Colouring and breaking sticks, pairwise coincidence losses, and clustering expression profiles

Peter Green and John Lau

University of Bristol
P.J.Green@bristol.ac.uk

Isaac Newton Institute, 11 December 2006

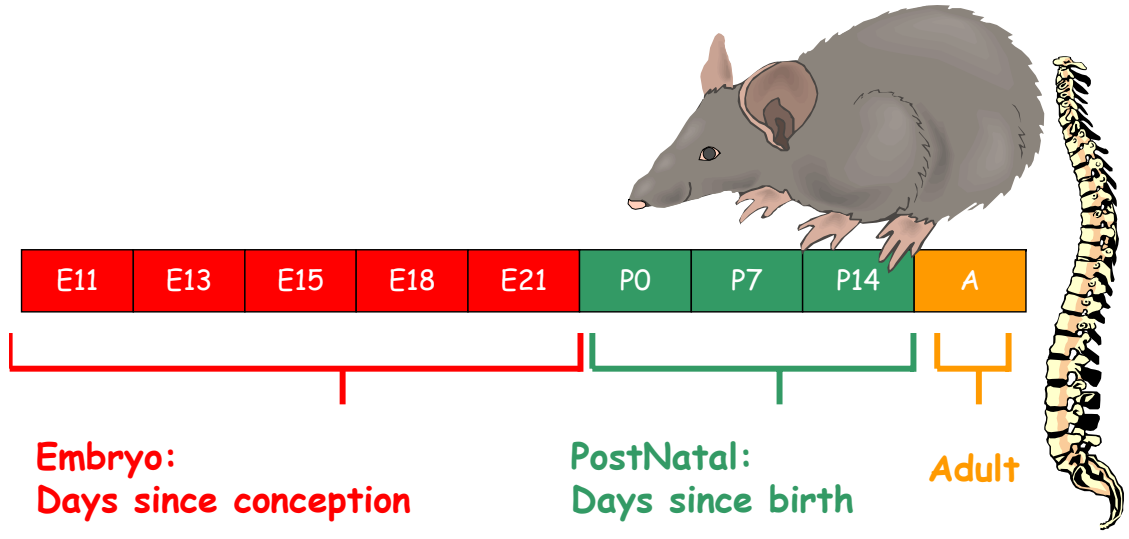
Gene expression data

We work with possibly replicated gene expression measures, often from Affymetrix gene chips. Data are $\{Y_{gsr}\}$, indexed by

- genes $g = 1, 2, \ldots, n$
- conditions $s = 1, 2, \ldots, S$, and
- replicates $r = 1, 2, \ldots, R_s$

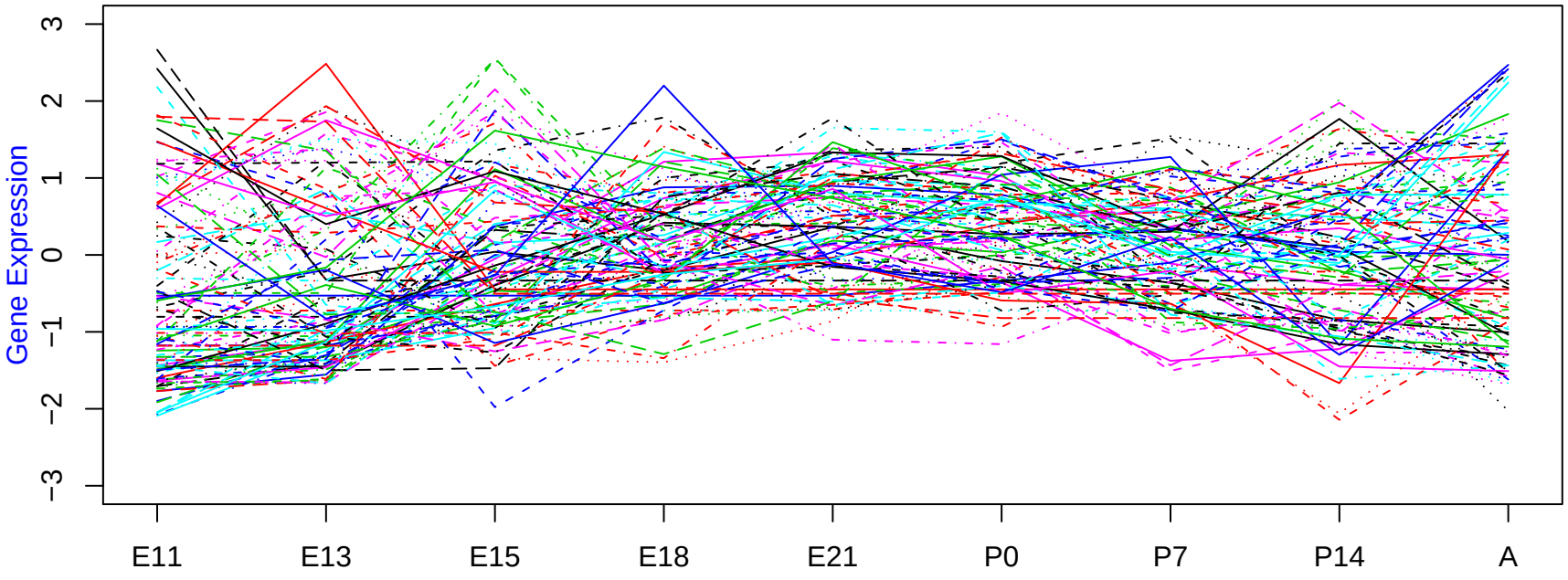
Typically R_s is very small, S is much smaller than n, and the 'conditions' represent different subjects, different treatments, or different experimental settings.

Rats CNS development



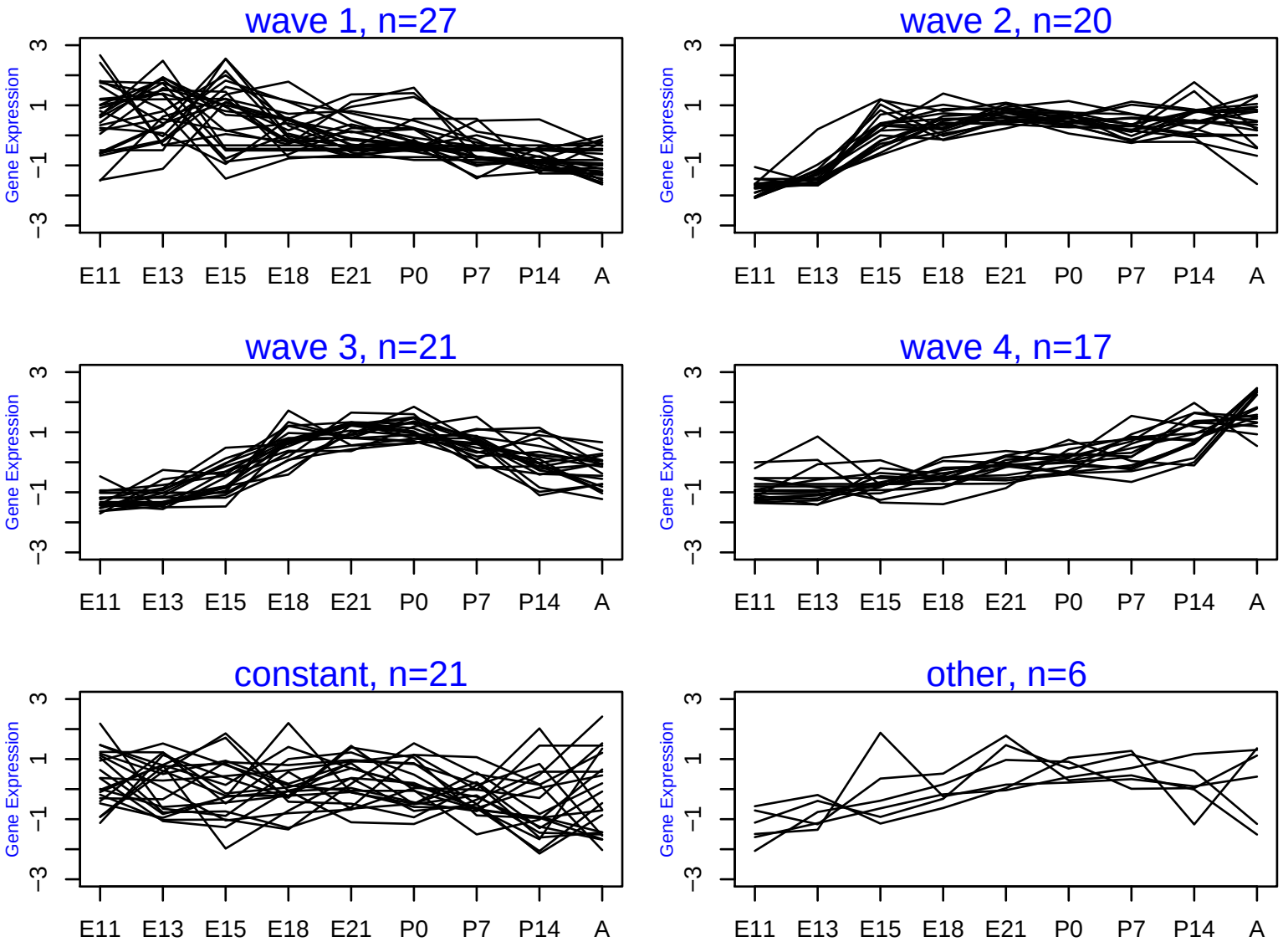
Wen et al (*PNAS*, 1998) studied development of central nervous system in rats: mRNA expression levels of 112 genes at 9 time points.

Rats data, normalised



Wen et al found clusters (waves) characterising distinct phases of development. . .

Rats data: Wen's clustering of genes

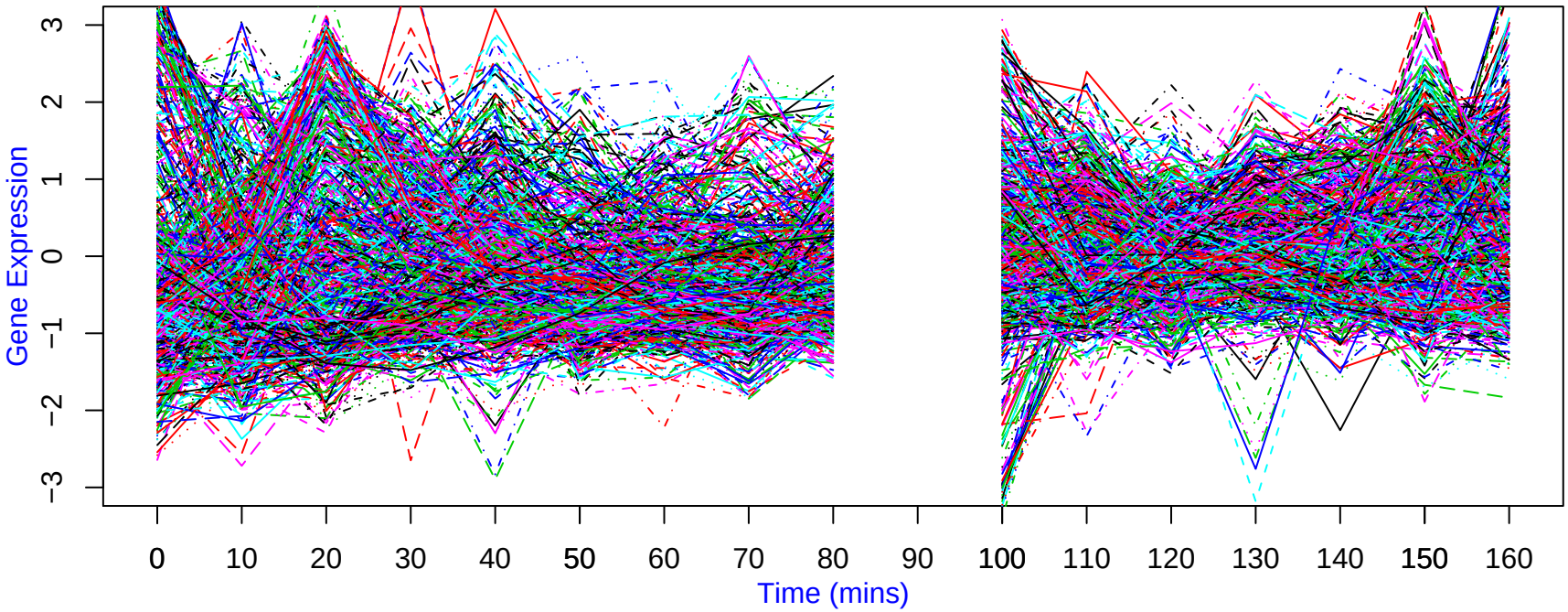


Yeast cell cycle data

Data from Cho et al (*Mol. Cell*, 1998) (can also be found in **R** `som` package). Yeast culture synchronised in G_1, then released and RNA collected at 10 minute intervals over 160 minutes ($\approx$ two cell cycles). 6601 genes $\times$ 17 time points. $t = 90$ excluded because of scaling difficulties.

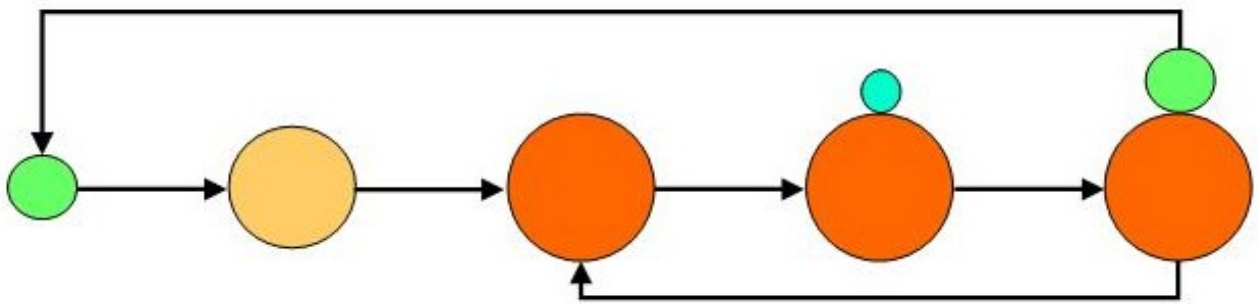
The biological interest is in identifying genes that are up- or down-regulated during the key phases of the cell cycle (early G_1, late G_1, S, G_2 and M), some of which may be involved in controlling the cycle itself.

Yeast data, normalised

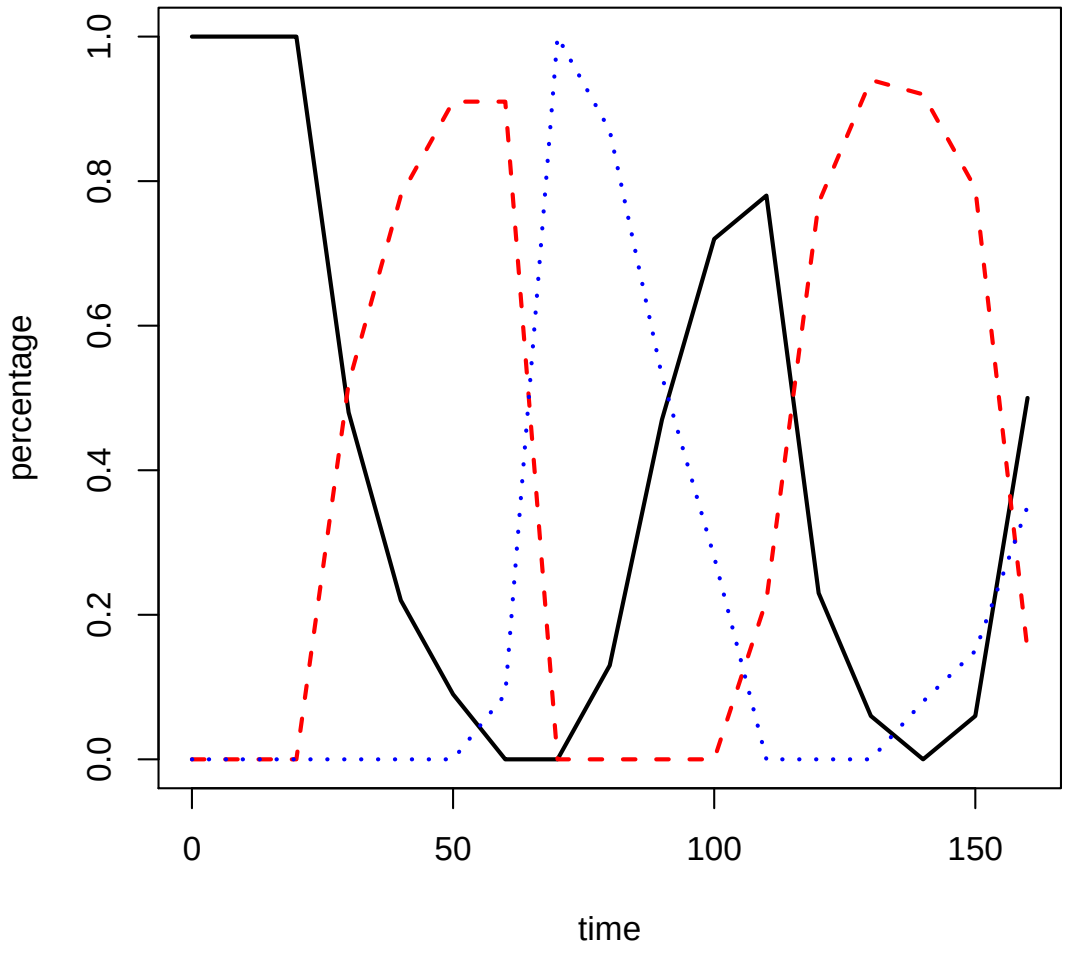


Yeast cell cycle

We have data on percentages of cells in each of three phases of growth (unbudded/small-budded/large-budded).



Yeast data, cell phase statistics



Keywords

- Gene expression profiles
- Time course experiments, other numerical covariates
- Dirichlet process and other mixtures
- Heterogeneous mixtures, by 'colouring and breaking sticks'
- MCMC samplers for partition models
- Loss functions and optimal clustering

1. Parametric expression profiles

We suppose there is a k-dimensional ($k \leq S$) covariate vector x_s describing each condition, and model parametric dependence of Y on x, whilst regarding genes as *a priori* exchangeable, seeking common patterns across s under a nonparametric model for clustering.

Although other variants are easily envisaged (and we see a generalisation later), we suppose initially that

$$Y_{gsr} \sim N(x_s' \beta_g, \tau_g^{-1}), \quad \text{independently}$$

where $\theta_g = (\beta_g, \tau_g) \in \mathcal{R}^{k+1}$ are drawn i.i.d. from a distribution G, where in turn G has a Dirichlet process prior:

$$G \sim DP(\alpha, G_0)$$

The Dirichlet process - view 0

Given a probability distribution G_0 on an arbitrary measure space Ω, and a positive real α, we say the random distribution G on Ω follows a Dirichlet process,

$$G \sim DP(\alpha, G_0)$$

if for all partitions $\Omega = \bigcup_{j=1}^{m} B_j$ ($B_j \cap B_k = \emptyset$ if $j \neq k$), and for all m,

$$(G(B_1), \ldots, G(B_m)) \sim \text{Dirichlet}(\alpha G_0(B_1), \ldots, \alpha G_0(B_m))$$

The Dirichlet process - view 0 (ctd.)

The base measure G_0 gives the *expectation* of G:

$$E(G(B)) = G_0(B)$$

Even if G_0 is continuous, G is a.s. discrete, so i.i.d. draws $\{\theta_g, g = 1, 2, \ldots, n\}$ from G exhibit ties.

α measures (inverse) *concentration*: given i.i.d. draws $\{\theta_g, g = 1, 2, \ldots, n\}$ from G,

- As $\alpha \to 0$, all θ_g are equal, drawn from G_0.
- As $\alpha \to \infty$, the θ_g are drawn i.i.d. from G_0.

The Dirichlet process - view 1

Sethuraman's 'stick-breaking' construction of G:

- draw $\theta_j^\star \sim G_0$, i.i.d., $j = 1, 2, \ldots$
- draw $V_j \sim \text{Beta}(1, \alpha)$, i.i.d., $j = 1, 2, \ldots$
- define G to be the discrete distribution putting probability $(1 - V_1)(1 - V_2) \ldots (1 - V_{j-1})V_j$ on $\theta_j^\star$
- draw θ_g i.i.d from G, $g = 1, 2, \ldots, n$.

The Dirichlet process - view 2

Finite mixture model $\sum_j w_j g_0(\cdot|\theta_j^\star)$ with Dirichlet weights:

- Draw $(w_1, w_2, \ldots, w_k) \sim$ Dirichlet$(\delta, \ldots, \delta)$
- Draw $c_g \in \{1, 2, \ldots, k\}$ with $P\{c_g = j\} = w_j$, i.i.d., $g = 1, \ldots, n$
- Draw $\theta_j^\star \sim G_0$, i.i.d., $j = 1, \ldots, k$
- Set $\theta_g = \theta_{c_g}^\star$

Let $k \to \infty$, $\delta \to 0$ such that $k\delta \to \alpha$.

G is invisible in view 2.

The Dirichlet process - view 3

Partition model: partition $\{1, 2, \ldots, n\} = \bigcup_{j=1}^{d} C_j$ at random, so that

$$p(C_1, C_2, \ldots, C_d) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^d \prod_{j=1}^{d} (n_j - 1)!$$

where $n_j = \#C_j$. (NB preference for unequal cluster sizes!) Draw $\theta_j^\star \sim G_0$, i.i.d., $j = 1, \ldots, d$, and set $\theta_g = \theta_j^\star$ if $g \in C_j$.

This model is the same as that leading to Ewens' sampling formula (which brings in multiplicities).

G is also invisible in view 3.

The Dirichlet process - reprise

Genes are clustered, according to a tractable distribution parameterised by $\alpha > 0$, and within each cluster the regression parameter/precision pair $\theta = (\beta, \tau)$ is drawn i.i.d. from G_0.

How nonparametric is that?

We take a standard normal–inverse Gamma model: $\theta = (\beta, \tau) \sim G_0$ means

$$\tau \sim \Gamma(a_0, b_0) \quad \text{and} \quad \beta|\tau \sim \mathsf{N}_k(m_0, (\tau t_0)^{-1} I)$$

This is a conjugate set-up, so that (β, τ) can be integrated out *in each cluster*.

Multiple notations for partitions

- $\boldsymbol{c}$ is a partition of $\{1, 2, \ldots, n\}$
- clusters of partition are $C_1, C_2, \ldots, C_d$
  (d is the *degree* of the partition):
  $\bigcup_{j=1}^{d} C_j = \{1, 2, \ldots, n\}$, $C_j \cap C_{j'} = \emptyset$ if $j \neq j'$
- c is the allocation vector: $c_g = j$ if and only if $g \in C_j$

!Take care with multiplicities, and distinction between allocations and partitions: labelling of C_j is arbitrary, likewise values of $\{c_g\}$.

The incremental algorithm (Gibbs sampler/Pólya urn/ Weighted Chinese restaurant process)

Exploiting conjugacy, here we only consider posterior sampling of partition alone.

MCMC on posterior for partition, limited to re-allocating single gene at a time (single-variable Gibbs sampler for c_g).

We allocate Y_g to a new cluster $C_\star$ with probability

$$\propto p(\boldsymbol{c}^{g\to\star}|\alpha) \times p(Y_g|\psi),$$

$\boldsymbol{c}^{g\to\star}$ denotes the current partition $\boldsymbol{c}$ with g moved to $C_\star$.

and to cluster C_j^{-g} with probability

$$\propto p(\boldsymbol{c}^{g \to j}|\alpha) \times (Y_{C_j^{-g} \cup \{g\}}|\psi)/p(Y_{C_j^{-g}}|\psi).$$

$\boldsymbol{c}^{g \to j}$ denotes the partition $\boldsymbol{c}$, with g moved to cluster C_j.

The ratio of marginal likelihoods $p(Y|\psi)$ can be interpreted as the posterior predictive distribution of Y_g given those observations already allocated to the cluster, i.e. $p(Y_g|Y_{C_j^{-g}}, \psi)$ (= multivariate t for NIG setup).

For Dirichlet mixtures, we have

$$p(\boldsymbol{c}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}\alpha^d \prod_{j=1}^{d}(n_j-1)!$$

where $n_j = \#C_j$ and $\boldsymbol{c} = (C_1, C_2, \ldots, C_d)$, so the re-allocation probabilities are explicit and simple in form.

But the same sampler can be used for many other partition models.

And the idea is not limited to moving one item at a time.

When the incremental sampler applies

All we require of the model are that

(a) A partition $\boldsymbol{c}$ of $\{1, 2, \ldots, n\}$ is drawn from a distribution with parameter α

(b) Conditionally on $\boldsymbol{c}$, parameters $(\theta_1, \theta_2, \ldots, \theta_d)$ are drawn independently from a distribution G_0 (possibly with a hyperparameter ψ)

(c) Conditional on $\boldsymbol{c}$ and on $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$, $\{y_1, y_2, \ldots, y_n\}$ are drawn independently, from not necessarily identical distributions $p(y_i|\boldsymbol{c}, \theta) = f_i(y_i|\theta_j)$ for $i \in C_j$.

Examples

$p(\boldsymbol{c}^{g\to\star}|\alpha)$ and $p(\boldsymbol{c}^{g\to j}|\alpha)$ are simply proportional to

- α and $\#C_j^{-g}$ for the DP mixture model
- $(k - d(\boldsymbol{c}^{-g}))\delta$ and $\#C_j^{-g} + \delta$ for the Dirichlet-multinomial finite mixture model
- $\theta + \alpha d(\boldsymbol{c}^{-g})$ and $\#C_j^{-g} - \alpha$ for the Pitman–Yor two-parameter Poisson–Dirichlet process
- etc., etc.

So the ease of using the Pólya urn/Gibbs sampler is not a reason to use DPM!

2. A Coloured Dirichlet process

To define a variant on the DP in which not all clusters are exchangeable:

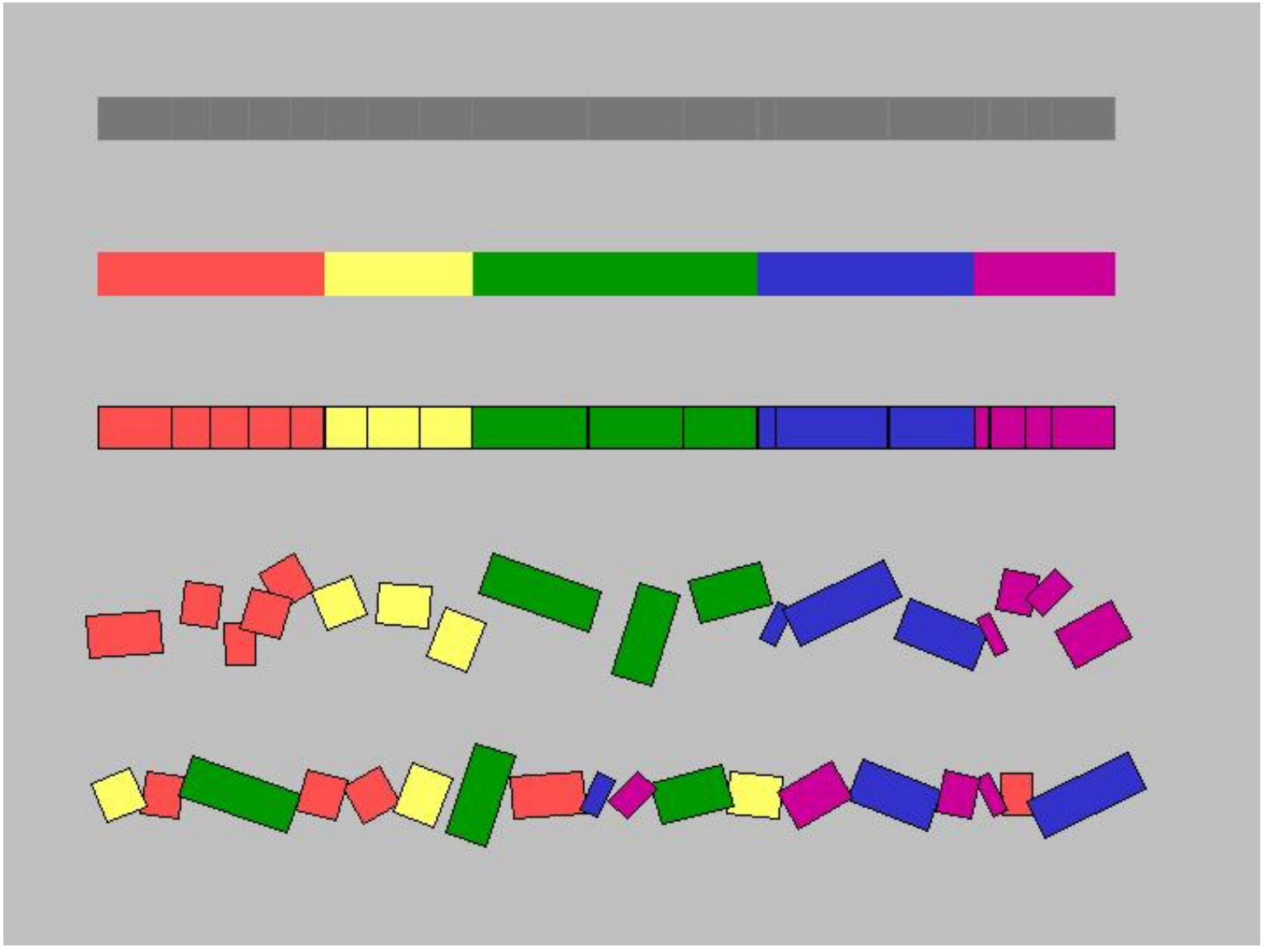
- for each 'colour' $k = 1, 2, \ldots$, draw G_k from a Dirichlet process $\mathrm{DP}(\alpha_k, G_{0k})$, independently for each k
- draw weights (w_k) from the Dirichlet distribution $\mathrm{Dir}(\gamma_1, \gamma_2, \ldots)$, independently of the G_k.
- define G on $\{k\} \times \Omega$ by $G(k, B) = w_k G_k(B)$.
- draw colour–parameter pairs (k_g, θ_g) i.i.d from G, $g = 1, 2, \ldots, n$

This process, denoted CDP$(\{(\gamma_k, \alpha_k, G_{0k})\})$, is a Dirichlet mixture of Dirichlet processes (with different base measures), $\sum_k w_k \mathrm{DP}(\alpha_k, G_{0k})$, with the added feature that the the colour of each cluster is identified (and indirectly observed), while labelling of clusters within colours is arbitrary.

It can be defined by a 'stick-breaking-and-colouring' construction:

- colour segments of the stick using the Dirichlet-distributed weights
- break each coloured segment using an infinite sequence of independent Beta$(1, \alpha_k)$ variables

Colouring and breaking sticks



Coloured partition distribution

The CDP generates the following partition model: partition $\{1, 2, \ldots, n\} = \bigcup_k \bigcup_{j=1}^{d_k} C_{kj}$ at random, so that

$$p(C_{11}, C_{12}, \ldots, C_{1d_1}; C_{21}, \ldots, C_{2d_2}; C_{31}, \ldots) =$$

$$\frac{\Gamma(\sum_k \gamma_k)}{\Gamma(n + \sum_k \gamma_k)} \prod_k \left(\frac{\Gamma(\alpha_k)\Gamma(n_k + \gamma_k)}{\Gamma(n_k + \alpha_k)\Gamma(\gamma_k)} \alpha_k^{d_k} \prod_{j=1}^{d_k} (n_{kj} - 1)! \right)$$

where $n_{kj} = \#C_{kj}$, $n_k = \sum_j n_{kj}$.

Note that the clustering remains exchangeable over items (genes). For $g \in C_{kj}$, set $k_g = k$ and $\theta_g = \theta_j^\star$, where $\theta_j^\star$ are drawn i.i.d. from G_{0k}.

Pólya urn sampler for the CDP

The explicit availability of the (coloured) partition distribution immediately allows generalisation of the Pólya urn Gibbs sampler to the CDP.

In reallocating item g, let n_{kj}^{-g} denote the number *among the remaining items* currently allocated to C_{kj}, and define n_k^{-g} accordingly. Then reallocate g to

- a new cluster of colour k, with probability
  $\propto \alpha_k \times (\gamma_k + n_k^{-g})/(\alpha_k + n_k^{-g}) \times p(Y_g|\psi)$
- the existing cluster C_{kj}, with probability
  $\propto n_{kj}^{-g} \times (\gamma_k + n_k^{-g})/(\alpha_k + n_k^{-g}) \times p(Y_g|Y_{C_{kj}^{-g}}, \psi)$

A Dirichlet process mixture with a background cluster ('top table' model)

In gene expression, natural to suppose a 'background' cluster that is not *a priori* exchangeable with the others.

Take 'limit of finite mixture' view, and adapt it:

- Draw $(w_0, w_1, w_2, \ldots, w_k) \sim$ Dirichlet$(\gamma, \delta, \ldots, \delta)$
- Draw $c_g \in \{0, 1, \ldots, k\}$ with $P\{c_g = j\} = w_j$, i.i.d., $g = 1, \ldots, n$
- Draw $\theta_0^\star \sim H_0$, $\theta_j^\star \sim G_0$, i.i.d., $j = 1, \ldots, k$
- Set $\theta_g = \theta_{c_g}^\star$

Let $k \to \infty$, $\delta \to 0$ such that $k\delta \to \alpha$, but leave γ fixed.

Top table model as a CDP

The top table model is a special case of the CDP, specifically CDP($\{(\gamma, 0, H_0), (\alpha, \alpha, G_0)\}$).

The two colours correspond to the background and regular clusters.

The limiting-case DP$(0, H_0)$ is a point mass, randomly drawn from H_0.

We can go a little further in our regression setting, and allow different regression models for each colour.

Top-table Dirichlet process incremental sampler

When re-allocating gene g, there are three kinds of choice: a new cluster $C_\star$, the 'top table' C_0, or a regular cluster $C_j, j \neq 0$: the corresponding prior probabilities

$p(\boldsymbol{c}^{g\to\star}|\alpha)$, $p(\boldsymbol{c}^{g\to 0}|\alpha)$ and $p(\boldsymbol{c}^{g\to j}|\alpha)$

are proportional to

α, $(\gamma + \#C_0^{-g})$ and $\#C_j^{-g}$

for the asymmetric DP mixture model.

3. Bayesian inference about partitions

The full posterior distribution – computed by sampling – tells us all about the partition and parameters: how to report a point estimate of the partition alone?

The posterior mode (MAP) partition is a common choice: but why? We would usually shy away from using posterior modes in such a high-dimensional problem.

Here we consider going the extra mile – and obtaining optimal Bayesian clustering under a pairwise coincidence loss function.

Loss functions for clustering

So long as our formulation is exchangeable with respect to labelling of both items and clusters, we are confined to loss functions with the same invariances.

These constraints, and issues of tractability, lead us to a pairwise coincidence loss function: for each pair of genes (g, g'), you incur a loss of

- a if $c_g = c_{g'}, \widehat{c}_g \neq \widehat{c}_{g'}$ (false negative)
- b if $c_g \neq c_{g'}, \widehat{c}_g = \widehat{c}_{g'}$ (false positive)

Loss functions for clustering (2)

After a little manipulation, we find minimising expected loss is the same as maximising

$$\ell(\widehat{\boldsymbol{c}}) = \sum_{1 \le g < g' \le G} \{(\rho_{gg'} - K) I[\widehat{c}_g = \widehat{c}_{g'}]\}$$

$$= \sum_{j} \sum_{g,g' \in \widehat{C}_j} (\rho_{gg'} - K)$$

where $K = b/(a+b) \in [0,1]$ and $\rho_{gg'} = P(c_g = c_{g'}|\boldsymbol{y})$. Note this requires only saving the posterior pairwise coincidence probabilities from the MCMC run.

How to optimise this?

Toy example

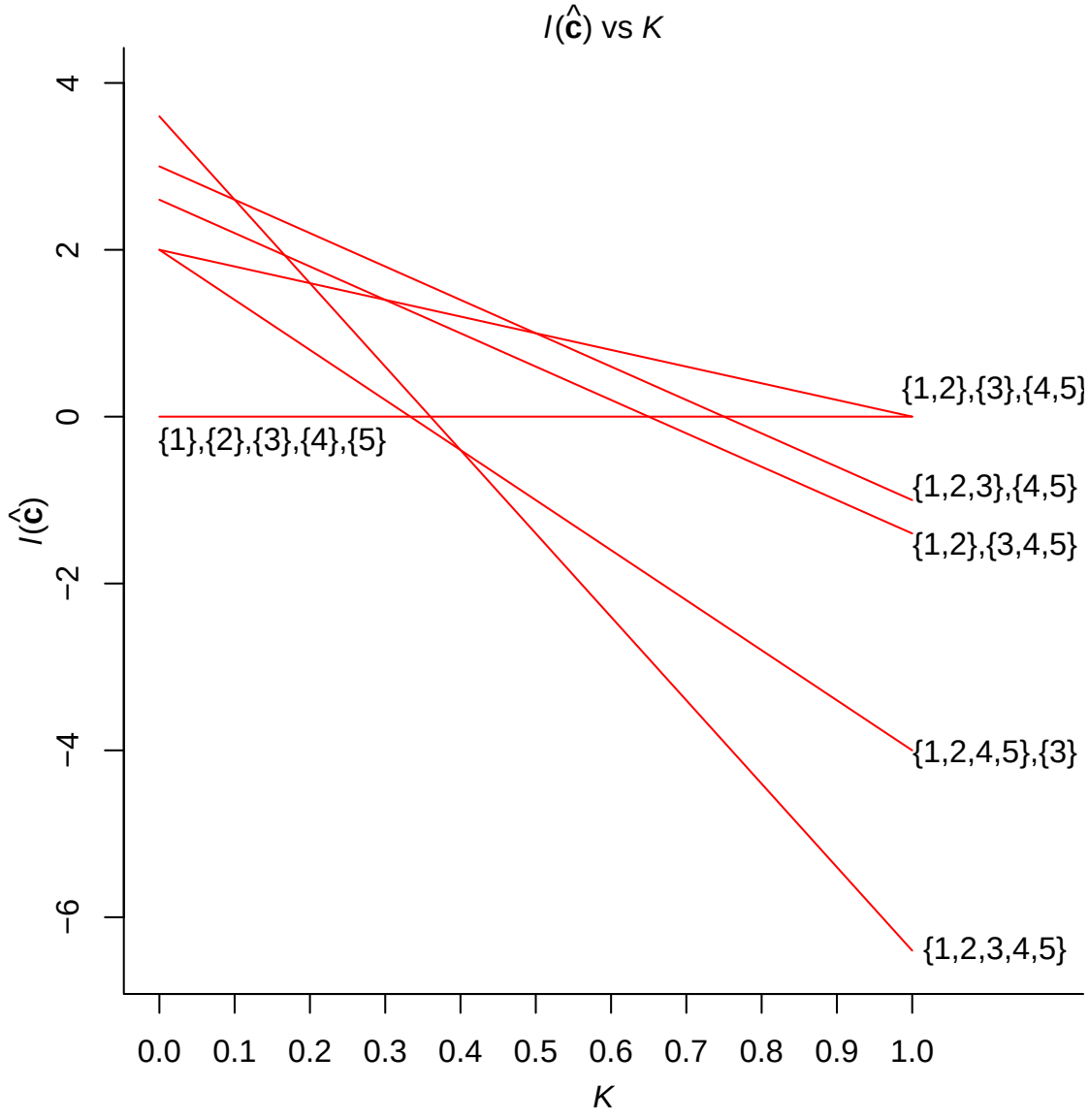
Suppose there are $n = 5$ items/elements, and that the partitions and corresponding (estimated) probabilities are

$$\begin{aligned} \boldsymbol{c}_1 &= \{\{1,2,3\},\{4,5\}\} & P(\boldsymbol{c}_1) &= 0.5 \\ \boldsymbol{c}_2 &= \{\{1,2\},\{3\},\{4,5\}\} & P(\boldsymbol{c}_2) &= 0.2 \\ \boldsymbol{c}_3 &= \{\{1,2\},\{3,4,5\}\} & P(\boldsymbol{c}_3) &= 0.3 \end{aligned}$$

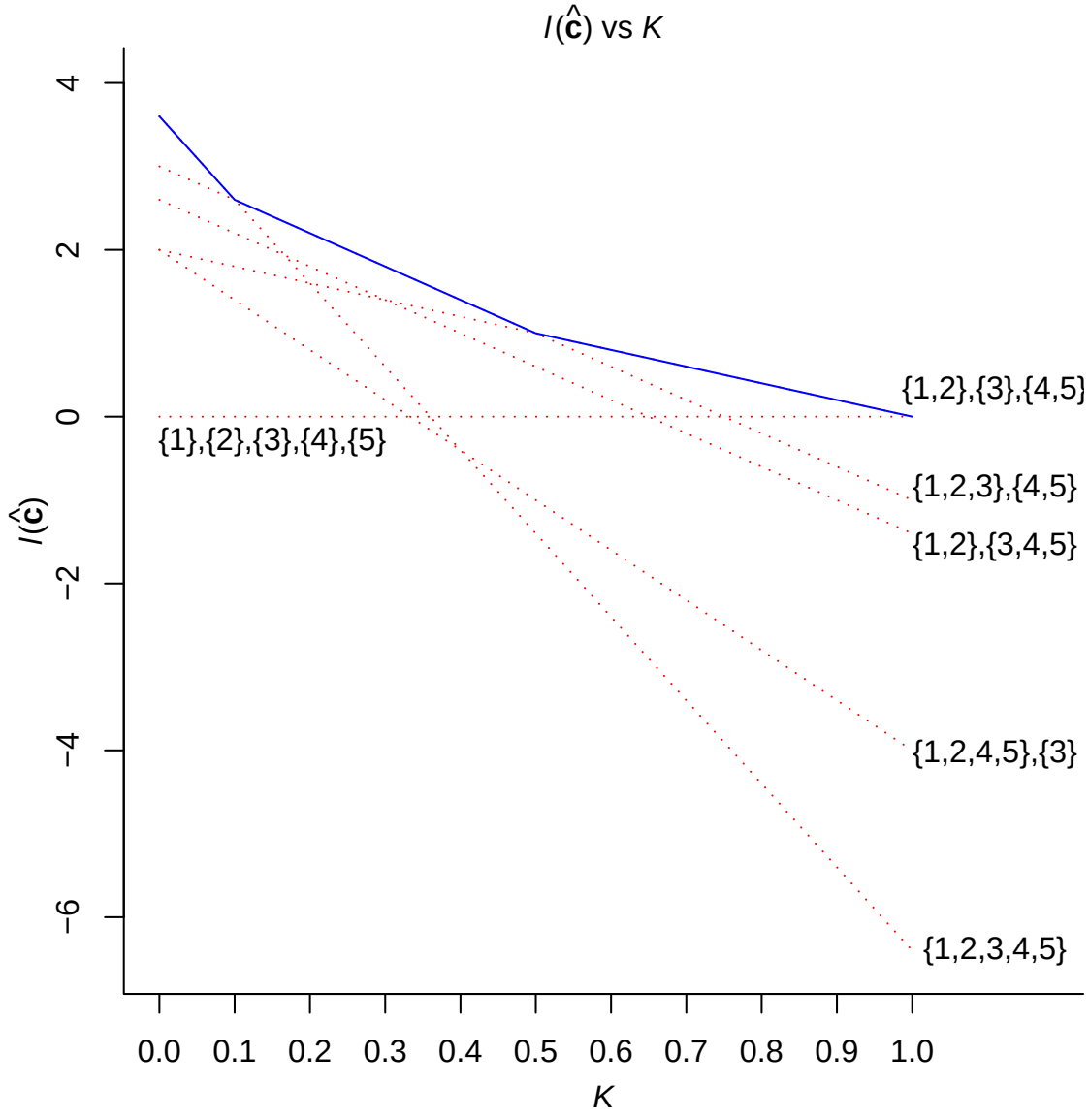
The ρ_{ij} matrix is

	1	2	3	4	5
1	–	1	0.5	0	0
2	–	–	0.5	0	0
3	–	–	–	0.3	0.3
4	–	–	–	–	1
5	–	–	–	–	–

Toy example


$l(\hat{\mathbf{c}})$ vs K
$l(\hat{\mathbf{c}})$
4
2
0
−2
−4
−6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
K
{1,2},{3},{4,5}
{1},{2},{3},{4},{5}
{1,2,3},{4,5}
{1,2},{3,4,5}
{1,2,4,5},{3}
{1,2,3,4,5}

Toy example


$l(\hat{\mathbf{c}})$ vs K
$l(\hat{\mathbf{c}})$
4
2
0
−2
−4
−6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
K
{1,2},{3},{4,5}
{1},{2},{3},{4},{5}
{1,2,3},{4,5}
{1,2},{3,4,5}
{1,2,4,5},{3}
{1,2,3,4,5}

Binary integer programming

Minimising the posterior expected loss can be cast as a standard linear integer programme in binary variables:

maximise

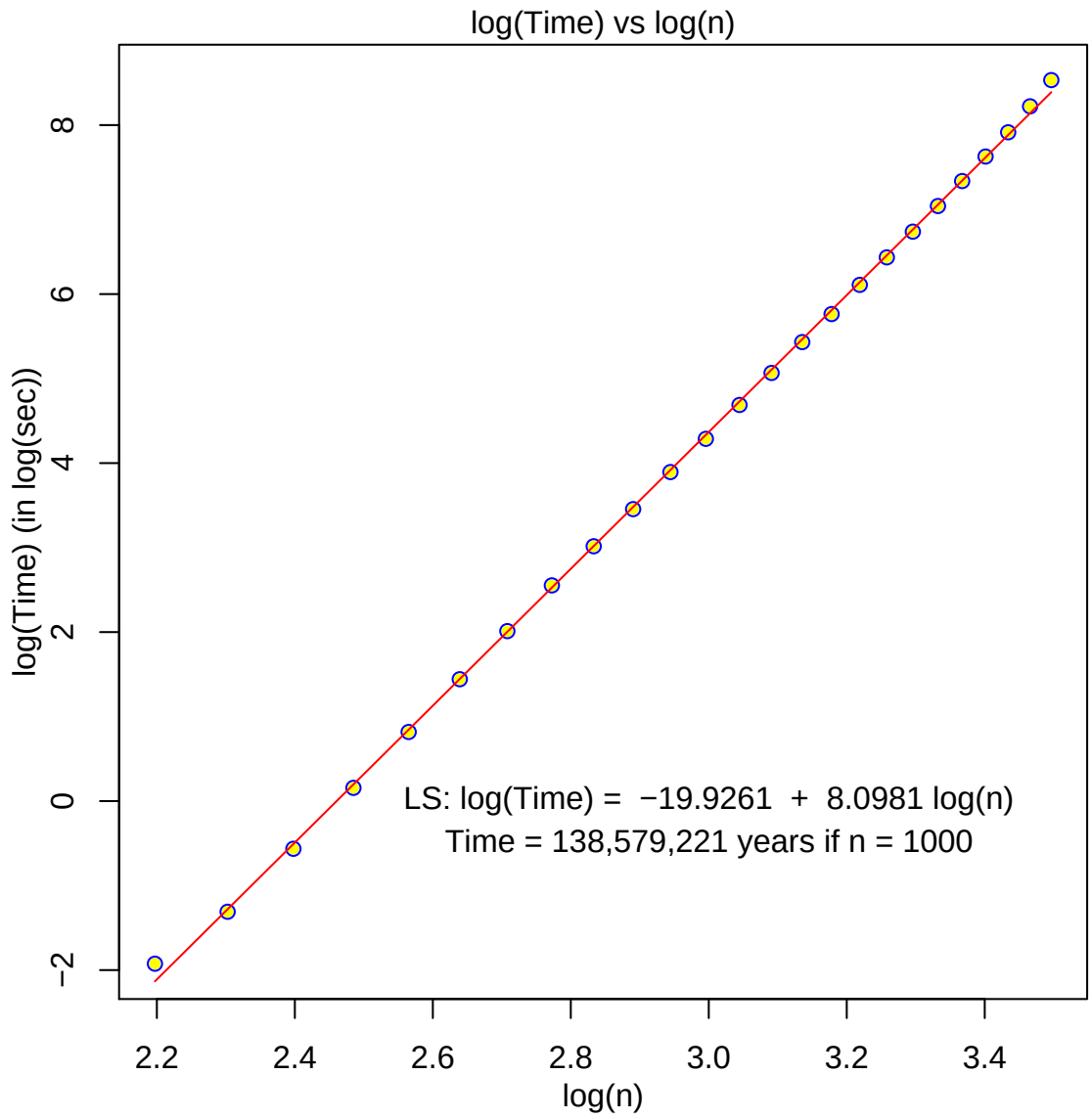
$$\ell(\widehat{\boldsymbol{c}}) = \sum (\rho_{gg'} - K) x_{gg'}$$

subject to all

$x_{gg'} \in \{0, 1\}$ and $x_{gg'} + x_{gg''} - x_{g'g''} \leq 1$ for all $\{g, g', g''\}$,

(transitivity of cluster membership), and on a small scale can easily be solved with standard (free) software.

This solution scales very badly with number of items (genes)! In fact, the problem is known to be NP hard.


log(Time) vs log(n)
log(Time) (in log(sec))
log(n)
LS: log(Time) = −19.9261 + 8.0981 log(n)
Time = 138,579,221 years if n = 1000
−2
0
2
4
6
8
2.2
2.4
2.6
2.8
3.0
3.2
3.4

A simple heuristic

We have had some success with a very simple heuristic – iteratively removing items (genes) from the partition one-by-one and reallocating them so as to maximise the objective function $\ell(\widehat{\boldsymbol{c}}) = \sum(\rho_{gg'} - K)x_{gg'}$ at each step.

Simultaneous approximate optimisation

We're interested in finding the optimal partition for all $K = b/(a+b)$. The optimum varies with K but remains constant on intervals of K. (There need be no monotonicity of the optimal partition with respect to K).

$\sum(\rho_{gg'} - K)x_{gg'}$ is a non-increasing linear function of K, and so the maximum of such functions over any candidate set $\mathcal{C}$ of partitions $\widehat{\boldsymbol{c}}$ is a non-increasing convex polygonal function of K, that is non-decreasing in $\mathcal{C}$ with respect to set inclusion.

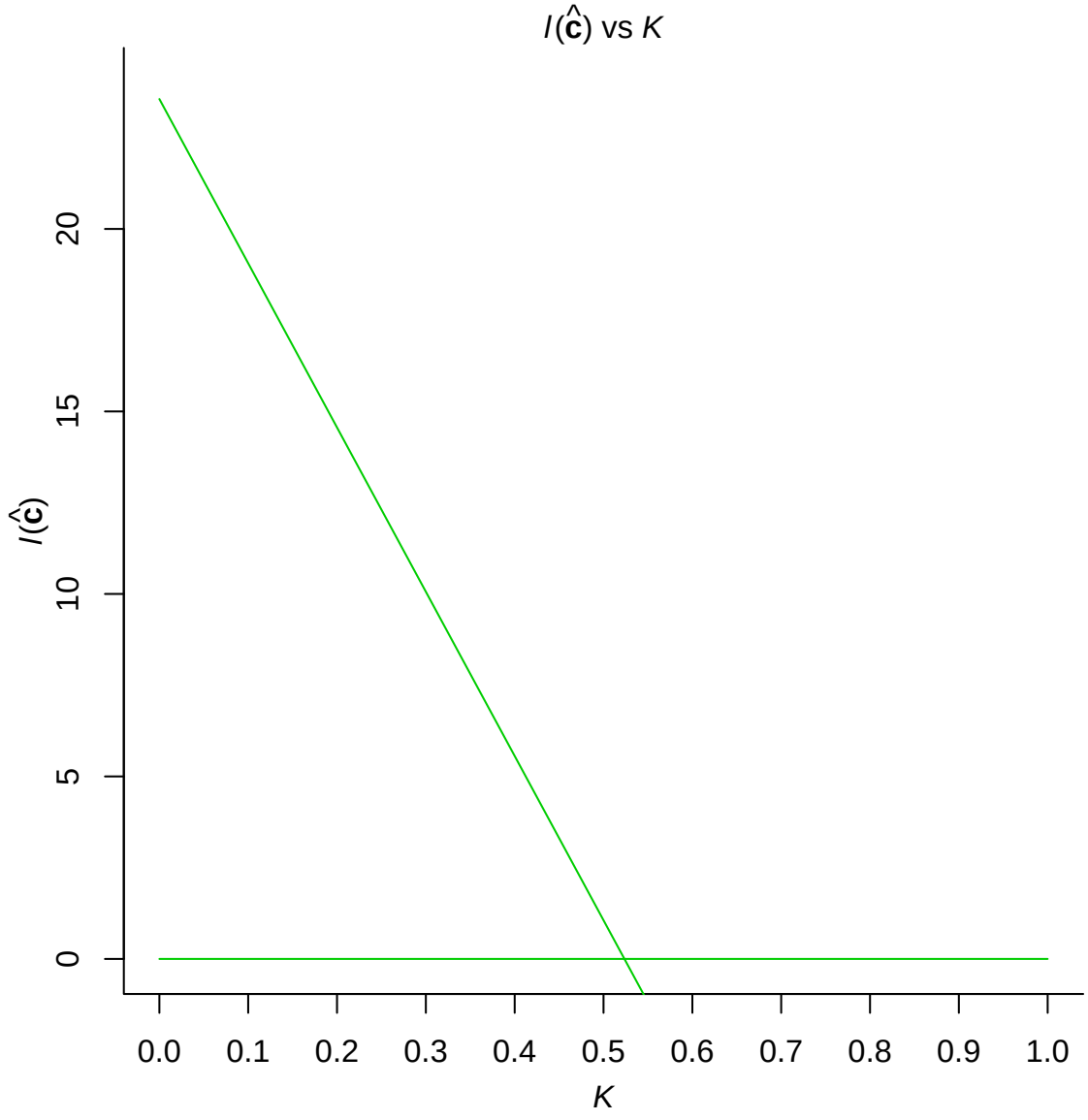
For any $\mathcal{C}$, $\sup_{\widehat{\boldsymbol{c}} \in \mathcal{C}} \sum(\rho_{gg'} - K)x_{gg'}$ is characterised by a smaller subset $\partial\mathcal{C} \in \mathcal{C}$ of 'active' partitions that define the convex hull, and our algorithm iteratively updates this active subset.

A 10-item example

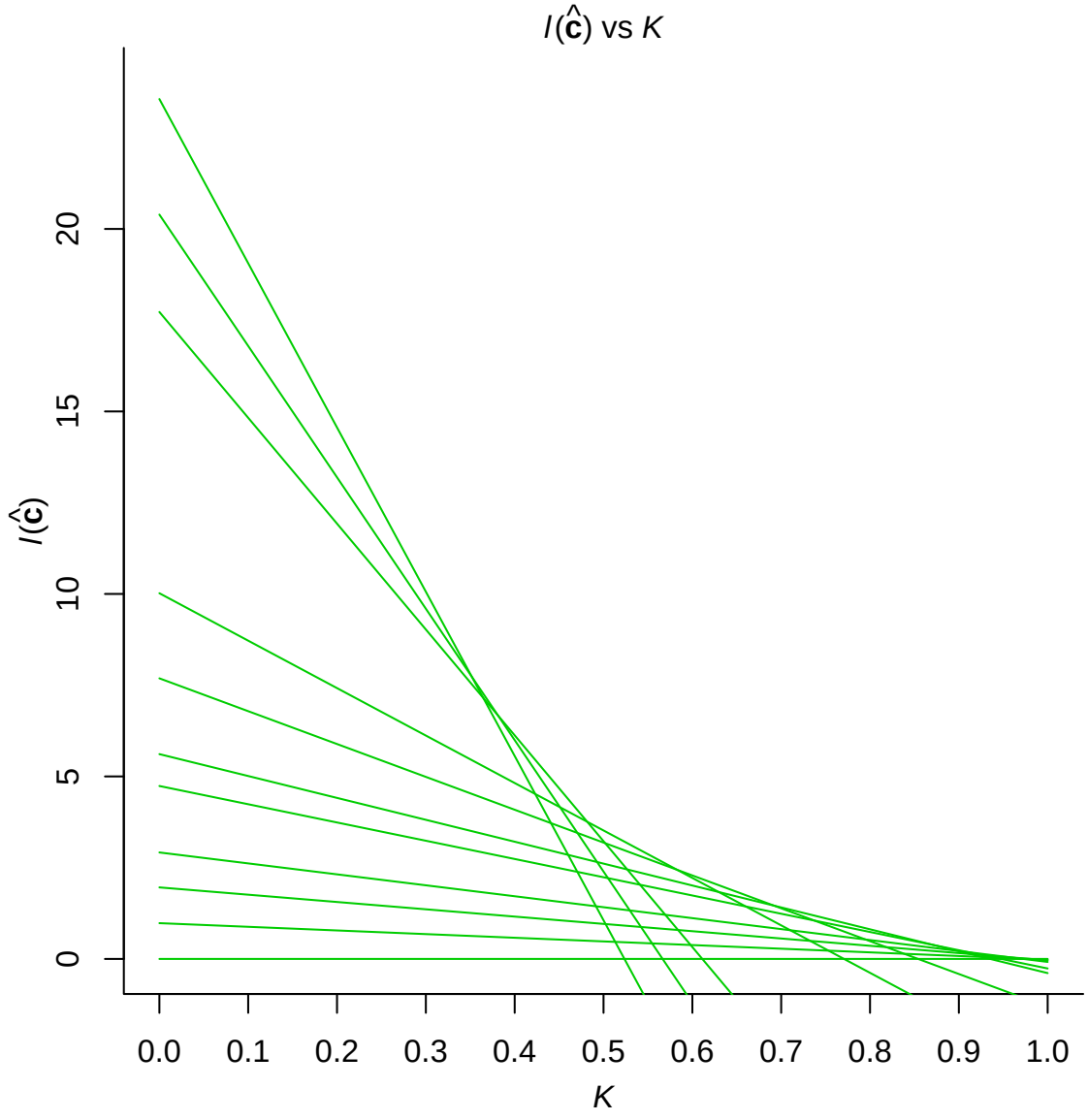
The ρ_{ij} matrix is

	1	2	3	4	5	6	7	8	9	10
1	–	0.1551	0.7426	0.2610	0.3069	0.9572	0.0990	0.6533	0.3563	0.4436
2	–	–	0.8502	0.5840	0.8307	0.8146	0.4835	0.8732	0.1441	0.3150
3	–	–	–	0.9333	0.4686	0.5256	0.0269	0.3754	0.4597	0.9801
4	–	–	–	–	0.5194	0.7063	0.9799	0.4827	0.9398	0.0252
5	–	–	–	–	–	0.8931	0.7275	0.3576	0.8293	0.3555
6	–	–	–	–	–	–	0.4834	0.0305	0.4134	0.6203
7	–	–	–	–	–	–	–	0.1808	0.8813	0.6506
8	–	–	–	–	–	–	–	–	0.0306	0.1819
9	–	–	–	–	–	–	–	–	–	0.6311

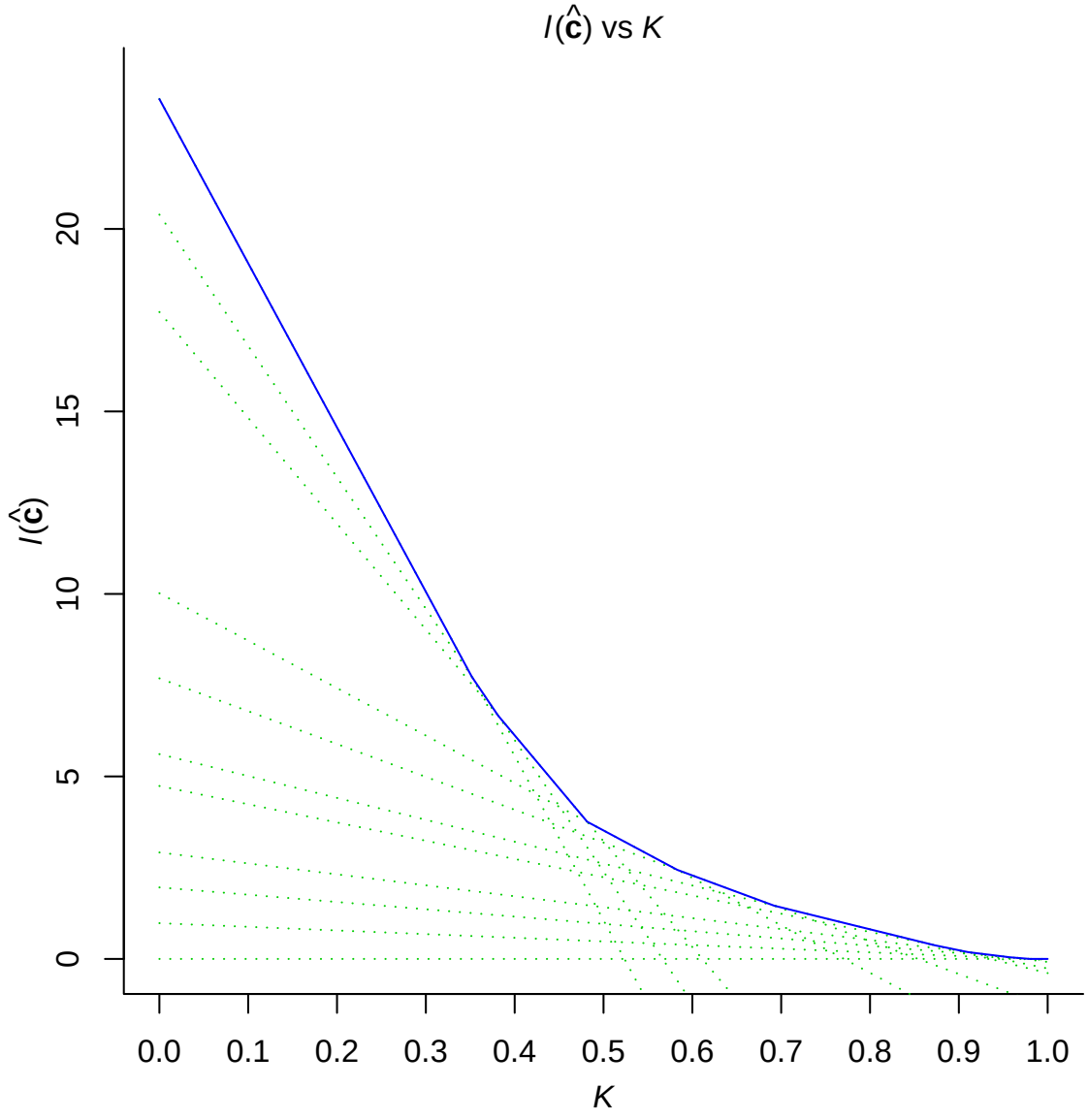
A 10-item example



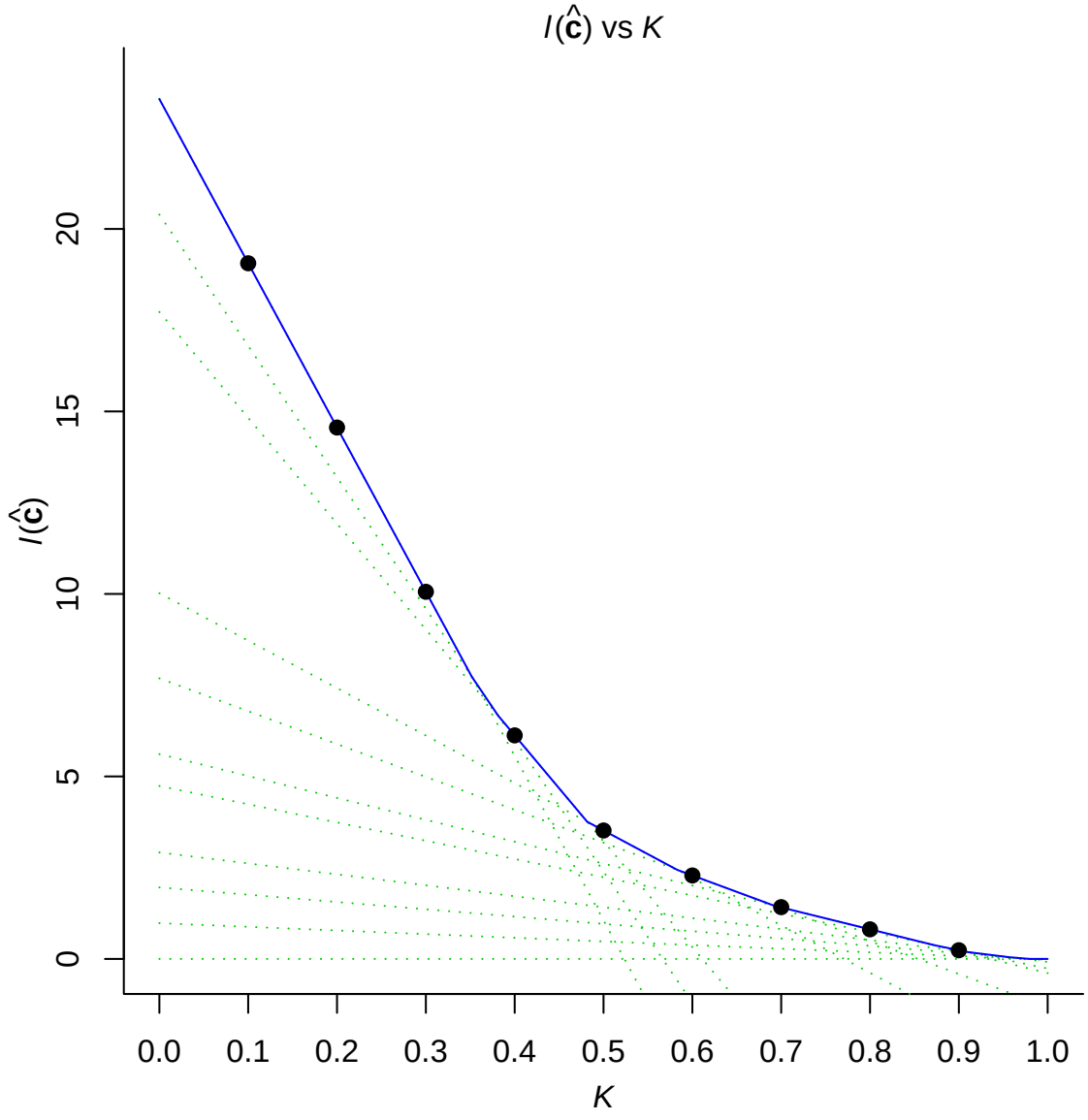
A 10-item example



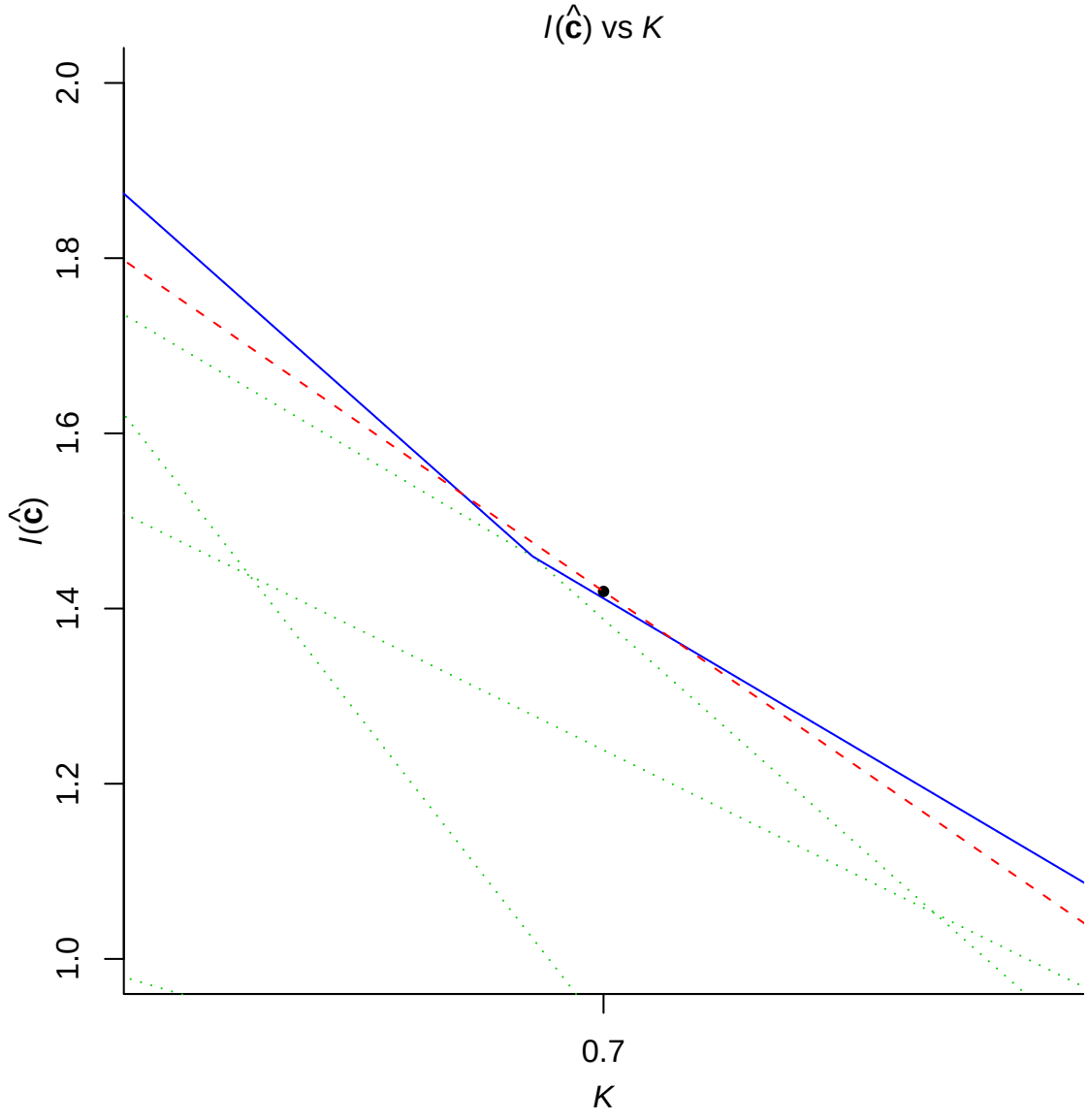
A 10-item example



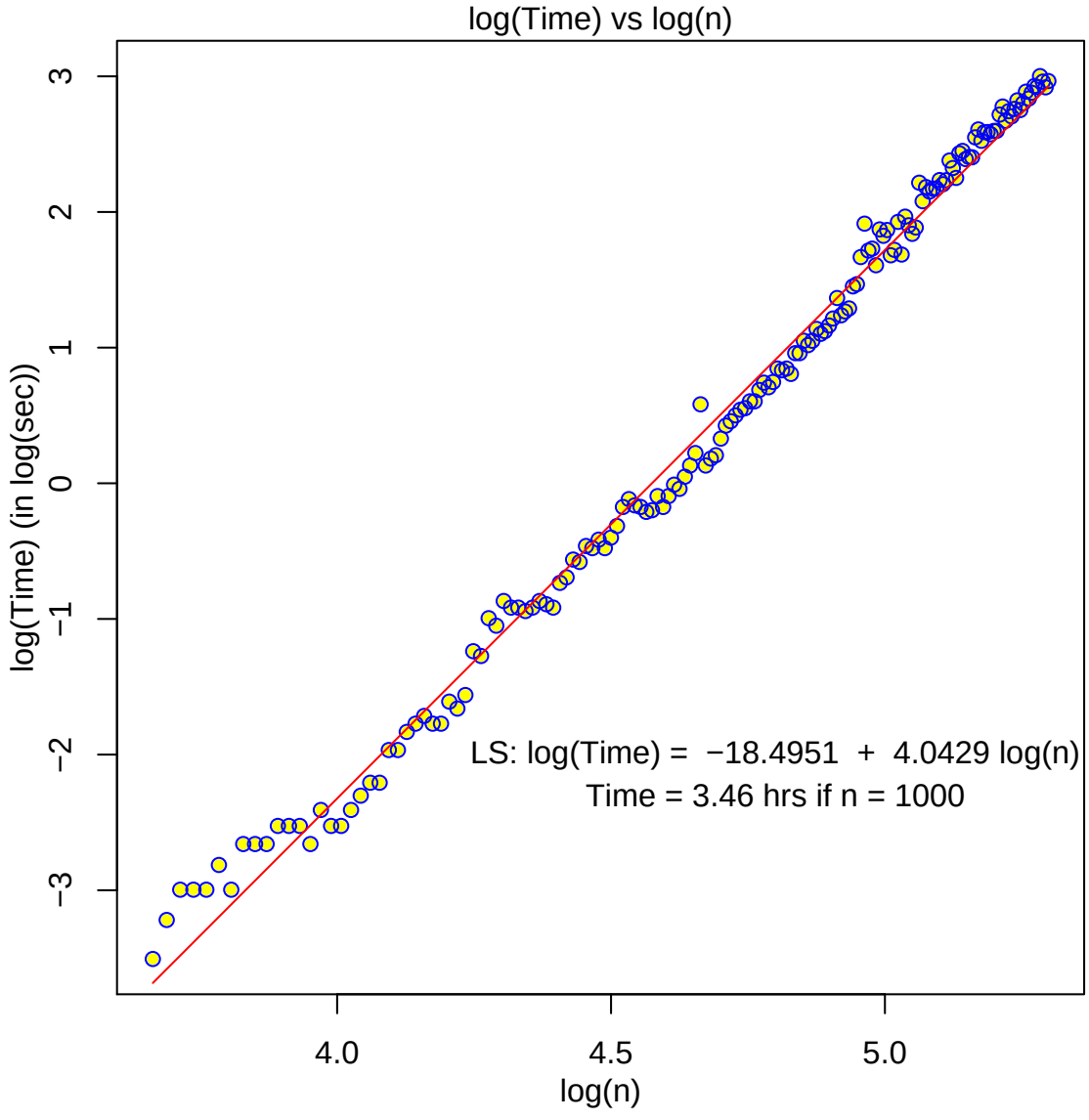
A 10-item example



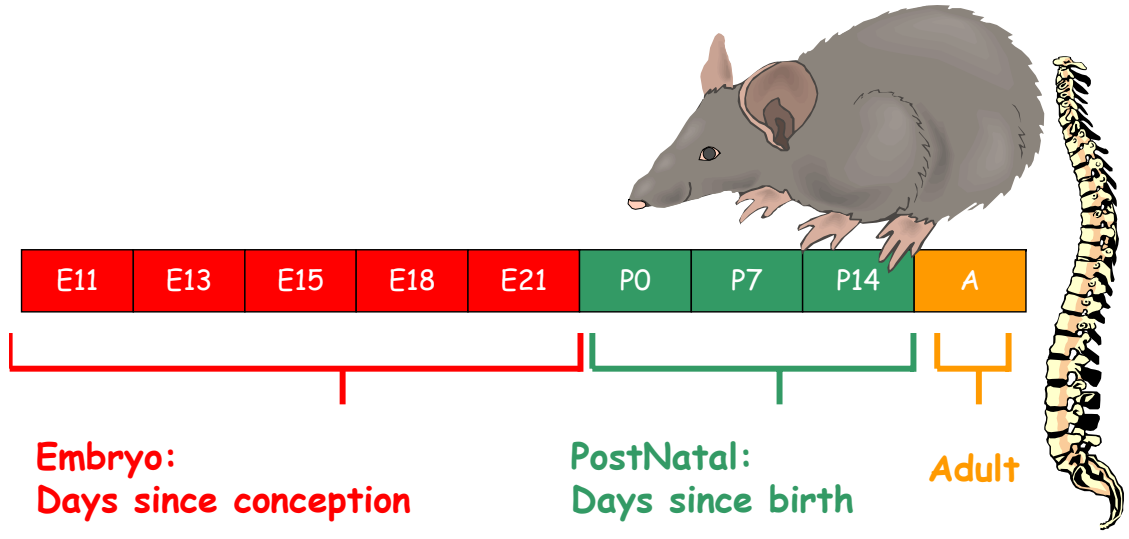
A 10-item example



Now we could afford to cluster 1000 items

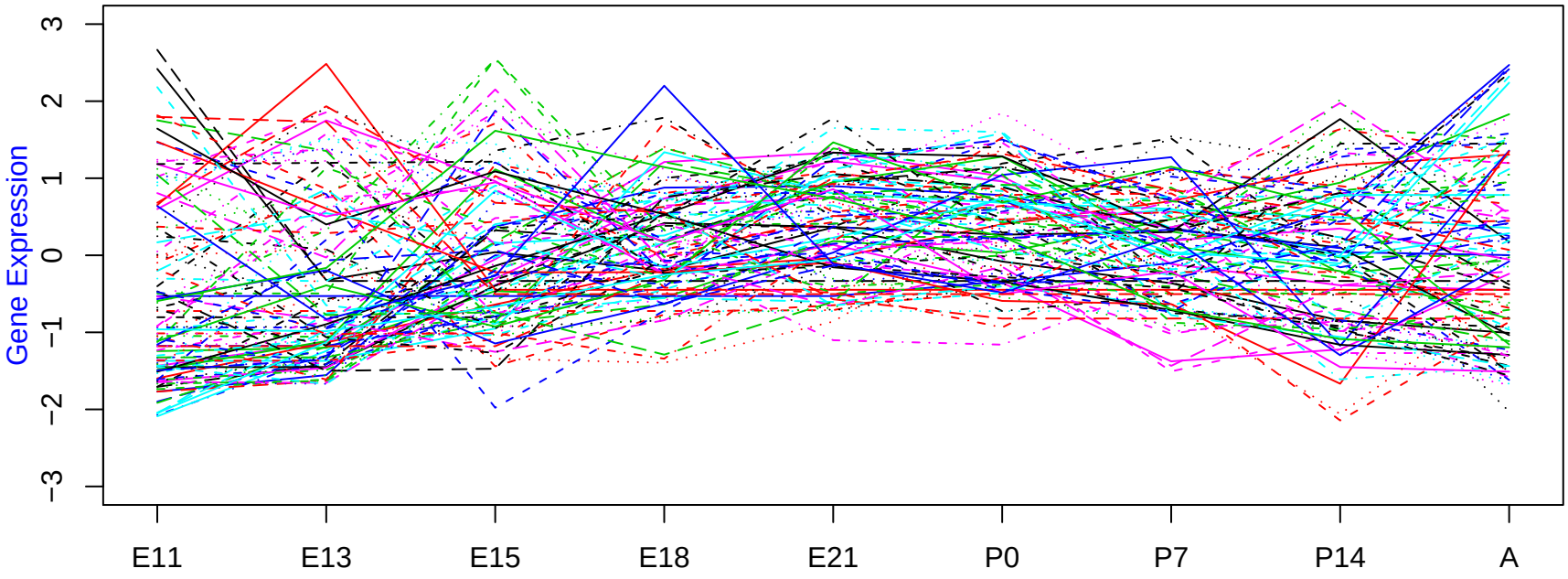


4. Rats CNS development



Wen et al (*PNAS*, 1998) studied development of central nervous system in rats: mRNA expression levels of 112 genes at 9 time points.

Rats data, normalised

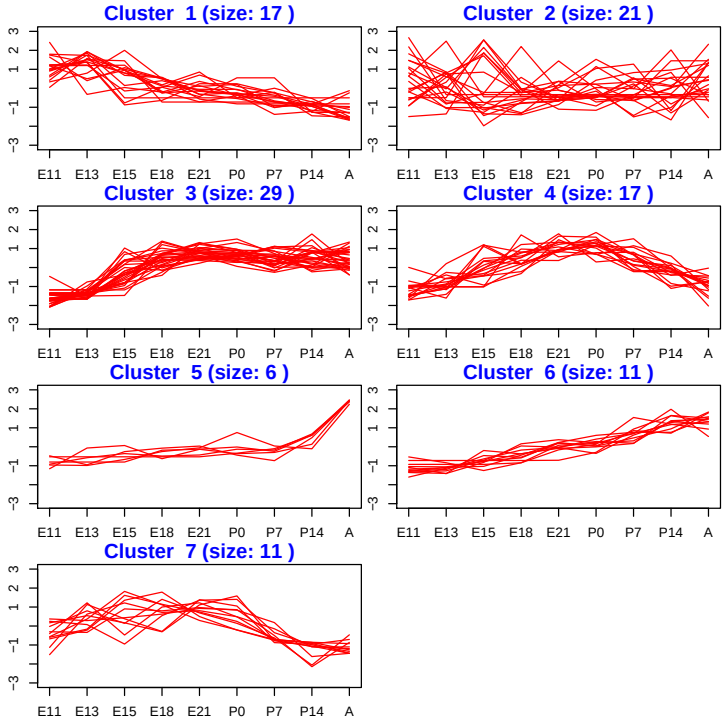


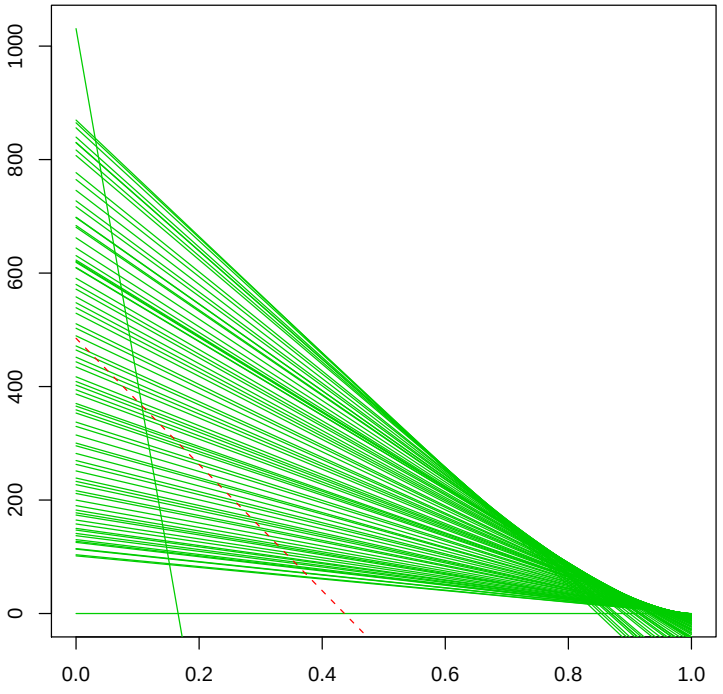
Rats: `stage+stage:day` model

Piecewise linear time dependence:

$$X = \begin{pmatrix} 1 & 11 & 0 & 0 & 0 \\ 1 & 13 & 0 & 0 & 0 \\ 1 & 15 & 0 & 0 & 0 \\ 1 & 18 & 0 & 0 & 0 \\ 1 & 21 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 7 & 0 \\ 0 & 0 & 1 & 14 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

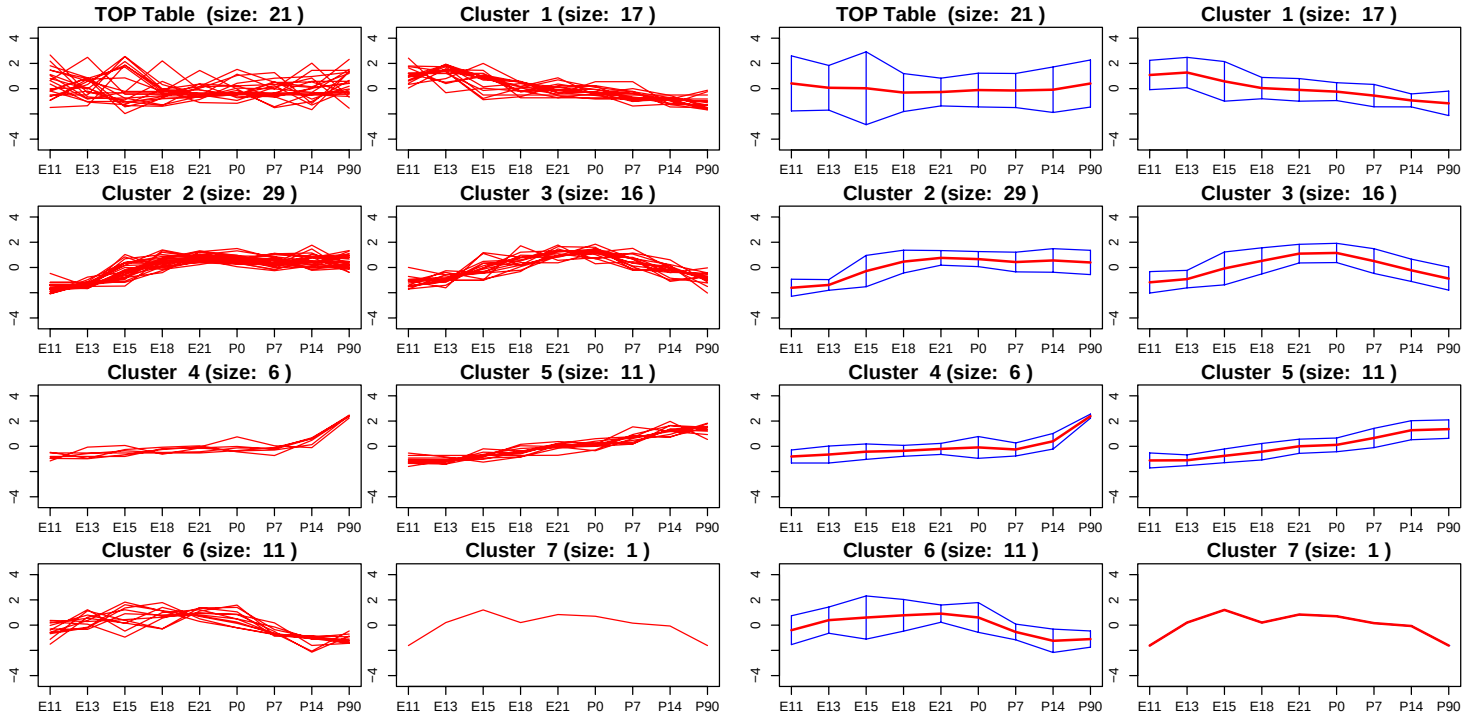
Rats: `stage+stage:day` model





Ordinary DP model: Wen's partition is substantially worse than optimal for any K.

Rats: `stage+stage:day` model



Clustered profiles using top-table CDP model.

Rats: `stage+stage:day` model

We are starting to assess inferred clusters against functional classes of genes:

	General gene class				Neurotransmitter receptors					
					Ligand class				Sequence class	
	% peptide signaling	% neurotr. receptors	% neuroglial markers	% diverse	% ACh	% GABA	% Glu	% 5HT	% ion channel	% G protein coupled
1	**41% (7)**	6% (1)	24% (4)	29% (5)	**100% (1)**				**100% (1)**	
2	3% (1)	**62% (18)**	31% (9)	3% (1)	17% (3)	**39% (7)**	33% (6)	11% (2)	**61% (11)**	39% (7)
3	6% (1)	**63% (10)**	19% (3)	13% (2)	20% (2)	20% (2)	**40% (4)**	20% (2)	**50% (5)**	**50% (5)**
4		33% (2)	17% (1)	**50% (3)**			**100% (2)**		**50% (1)**	**50% (1)**
5	18% (2)	27% (3)	**45% (5)**	9% (1)	33% (1)	**67% (2)**			**67% (2)**	33% (1)
6	27% (3)	18% (2)	18% (2)	**36% (4)**	**50% (1)**		**50% (1)**		**100% (2)**	
7		**100% (1)**			**100% (1)**				**100% (1)**	
top	**62% (13)**	10% (2)	5% (1)	24% (5)	**100% (2)**				**100% (2)**	

5. MAP vs. optimal clustering

How optimal is MAP partition?

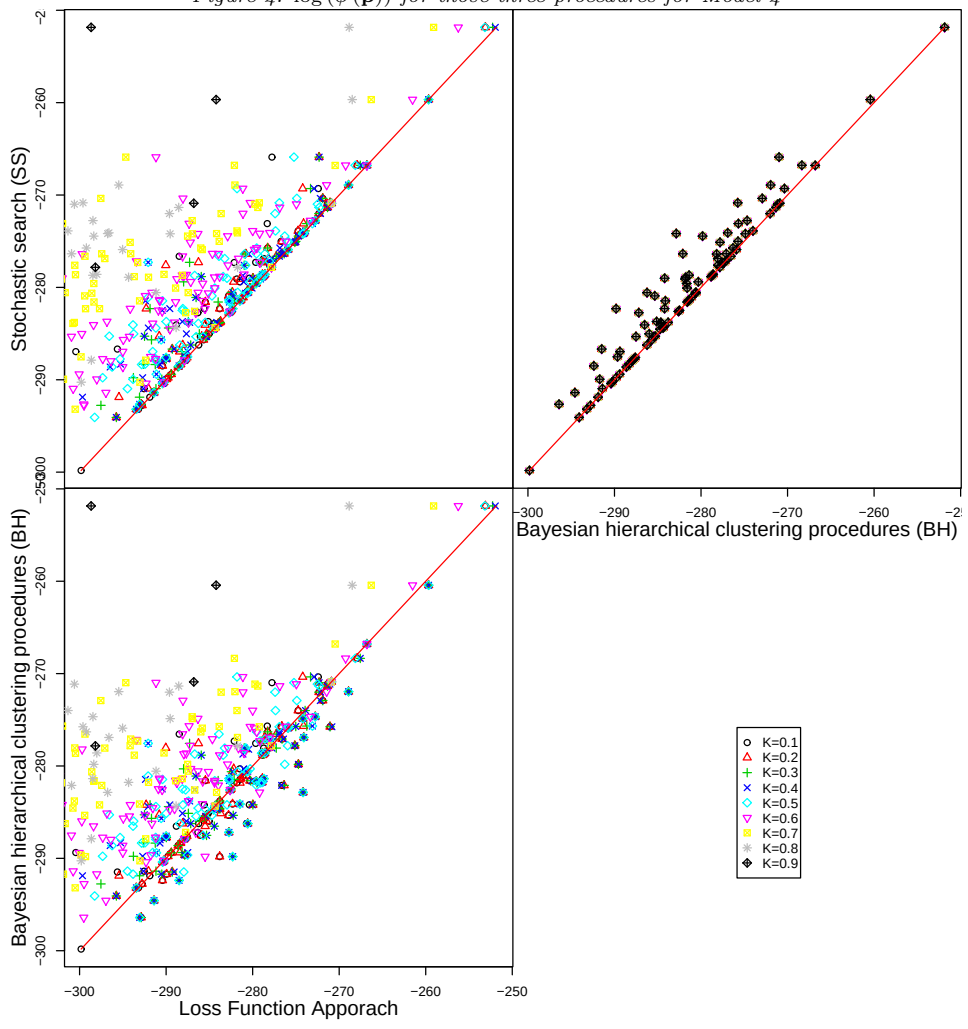
How probable is optimal partition?

Simulation (100 replicates) of samples of size $n = 100$ from 4–component bivariate normal mixture, no covariates, $S = k = 2$. We use DPM prior.

MAP approximated by a naive stochastic search (SS) and by (deterministic) Bayesian hierarchical clustering procedure of Heard, Holmes and Stephens (BH).

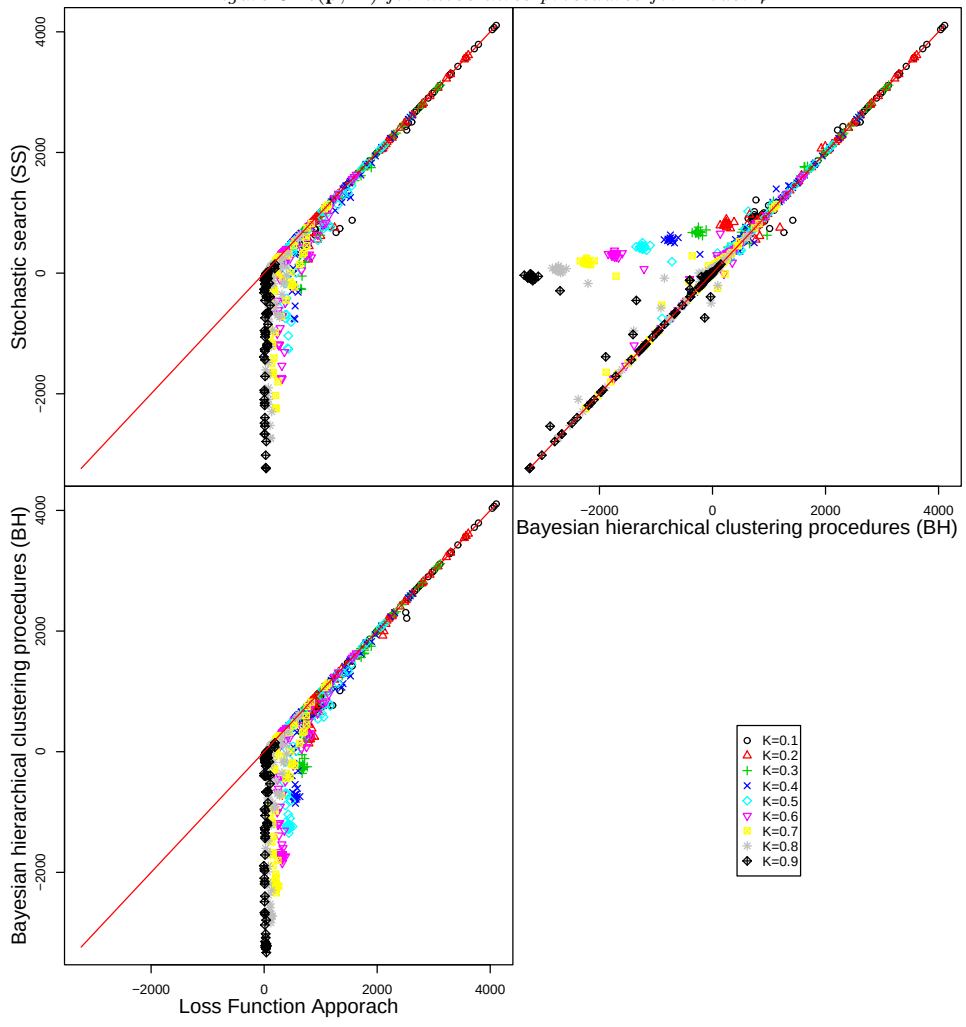
MAP vs. optimal clustering: log posterior

Figure 4: $\log(\phi(\mathbf{p}))$ *for those three procedures for Model 4*



MAP vs. optimal clustering: loss function

Figure 8: $\ell(\mathbf{p}, K)$ for those three procedures for Model 4



Summary

- (C)DP+regression is a flexible model that combines
  - parametric dependence on condition-specific covariates
  - non-parametric clustering of genes, allowing baseline category or other 'colours'
- conjugate specification greatly facilitates computation
- wider applicability of 'incremental' samplers
- possibility to approximate optimal clustering for certain loss functions