Multi-level functional principal components analysis models for replicated genomics time series data sets

Maurice Berk

Imperial College London

September 11th, 2012

Talk outline

A motivating genomics experiment

- Want to understand the immune response to infection with BCG

A motivating genomics experiment

- Want to understand the immune response to infection with BCG
- Which genes are switched on (or off) by BCG infection?

A motivating genomics experiment

- Want to understand the immune response to infection with BCG
- Which genes are switched on (or off) by BCG infection?
- Collect blood samples from healthy human volunteers (replicates)

A motivating genomics experiment

- Want to understand the immune response to infection with BCG
- Which genes are switched on (or off) by BCG infection?
- Collect blood samples from healthy human volunteers (replicates)
- Infect the blood samples with BCG

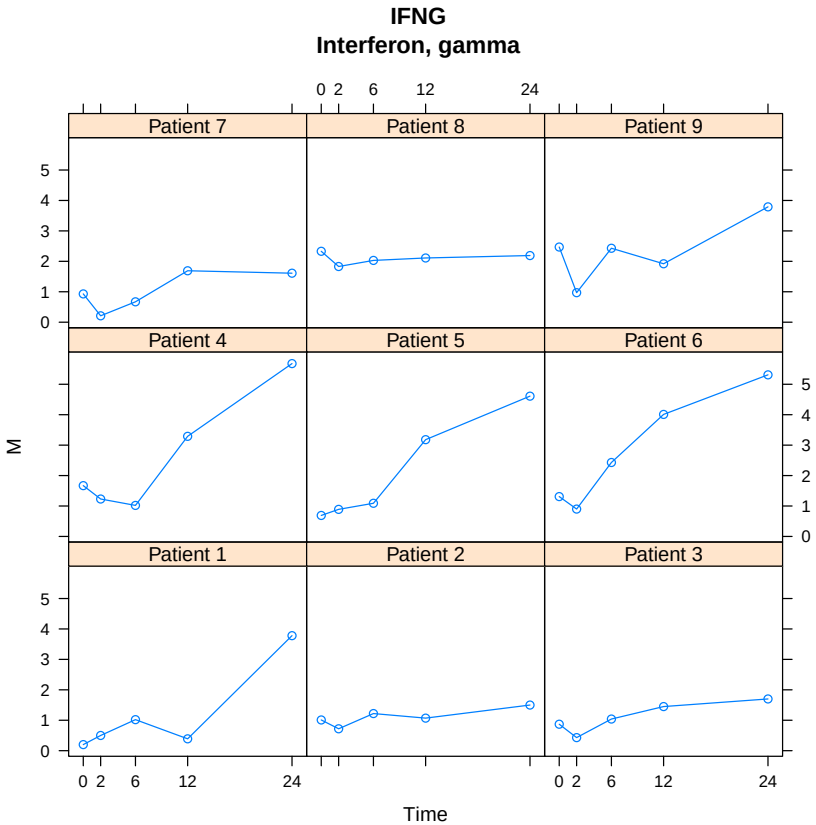
A motivating genomics experiment

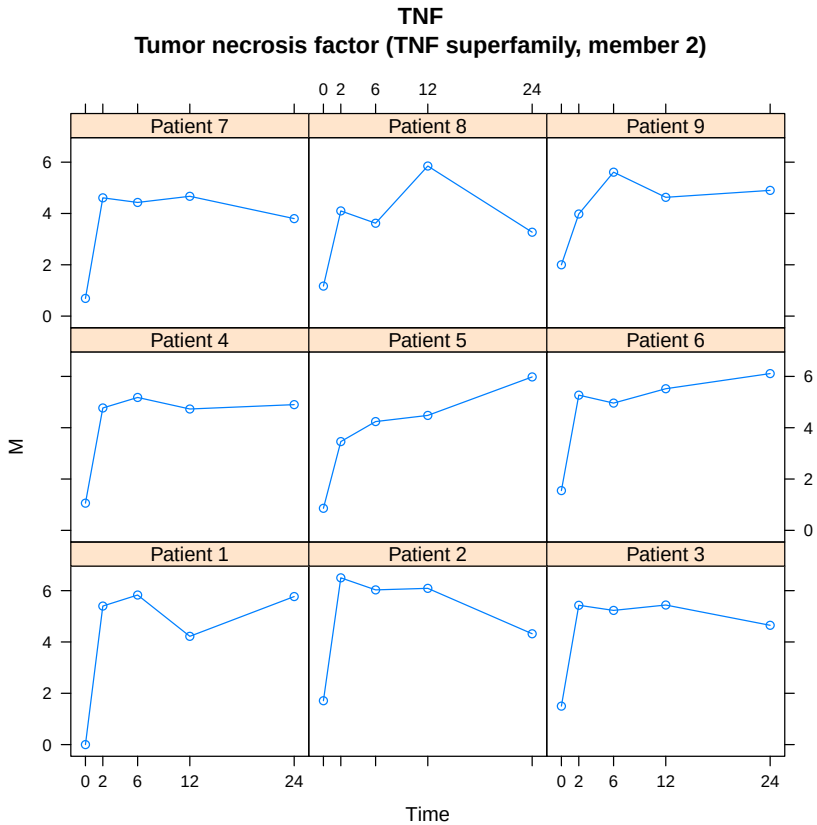
- Want to understand the immune response to infection with BCG
- Which genes are switched on (or off) by BCG infection?
- Collect blood samples from healthy human volunteers (replicates)
- Infect the blood samples with BCG
- Perform microarray hybridizations at several time points after infection

A motivating genomics experiment

- Want to understand the immune response to infection with BCG
- Which genes are switched on (or off) by BCG infection?
- Collect blood samples from healthy human volunteers (replicates)
- Infect the blood samples with BCG
- Perform microarray hybridizations at several time points after infection
- For every gene, end up with one time series per volunteer

Raw Data


IFNG
Interferon, gamma
Patient 7
Patient 8
Patient 9
Patient 4
Patient 5
Patient 6
Patient 1
Patient 2
Patient 3
M
Time
0 2 6 12 24


TNF
Tumor necrosis factor (TNF superfamily, member 2)
Patient 7
Patient 8
Patient 9
Patient 4
Patient 5
Patient 6
Patient 1
Patient 2
Patient 3
M
Time
0 2 6 12 24

Current State of Play

- FDA is a popular modelling choice for genomics time series data sets (see Coffey and Hinde, 2011)

Current State of Play

- FDA is a popular modelling choice for genomics time series data sets (see Coffey and Hinde, 2011)
- Replicated data sets have received much less attention — but see Storey et al (2005), Liu & Yang (2009) and Berk et al (2011)

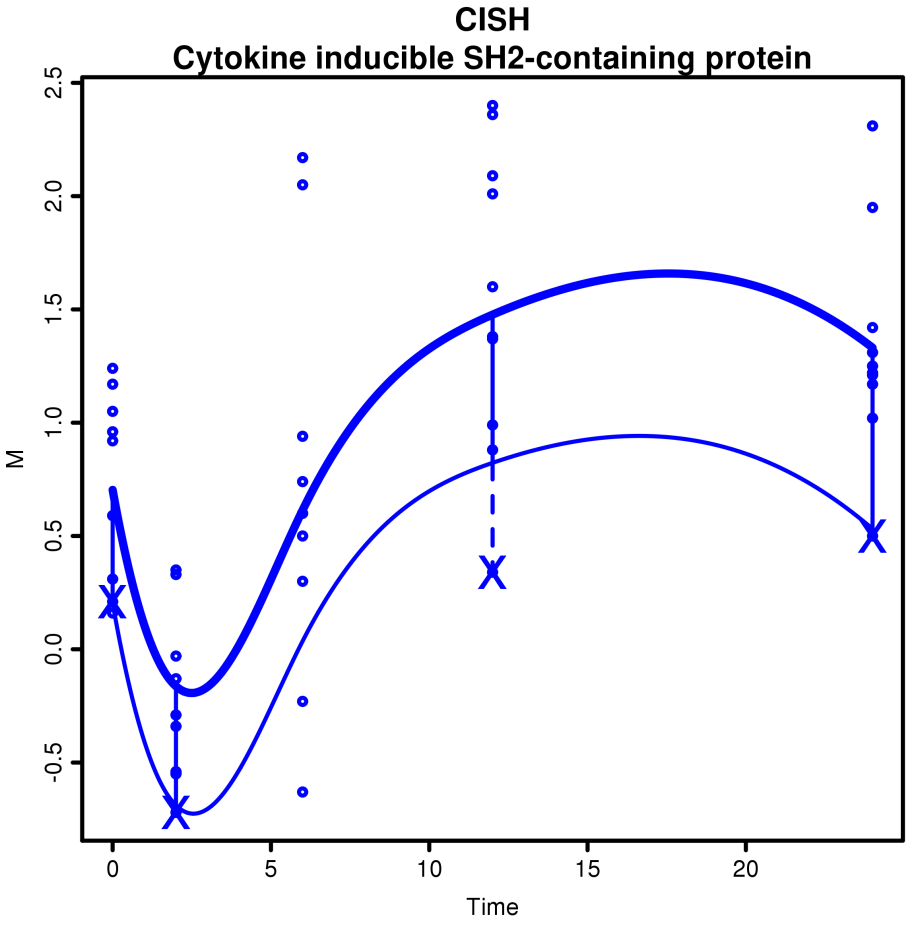
Current State of Play

- FDA is a popular modelling choice for genomics time series data sets (see Coffey and Hinde, 2011)
- Replicated data sets have received much less attention — but see Storey et al (2005), Liu & Yang (2009) and Berk et al (2011)
- But these methods model each gene independently!

Current State of Play

- FDA is a popular modelling choice for genomics time series data sets (see Coffey and Hinde, 2011)
- Replicated data sets have received much less attention — but see Storey et al (2005), Liu & Yang (2009) and Berk et al (2011)
- But these methods model each gene independently!
- Multi-level approaches exist in other domains — Di et al (2009) and Zhou et al (2010)

The functional mixed-effects model



- $y_i(t) = \mu(t) + v_i(t) + \epsilon(t)$

The functional mixed-effects model

We can represent functions $\mu(t)$ and $v_i(t)$ using splines:

LMM Representation

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\mu} + \mathbf{X}_i\mathbf{v}_i + \boldsymbol{\epsilon}_i$$

and have a linear mixed-effects model

- Impose distributions on $\mathbf{v}_i$ and $\boldsymbol{\epsilon}_i$
- Treat random-effects $\mathbf{v}_i$ as missing data in the EM algorithm

A multi-level functional mixed-effects model

Definition

$$y_{ij}(t) = \mu(t) + f_i(t) + g_{ij}(t) + \epsilon_{ij}(t)$$

- $\mu(t)$ is the grand mean across all genes
- $f_i(t)$ is the gene specific deviation from that grand mean
- $g_{ij}(t)$ is the replicate specific deviation from the gene mean

Could represent the functions $\mu(t)$, $f_i(t)$ and $g_{ij}(t)$ using splines as before ...

Functional Principal Components Analysis

. . . but as all gene means are now being estimated simultaneously, it would be better to do a functional PCA

Functional Principal Components Analysis

. . . but as all gene means are now being estimated simultaneously, it would be better to do a functional PCA

Karhunen-Loève Decomposition

$$y_{ij}(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k(t)\alpha_{ik} + \sum_{l=1}^{\infty} \zeta_{il}(t)\beta_{ijl} + \epsilon_{ij}(t)$$

- $\xi_k(t)$ is the k-th PC function at the gene level
- α_{ik} is the loading on PC function k for gene i
- $\zeta_{il}(t)$ is the l-th PC function at the replicate level, for gene i
- β_{ijl} is the loading on PC function l for replicate j for gene i

Reduced Rank Functional Principal Components Analysis

Truncated Decomposition

$$y_{ij}(t) = \mu(t) + \sum_{k=1}^{K} \xi_k(t)\alpha_{ik} + \sum_{l=1}^{L_i} \zeta_{il}(t)\beta_{ijl} + \epsilon_i(t)$$

Reduced Rank Functional Principal Components Analysis

Truncated Decomposition

$$y_{ij}(t) = \mu(t) + \sum_{k=1}^{K} \xi_k(t)\alpha_{ik} + \sum_{l=1}^{L_i} \zeta_{il}(t)\beta_{ijl} + \epsilon_i(t)$$

LMM Representation

$$\mathbf{y}_{ij} = \mathbf{B}_{ij}\boldsymbol{\theta}_{\mu} + \sum_{k=1}^{K} \mathbf{B}_{ij}\boldsymbol{\theta}_{\alpha_k}\alpha_{ik} + \sum_{l=1}^{L_i} \mathbf{B}_{ij}\boldsymbol{\theta}_{\beta_{i_l}}\beta_{ijl} + \boldsymbol{\epsilon}_{ij}$$

Reduced Rank Functional Principal Components Analysis

Truncated Decomposition

$$y_{ij}(t) = \mu(t) + \sum_{k=1}^{K} \xi_k(t)\alpha_{ik} + \sum_{l=1}^{L_i} \zeta_{il}(t)\beta_{ijl} + \epsilon_i(t)$$

LMM Representation

$$\mathbf{y}_{ij} = \mathbf{B}_{ij}\boldsymbol{\theta}_{\mu} + \sum_{k=1}^{K} \mathbf{B}_{ij}\boldsymbol{\theta}_{\alpha_k}\alpha_{ik} + \sum_{l=1}^{L_i} \mathbf{B}_{ij}\boldsymbol{\theta}_{\beta_{i_l}}\beta_{ijl} + \boldsymbol{\epsilon}_{ij}$$

Distributional Assumptions and Constraints

$$\boldsymbol{\alpha}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_{\alpha}) \quad \boldsymbol{\beta}_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_{\beta_i}) \quad \boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(\mathbf{0}, \sigma_i^2 \mathbf{I}_{N_{ij}}) \quad \boldsymbol{\alpha}_\mathbf{i} \perp \boldsymbol{\beta}_\mathbf{ij} \perp \boldsymbol{\epsilon}_\mathbf{ij}$$

$$\mathbf{B}^T\mathbf{B} = \mathbf{I} \quad \boldsymbol{\Theta}_{\alpha}^T\boldsymbol{\Theta}_{\alpha} = \mathbf{I} \quad \boldsymbol{\Theta}_{\beta_i}^T\boldsymbol{\Theta}_{\beta_i} = \mathbf{I}$$

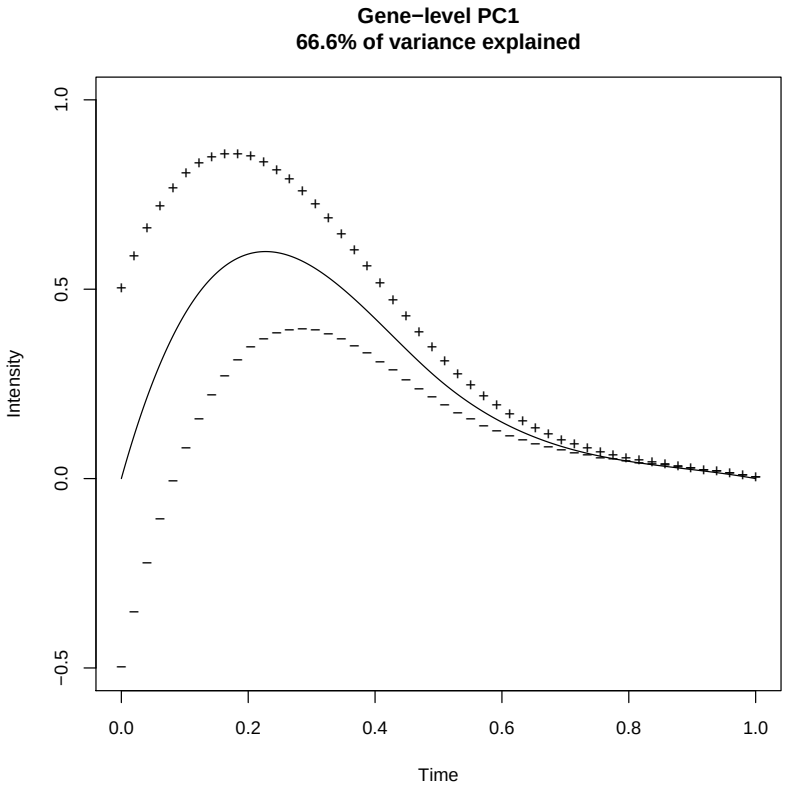
Comparison with Zhou et al (2010)

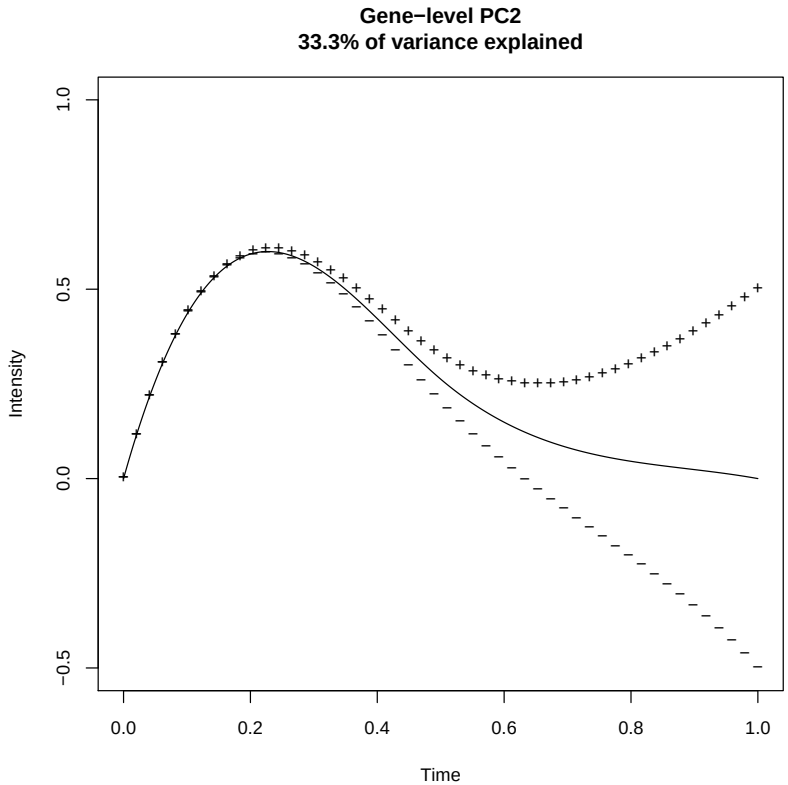
- Considered spatial correlations between the variables
- They assume second-level variance is the same for all variables
- They assume the error variance is not variable dependent
- Illustrated on a data set with three variables

Simulation setting

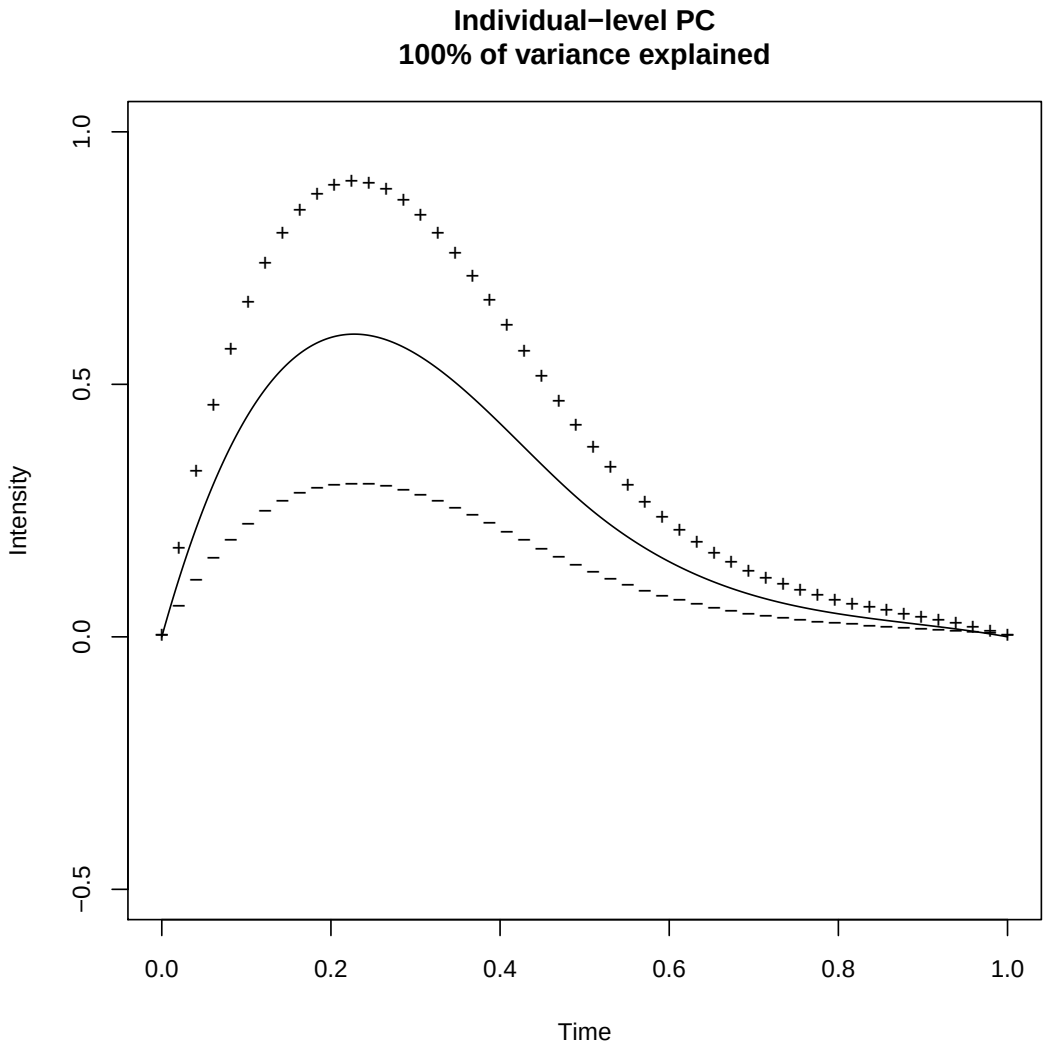
- Varied number of replicates (5, 10, 20)
- Varied number of genes (100, 1000, 10000)
- Fixed number of time points (5)
- Fixed basis (B-splines, 1 knot)
- Fixed grand mean
- Fixed 2 PCs at the gene level
- Fixed 1 PC at the replicate level
- Fixed variance components D_α and D_{β_i}
- Generated 1000 data sets for each condition

Simulation setting





Simulation setting



Simulation setting results

Mean-squared error of gene-level curves

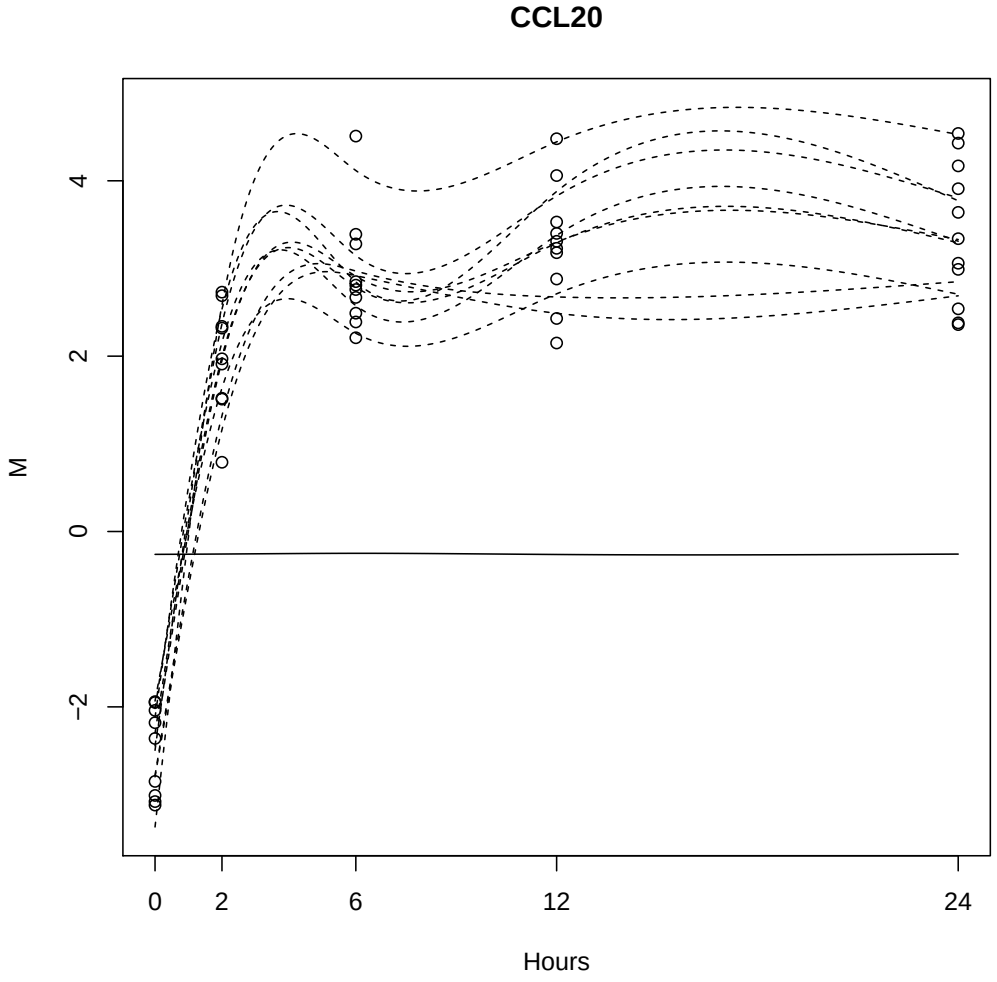
Multi-level model

		Number of Replicates		
		5	10	20
Number of Genes	100	0.000406	0.000195	0.0000953
	1000	0.000348	0.000167	0.0000813
	10000	0.000342	0.000164	0.0000799

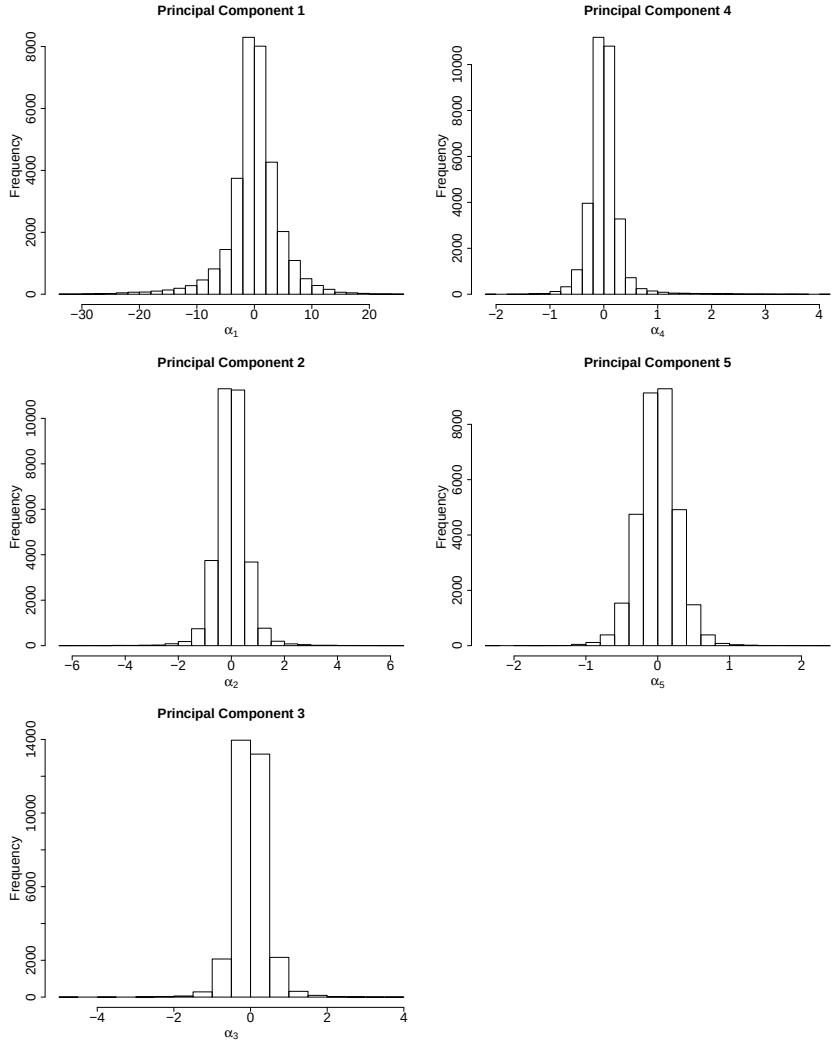
Gene-at-a-time model

		Number of Replicates		
		5	10	20
Number of Genes	100	0.00226	0.00113	0.000562
	1000	0.00225	0.00113	0.000564
	10000	0.00226	0.00113	0.000564

Real data fits



Initialised gene-level PC loadings



The solution

LMM Representation

$$\mathbf{y}_{ij} = \mathbf{B}_{ij}\boldsymbol{\theta}_{\mu} + \sum_{k=1}^{K} \mathbf{B}_{ij}\boldsymbol{\theta}_{\alpha_k}\alpha_{ik} + \sum_{l=1}^{L_i} \mathbf{B}_{ij}\boldsymbol{\theta}_{\beta_{i_l}}\beta_{ijl} + \boldsymbol{\epsilon}_{ij}$$

Distributional Assumptions and Constraints

$$\alpha_{ik} \sim \mathcal{S}t\mathcal{N}(\xi_k, \sigma^2_{\alpha_k}, \lambda_k, \nu_k)$$

$$\boldsymbol{\beta}_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_{\beta_i}) \quad \boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(\mathbf{0}, \sigma^2_i \mathbf{I}_{N_{ij}}) \quad \boldsymbol{\alpha}_{\mathbf{i}} \perp \boldsymbol{\beta}_{\mathbf{ij}} \perp \boldsymbol{\epsilon}_{\mathbf{ij}}$$

$$\mathbf{B}^T\mathbf{B} = \mathbf{I} \quad \boldsymbol{\Theta}^T_{\alpha}\boldsymbol{\Theta}_{\alpha} = \mathbf{I} \quad \boldsymbol{\Theta}^T_{\beta_i}\boldsymbol{\Theta}_{\beta_i} = \mathbf{I}$$

The skew-t-normal distribtuion

$$f(\alpha_{ik}|\xi_k, \sigma^2_{\alpha_k}, \lambda_k, \nu_k) = 2t_{\nu_k}(\alpha_{ik}|\xi_k, \sigma^2_{\alpha_k})\Phi\left(\frac{\alpha_{ik}-\xi_k}{\sigma_{\alpha_k}}\right)$$

Implications

- $\mathbf{y}_{ij}$ now has an unknown distribution

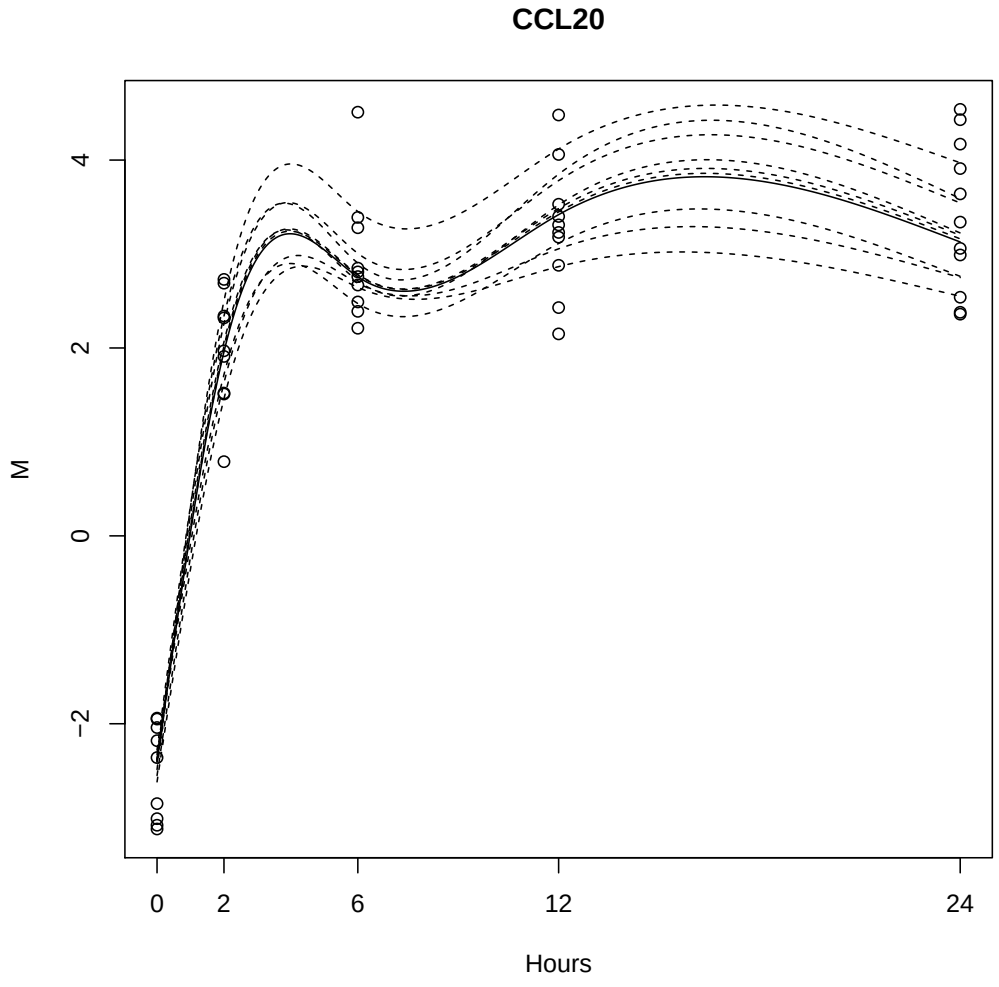
Implications

- $\mathbf{y}_{ij}$ now has an unknown distribution
- If all α_{ik}, the β_{ikl} and $\boldsymbol{\epsilon}_{ij}$ were skew-t-normally distributed *with common degrees of freedom* then the distribution would be known

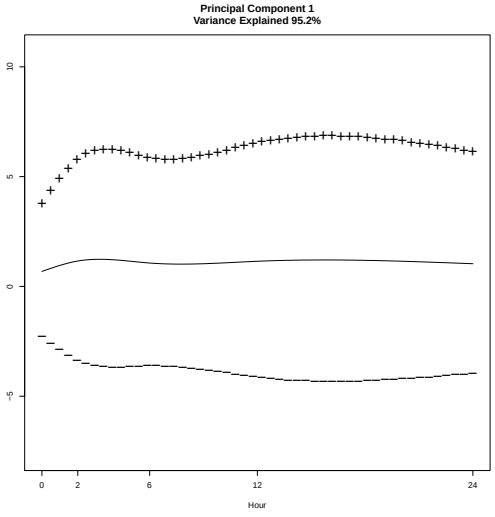
Implications

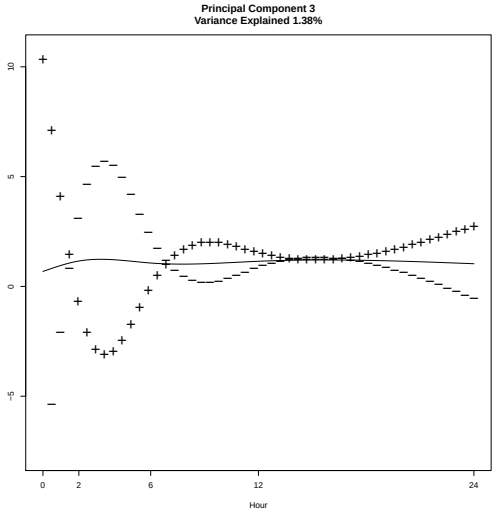
- $\mathbf{y}_{ij}$ now has an unknown distribution
- If all α_{ik}, the β_{ikl} and $\boldsymbol{\epsilon}_{ij}$ were skew-t-normally distributed *with common degrees of freedom* then the distribution would be known
- Can use the Monte-Carlo EM algorithm to *approximate* the required conditional expectations at the E-step

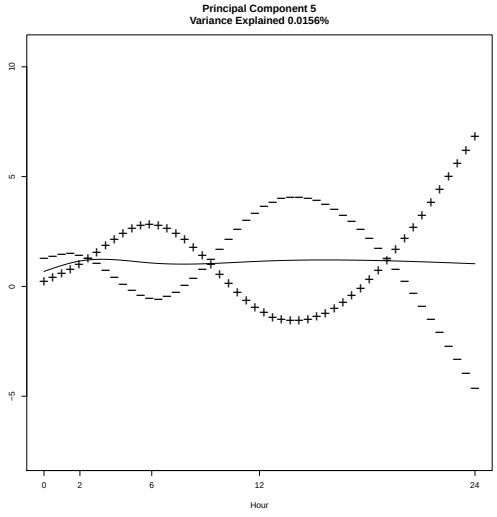
It works

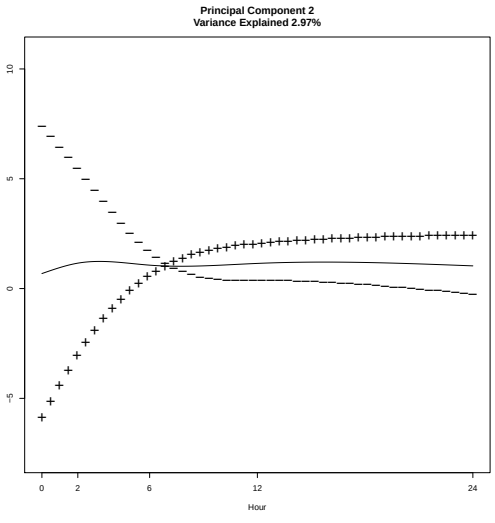


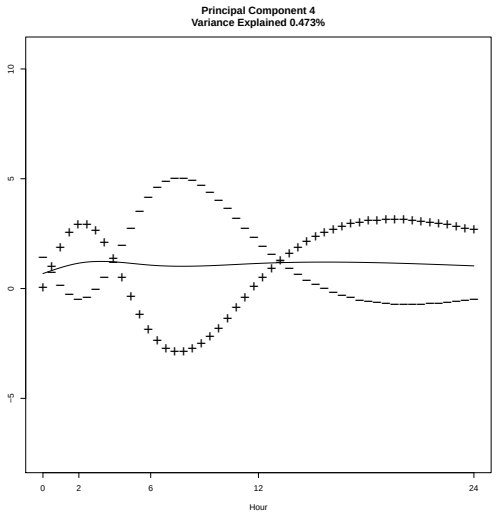
It works


Principal Component 1
Variance Explained 95.2%
Hour


Principal Component 3
Variance Explained 1.38%
Hour


Principal Component 5
Variance Explained 0.0156%
Hour


Principal Component 2
Variance Explained 2.97%
Hour


Principal Component 4
Variance Explained 0.473%
Hour

Conclusions

- Due to the technology involved, ethical constraints and the nature of the genome, replicated genomics time series data sets really do call for specific, sophisticated models
- The approach presented here is currently completely impractical
- A multi-level model is worth pursuing
- Functional PCA is hard for non-FDA specialists to understand
- Have not considered correlations amongst genes

Acknowledgements

Acknowledgements:

- PhD supervisors Giovanni Montana and Michael Levin
- Cheryl Hemingway for providing access data presented here
- Wellcome Trust

Additional Resources:

- My website: `http://www2.imperial.ac.uk/~mab201`
- `sme` R package availabe on CRAN