Two-level infinite mixture for multi-domain data

Simon Rogers\textsuperscript{1}, Janne Sinkkonen\textsuperscript{2}, Arto Klami\textsuperscript{2}, Samuel Kaski\textsuperscript{2} and Mark Girolami\textsuperscript{1}

December 13th 2008

\textsuperscript{1} University of Glasgow, UK

\textsuperscript{2} TKK, Finland
Outline

Introduction

The model

Inference

Experimental results
  Synthetic data
  HMEC data

Conclusions
Reasonable amount is known about transcription (lots of mRNA data)

Less is known about translation (not much protein data)

mRNA and protein levels provide 2 views of the same gene

To understand more about translation, the two views must be analysed together
Reasonable amount is known about transcription (lots of mRNA data)
- Reasonable amount is known about transcription (lots of mRNA data)
- Less is known about translation (not much protein data)
Reasonable amount is known about transcription (lots of mRNA data)
Less is known about translation (not much protein data)
mRNA and protein levels provide 2 views of the same gene
To understand more about translation, the two views must be analysed together
HMEC Dataset

- Dataset described in [Waters et al., 2008] and previously analysed in [Rogers et al., 2009].
- mRNA and protein time-series for \( \sim 500 \) genes
- Measurements taken from 0 to 24hr after stimulation with growth factor
- Example - TLN1
How should we analyse this data?

- Both data sources have strong cluster structure
- But...clusterings are different!
How should we analyse this data?

- Both data sources have strong cluster structure
- But...clusterings are different!

Previous analysis:

Required manual tuning (number of components) and mining for interesting biology.
More previous analysis
Objectives

- Biologists like clusters!
- We would like to model the relationship between clusters in the two views.
- Genes that cluster differently are as interesting as genes that cluster the same.
Outline

Introduction

The model

Inference

Experimental results
  Synthetic data
  HMEC data

Conclusions
A mixture over contingency tables

- Assume a $K$-component mixture for mRNA data
- $J$-component mixture for protein data
- Each gene is assigned to a $(k, j)$ pair
- Prior on $(k, j)$
  - Previously $→ p(k, j) = p(k)p(j|k)$

Top-level $(i)$ components link together groups of $j$s and $k$s and will hopefully have biological interpretation.
A mixture over contingency tables

- Assume a $K$-component mixture for mRNA data
- $J$-component mixture for protein data
- Each gene is assigned to a $(k, j)$ pair
- Prior on $(k, j)$
  - Previously → $p(k, j) = p(k)p(j|k)$
  - Now → $p(k, j) = \sum_i p(i)p(j|i)p(k|i)$

Top-level ($i$) components link together groups of $j$s and $k$s and will hopefully have biological interpretation.
A mixture over contingency tables

- Assume a $K$-component mixture for mRNA data
- $J$-component mixture for protein data
- Each gene is assigned to a $(k, j)$ pair
- Prior on $(k, j)$
  - Previously $\rightarrow p(k, j) = p(k)p(j|k)$
  - Now $\rightarrow p(k, j) = \sum_i p(i)p(j|i)p(k|i)$

Top-level $(i)$ components link together groups of $j$s and $k$s and will hopefully have biological interpretation.
Contingency table representation

Potentially infinite number of components in all three mixtures.
The infinite Chinese restaurant franchise

- Restaurants → top-level components.
- Tables → restaurant specific marginal instances.
- Dishes → marginal mRNA and protein clusters.
Outline

Introduction

The model

Inference

Experimental results
  Synthetic data
  HMEC data

Conclusions
Inference

- **Gibbs sampling**
  - Processes and margin cluster parameters integrated out.
- Similar to franchise scheme suggested for HDPs
- Top-level (restaurant) assignment is obtained by marginalizing over the potential table assignments within that restaurant.
- Concentration parameters (3-levels) sampled using Metropolis-Hastings scheme.
Collapsed Gibbs sampler

- Remove gene (customer) \( n \) from the current assignments
- Stage 1: Genes (People) → Top-level (Restaurant) assignments:

\[
p(z_n = i) \propto \begin{cases} 
C^{-n}_i p(x_n|z_n = i, \Delta^x) p(y_n|z_n = i, \Delta^y) & \text{for an existing } i, \\
\alpha p(x_n|t^*, \Delta^x) p(y_n|u^*, \Delta^y) & \text{for a new } i.
\end{cases}
\]

- Restaurant likelihood, \( p(x_n|z_n = i, \Delta^x) \), computed by marginalising tables in \( i \) (plus possible new table)
- New restaurant likelihood, \( p(x_n|t^*, \Delta^x) \), computed by marginalising over dishes in the model (plus possible new dish)
Collapsed Gibbs sampler

- **Stage 2: Genes (People) → Table assignments:**

\[
p(v_n = t) \propto \begin{cases} 
  c_{it}^{-n} p(x_n|X_{it}^{-n}, \Delta) & \text{for a table } t \text{ in the restaurant } i, \\
  \beta p(x_n|t^*, \Delta) & \text{for a new table,}
\end{cases}
\]

- **Stage 3: Table → Marginal component (Dish) assignments:**

\[
p(w_{it^*} = j) \propto \begin{cases} 
  d_j^{-n} p(x_n|X_j^{-n}, \Delta) & \text{for an existing component (dish) } j, \\
  \gamma p(x_n|\Delta) & \text{for a new component (dish).}
\end{cases}
\]

Only x marginal shown.
Hyper-parameter sampling

- $\alpha$ - Top-level components (restaurants)
- $\gamma^x, \gamma^y$ - Marginal components (dishes)
- $\beta^x, \beta^y$ - Tables (could be restaurant specific)
- Generally, for $I$ components:

$$p(z|\alpha) \propto \frac{\alpha^I \Gamma(\alpha)}{\Gamma(N + \alpha)}$$

- Sampling scheme depends on prior
  - Gamma $\rightarrow$ Gibbs sampling (West 1992)
  - Inverse-Gamma $\rightarrow$ Adaptive importance sampling
- In general, Metropolis-Hastings proved adequate.
Outline

Introduction

The model

Inference

Experimental results
  Synthetic data
  HMEC data

Conclusions
Synthetic data

- 3 top-level components
- Each corresponds to two $x$ components and one $y$ component in the marginals
Synthetic data

- Posterior mode at correct number of components.
- High weight at higher values of $I$ partly due to unstable components.
Synthetic data

Typical posterior sample

Top: contingency tables. Bottom: assigned data
HMEC Dataset

- Dataset described in [Waters et al., 2008] and previously analysed in [Rogers et al., 2009].
- mRNA and protein time-series for ~ 500 genes
- Measurements taken from 0 to 24hr after stimulation with growth factor
- Example - TLN1
HMEC - Global results

Small number of marginal components, large number of top-level components. \( I \gg J, K \).

Lots of restaurants, very few dishes!
Small modules (3 of many)

Ribosomes and Proteosome are large complexes - not surprising to see a high degree of co-regulation at both transcriptional and translational level.
Conclusions

▶ Non-parametric model capable of finding shared and unshared variability.
▶ Non-linear (cluster) relationships.
▶ Interpretable results.
▶ Verified on recent biological data consisting of mRNA and protein profiles
▶ Broad agreement between this analysis and [Rogers et al., 2009]...
  ▶ ...requires less tuning and manual mining
  ▶ Translational picture is highly complex - many small modules
Acknowledgements

- EPSRC
- H. Steven Wiley et al. at PNL for use of data
- Walter Kolch for biological input
- PASCAL short visit funding

Data is available! Download at

http://www.dcs.gla.ac.uk/inference/genomic_integration.html
References
