Hierarchical Bayesian Modelling Identifies Shared Gene Function

Peter Sykacek[12] & Gos Micklem[13]

Departments of Genetics[1], Pathology[2] & CCBI[3]
University of Cambridge
ps408@cam.ac.uk

Data & Biology: Richard Clarkson, Cris Print
Methodological Discussions: David J. C. MacKay & Inference Group

Problem Statement

- 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

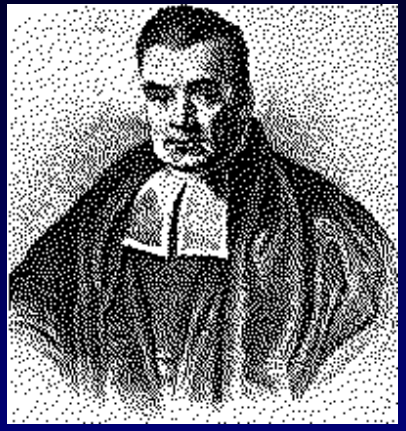
Mammary Gland tc. (lact. day & hours of involution)

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	+	-	-

Probabilistic Approach



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

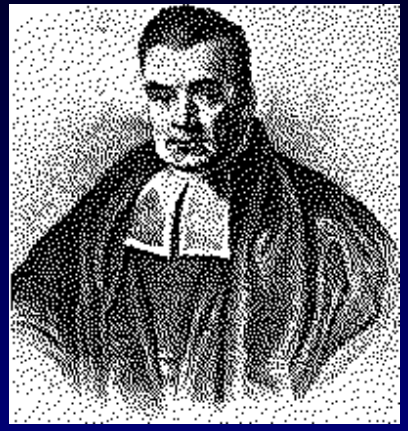
Probabilistic Approach

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 Approach



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

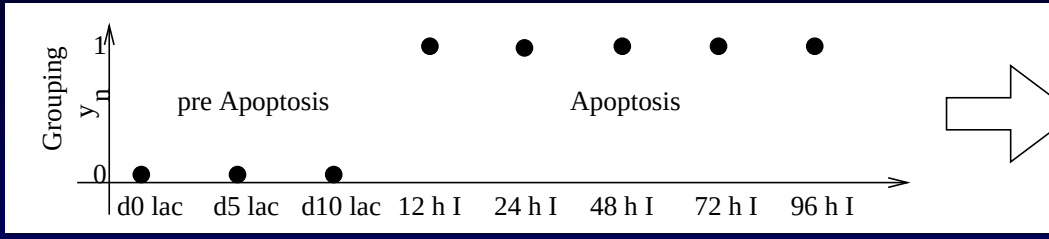
$$\alpha_{opt} = \text{argmax}_{\alpha} < u(\alpha) >$$, where

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

Second consequence: Decisions by maximising expected utilities

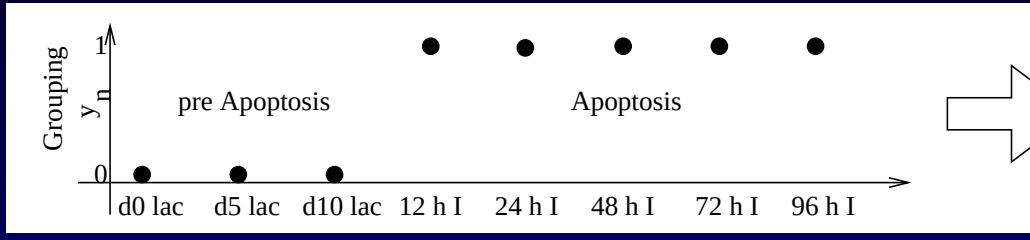
Probabilistic Gene Ranking

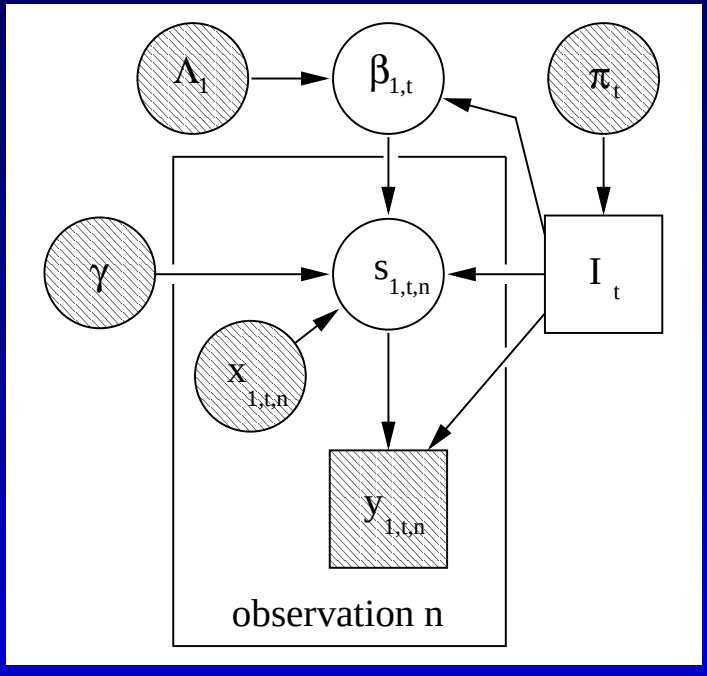
Apoptosis (lac. vs. inv.!) in the Mouse Mammary Gland



Probabilistic Gene Ranking

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





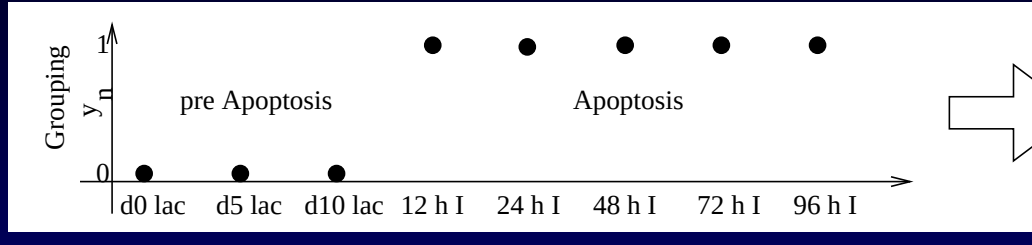
Latent variable probit GLM.

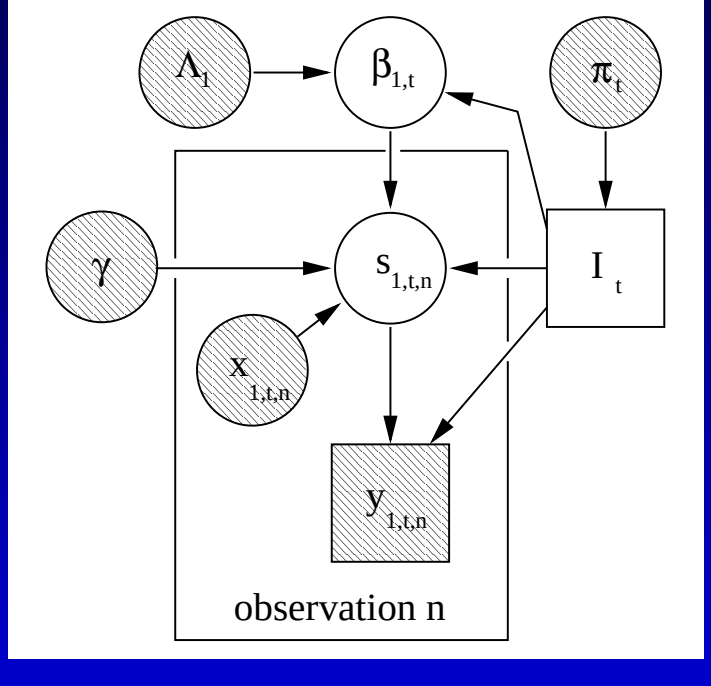
$$\text{if } I_t = \begin{cases} 1 : s_{1,t,n} \sim 1 + x_{t,n} \\ 0 : s_{1,t,n} \sim 1 \end{cases}$$

$s_{1,t,n}$ 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.

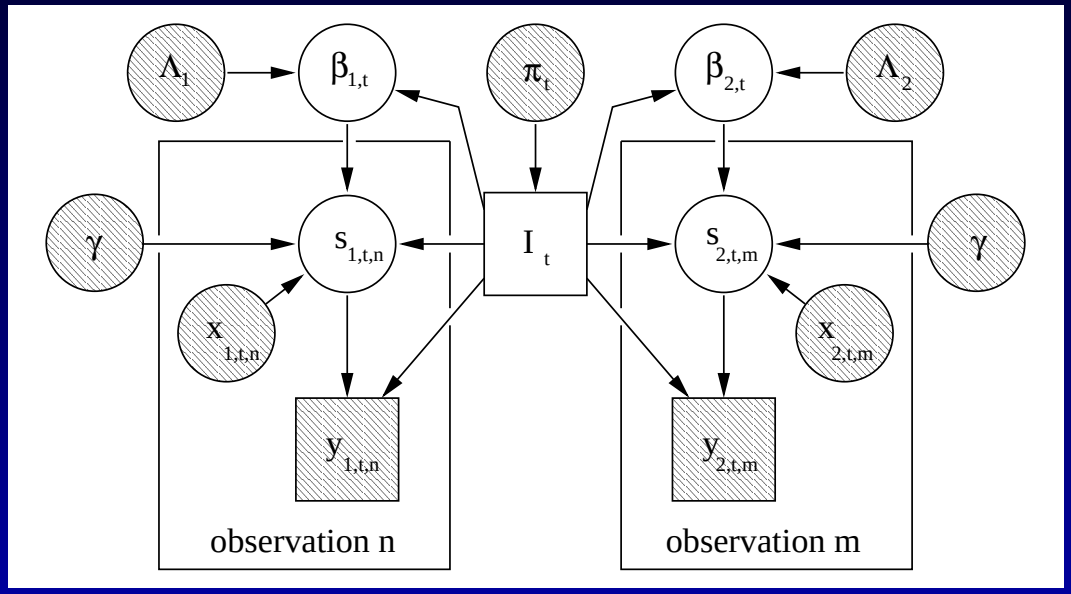
$$\text{if } I_t = \begin{cases} 1 : s_{1,t,n} \sim 1 + x_{t,n} \\ 0 : s_{1,t,n} \sim 1 \end{cases}$$

$s_{1,t,n}$ 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, the posterior $P(I_t|\mathcal{D}_1)$, serves as a probabilistic rank measure. (VB-eqns.)

Shared Gene Function

Include Information about Endothelial Cell Death

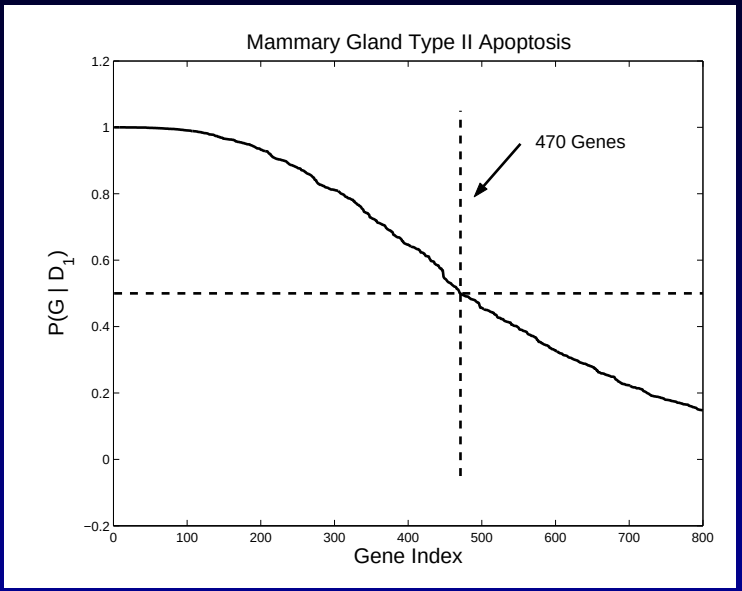


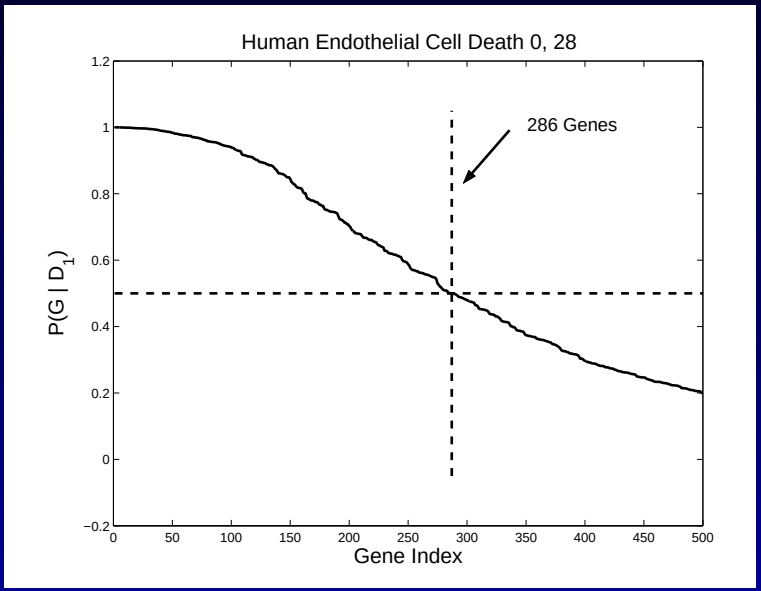
Model 0 hrs. vs. 28 hrs. as latent variable probit GLM. Calculate $P(\mathcal{D}_2|I_t)$, the marginal likelihood.

Bayes theorem gives a *principled* measure for ranking

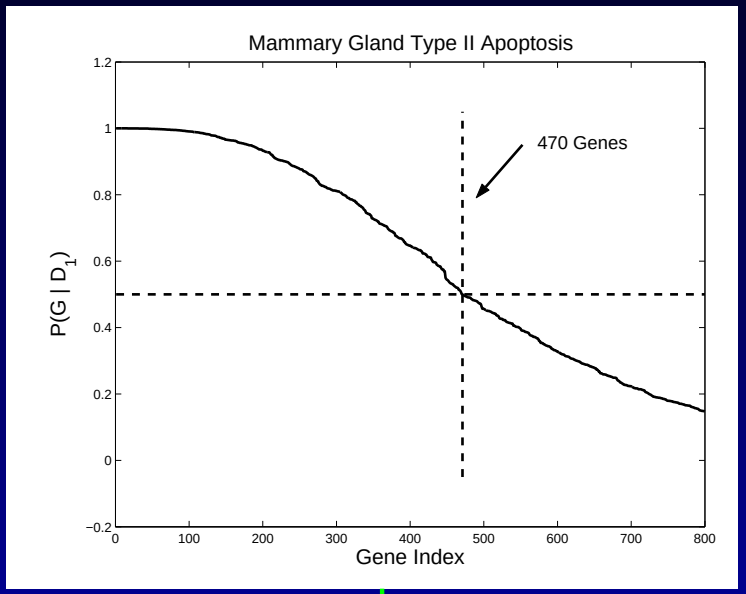
$$P(I_t|\mathcal{D}_1,\mathcal{D}_2) = \frac{P(I_t|\mathcal{D}_1)p(\mathcal{D}_2|I_t)}{p(\mathcal{D}_2|\mathcal{D}_1)}$$

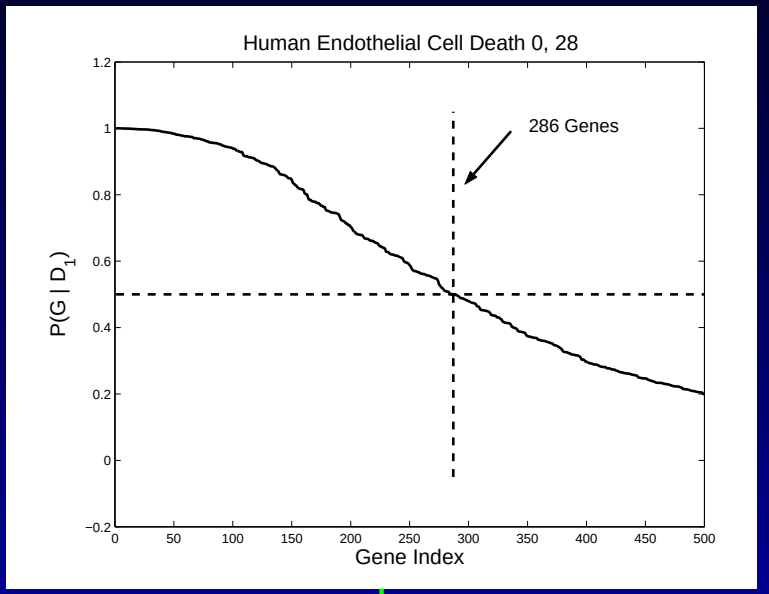
Don’t Do That at Home!

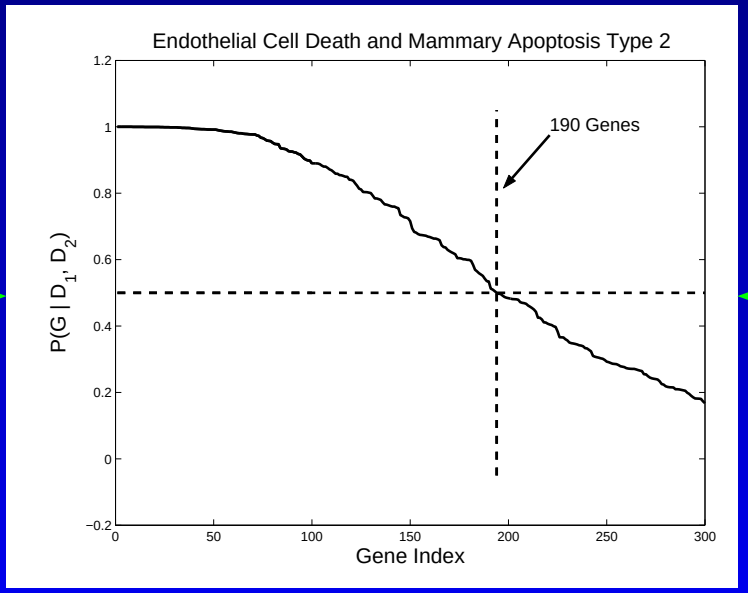




Don’t Do That at Home!







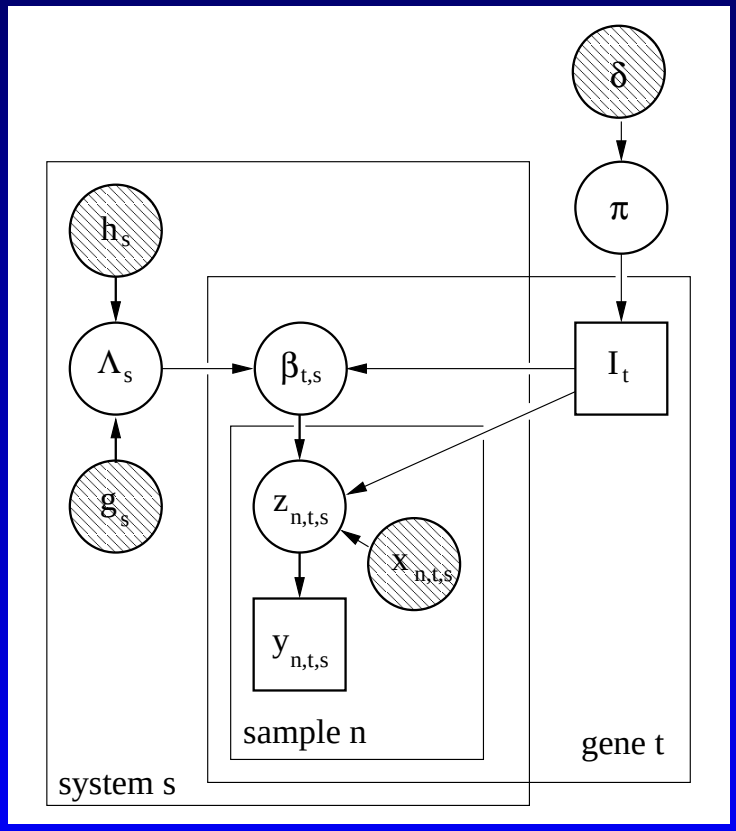
Improving on Previous Model

- 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

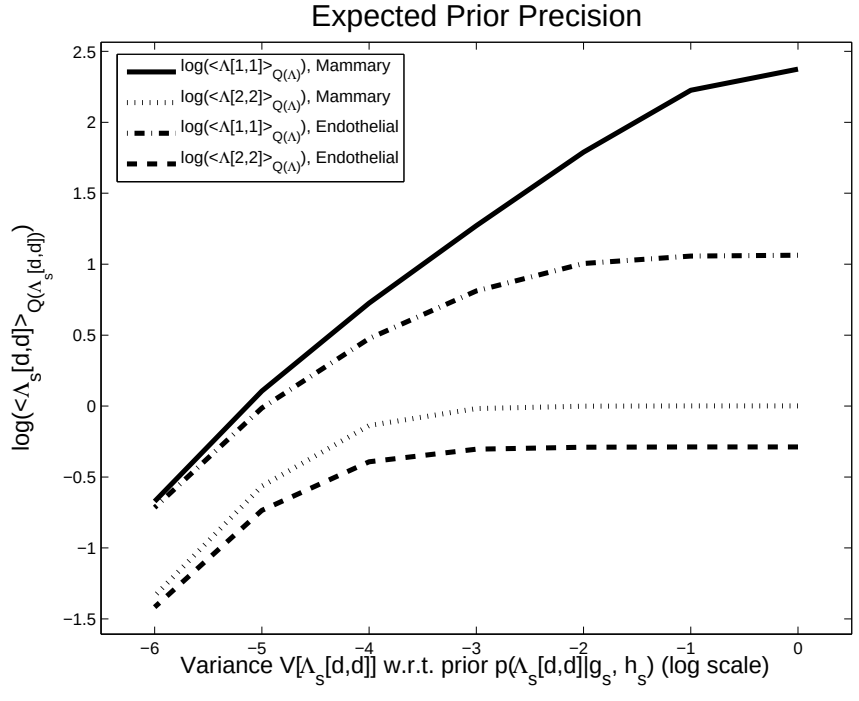
- Hyper parameters (π_t, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$) influence probability measure $P(I_t|\mathcal{D}_1, \mathcal{D}_2)$.
- Less critical for $P(I_t = 1|\pi_t)$ (e.g. 0.5 for ignorance). However even a pragmatic approach for adjusting $\mathbf{\Lambda}$ like $\min_t p(\hat{\boldsymbol{\beta}}_t|\mathbf{\Lambda})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

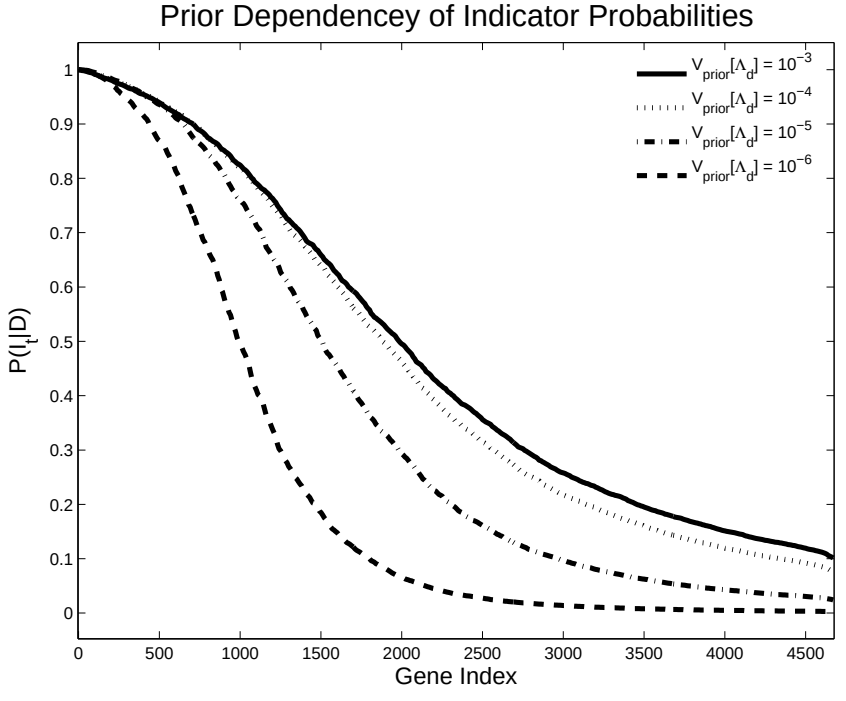
Better solution uses hierarchical priors



- 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 Check

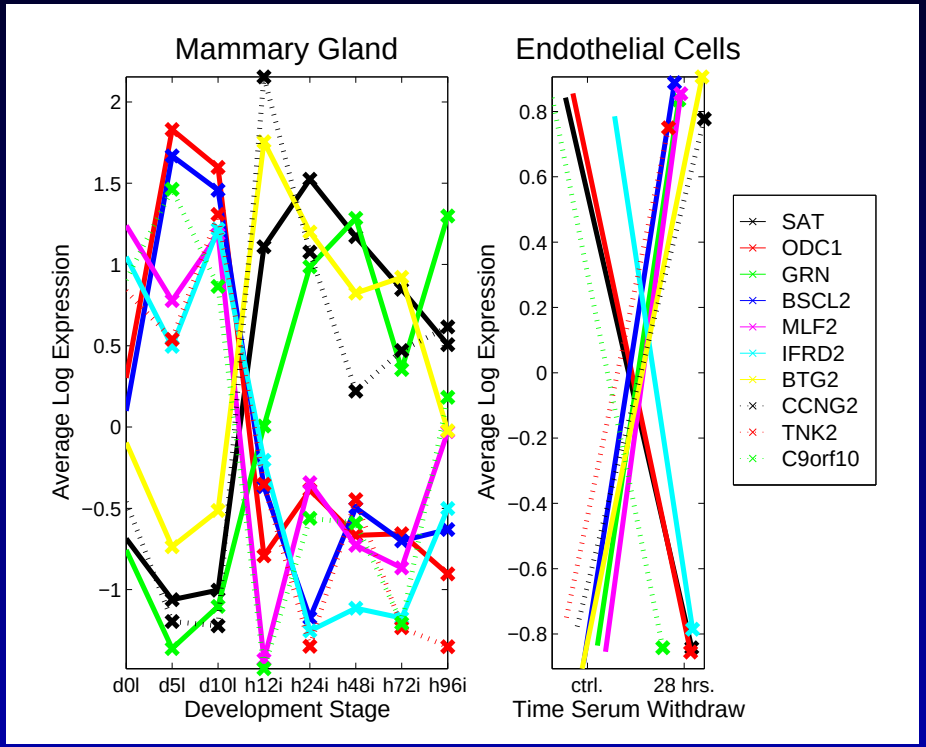




For the hyper parameters this suggests $g \leq 0.01$ and $h \leq 1$.

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

Top Ten



Top 10 $P(I_t = 1|\mathcal{D}_1, \mathcal{D}_2)$ for Mammary lactation vs. involution *and* Endothelial cell death (result updated 01 2007).

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

Summary

- A realtively straight forward approach allows inference of shared gene function.
- Beware of non hierarchical models - arbitrary gene measures can be adjusted for using the “right” prior.
- Variational methods provide a rather efficient tool to explore models before deciding for a final possibly MCMC based implementation.

Table of Contents

Variational Bayes

Mean field ansatz plus Jensens 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

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 maximize 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