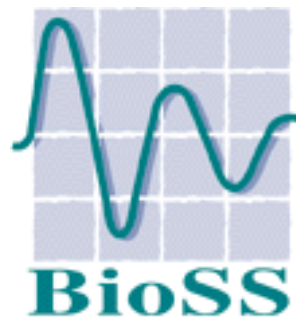


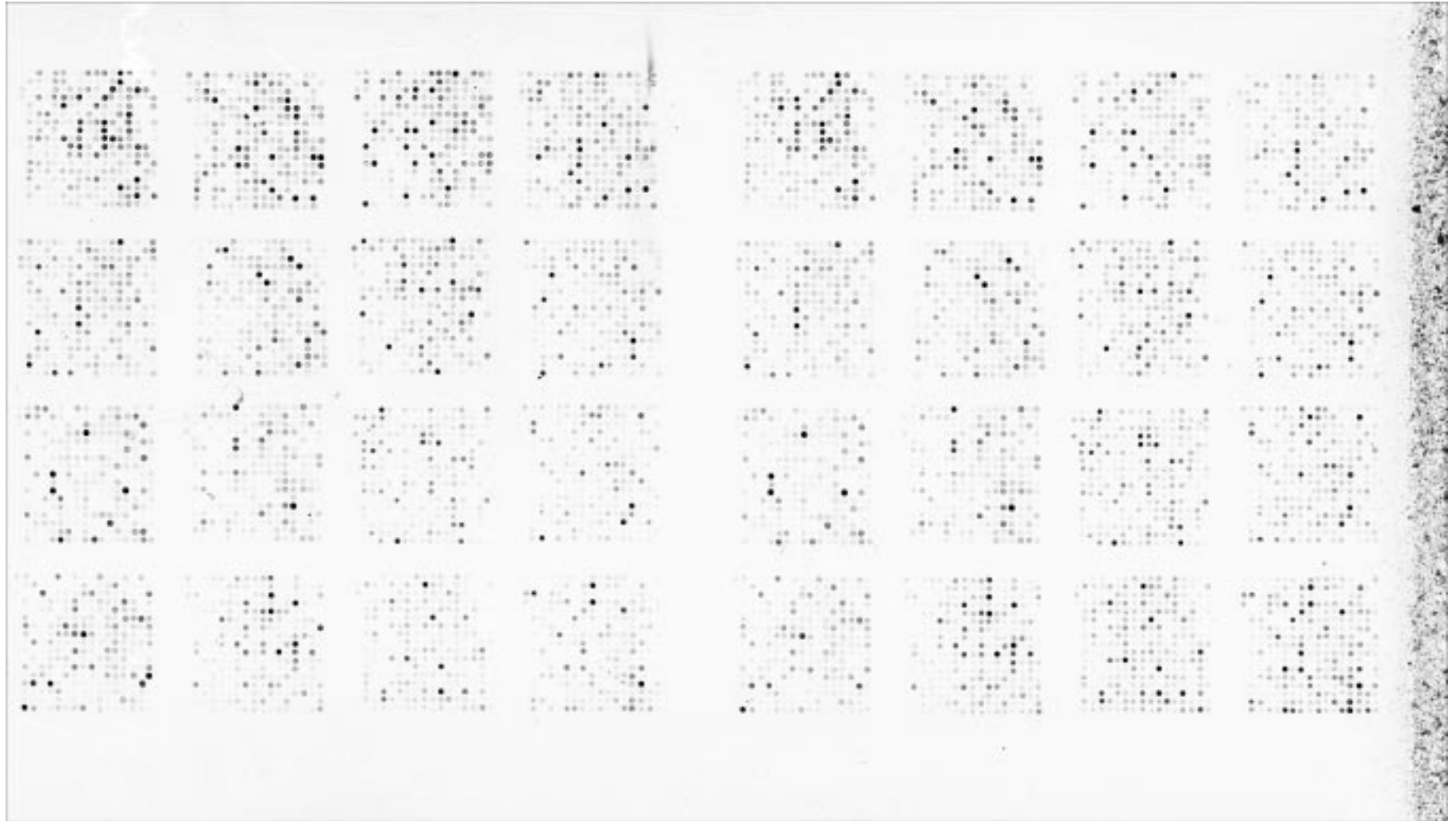
EFFICIENCY OF FUNCTIONAL REGRESSION ESTIMATORS FOR COMBINING MULTIPLE LASER SCANS OF cDNA MICROARRAYS

Chris Glasbey & Mizan Khondoker

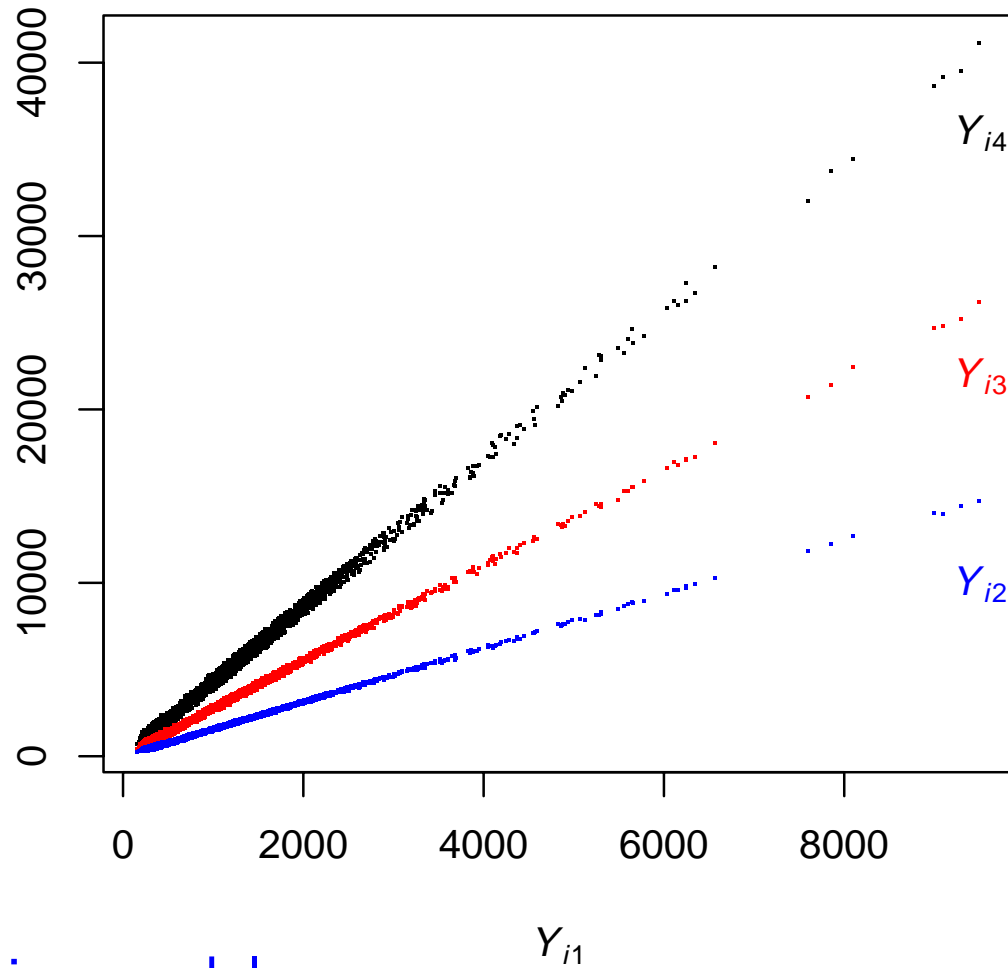


Biomathematics and Statistics Scotland

cDNA array to study ingestion of apoptotic cells by rodent macrophages



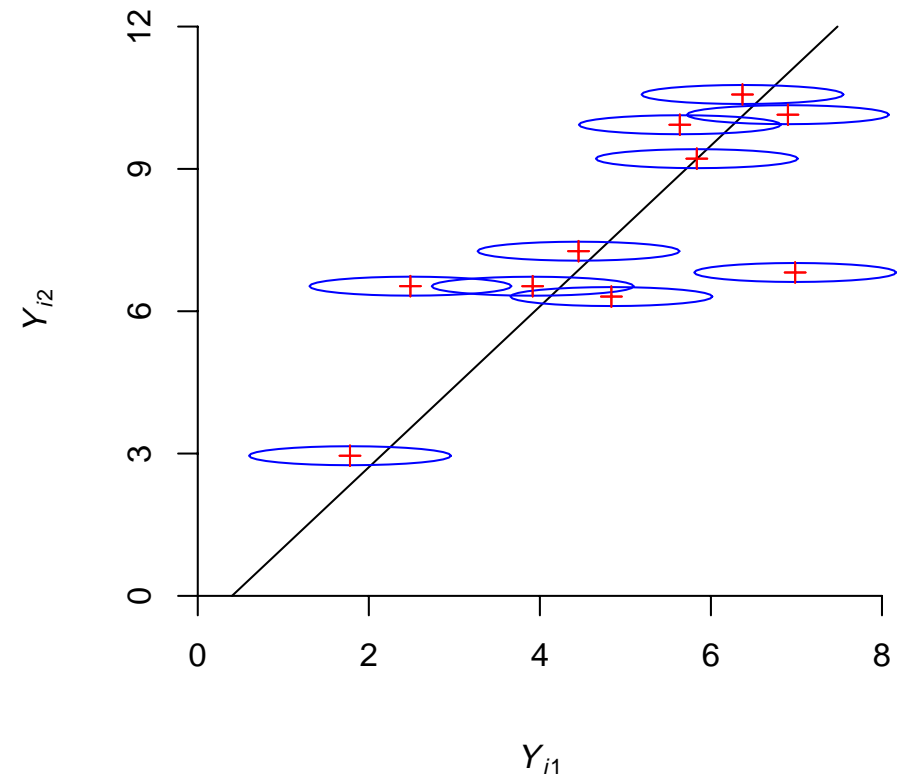
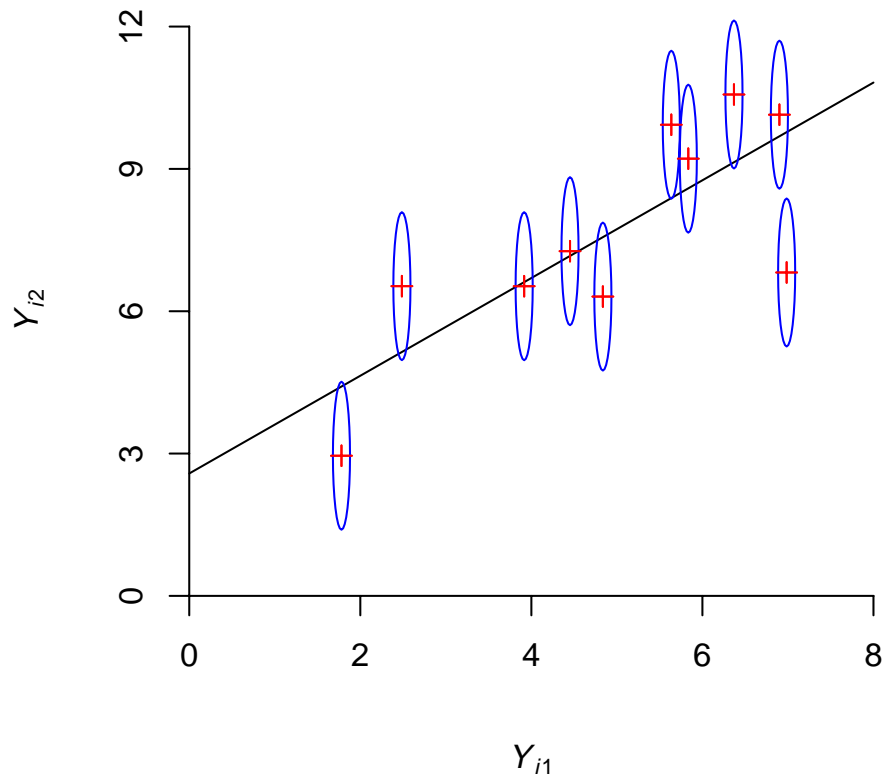
Spot summary data from 4 laser scans



Functional regression model

$$Y_{ij} \sim N(\mu_i \beta_j, \sigma_j^2) \quad \text{for gene } i = 1, \dots, n, \quad \text{scan } j = 1, \dots, m$$

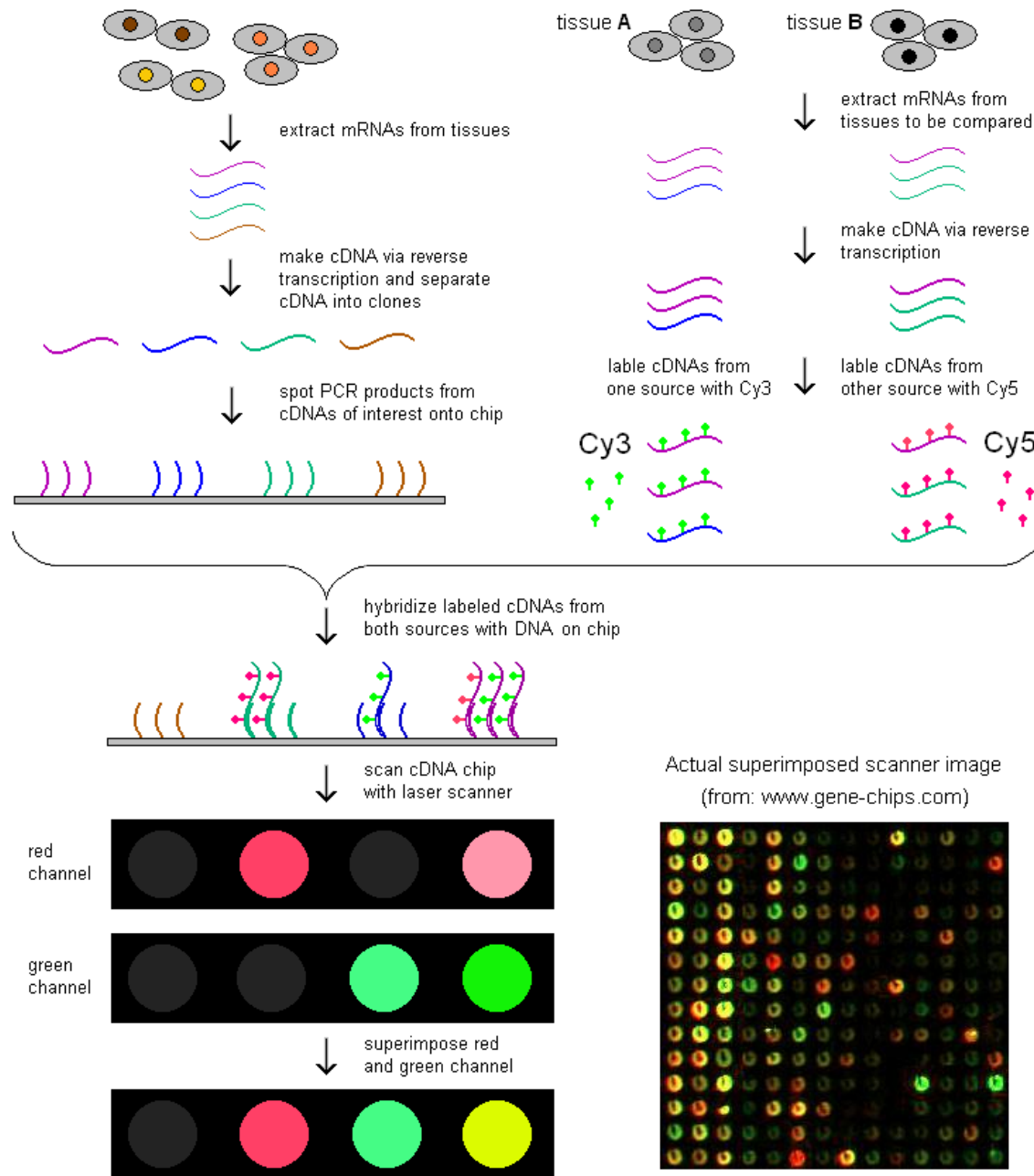
Functional regression is intractable if there are 2 variables and an intercept term, because there is insufficient information to distinguish σ_1 from σ_2



But what if there are > 2 variables and no intercept term?

Functional regression appears in many guises:

- Ultrastructural relationship
- One-factor factor analysis model
- Instrumental- or latent-variable model
- Errors-in-variables or measurement error model



Stages in analysing microarray data:

- A. Estimation of gene expression level from scan data
- B. Normalisation to standardise means and variances
- C. Detection of differential expression
- D. Multivariate methods
- E. Dynamic models

Model: $Y_{ij} \sim N(\mu_i \beta_j, \sigma_j^2)$ (with $\beta_1 \equiv 1$ for identifiability)

The minimum variance unbiased estimator of gene i expression level is

$$\hat{\mu}_i = \sum_j w_j \frac{Y_{ij}}{\beta_j}, \quad w_j = \frac{\beta_j^2 / \sigma_j^2}{\sum_k \beta_k^2 / \sigma_k^2}$$

But how do we estimate β s and σ s?

Maximum likelihood fails, because likelihood $\rightarrow \infty$ as $\sigma_j^2 \rightarrow 0$ for any j

However, for $m \geq 3$ there is sufficient information in second moments to estimate β s and σ s, for example by minimising:

$$\text{trace}\{(S-V)^2\} : \quad S = \frac{Y^T Y}{n}, \quad V = E(S) = \overline{\mu^2} \beta \beta^T + \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\}$$

2nd moment estimates for $Y_{ij} \sim N(\mu_i\beta_j, \sigma_j^2)$

j	$\hat{\beta}_j$	$\hat{\sigma}_j$	$sd(\hat{\mu})$
1	1.00	27.0	27.0
2	1.56	30.6	19.6
3	2.75	49.6	18.1
4	4.29	149.5	34.8
all			11.3

These estimators are consistent, but are they efficient?

We simulated 1000 datasets similar to the observed data, then fitted the model

Estimation method	$10^6 \times \text{rmse}$			$10^3 \times \text{rmse}$			
	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$
2nd moments	562	961	2110	344	507	920	1389
centred 2nd moments	826	1406	3003	344	507	920	1389
1st and 2nd moments	768	1316	2858	13002	18210	30121	31226
1st and centred 2nd moments	768	1316	2858	12907	18216	30045	31024

Are these the smallest possible root-mean-square errors?

Maximum likelihood theory fails because the number of parameters increases with sample size

One way around this is to eliminate the μ s. For example:

$$\left(Y_{ij} - \frac{\beta_j}{\beta_k} Y_{ik} \right) \equiv Z_{ijk} \sim N(0, \omega_{jk}^2), \quad \text{where } \omega_{jk}^2 \equiv \sigma_j^2 + \frac{\beta_j^2}{\beta_k^2} \sigma_k^2$$

We can estimate the β s and σ s by maximising the product of likelihoods of all such pairwise differences:

$$\prod_{j \neq k} \prod_i \frac{1}{\omega_{jk}} \exp \left[-\frac{Z_{ijk}^2}{2\omega_{jk}^2} \right]$$

More simulation results:

Estimation method	$10^6 \times \text{rmse}$			$10^3 \times \text{rmse}$			
	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$
2nd moments	562	961	2110	344	507	920	1389
Pairwise differences or $(\beta_k Y_{ij} - \beta_j Y_{ik})$	601	1073	2351	270	398	707	1392
	2856	5075	7516	274	398	706	1392

None of these methods is efficient!

Chan & Mak (*Biometrika*, 1983) proposed consistent estimators for the more general model

$$Y_{ij} \sim N(\alpha_j + \mu_i \beta_j, \sigma_j^2),$$

which they showed to be identical to maximum likelihood estimators for the Gaussian structural model

$$\mu_i \sim N(\nu, \tau^2) \quad \Rightarrow \quad Y_{i.} \sim N_m(\alpha + \nu \beta, \tau^2 \beta \beta^T + \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\})$$

We omit α s

We could also omit ν , because $Y_{i.}$ and $-Y_{i.}$ are equivalent in the functional regression model, and minimise $\log |V| + \text{trace}(SV^{-1})$

If the set of values of μ is compatible with a Gaussian distribution, we would expect these estimators to be reasonably efficient. But, what if they are far from Gaussian?

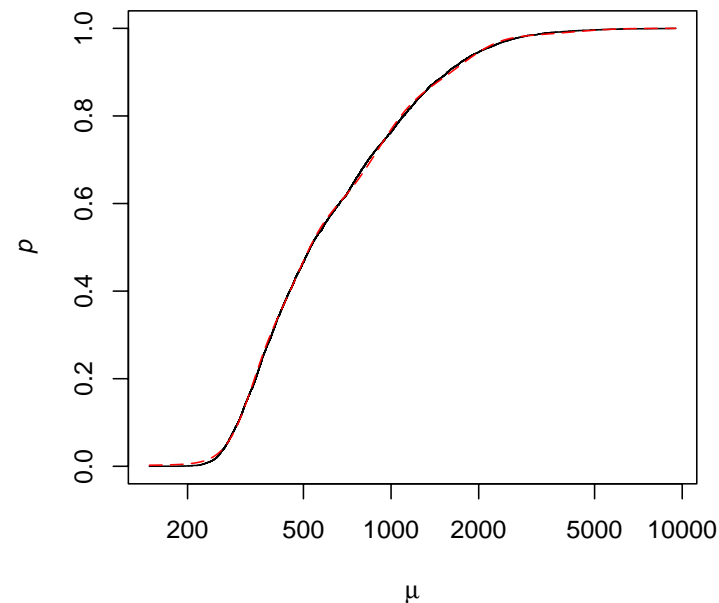
More simulation results:

Estimation method	$10^6 \times \text{rmse}$			$10^3 \times \text{rmse}$			
	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$
2nd moments	562	961	2110	344	507	920	1389
Pairwise differences	601	1073	2351	270	398	707	1392
Gaussian	562	961	2109	269	396	707	1389
Gaussian (with $\nu \equiv 0$)	562	962	2110	269	396	707	1390

Are the Gaussian methods efficient?

If we assume the μ s are drawn from a distribution, then this can be approximated by a mixture of Gaussians

Empirical and **fitted** cumulative distribution function of μ



Where $\mu \sim N(\nu_k, \tau_k^2)$ with probability π_k for $k = 1, \dots, K$

k	$\hat{\nu}_k$	$\hat{\tau}_k$	$\hat{\pi}_k$
1	320	48	0.25
2	453	90	0.27
3	808	241	0.28
4	1530	536	0.17
5	3379	1632	0.03

If μ is distributed as a mixture of Gaussians, then $Y_{i.}$ is a mixture of m -dimensional Gaussians:

$$Y_{i.} \sim N_m(\nu_k \beta, \tau_k^2 \beta \beta^T + \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\}) \quad \text{prob. } \pi_k \quad \text{for } k = 1, \dots, K$$

We can estimate β s and σ s by maximising the likelihood of this mixture distribution

If we assume that the distribution of μ is known, the remaining parameters should be estimated with super efficiency

We simulated 1000 datasets from the mixture model, then fitted by the five methods:

Estimation method	$10^6 \times \text{rmse}$			$10^3 \times \text{rmse}$				rmse
	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$	$\hat{\mu}$
2nd moments	562	961	2110	344	507	920	1389	11.3
Pairwise differences	601	1073	2351	270	398	707	1392	11.3
Gaussian	562	961	2109	269	396	707	1389	11.3
Gaussian (with $\nu \equiv 0$)	562	962	2110	269	396	707	1390	11.3
Known Gaussian mixture	563	961	2115	269	395	701	1387	11.3

Using a known Gaussian mixture should give a lower bound on what is achievable by any other method, and hence a lower bound on efficiencies

We see that the Gaussian methods, with or without ν , are efficient

Still true if we include intercept terms (α)

However, all methods estimate μ s efficiently

To see what happens with noisier data, we repeated the simulations with $10 \times \sigma$ s:

Estimation method	$10^5 \times \text{rmse}$			$10^2 \times \text{rmse}$				rmse
	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$	$\hat{\mu}$
2nd moments	579	998	2145	344	507	920	1389	113
Pairwise differences	2292	4742	8972	271	393	701	1565	114
Gaussian	571	979	2128	268	392	696	1382	113
Gaussian (with $\nu \equiv 0$)	571	979	2126	269	396	706	1390	113
Known Gaussian mixture	537	909	2071	254	344	609	1345	113

Gaussian methods are less efficient in this case

Summary

Combining multiple laser scans of cDNA microarrays can be formulated as a functional regression problem

Maximum likelihood cannot be used to fit such models, even though in ≥ 2 dimensions there is sufficient information in the data to estimate all parameters

However, we have shown that fitting a seemingly inappropriate Gaussian structural model leads to efficient estimators in our region of parameter space

Matching second moments and using pairwise differences are less efficient, though better than several other methods we have used

For further details, see paper on <http://www.bioss.ac.uk/~chris>