# Mathematics for Cancer Research

making optimal use of cancer data

ACC Coolen

King's College London

1. Quality of raw data
   - Bayesian analysis of imaging data
   - Proteome data decontamination

2. Complex signalling processes
   - Many-variable systems in biology
   - Signalling in the proteome
   - Cytokine signalling in adaptive immune system

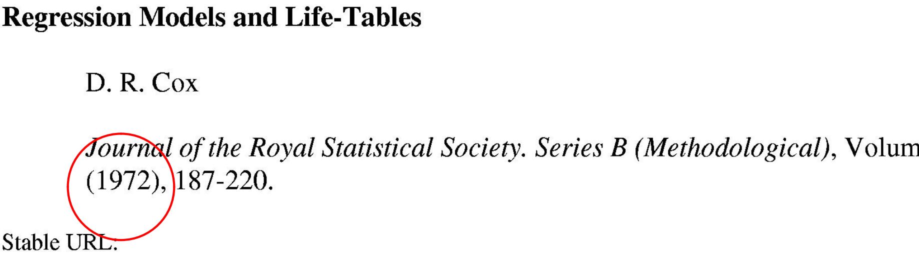
3. Risk associations and outcome prediction
   - Overfitting in clinical outcome prediction
   - Bayesian latent variable analysis
   - Prediction from high dimensional covariates
   - Heterogeneity and competing risks

*a selection of past and present projects ...*

## biomedical research in 21st century

biology,
medicine,
chemistry,
physics,
engineering,
computer science,
mathematics,
....

‘next generation’ data ...
... previous generation analysis



**Regression Models and Life-Tables**

D. R. Cox

*Journal of the Royal Statistical Society. Series B (Methodological)*, Volum
(1972), 187-220.

Stable URL:

1. Quality of raw data
   - Bayesian analysis of imaging data
   - Proteome data decontamination
2. Complex signalling processes
   - Many-variable systems in biology
   - Signalling in the proteome
   - Cytokine signalling in adaptive immune system
3. Risk associations and outcome prediction
   - Overfitting in clinical outcome prediction
   - Bayesian latent variable analysis
   - Prediction from high dimensional covariates
   - Heterogeneity and competing risks

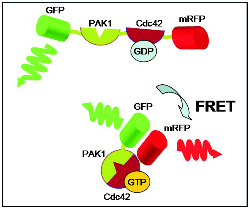
# Bayesian analysis of imaging data

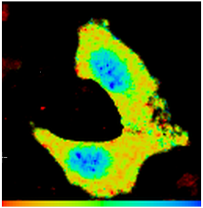
Fluorescence Lifetime Imaging
data: arrival times of photons

- **goal**

  emission lifetime of
  light emitting molecules

  fast processes:
  small nr of photons

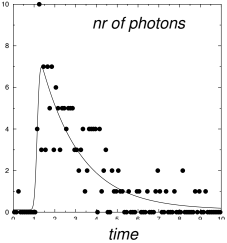




- **problem with small photon nrs**

  – to fit to decay curve,
  need histogram of arrival times

  – large bins: time resolution poor ...
  small bins: vertical resolution poor ...



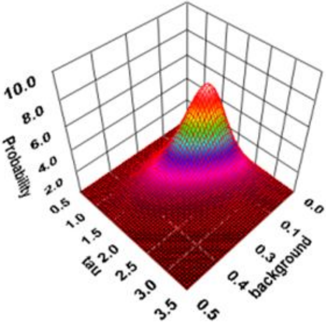
## Bayesian analysis

photon detection = emission physics + instrument + noise
parameters $\boldsymbol{\theta}$

$$\textit{forward model}: \quad p(\textit{data}|\boldsymbol{\theta}), \qquad \textit{prior}: \quad p(\boldsymbol{\theta})$$

- calculate $p(\textit{data}|\boldsymbol{\theta})$
- Bayesian identity:

$$p(\boldsymbol{\theta}|\textit{data}) = \frac{p(\textit{data}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{\theta}'\ p(\textit{data}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}$$



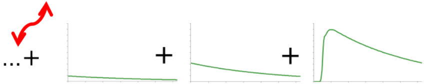
### benefits

– exact, statistically optimal
– estimates *with error bars*

## ‘**forward modelling**’

$$p(\Delta t)=\theta(\Delta t)\theta(T-\Delta t)\left\{\overset{\text{Background photons}}{\frac{w_0}{T}}+(1-w_0)\overbrace{\frac{\iint_0^\infty dtdu\; p(t)\Gamma(u)\delta\left(\Delta t-t-u+T_m.\text{int}\left(\frac{t+u}{T_m}\right)\right)}{\int_0^T d\Delta t' \iint_0^\infty dtdu\; p(t)\Gamma(u)\delta\left(\Delta t'-t-u+T_m.\text{int}\left(\frac{t+u}{T_m}\right)\right)}}^{\text{Decay photons}}\right\}$$

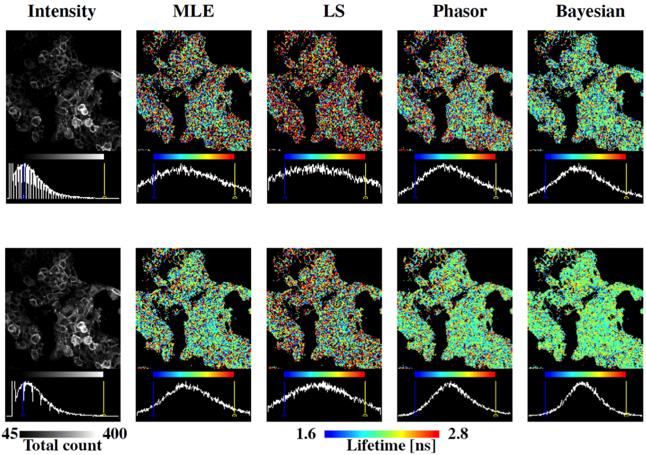
$$=\theta(\Delta t)\theta(T-\Delta t)\left\{\frac{w_0}{T}+\frac{1-w_0}{\Lambda(T,T_m)}\int_0^\infty dt\; p(t)\sum_{\ell\geq 0}\Gamma(\Delta t-t+\ell T_m)\right\}$$



includes:

– instrument response function
– artifacts of repetitive excitation
– multi-exponential delay distributions
– Bayesian model selection

example:
human epithelial cancer cells



*compared to existing methods:*
*half nr photons needed for same accuracy*

# Protein interaction networks

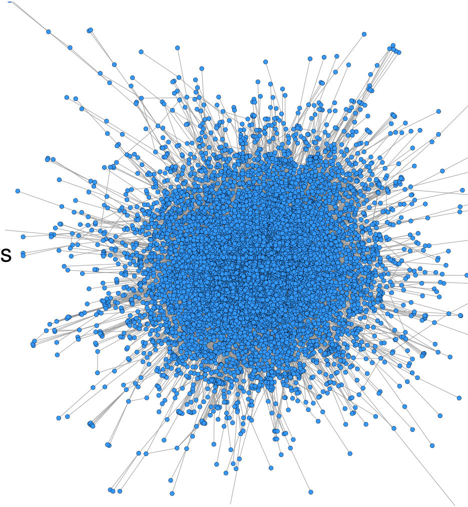
Quantify topology:

- $p(k)$:

  fraction of *nodes* that
  have $k$ neighbours (degree distr)

- $W(k, k')$:

  fraction of *links* that
  connect nodes with $k$ and $k'$ neighbours

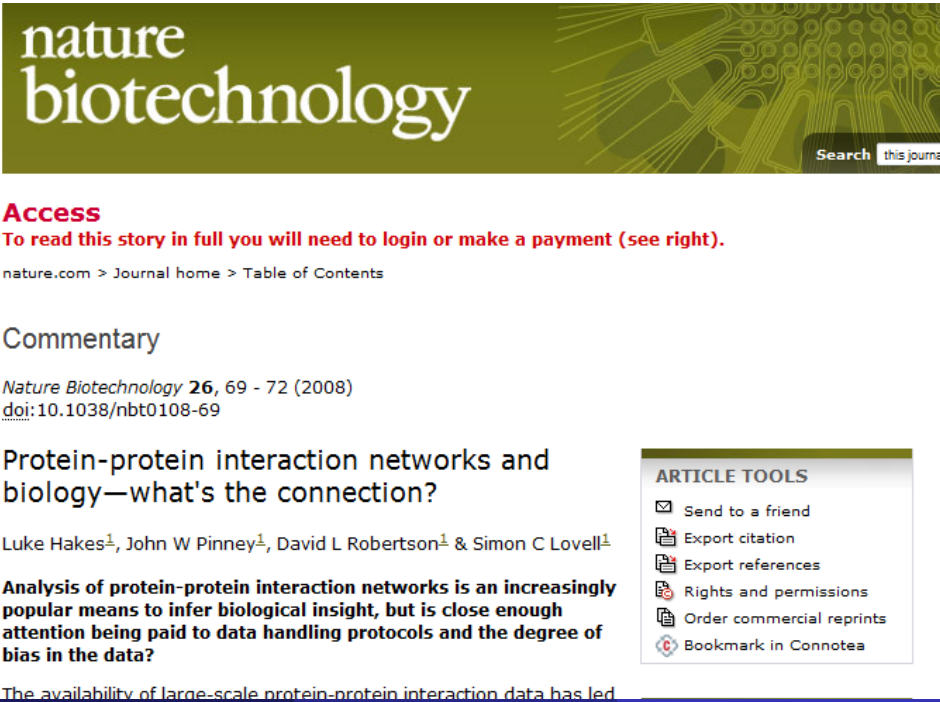


**Mathematical tools**

graph theory, information-theory, and statistical physics

tailored random graph families, characterised by $\{p, W\}$:

*quantify complexity, appropriate network null models, algorithms for correct randomisation, proxies for process modelling, network dissimilarity measures, ...*



nature biotechnology

**Access**
**To read this story in full you will need to login or make a payment (see right).**

nature.com > Journal home > Table of Contents

Commentary

*Nature Biotechnology* **26**, 69 - 72 (2008)
doi:10.1038/nbt0108-69

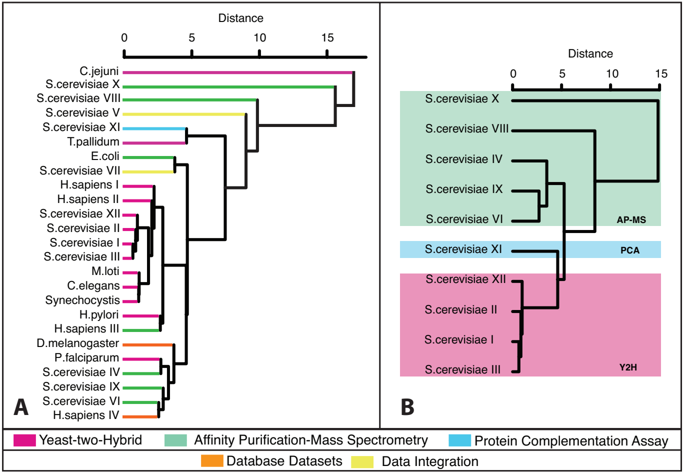
## Protein-protein interaction networks and biology—what's the connection?

Luke Hakes[1], John W Pinney[1], David L Robertson[1] & Simon C Lovell[1]

**Analysis of protein-protein interaction networks is an increasingly popular means to infer biological insight, but is close enough attention being paid to data handling protocols and the degree of bias in the data?**

The availability of large-scale protein-protein interaction data has led

Quantify network dissimilarity using information theory

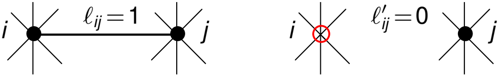


- PPINs of same species are similar only if measured via same method
- strong **bias** in PPIN data, induced by **experimental method**, that overrules species information

analysis of
## data contamination by experimental bias

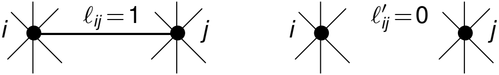
- node undersampling:

  $x(k_i)$: prob to detect protein $i$



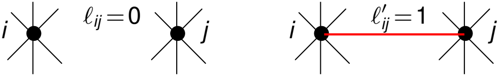
- link undersampling:

  $y(k_i, k_j)$: prob to detect interaction $(i, j)$



- link oversampling:

  $z(k_i, k_j)/N$: prob to report false positive interaction

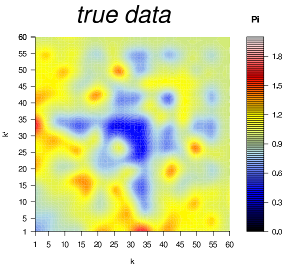


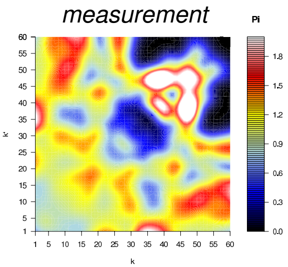
methods from statistical physics:

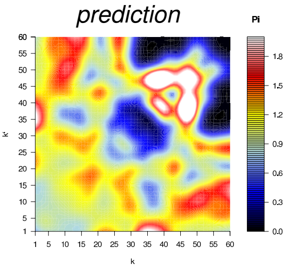
*relation between measured* $p(k)$ *and* $W(k, k')$
*and true* $p(k)$ *and* $W(k, k')$

*in terms of* $x(k), y(k), z(k, k')$

colour plots of $W(k,k')/W(k)W(k')$:


*true data*


*measurement*

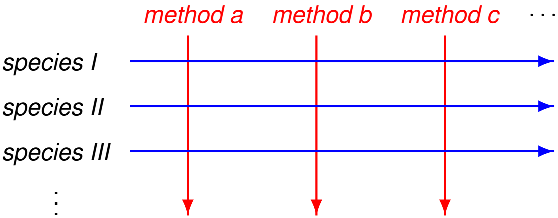

*prediction*

# Bayesian decontamination of PPIN data

– protein species $\ell = 1 \ldots L$
unknown networks $\mathbf{c}^\ell$

– experimental methods $\alpha = 1 \ldots M$ (Y2H, PCA, MS, ...)
unknown error parameters $\theta^\alpha = \{x^\alpha(k), y^\alpha(k,k'), z^\alpha(k,k')\}$

matrix of $M \times L$
observed networks $\mathbf{c}^{\ell,\alpha}$:



*method a* *method b* *method c* $\cdots$

*species I*

*species II*

*species III*

$\vdots$

**recover**:

true PINs $\{\mathbf{c}^1, \ldots, \mathbf{c}^L\}$
sampling pars $\{\theta^1, \ldots, \theta^M\}$

1. Quality of raw data
   - Bayesian analysis of imaging data
   - Proteome data decontamination

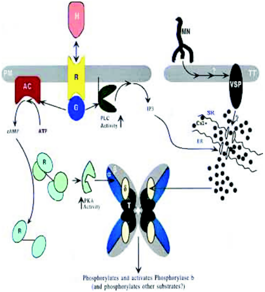
2. Complex signalling processes
   - Many-variable systems in biology
   - Signalling in the proteome
   - Cytokine signalling in adaptive immune system

3. Risk associations and outcome prediction
   - Overfitting in clinical outcome prediction
   - Bayesian latent variable analysis
   - Prediction from high dimensional covariates
   - Heterogeneity and competing risks

# Analysis of signalling processes

proteome:

usual description
reaction equations


H
MN
PM
R
TT
AC
VSP
G
IP3
PLC Activity
cAMP
ATP
SR
$Ca^{2+}$
ER
R
PKA Activity
α
β
δ
T
R
Phosphorylates and activates Phosphorylase b
(and phosphorylates other substrates?)

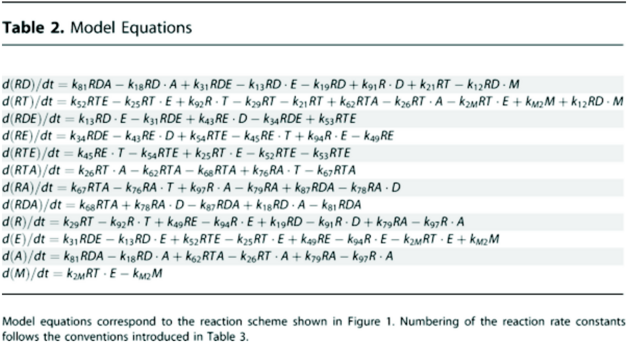


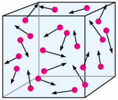
**Table 2.** Model Equations

$$
\begin{aligned}
d(RD)/dt &= k_{81}RDA - k_{18}RD \cdot A + k_{31}RDE - k_{13}RD \cdot E - k_{19}RD + k_{91}R \cdot D + k_{21}RT - k_{12}RD \cdot M \\
d(RT)/dt &= k_{52}RTE - k_{25}RT \cdot E + k_{92}R \cdot T - k_{29}RT - k_{21}RT + k_{62}RTA - k_{26}RT \cdot A - k_{2M}RT \cdot E + k_{M2}M + k_{12}RD \cdot M \\
d(RDE)/dt &= k_{13}RD \cdot E - k_{31}RDE + k_{43}RE \cdot D - k_{34}RDE + k_{53}RTE \\
d(RE)/dt &= k_{34}RDE - k_{43}RE \cdot D + k_{54}RTE - k_{45}RE \cdot T + k_{94}R \cdot E - k_{49}RE \\
d(RTE)/dt &= k_{45}RE \cdot T - k_{54}RTE + k_{25}RT \cdot E - k_{52}RTE - k_{53}RTE \\
d(RTA)/dt &= k_{26}RT \cdot A - k_{62}RTA - k_{68}RTA + k_{76}RA \cdot T - k_{67}RTA \\
d(RA)/dt &= k_{67}RTA - k_{76}RA \cdot T + k_{97}R \cdot A - k_{79}RA + k_{87}RDA - k_{78}RA \cdot D \\
d(RDA)/dt &= k_{68}RTA + k_{78}RA \cdot D - k_{87}RDA + k_{18}RD \cdot A - k_{81}RDA \\
d(R)/dt &= k_{29}RT - k_{92}R \cdot T + k_{49}RE - k_{94}R \cdot E + k_{19}RD - k_{91}R \cdot D + k_{79}RA - k_{97}R \cdot A \\
d(E)/dt &= k_{31}RDE - k_{13}RD \cdot E + k_{52}RTE - k_{25}RT \cdot E + k_{49}RE - k_{94}R \cdot E - k_{2M}RT \cdot E + k_{M2}M \\
d(A)/dt &= k_{81}RDA - k_{18}RD \cdot A + k_{62}RTA - k_{26}RT \cdot A + k_{79}RA - k_{97}R \cdot A \\
d(M)/dt &= k_{2M}RT \cdot E - k_{M2}M
\end{aligned}
$$

Model equations correspond to the reaction scheme shown in Figure 1. Numbering of the reaction rate constants follows the conventions introduced in Table 3.

- cannot solve eqns analytically ...
- uncertain pathways and parameters ...
- too many components for numerical exploration ...

## statistical physics



$\sim 10^{24}$ positions, velocities

$(\vec{x}_1, \vec{v}_1),\ (\vec{x}_2, \vec{v}_2),\ \ldots$

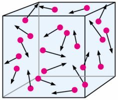
Newton's equations

$\frac{d}{dt}(\vec{x}_1, \vec{v}_1) = \ldots,\ \frac{d}{dt}(\vec{x}_2, \vec{v}_2) = \ldots$ $\leftarrow$ don't try to solve these!

*macroscopic description:*

densities, correlation functions,
perturbation response functions,
phase transitions ...

## statistical physics



$\sim 10^{24}$ positions, velocities

$(\vec{x}_1, \vec{v}_1),\ (\vec{x}_2, \vec{v}_2),\ \ldots$

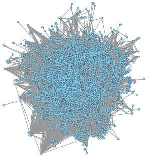
Newton's equations

$\frac{d}{dt}(\vec{x}_1, \vec{v}_1) = ...,\ \frac{d}{dt}(\vec{x}_2, \vec{v}_2) = ...$

*macroscopic theory:*

densities, correlation functions,
response functions (to perturbations),
phase transitions ...

## statistical biology



$\sim 10^{4}$ concentr of proteins & complexes

$\vec{x}_1,\ \vec{x}_2,\ \vec{x}_3,\ \ldots$

reaction equations

$\frac{d}{dt}\vec{x}_1 = ...,\ \frac{d}{dt}\vec{x}_2 = ...,\ \frac{d}{dt}\vec{x}_3 = ...$

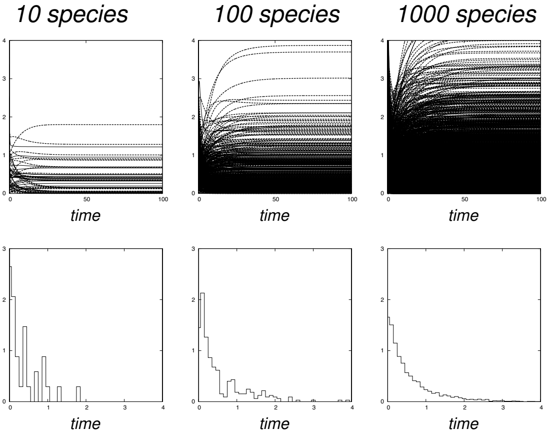
*macroscopic theory:*

???

## numerical illustration

2 post-transl states/protein,
binary complexes,
random topologies & rates,
7 partners on average

*dashed: complexes*
*solid: unbound proteins*



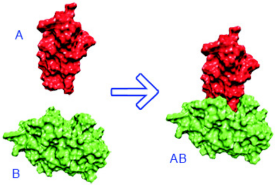
*individual concentrations*

*stationary state distribution of concentrations*

depends only on param & network *statistics*!

# Signalling dynamics in the proteome

from many-particle physics
to *many-particle biology*



- notation:

  $i = 1 \ldots N$ labels proteins
  $x_i^\alpha$: concentr of protein $i$ in state $\alpha$
  $x_{ij}$: concentration of dimer $i \asymp j$

- events:

| | | *rate:* |
|---|---|---|
| complex formation: | $(i,\alpha)+(j,\beta) \to (i \asymp j)$ | $k_{ij}^{\alpha\beta+} x_i^\alpha x_j^\beta$ |
| complex dissociation: | $(i \asymp j) \to (i,\alpha)+(j,\beta)$ | $k_{ij}^{\alpha\beta-} x_{ij}$ |
| conformation change: | $(i,\alpha) \to (i,\beta)$ | $\lambda_i^{\alpha\beta} x_i^\alpha$ |
| protein degradation: | $(i,\alpha) \to \emptyset$ | $\gamma_i^\alpha x_i^\alpha$ |
| protein synthesis: | $\emptyset \to (i,\alpha)$ | $\theta_i^\alpha$ |

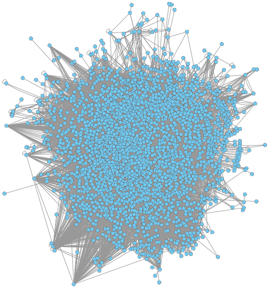
- reaction eqns:

$$\frac{\mathrm{d}}{\mathrm{d}t}x_i^\alpha=\sum_j c_{ij}\overbrace{\sum_\beta[k_{ij}^{\alpha\beta-}x_{ij}-k_{ij}^{\alpha\beta+}x_i^\alpha x_j^\beta]}^{complex\ \ formation\ \&\ \ dissociation}+\overbrace{\sum_\beta[\lambda_i^{\beta\alpha}x_i^\beta-\lambda_i^{\alpha\beta}x_i^\alpha]}^{post-transl\ \ modification}-\overbrace{\gamma_i^\alpha x_i^\alpha}^{decay}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}x_{ij}=c_{ij}\sum_{\alpha\beta}[k_{ij}^{\alpha\beta+}x_i^\alpha x_j^\beta-k_{ij}^{\alpha\beta-}x_{ij}]$$

- tailored random PPIN (prescribed degrees)
$c_{ij}=0,1$

$$p(\mathbf{c})=\frac{\prod_i\delta_{k_i,\sum_{j\neq i}c_{ij}}}{Z}\prod_i\Big[c_0\delta_{c_{ii},1}+(1-c_0)\delta_{c_{ii},0}\Big]$$



- draw reaction rates randomly
from realistic distributions $P(k^+,k^-)$, $P(\lambda,\gamma)$

## generating functional analysis

calculate correlations, response functions etc ...
in heterogeneous many-variable systems
without solving microscopic equations!

- **after calculations** ...
  (path integral techniques, saddle-point integration, etc)

  for $N \to \infty$: exact
  macroscopic equations

$$W = \mathcal{G}_1[W], \quad D = \mathcal{G}_2[W], \qquad \mathcal{G}_{1,2}: \textit{ complicated but \underline{exact} formulas}$$

macroscopic
quantities:

$$D[\{x\}|\{y\}], \quad W[\{x\}|\{y\}]$$

$\{x\}$ : *trajectories* $x_\alpha(t)$

$\{y\}$ : *time dependent production rates* $y_\alpha(t)$

$D[\{x\}|\{y\}]$ describes response
to single-node perturbations

motivation:
immune cancer therapies

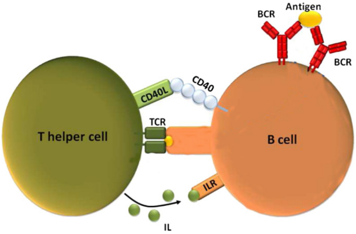
## Cytokine signalling in adaptive immune system

- B-*clones* $b_\mu$

  each can recognise *specific* antigen $a_\mu$

- T-*clones* $\sigma_i$

  coordinate B-clones via cytokines $\xi_i^\mu = -1, 0, 1$
  ($\xi_i^\mu = -1$: contract, $\xi_i^\mu = +1$: expand)

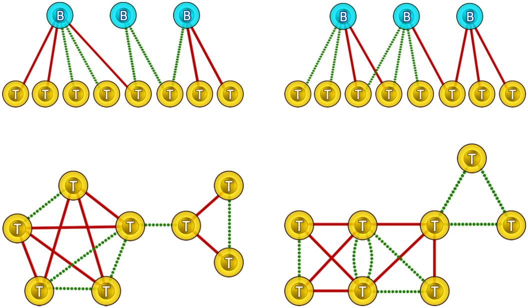


model of
Barra and Agliari:

$$p(\boldsymbol{\sigma}, \mathbf{b}) = \frac{\mathrm{e}^{-\sqrt{\beta} H(\boldsymbol{\sigma}, \mathbf{b})}}{Z} \qquad H(\boldsymbol{\sigma}, \mathbf{b}) = \frac{1}{2\sqrt{\beta}} \sum_{\mu=1}^{n_B} b_\mu^2 - \sum_{\mu=1}^{n_B} b_\mu \overbrace{\Big( \sum_{i=1}^{n_T} \xi_i^\mu \sigma_i + \lambda_\mu a_\mu \Big)}^{\text{expansion force on clone } \mu}$$

‘integrate out’ the B-clones,
results in model of interacting T-clones:

$$p(\boldsymbol{\sigma}) = \frac{e^{-\beta H(\boldsymbol{\sigma})}}{Z_T} \qquad H(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j=1}^{n_T} \sigma_i\sigma_j \sum_{\mu=1}^{n_B} \xi_i^\mu \xi_j^\mu - \sum_{i=1}^{n_T} \sigma_i \sum_{\mu=1}^{N_B} \lambda_\mu g_\mu \xi_i^\mu$$



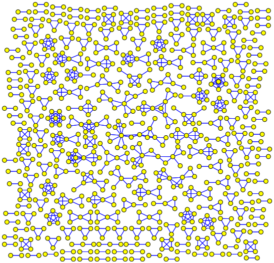
$n_B \sim 10^8$

$n_T \sim 2.10^8$

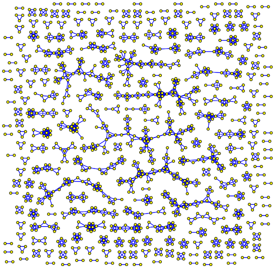
*how can promiscuous T-clones coordinate an extensive number of B-clones simultaneously?*

relevant parameters in
$T{-}T$ network:

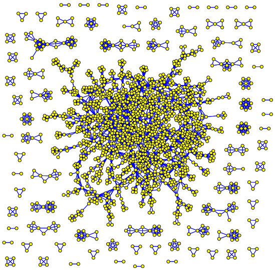
$c$: $T$-clone promiscuity
$\alpha$: $n_B/n_T$



$\alpha c^2 < 1$



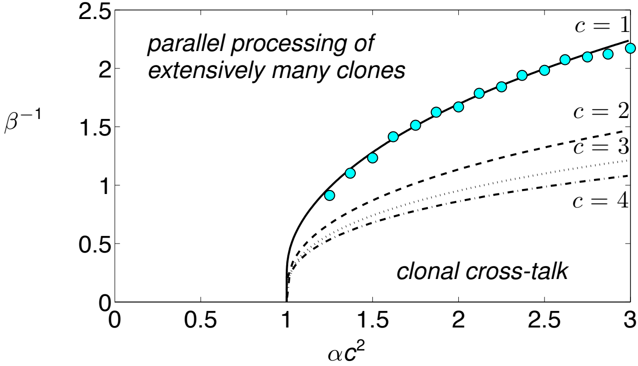
$\alpha c^2 = 1$



$\alpha c^2 > 1$

solve model as a statistical mechanics one
(i.e. calculate asymptotic disorder-averaged free energy)

after calculation (finite connectivity replica analysis):
exact formula for clonal cross-talk transition lines

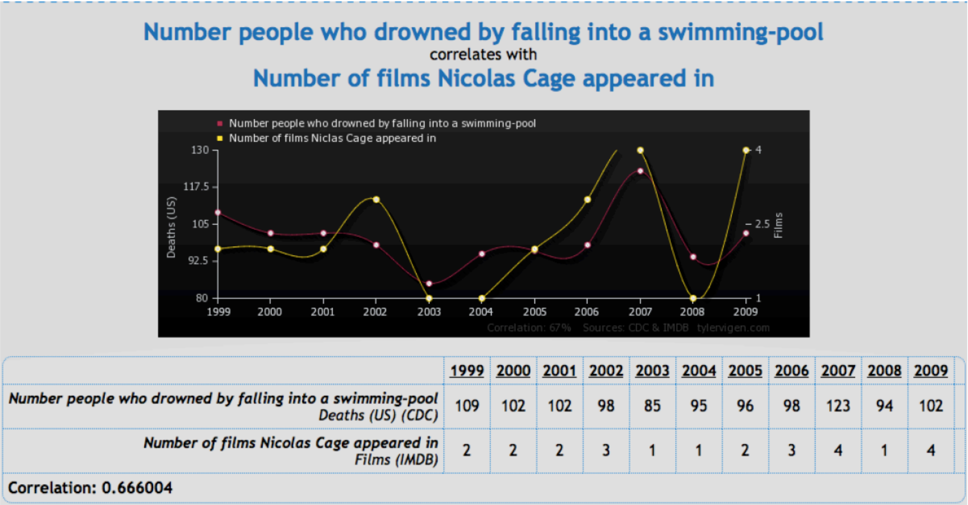


$\alpha$: $n_B/n_T$
$c$: $T$-cell promiscuity
$\beta^{-1}$: noise in clonal dynamics

1. Quality of raw data
   - Bayesian analysis of imaging data
   - Proteome data decontamination

2. Complex signalling processes
   - Many-variable systems in biology
   - Signalling in the proteome
   - Cytokine signalling in adaptive immune system

3. Risk associations and outcome prediction
   - Overfitting in clinical outcome prediction
   - Bayesian latent variable analysis
   - Prediction from high dimensional covariates
   - Heterogeneity and competing risks



## Number people who drowned by falling into a swimming-pool

correlates with

## Number of films Nicolas Cage appeared in

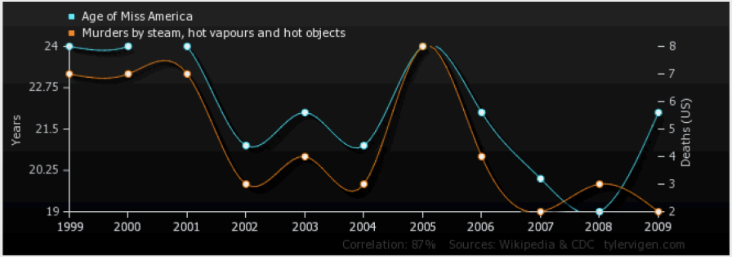
| | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Number people who drowned by falling into a swimming-pool*<br>*Deaths (US) (CDC)* | 109 | 102 | 102 | 98 | 85 | 95 | 96 | 98 | 123 | 94 | 102 |
| *Number of films Nicolas Cage appeared in*<br>*Films (IMDB)* | 2 | 2 | 2 | 3 | 1 | 1 | 2 | 3 | 4 | 1 | 4 |

**Correlation: 0.666004**

# Age of Miss America

correlates with

# Murders by steam, hot vapours and hot objects



| | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Age of Miss America***<br>*Years (Wikipedia)* | 24 | 24 | 24 | 21 | 22 | 21 | 24 | 22 | 20 | 19 | 22 |
| ***Murders by steam, hot vapours and hot objects***<br>*Deaths (US) (CDC)* | 7 | 7 | 7 | 3 | 4 | 3 | 8 | 4 | 2 | 3 | 2 |

**Correlation: 0.870127**

# Tools to combat overfitting
in covariate-to-outcome analysis

- **Pin down the problem**

  predict ‘safe’ ratio covariates/sample
  for Cox regression?

- **Eliminate redundant information**

  improve covariates/samples ratio
  latent vars (information theory), find ‘true’ dimension

- **Model (avoid?) overfitting effects**

  handle statistics of full parameter uncertainty,
  while keeping computations feasible

## Tools to combat overfitting

in covariate-to-outcome analysis

- **Pin down the problem**

  predict ‘safe’ ratio covariates/samples
  for Cox regression?

- **Eliminate redundant information**

  improve covariates/samples ratio
  latent vars (information theory), find ‘true’ dimension

- **Model (avoid?) overfitting effects**

  handle statistics of full parameter uncertainty,
  while keeping computations feasible

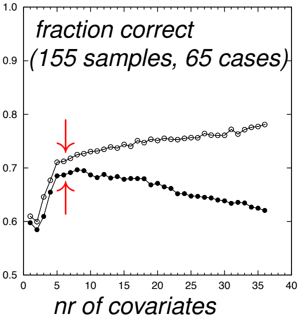
all based on
**Bayesian principles**

overfitting in

## Proportional hazards regression

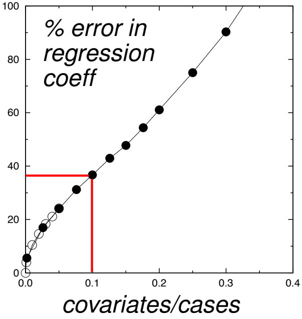
associations between covariates and risk
for time-to-event outcome data,
multivariate version for outcome prediction

p-values, confidence intervals
<u>don't</u> measure overfitting!



rule of thumb:
'10 samples per case'
too optimistic ...

uncorrelated covariates
o: *1000 samples & cases*
●: *500 samples & cases*



*developing analytical theory,*
*that predicts onset of overftting*
*in terms of statistics of covariates*
*and nr of samples and cases*

# Bayesian latent variable methods

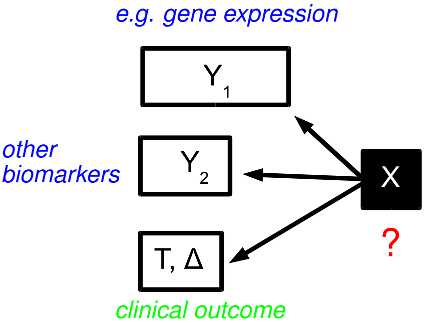
for survival analysis

Assume:

(a) data $Y_k \in \mathbb{R}^d$ are *high-dim windows* on *low dim* latent variables $X \in \mathbb{R}^q$

(b) $X$ actually drives outcome

(c) $q < d$

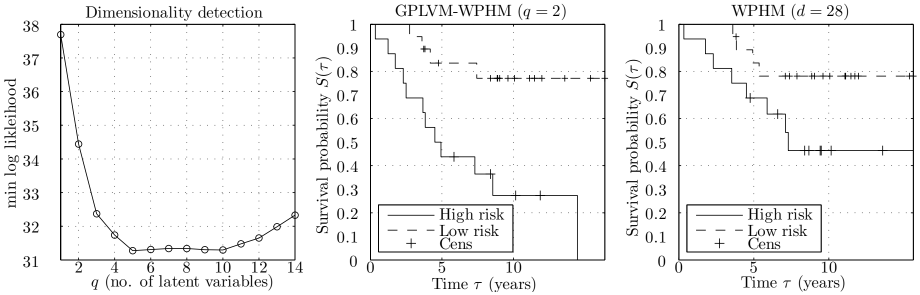


- nonlinear stochastic relations $Y_k = f_k(X) + noise$
- dimension detection: optimal $q$?
- find most probable latent variables $X$
- use $X$ to predict clinical outcome

*Gaussian process latent variable model (GPLVM) combined with Weibull proportional hazards model (WPHM)*

## Results from METABRIC gene signature data

*data $Y$: scores of 28 gene signatures*
*outcome: overall survival time*



left: $q \leq 5$, dimension of $X$ (predicted from training set, $n = 74$)

middle: predicted low/high risk groups, $q = 2$
(tested in validation set, $n = 74$)

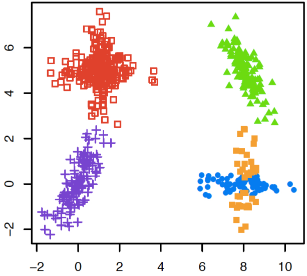
right: predicted low/high risk groups from $Y$
(tested in validation set, $n = 74$)

# Discriminant analysis

data: $\mathcal{D}=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_N,y_N)\}$

$\mathbf{x}_i$: covariates
$y_i$: class labels

goal:
class $y$ of new observation $\mathbf{x}$



**model based approaches**

parametrise $p(\mathbf{x}|y,\boldsymbol{\theta})$,
estimate $\boldsymbol{\theta}$ from data,
then use:

$$p(y|\mathbf{x},\boldsymbol{\theta})=\frac{p(\mathbf{x}|y,\boldsymbol{\theta})p(y)}{\sum_{y'}p(\mathbf{x}|y',\boldsymbol{\theta})p(y')}$$

popular method:

**mclustDA** (Fraley & Raftery)
MAP estimation of $\boldsymbol{\theta}$

high dim data, $d \sim 10^3, 10^4$:
optimise $\sim 10^3, 10^8$ pars ...

*serious overfitting,*
*CPU demands prohibitive*

# Bayesian multi-class outcome prediction

for high-dimensional data

1. in view of overfitting:
   *full Bayesian* parameter estimation,
   instead of MAP (e.g. mclustDA)

$$\textit{MAP}: \qquad p(y|\mathbf{x},\mathcal{D}) = p(y|\mathbf{x},\boldsymbol{\theta}_{\rm MAP}), \qquad \boldsymbol{\theta}_{\rm MAP} = \textit{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\mathcal{D})$$

$$\textit{Bayes}: \qquad p(y|\mathbf{x},\mathcal{D}) = \int \mathrm{d}\boldsymbol{\theta}\ p(y|\mathbf{x},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})$$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{\theta}'\ p(\boldsymbol{\theta}')p(\mathcal{D}|\boldsymbol{\theta}')}$$

2. computational feasibility:
   evaluate *d*-dimensional integrals *analytically*

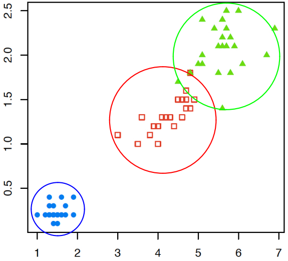
3. desirable:
   determine MAP-optimal hyper-pars *analytically*

## simplest model

Gaussian
covariate
distribution
for each class

$$p(\mathbf{x}|y,\boldsymbol{\theta}) = \frac{\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_y)^2/\alpha_y^2}}{(\alpha_y\sqrt{2\pi})^d}$$

$\boldsymbol{\mu}_y$ : *class signatures,*
*with Gaussian priors*



*generative*

all data assumed
informative

$$p(\mathbf{x},\mathbf{x}_1,\ldots,\mathbf{x}_n,y,y_1,\ldots,y_n|\boldsymbol{\theta}) = p(\mathbf{x},y|\boldsymbol{\theta})\prod_{i=1}^{n} p(\mathbf{x}_i,y_i|\boldsymbol{\theta})$$

*discriminative*

extract only link
between **x** and *y*

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n,y|\mathbf{x},y_1,\ldots,y_n,\boldsymbol{\theta}) = p(y|\mathbf{x},\boldsymbol{\theta})\prod_{i=1}^{n} p(\mathbf{x}_i|y_i,\boldsymbol{\theta})$$

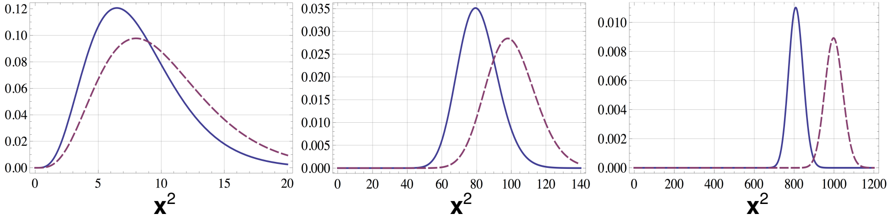
1. *full Bayesian* parameter estimation: ✓
2. evaluate *d*-dimensional integrals *analytically*: ✓
3. determine optimal hyper-pars *analytically*: ✓

## Signature- versus variability-based classification

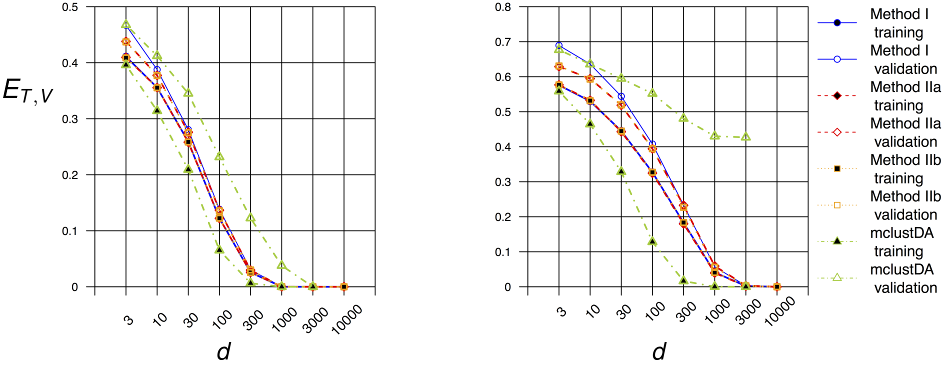
weak class ‘signatures’ in data:

classification still possible,
but will become variability-based:
(increasingly effective for large $d$)

$p(\mathbf{x}^2|y)$



$d = 10$ $\qquad$ $d = 100$ $\qquad$ $d = 1000$



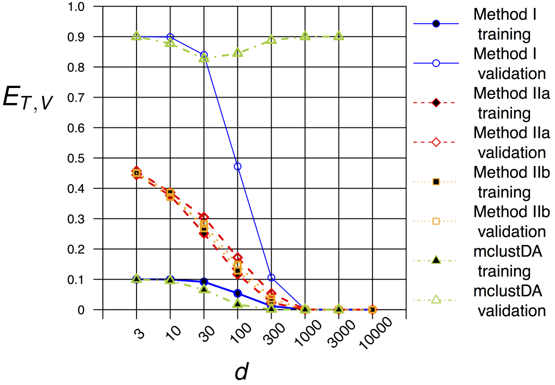
*LOOCV error curves, averaged over 100 data sets,*
*$n = 100$ samples with identical class centres*

*Left:*

| $f_1$ | $f_2$ | $\alpha_1$ | $\alpha_2$ |
|---|---|---|---|
| 0.5 | 0.5 | 0.24 | 0.28 |

*Right:*

| $f_1$ | $f_2$ | $f_3$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |
|---|---|---|---|---|---|
| 0.33 | 0.33 | 0.34 | 0.24 | 0.26 | 0.28 |



*Error curves (100 training/100 validation), averaged over 100 data sets,* $n\!=\!100$ *samples with identical class centres*

| | $f_1$ | $f_2$ | $\alpha_1$ | $\alpha_2$ |
|---|---|---|---|---|
| $T$ | 0.1 | 0.9 | 0.24 | 0.28 |
| $V$ | 0.9 | 0.1 | 0.24 | 0.28 |

*mclustDA* and method I struggle when training and validation sets differ in class membership balance

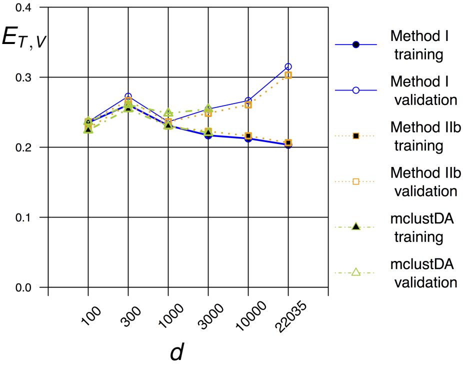
# Triple-negative breast cancer

prediction of survival
from gene expression

$y = 1$: BC death within 5 yrs
$y = 2$: survived for at least 5 yrs

$n = 165$, $d = 22{,}035$
$(f_1, f_2) = (0.25, 0.75)$

performance measured via LOOCV,
genes ranked by correlation with outcome



- all methods give similar results
- Bayesian methods can go to much larger $d$
- min $E_V \approx 0.24$ ($\sim$ going for largest class)

*either gene expression data confer no predictive information on 5 yr TNBC survival, or all methods suffer from model mismatch*

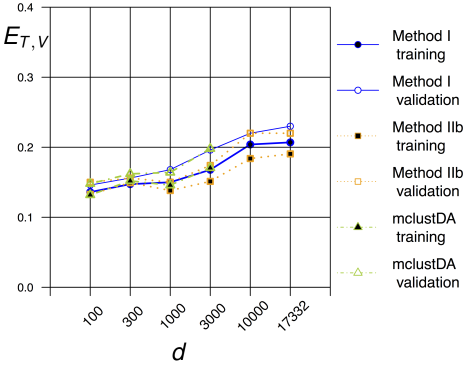
# TCGA Breast cancer data

prediction of receptor status

$y = 1$: ER-negative, HER2-negative
$y = 2$: ER-positive, HER2-negative
$y = 3$: ER-negative, HER2-positive
$y = 4$: ER-positive, HER2-positive

$n = 500$, $d = 17{,}332$
$(f_1, f_2, f_3, f_4) = (0.19, 0.66, 0.04, 0.11)$

performance measured via LOOCV,
genes ranked by correlation with outcome



- optimal predictive information in first 100 ranked genes
- Bayesian methods can go to much larger $d$
- min $E_V \approx 0.14$ (significant)

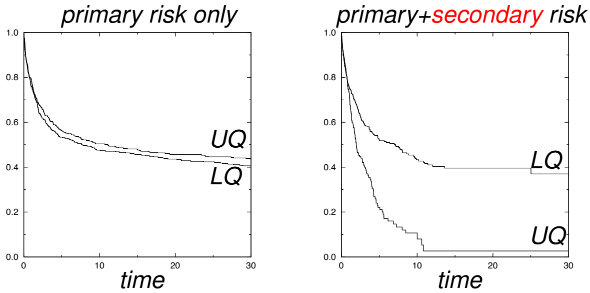
*gene expression profiles of breast cancer patients are reliable predictors of their ER and HER2 status*

## conventional methods

*Kaplan-Meier estimators*
*Cox regression*
*.....*

- cannot handle disease/host heterogeneity beyond variability in covariates
- assume different risks are uncorrelated
- dangerous when many censoring events ...

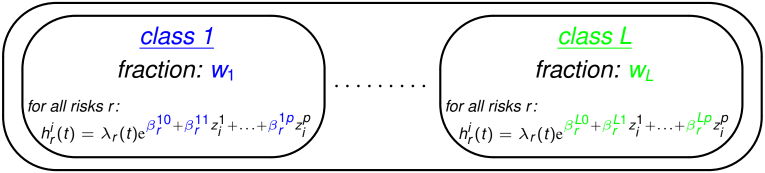
predicted survival probabilities can be badly wrong ...



## More advanced methods

- model all risks and their relations, at individual and cohort level
- event times assumed uncorrelated only at the level of *individuals*
- individuals with same covariates may have *distinct* risk profiles
- Bayesian analysis, so reliable error bars

Latent class heterogeneity:



*class 1*

*fraction:* $w_1$

*for all risks* $r$*:*

$h_r^i(t) = \lambda_r(t)e^{\beta_r^{10}+\beta_r^{11}z_i^1+\ldots+\beta_r^{1p}z_i^p}$

. . . . . . . . . .

*class L*

*fraction:* $w_L$

*for all risks* $r$*:*

$h_r^i(t) = \lambda_r(t)e^{\beta_r^{L0}+\beta_r^{L1}z_i^1+\ldots+\beta_r^{Lp}z_i^p}$

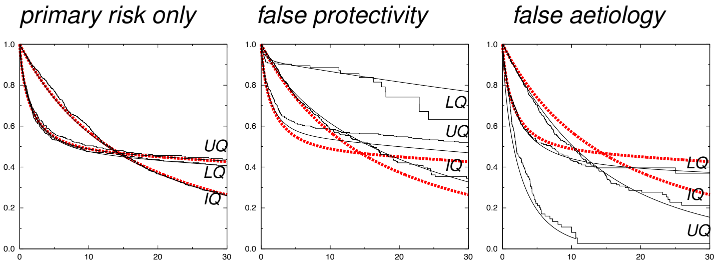
*prop hazards within sub classes* $\not\Longrightarrow$ *prop hazards at cohort level!*

*can account for:*

*association heterogeneity, non-proportional hazards, covariate interactions, competing risks, ...*

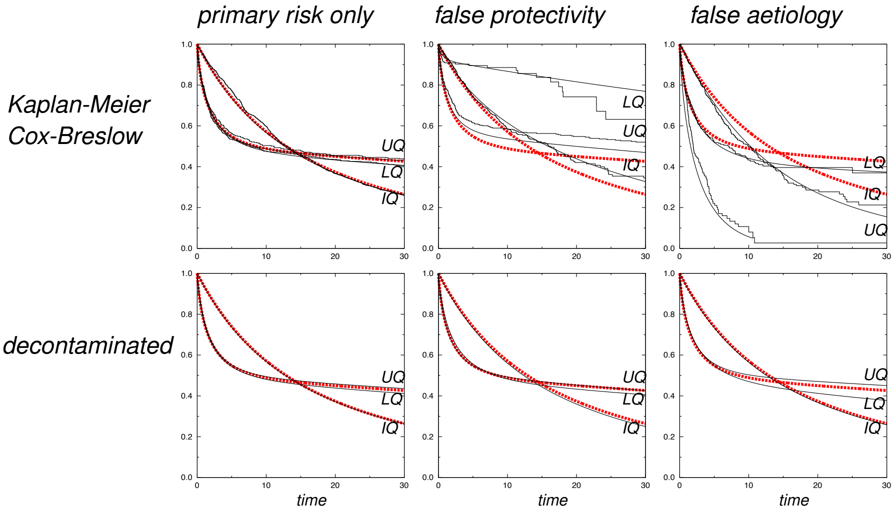
# synthetic data

*Kaplan-Meier*
*Cox-Breslow*



*red dashed: true survival curves*

# synthetic data



*red dashed: true survival curves*

## Bayesian retrospective class identification

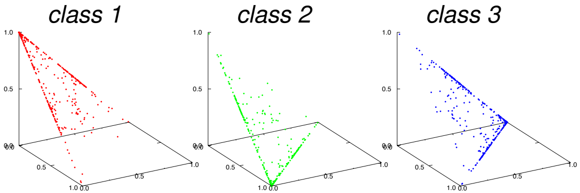
$$P(\ell|t,r,\mathbf{z}) \;=\; \frac{w_\ell\; \mathrm{e}^{\hat{\boldsymbol{\beta}}_r^{\ell}\cdot\mathbf{z}-\sum_{r'=1}^{R}\exp(\hat{\boldsymbol{\beta}}_{r'}^{\ell}\cdot\mathbf{z})\int_0^t \mathrm{d}s\;\hat{\lambda}_{r'}(s)}}{\sum_{\ell'=1}^{L} w_{\ell'}\; \mathrm{e}^{\hat{\boldsymbol{\beta}}_r^{\ell'}\cdot\mathbf{z}-\sum_{r'=1}^{R}\exp(\hat{\boldsymbol{\beta}}_{r'}^{\ell'}\cdot\mathbf{z})\int_0^t \mathrm{d}s\;\hat{\lambda}_{r'}(s)}}$$

Data:

3 classes,
$w_1 = w_2 = w_3 = \frac{1}{3}$
2 competing risks

$$\beta_1^1 = (0.5, 0.5, 0.5) + (2, 0, 2)$$
$$\beta_1^2 = (0.5, 0.5, 0.5) + (-2, -2, 0)$$
$$\beta_1^3 = (0.5, 0.5, 0.5) + (0, 2, -2)$$

each individual $i$:
point $(p_1^i, p_2^i, p_3^i)$ in $\mathbb{R}^3$
$p_\ell^i = P(\ell|t_i, r_i, \mathbf{z}_i)$


class 1

class 2

class 3

# Prostate cancer study on the ULSAM data set

|  |  |
|---|---|
| $N = 2047$ | |
| primary events: | 208 |
| death (non-PC ): | 910 |
| end of trial: | 929 |

hazard rates:
$HR_j = \mathrm{e}^{2\beta_j}$

|  | CLASSES | PRIMARY RISK |  |  |  |  | SECONDARY RISK |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | *BMI* | *selen* | *phys1* | *phys2* | *smok* | *BMI* | *selen* | *phys1* | *phys2* | *smok* |
| *Cox* |  | 0.14 | -0.15 | 0.20 | -0.09 | -0.08 |  |  |  |  |  |
| *new* | $w_1 = 0.51$ | 1.22 | -0.41 | 0.73 | -0.01 | 1.43 | 0.82 | -0.42 | -0.31 | -0.14 | 1.35 |
|  | $w_2 = 0.49$ | -0.07 | -0.16 | 0.19 | -0.10 | -0.27 | 0.10 | -0.07 | -0.07 | 0.04 | 0.18 |
|  | frailties: | $\beta^1_{10} - \beta^2_{10} = -4.61$ (HR 0.010) |  |  |  |  | $\beta^1_{20} - \beta^2_{20} = -4.06$ (HR 0.017) |  |  |  |  |

healthy group: strong effects of covariates,
BMI and smoking important risk factors

frail group: weak effects of covariates,
BMI and smoking weakly protective
(reverse causal effect?)

## Breast cancer study (AMORIS data base)

potential of serum lipids, measured prior to diagnosis, to predict risk of BC death

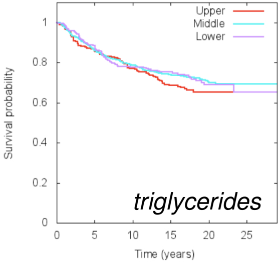
$N = 1798$, all BC diagnosed

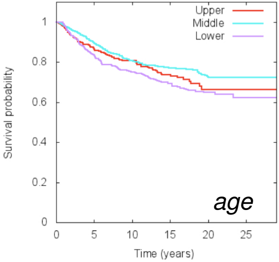
| | |
|---|---|
| primary events (BC death): | 259 |
| secondary events (CV death): | 179 |
| tertiary events (other death): | 423 |
| censoring: | 937 |

covariates:

triglycerides, cholesterol, glucose
age, 3 socio-economic variables

- Cox regression:
  no significant assoc
- risk-specific KM curves:
  no proportional hazards
  in primary risk
  (Cox invalid ...)
- KM curves themselves
  unreliable (competing
  risks 2 and 3?)



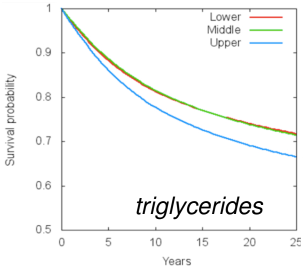


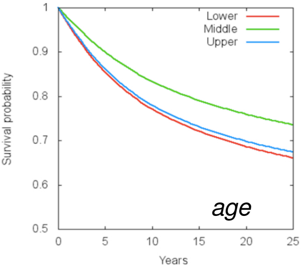
**heterogeneous model**
predicts three classes,
explains non-monotonic relations

- class 1, 57%:
  *triglycerides HR*$>$1
  *age HR*$>$1

  class 2, 37%:
  *age HR*$<$ 1

  class 3, 6%:
  *no significant assoc*

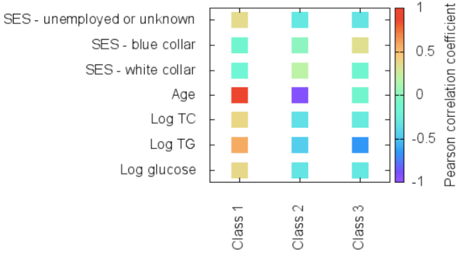


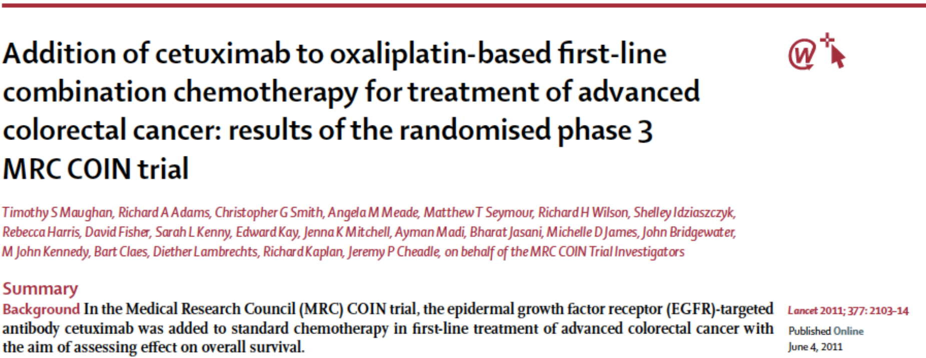


- correlations of class membership probabilities with covariates:

  Class 1, <u>older women</u>:
  *triglycerides HR*$>$1*, age HR*$>$1

  Class 2, <u>younger women</u>:
  *age HR*$<$1





# Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial

*Timothy S Maughan, Richard A Adams, Christopher G Smith, Angela M Meade, Matthew T Seymour, Richard H Wilson, Shelley Idziaszczyk, Rebecca Harris, David Fisher, Sarah L Kenny, Edward Kay, Jenna K Mitchell, Ayman Madi, Bharat Jasani, Michelle D James, John Bridgewater, M John Kennedy, Bart Claes, Diether Lambrechts, Richard Kaplan, Jeremy P Cheadle, on behalf of the MRC COIN Trial Investigators*

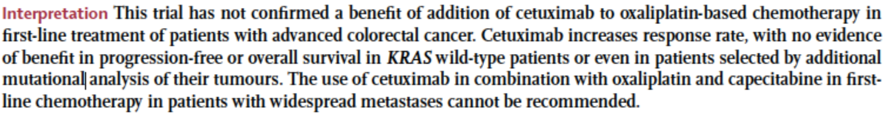
## Summary

**Background** In the Medical Research Council (MRC) COIN trial, the epidermal growth factor receptor (EGFR)-targeted antibody cetuximab was added to standard chemotherapy in first-line treatment of advanced colorectal cancer with the aim of assessing effect on overall survival.

*Lancet* 2011; 377: 2103–14

Published Online
June 4, 2011

outcome:



**Interpretation** This trial has not confirmed a benefit of addition of cetuximab to oxaliplatin-based chemotherapy in first-line treatment of patients with advanced colorectal cancer. Cetuximab increases response rate, with no evidence of benefit in progression-free or overall survival in *KRAS* wild-type patients or even in patients selected by additional mutational analysis of their tumours. The use of cetuximab in combination with oxaliplatin and capecitabine in first-line chemotherapy in patients with widespread metastases cannot be recommended.

# Bayesian latent class analysis of COIN data

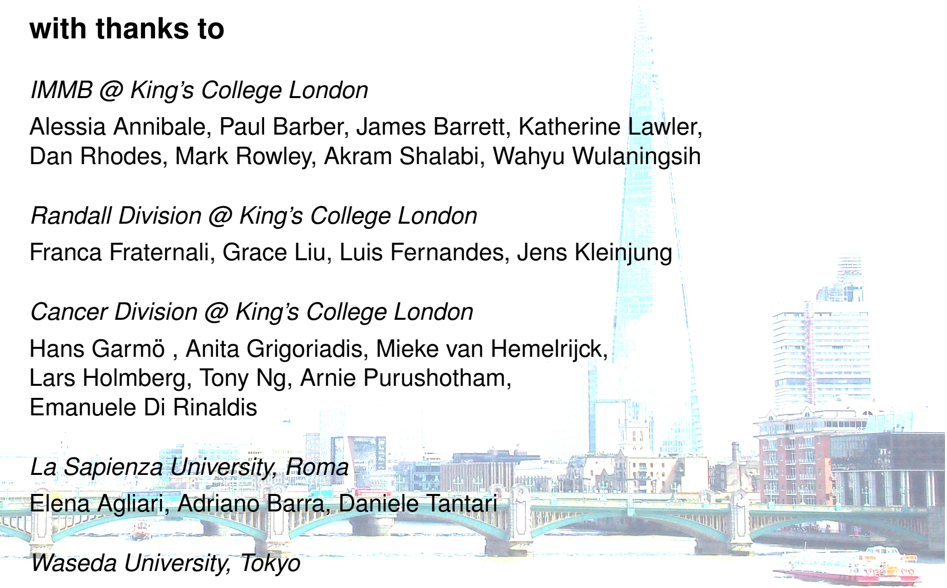
hazard ratios:

| | *FRET* | *Her3* | *Her2-Her3* | *Her2* | *Cetuximab* | *KRAS mut* |
|---|---|---|---|---|---|---|
| *Cox* | 0.5 | 1.0 | 1.8 | 1.1 | 0.7 | 1.7 |
| *new model:* | | | | | | |
| *class I, 40%* | 0.7 | 1.5 | 3.7 | 1.1 | 0.3 | 2.5 |
| *class II, 60%* | 0.6 | 1.2 | 0.7 | 0.9 | 1.1 | 1.4 |

*higher overall risk in class II*

- two sub-cohorts, with similar base hazard rates, but distinct overall frailties and associations.
- methods provides retrospective class assignment
- new tools to identify *a priori* the responders to Cetuximab?

## with thanks to



*IMMB @ King’s College London*

Alessia Annibale, Paul Barber, James Barrett, Katherine Lawler, Dan Rhodes, Mark Rowley, Akram Shalabi, Wahyu Wulaningsih

*Randall Division @ King’s College London*

Franca Fraternali, Grace Liu, Luis Fernandes, Jens Kleinjung

*Cancer Division @ King’s College London*

Hans Garmö , Anita Grigoriadis, Mieke van Hemelrijck,
Lars Holmberg, Tony Ng, Arnie Purushotham,
Emanuele Di Rinaldis

*La Sapienza University, Roma*

Elena Agliari, Adriano Barra, Daniele Tantari

*Waseda University, Tokyo*

Masato Inoue