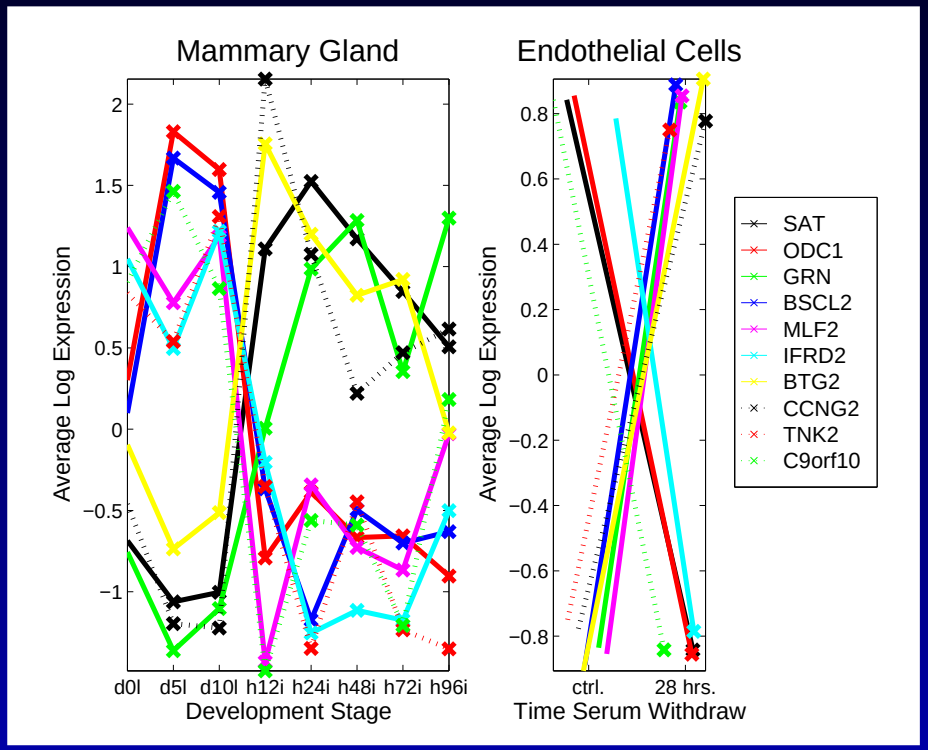




For the hyper parameters this suggests $g \leq 0.01$ and $h \leq 1$.

We also conclude that equal cost results in many potential candidate genes.

Top Ten



Top 10 $P(I_t = 1|\mathcal{D}_1, \mathcal{D}_2)$ for Mammary lactation vs. involution *and* Endothelial cell death.

Gene Symbol	$P(I_t\|\mathcal{D})$
SAT	0.99951
ODC1	0.99921
GRN	0.99921
BSCL2	0.99919
MLF2	0.99884
IFRD2	0.99867
BTG2	0.99843
CCNG2	0.99826
TNK2	0.99789
C9orf10	0.99783

Biological meaning of this list could provide an important sanity check of the approach!

Gene Ontology Assessment

Gene lists are difficult to assess for Biological meaning.

Compact summary by mapping to Gene ontology DAG:

- Reannotate the (always) inconsistent GO annotations.
- Use Fishers exact test to infer GO categories with a significant enrichment of active over inactive genes (FATIGO).

Result:

238 active GO categories, many related to metabolic processes. Several active GO categories from the “cell death” subgraph are in line with our biological hypothesis and an indirect benchmark of the ranking.

Summary

- A Bayesian approaches provide means for data integration, parameter inference and selecting appropriate model classes.
- Avoid non hierarchical models for combining information - arbitrary gene measures can be adjusted for using the “right” prior.
- Analysis results that go beyond gene lists allow for a more efficient communication with biologists and may provide indirect evidence for gene lists.
- Supplemental information for our recent Bioinformatics publication is available at `http://www.sykacek.net/research.html#mcabf` (also GPL MatLab code).

Acknowledgements

Collaborators:

University of Cambridge, UK:
Gos Micklem
Rob Furlong
David J. C. MacKay

University of Cardiff, UK:
Richard Clarkson

University of Auckland, NZ:
Cris Print

Funding:

WWTF
ARCS
Baxter AG
BBSRC

Postdocs wanted: http://www.biotec.boku.ac.at/bijobs.html!

Table of Contents

Variational Bayes I

Mean field Ansatz plus Jensen's inequality. For all pdfs $Q(\theta)$:

$$\log\left(\int_\theta p(D|\theta)p(\theta)d\theta\right) \geq$$
$$\int_\theta \big(\log(p(D|\theta)) + \log(p(\theta)) - \log(Q(\theta))\big)Q(\theta)d\theta$$
$$= \log(p(D)) + \int_\theta \big(\log(p(\theta|D)) - \log(Q(\theta))\big)Q(\theta)d\theta$$

the last integral is a negative Kullback Leibler divergence and thus smaller or equal zero.

\+ easy to compute; - systematic error as only an approximation.

back

jump 2 TOC

Variational Bayes II

Joint Distribution implied by the previous DAG

$$p(I_t, \boldsymbol{\beta}_{1,t}, S_{1,t}, D_{1,t} | \boldsymbol{\Lambda}_1, \pi_t, \gamma, X_{1,t}) = P(I_t|\pi_t) p(\boldsymbol{\beta}_{1,t}|\boldsymbol{\Lambda}_1, I_t)$$
$$\times \prod_n \Big(p(s_{1,t,n}|\boldsymbol{\beta}_{1,t}, \boldsymbol{x}_{1,t,n}, I_t, \gamma) P(y_{1,t,n}|s_{1,t,n}, I_t) \Big)$$

where $S_{1,t} = \{s_{1,t,1}, ..., s_{1,t,N}\}$ and $D_{1,t} = \{y_{1,t,1}, ..., y_{1,t,N}\}$.

- Approximate posterior by a mean field expansion $Q(\boldsymbol{\beta}_{1,t}|I_t) \prod_n Q(s_{1,t,n}|I_t)$.
- Write down negative free energy and maximise the functional iteratively w.r.t. all Q-distributions.
- The negative free energy $F_{\max}(Q)$ approximates the log marginal likelihood and thus $P(I_t|D_{1,t}, \boldsymbol{\Lambda}_1, \pi_t, \gamma, X_{1,t})$.

back