Non-parametric Bayesian Methods in Machine Learning

Dr. Simon Rogers
School of Computing Science
University of Glasgow
simon.rogers@glasgow.ac.uk
@sdrogers

May 10, 2014

Outline

- (My) Bayesian philosophy
- Gaussian Processes for Regression and Classification (Monday)
    - GP preliminaries
    - *Application 1*: typing on touch-screens
    - Classification (including semi-supervised)
    - *Application 2*: clinical (dis)-agreement
- Dirichlet Process flavoured Cluster Models (Tuesday)
    - DP preliminaries
    - *Application 3*:Idenfitying metabolites
    - *Application 4*:Cluster models for multiple data views
- Summary

Relevant publications

- The four applications are described in the following papers:
  - Uncertain Text Entry on Mobile Devices Weir et. al, CHI 2014
  - Investigating the Disagreement Between Clinicians' Ratings of Patients in ICUs Rogers et. al 2013, IEEE Trans Biomed Health Inform
  - MetAssign: Probabilistic annotation of metabolites from LC–MS data using a Bayesian clustering approach Daly et. al, Bioinformatics, under review
  - Infinite factorization of multiple non-parametric views Rogers et. al, Machine Learning 2009

About me

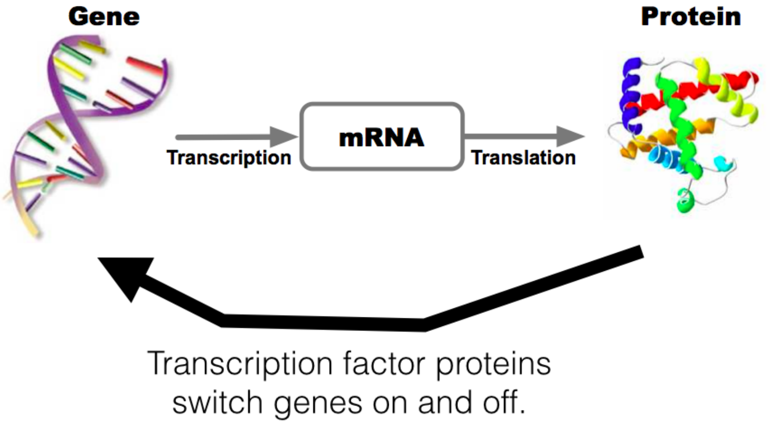
- I'm not a statistican by training (don't ask me to prove anything!).
- Education:
  - Undergraduate Degree: Electrical and Electronic Engineering (Bristol)
  - PhD: Machine Learning Techniques for Microarray Analysis (Bristol)
- Currently:
  - Lecturer: Computing Science
  - Research Interests: Machine Learning and Applied Statistics in Computational Biology and Human-Computer Interaction (HCI)

Lecture 9: Infinte factorisaion of multiple non-parametric views

Dr. Simon Rogers
School of Computing Science
University of Glasgow
simon.rogers@glasgow.ac.uk
@sdrogers

May 10, 2014

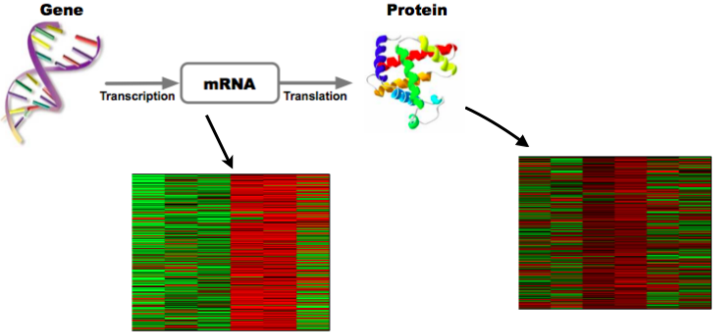
Introduction



- How related are mRNA and protein?
- **Where is the external control?**

Data

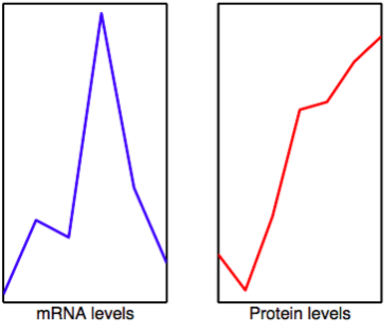
- mRNA and protein time series for ~ 500 genes

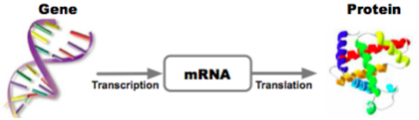


mRNA & protein for ~500 genes

(rows in matrix correspond to one another)

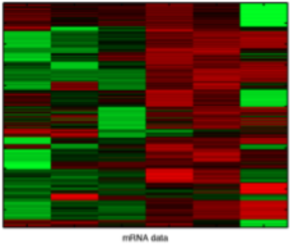
Most don't look correlated



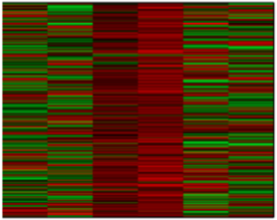


- Most don't look correlated.
  - Time delays? Saturation? Decay rates? Post-transcriptional control?

Non-parametric relationships



Cluster genes by mRNA



Cluster genes by protein

- Use clustering to define similarity.
- If genes A,B,C and D cluster together on both sides, then profiles are similar
  - They are controlled by similar processes

Generative coupled mixture model

- In Rogers et. al 2008 we developed a generative linked cluster model
- Prior membership of gene n in proteomic cluster j was dependent on assignment of mRNA to cluster k:

$$P(z_{nj} = 1, z_{nk} = 1) = P(z_{nj} = 1 | z_{nk} = 1) P(z_{nk} = 1)$$

Generative coupled mixture model

- In Rogers et. al 2008 we developed a generative linked cluster model
- Prior membership of gene n in proteomic cluster j was dependent on assignment of mRNA to cluster k:

$$P(z_{nj} = 1, z_{nk} = 1) = P(z_{nj} = 1 | z_{nk} = 1) P(z_{nk} = 1)$$

- Note that another obvious way to factorise this joint it to assume independence:

$$P(z_{nj} = 1, z_{nk} = 1) = P(z_{nj} = 1) P(z_{nk} = 1)$$

- I.e. cluster them separately

Generative coupled mixture model

- In Rogers et. al 2008 we developed a generative linked cluster model
- Prior membership of gene n in proteomic cluster j was dependent on assignment of mRNA to cluster k:

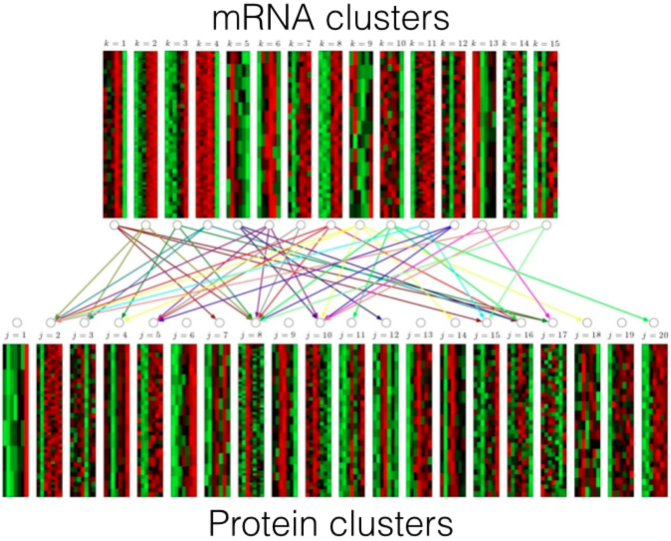
$$P(z_{nj} = 1, z_{nk} = 1) = P(z_{nj} = 1 | z_{nk} = 1) P(z_{nk} = 1)$$

- Note that another obvious way to factorise this joint it to assume independence:

$$P(z_{nj} = 1, z_{nk} = 1) = P(z_{nj} = 1) P(z_{nk} = 1)$$

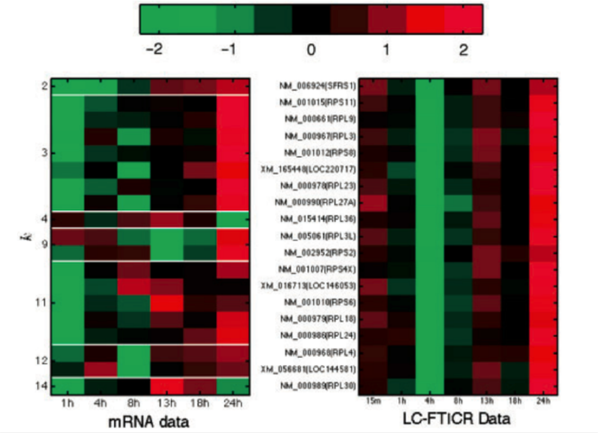
- I.e. cluster them separately
- When we do inference, we can find the links between clusters

Lots of links



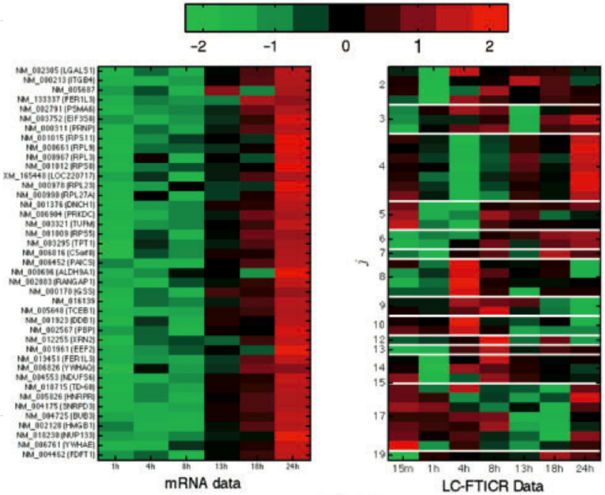
- If all control before transcription, we would expect sparse links.
- That doesn't happen!

Ribosomes



- Some strong links
- These are all ribosomal proteins
- The ribosome is where proteins are constructed
- Makes sense for them to be tightly transcriptionally controlled

Crazy genes



- In some cases, profiles were all over the place
- Here, highly conserved mRNA profiles, diverse protein profiles

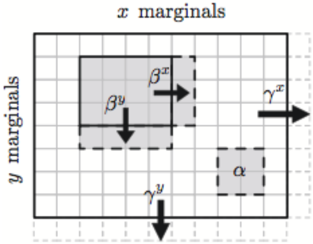
More flexible models

- At this point, we thought about more flexible models (!)
- In particular, how about decomposing the joint density as:

$$P(z_{nk} = 1, z_{nj} = 1) = \sum_i P(z_{nk} = 1|i)P(z_{nj} = 1|i)P(i)$$

- Each latent factor (i) defines a distribution over mRNA and protein clusters
- Use DP priors on i and the clusters in the two *views*

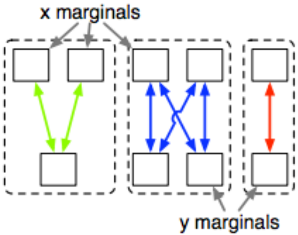
Contingency tables



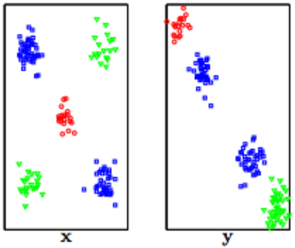
- Can visualise $P(j, k)$ as a table
- Each i is a block (if the clusters are ordered nicely)
- Numbers of rows, columns and blocks can all vary
- Restaurant analogies are possible but unhelpful

- Inference can be done with Gibbs sampling
- Details in Rogers et. al 2009

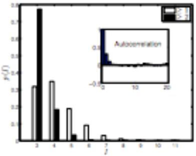
Synthetic example



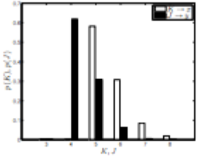
(a) Structure of Gaussian synthetic example. Top boxes represent x marginal components, bottom y. Arrows and dashed lines represent the block structure.



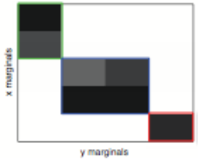
(b) Synthetic dataset for **x** (left) and **y** (right). Symbols/colors represent top-level clustering.



(c) Marginal posterior distribution over the number of top-level components, I. White bars show the histogram over all clusters, whereas black bars ignore singleton clusters. (Inset: autocorrelation)



(d) Posterior distribution over the number of marginal components. For both margins the mode corresponds with the true solution.



(e) Example contingency table from the sampler. Gray shades denote the count of samples in each cell, and the block structure corresponds exactly to that shown in subfigure (a).

Interpretation

- Interpretation is hard
- Ultimately we're interested in biological processes present in the data

Interpretation

- Interpretation is hard
- Ultimately we're interested in biological processes present in the data
- For each gene GO annotations are available
  - These tell us what the gene is known to be involved in

Interpretation

- Interpretation is hard
- Ultimately we're interested in biological processes present in the data
- For each gene GO annotations are available
  - These tell us what the gene is known to be involved in
- We can compute whether or not a GO term is *enriched* in a cluster

Interpretation

- Interpretation is hard
- Ultimately we're interested in biological processes present in the data
- For each gene GO annotations are available
  - These tell us what the gene is known to be involved in
- We can compute whether or not a GO term is *enriched* in a cluster
- Can do this at each posterior sample for all three types of cluster

Interpretation

- Interpretation is hard
- Ultimately we're interested in biological processes present in the data
- For each gene GO annotations are available
  - These tell us what the gene is known to be involved in
- We can compute whether or not a GO term is *enriched* in a cluster
- Can do this at each posterior sample for all three types of cluster
- Average the p-values across samples to obtain average values for each gene in the three cluster types.

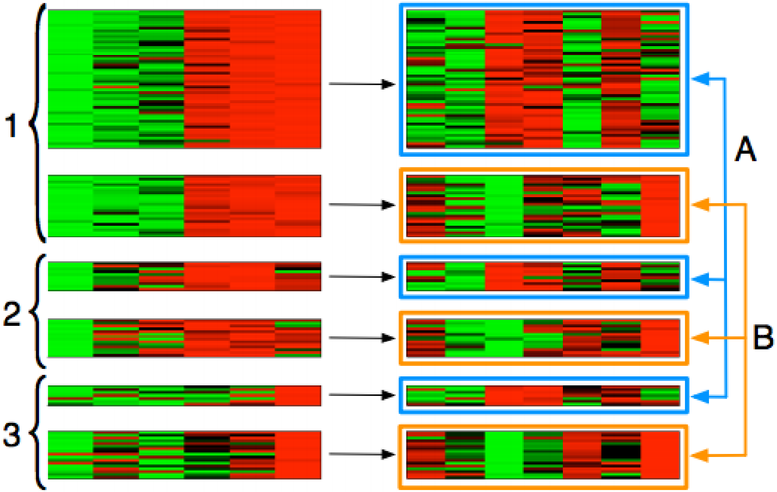
Interpretation

- Interpretation is hard
- Ultimately we're interested in biological processes present in the data
- For each gene GO annotations are available
  - These tell us what the gene is known to be involved in
- We can compute whether or not a GO term is *enriched* in a cluster
- Can do this at each posterior sample for all three types of cluster
- Average the p-values across samples to obtain average values for each gene in the three cluster types.
- Tells us if that gene is involved in that process according to mRNA, Protein, or both

Interpretation

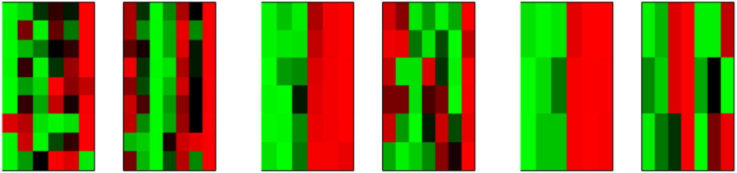
- Interpretation is hard
- Ultimately we're interested in biological processes present in the data
- For each gene GO annotations are available
  - These tell us what the gene is known to be involved in
- We can compute whether or not a GO term is *enriched* in a cluster
- Can do this at each posterior sample for all three types of cluster
- Average the p-values across samples to obtain average values for each gene in the three cluster types.
- Tells us if that gene is involved in that process according to mRNA, Protein, or both
- As in previous example, the clustering is not our final goal!

Results 1: what kind of blocks are present



- Highly inter-connected. Clusters on left shown by numbers, on right by letters.

Results 2: what's enriched?



(a) GO:0003735

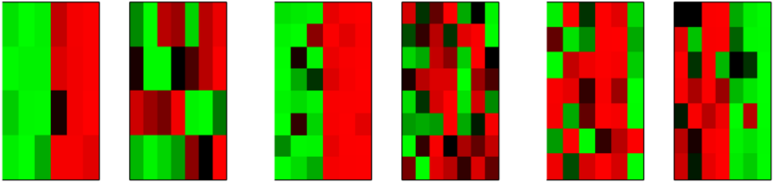
(b) GO:0008380

(c) GO:0000502

Fig. 8 3 examples of gene ontology terms significantly enriched in top level components. In all cases, left heat map is mRNA data, right is protein data.

- Some terms (and genes) enriched in the top components (i.e. for mRNA and protein)

Results 2: what's enriched?



(a) GO:0006281 (b) GO:0006457 (c) GO:0007155

- Some terms (and genes) enriched in one component and not the other (a,b: enriched in mRNC; c: enriched in protein)

Conclusions

- Flexible model can pull our interesting biology
- Actually equivalent to Rich Savage's model from Savage et. al 2010 for integrating mRNA and Transcription Factor (TF) binding data
- But, it's hard to use and interpret

Lecture 10: Summary

Dr. Simon Rogers
School of Computing Science
University of Glasgow
simon.rogers@glasgow.ac.uk
@sdrogers

May 10, 2014

Summary: GPs

- Flexible method for regression
- Auxiliary variable trick allows:
  - Binary classification
  - Multi-class classification
  - Semi-supervised classification
  - Ordinal regression
- Applications:
  - Modelling uncertainty in text entry
  - Investigating the disagreement between A&E clinicians

Summary: DPs

- Infinite prior for mixture models
- Infer the number of clusters
- No magic: The cost is that we have to define what a cluster is quite precisely
- Applications:
  - Identifying metabolites
  - Finding patterns across different datasets

Summary: all

- GPs and DPs allow computationally tractable Bayesian inference
- In all applications, inference of the curve, or clustering was only an intermediate step:
  - *Typing*: propogate uncertainty in key predictions
  - *Clinicians*: average over curve to infer characteristics of clinicians
  - *Metabolomics*: average over clustering to annotate peaks
  - *Multiview*: average over clusterings to extract biological information
- Thinking about what the results of regression / clustering will be used for is useful.
- GP and DP provide access to the posterior and therefore the ability to average over it