Bayesian clinical classification from high-dimensional data:

signatures versus variability

ACC Coolen

King's College London

1. Discriminant analysis in high dimensional spaces
2. Bayesian multi-class outcome prediction
3. Application to synthetic data
4. Application to cancer data
5. Summary

1 Discriminant analysis in high dimensional spaces

2 Bayesian multi-class outcome prediction

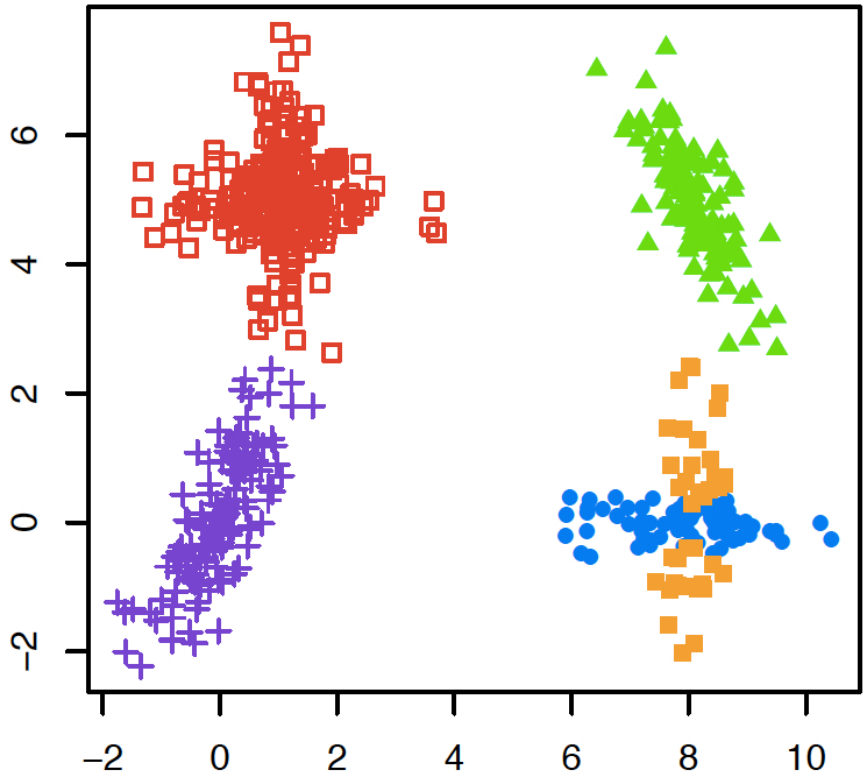
3 Application to synthetic data

4 Application to cancer data

5 Summary

Discriminant analysis

in high-dimensional spaces



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

$\mathbf{x}_i \in \mathbb{R}^d$: observations
$y_i \in \{1, \ldots, c\}$: class labels

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

model based approaches

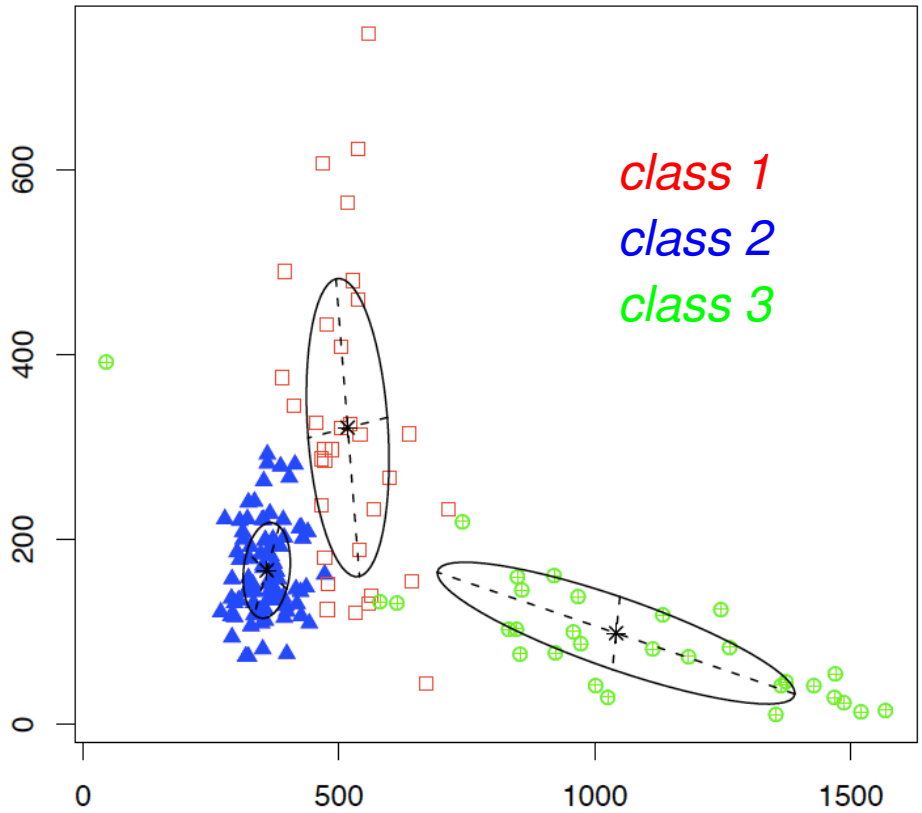
parametrisation $p(\mathbf{x}|y, \theta)$,

estimate θ from $\mathcal{D}$,
all $(\mathbf{x}_i, y_i)$ assumed indep

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

Fraley and Raftery – mclustDA
(1998, 2002, 2012)

- Gaussian mixtures for $p(\mathbf{x}|y,\theta)$ of each y
- parameters θ: EM algorithm
- model selection (mixture): BIC



high dim data $(d \sim 10^4)$

serious overfitting issues ...
CPU demands prohibitive ...
(impossible for $d > 3000$)

the problem:
optimising $\mathcal{O}(d^2)$ model parameters!
(moments of Gaussian in d dim)

1 Discriminant analysis in high dimensional spaces

2 Bayesian multi-class outcome prediction

3 Application to synthetic data

4 Application to cancer data

5 Summary

Bayesian multi-class outcome prediction

in high-dimensional spaces

1. in view of overfitting:

   full Bayesian parameter estimation,
   instead of MAP (e.g. mclustDA)

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

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

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

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

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

I: generative classification

all data in $\mathcal{D}$ assumed informative,

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

$$p(y|\mathbf{x}, \mathcal{D}) = \frac{\pi_y \int \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{\theta}) p(\mathbf{x}|y, \boldsymbol{\theta}) \prod_{i=1}^{n} p(\mathbf{x}_i|y_i, \boldsymbol{\theta})}{\sum_{y'=1}^{c} \pi_{y'} \int \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{\theta}) p(\mathbf{x}|y', \boldsymbol{\theta}) \prod_{i=1}^{n} p(\mathbf{x}_i|y_i, \boldsymbol{\theta})}$$

II: discriminative classification

extract from $\mathcal{D}$ only link between $\mathbf{x}$ and y,
class labels $\{y_1, \ldots, y_n\}$ non-informative

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

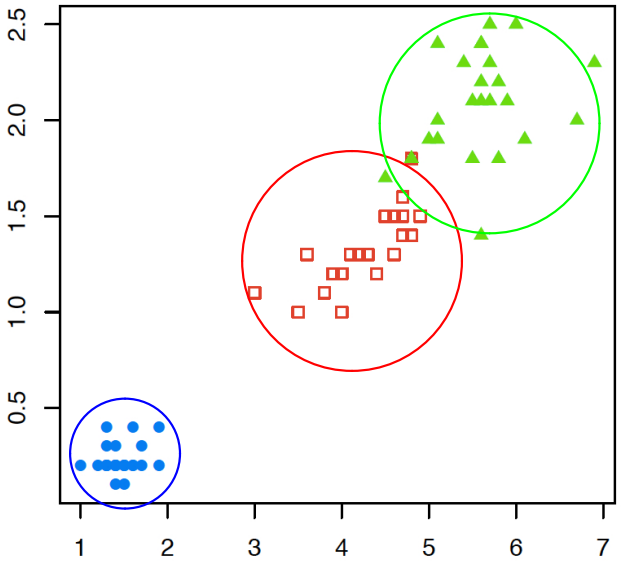
$$p(y|\mathbf{x}, \mathcal{D}) = \frac{\int \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \left(\frac{\pi_y p(\mathbf{x}|y, \boldsymbol{\theta})}{\sum_{y'=1}^{c} \pi_{y'} p(\mathbf{x}|y', \boldsymbol{\theta})} \right) \prod_{i=1}^{n} p(\mathbf{x}_i|y_i, \boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(\mathbf{x}_i|y_i, \boldsymbol{\theta})}$$

simplest model

- Gaussian class-dep covariate distributions

$$p(\mathbf{x}, y|\boldsymbol{\theta}) = p(\mathbf{x}|y, \boldsymbol{\theta})\ \pi_y$$

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



- $\boldsymbol{\mu}_y \in \mathbb{R}^d$: *class signatures* with priors

$$p(\boldsymbol{\mu}_y|\beta_y) = \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{\mu}_y^2/\beta_y^2}}{(\beta_y\sqrt{2\pi})^d}$$

- parameters:

$$cd \text{ micro pars}: \quad \boldsymbol{\theta} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c\}, \qquad \boldsymbol{\mu}_y \in \mathbb{R}^d$$

$$\text{hyperpars}: \quad H = \{(\alpha_1, \beta_1, \pi_1), \ldots, (\alpha_c, \beta_c, \pi_c)\}, \qquad \sum_{y \le c} \pi_y = 1$$

generative
(Gaussian $\boldsymbol{\theta}$-integrals)

$$p(y|\mathbf{x},\mathcal{D},H) = \frac{(\pi_y/S_y^d)\,\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_y)^2/S_y^2}}{\sum_{z=1}^{c}(\pi_z/S_z^d)\,\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_z)^2/S_z^2}}$$

with

$$\mathbf{m}_y = \langle\mathbf{x}\rangle_y\,\frac{nf_y\beta_y^2}{nf_y\beta_y^2+\alpha_y^2}, \qquad S_y^2 = \alpha_y^2\,\frac{\alpha_y^2+(nf_y+1)\beta_y^2}{\alpha_y^2+nf_y\beta_y^2}, \qquad f_y = \frac{1}{n}\sum_{i=1}^{n}\delta_{yy_i}$$

discriminative
(non-Gaussian $\boldsymbol{\theta}$-integrals)

$$p(y|\mathbf{x},\mathcal{D},H) =$$

$$\int\Big(\prod_{z=1}^{c}\frac{\mathrm{d}u_z\mathrm{d}v_z}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}u_z^2}\,\mathcal{P}(v_z)\Big)\left\{\frac{(\pi_y/\alpha_y^d)\,\mathrm{e}^{-\frac{1}{2\alpha_y^2}\left(|\mathbf{x}-\mathbf{m}_y|-\frac{\alpha_y\beta_y u_y}{\sqrt{\alpha_y^2+nf_y\beta_y^2}}\right)^2-\frac{1}{2}\frac{\beta_y^2 v_y}{\alpha_y^2+nf_y\beta_y^2}}}{\sum_{z=1}^{c}(\pi_z/\alpha_z^d)\,\mathrm{e}^{-\frac{1}{2\alpha_z^2}\left(|\mathbf{x}-\mathbf{m}_z|-\frac{\alpha_z\beta_z u_z}{\sqrt{\alpha_z^2+nf_z\beta_z^2}}\right)^2-\frac{1}{2}\frac{\beta_z^2 v_z}{\alpha_z^2+nf_z\beta_z^2}}}\right\}$$

$$\mathcal{P}(v) = \frac{(\frac{1}{2}v)^{\frac{1}{2}(d-3)}\mathrm{e}^{-\frac{1}{2}v}}{2\Gamma(\frac{1}{2}(d-1))} \qquad (\textit{chi sqr dist},\ d-1\ \textit{d.o.f.})$$

MAP hyperparameter determination

non-informative hyper-priors,
$\hat{H} = \text{argmax}_H \log p(\mathcal{D}|H)$

generative model

$$\hat{H} = \text{argmax}_{\{\alpha_y,\beta_y,\pi_y\}} \sum_{z=1}^{c} \Big\{ \frac{nf_z}{d} \log \pi_z - (nf_z - 1) \log \alpha_z - \frac{1}{2} \log(\alpha_z^2 + nf_z\beta_z^2) - \frac{1}{2} nf_z \Big(\frac{\Sigma_z^2}{\alpha_z^2} + \frac{X_z^2}{\alpha_z^2 + nf_z\beta_z^2} \Big) \Big\}$$

$$X_y^2 = \frac{1}{d}\langle \mathbf{x} \rangle_y^2, \qquad \Sigma_y^2 = \frac{1}{d}(\langle \mathbf{x}^2 \rangle_y - \langle \mathbf{x} \rangle_y^2)$$

discriminative model

$$\hat{H} = \text{argmax}_{\{\alpha_y,\beta_y,\pi_y\}} \sum_{z=1}^{c} \Big\{ - (nf_z - 1) \log \alpha_z - \frac{1}{2} \log(\alpha_z^2 + nf_z\beta_z^2) - \frac{1}{2} nf_z \Big(\frac{\Sigma_z^2}{\alpha_z^2} + \frac{X_z^2}{\alpha_z^2 + nf_z\beta_z^2} \Big) \Big\}$$

$\hat{\alpha}_y$: width of $p(\mathbf{x}|y,\boldsymbol{\theta})$

$\hat{\beta}_y$: width of prior $p(\boldsymbol{\mu}_y)$

$\hat{\pi}_y$: class frequency

$$\forall y: \qquad \hat{\alpha}_y^2 = \Sigma_y^2 + X_y^2 - \hat{\beta}_y^2, \qquad \hat{\beta}_y^2 = \left(X_y^2 - \frac{\Sigma_y^2}{nf_y - 1}\right)\theta\left[X_y^2 - \frac{\Sigma_y^2}{nf_y - 1}\right]$$

$$\textit{generative}: \qquad \hat{\pi}_y = f_y$$

$$\textit{discriminative}: \quad \hat{\pi}_y = \textit{undetermined} \;\; (\textit{choose } c^{-1})$$

Weak signals:

$$X_y^2 < \Sigma_y^2/(nf_y - 1): \qquad \hat{\beta}_y = 0, \;\; \textit{no class } y \textit{ signature} \;\; (\textit{Occam's razor})$$

So:

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

Signature- versus variability-based classification

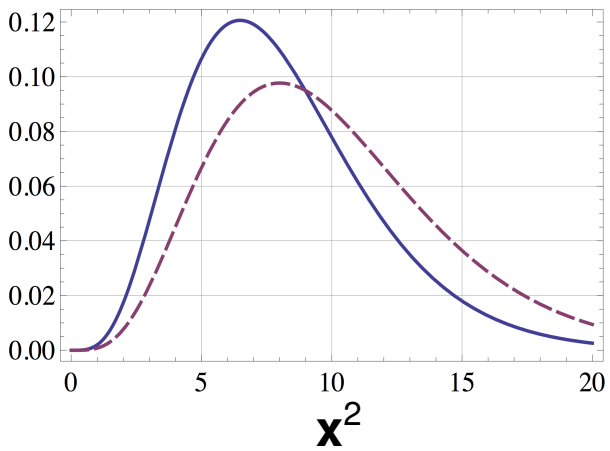
weak class 'signatures' in data:

$\hat{\beta}_y = 0 \;\rightarrow\; \hat{\mathbf{m}}_y = \mathbf{0}$

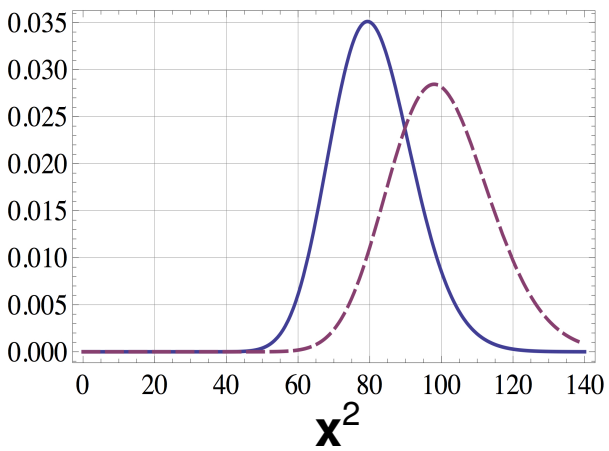
classification still possible,
but will become variability-based:
(increasingly effective for large d)

two classes with identical class averages

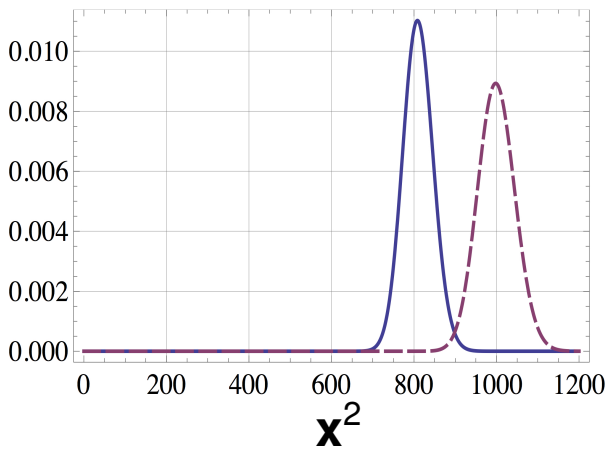
$p(\mathbf{x}^2|y)$



$d = 10$



$d = 100$



$d = 1000$

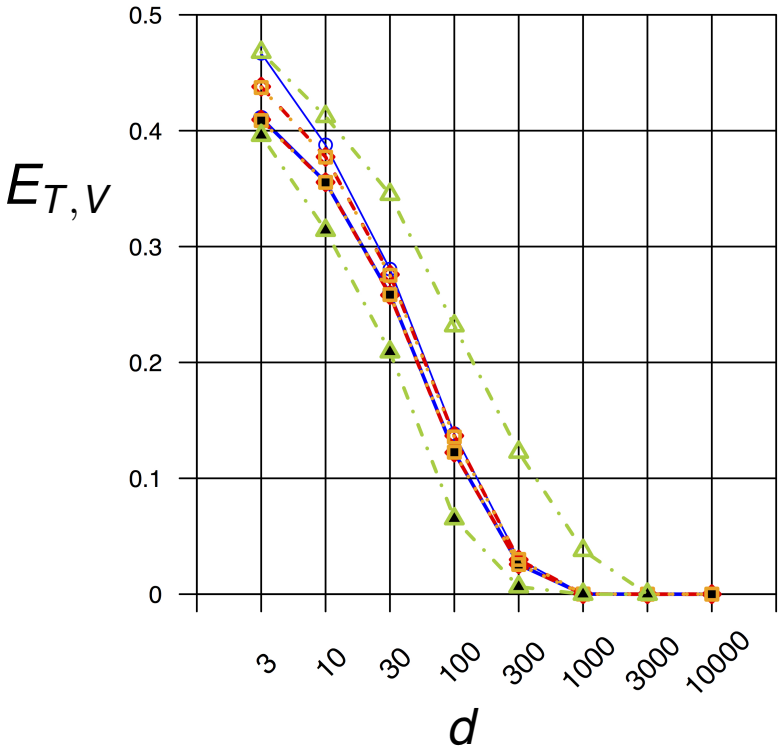
1 Discriminant analysis in high dimensional spaces

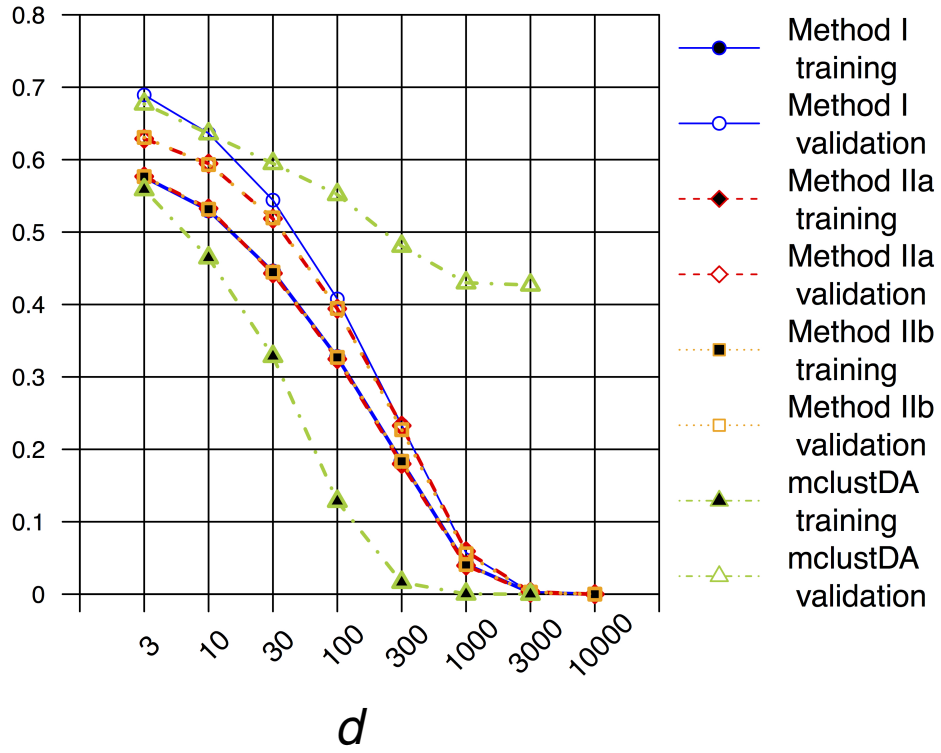
2 Bayesian multi-class outcome prediction

3 Application to synthetic data

4 Application to cancer data

5 Summary





Error curves (LOOCV), averaged over 100 data sets

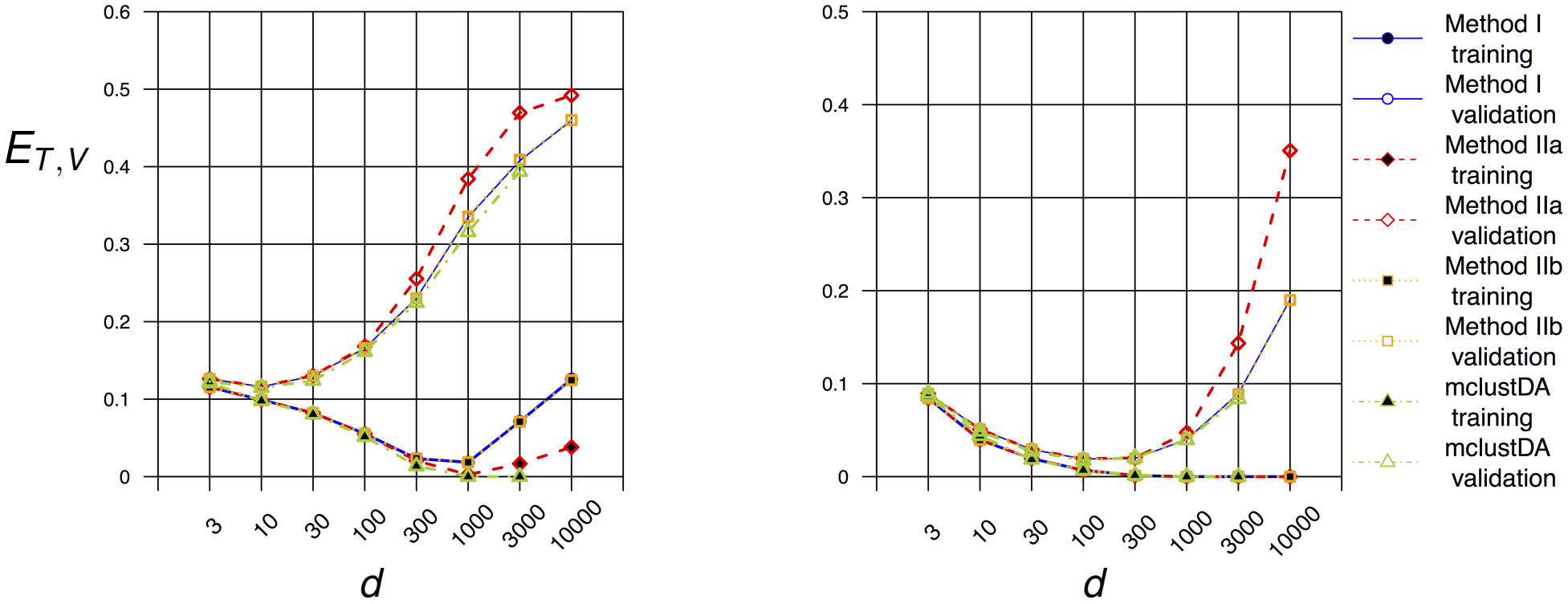
Left:

n	f_1	f_2	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	α_1	α_2
100	0.5	0.5	**0**	**0**	0.24	0.28

Right:

n	f_1	f_2	f_3	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$	α_1	α_2	α_3
100	0.33	0.33	0.34	**0**	**0**	**0**	0.24	0.26	0.28

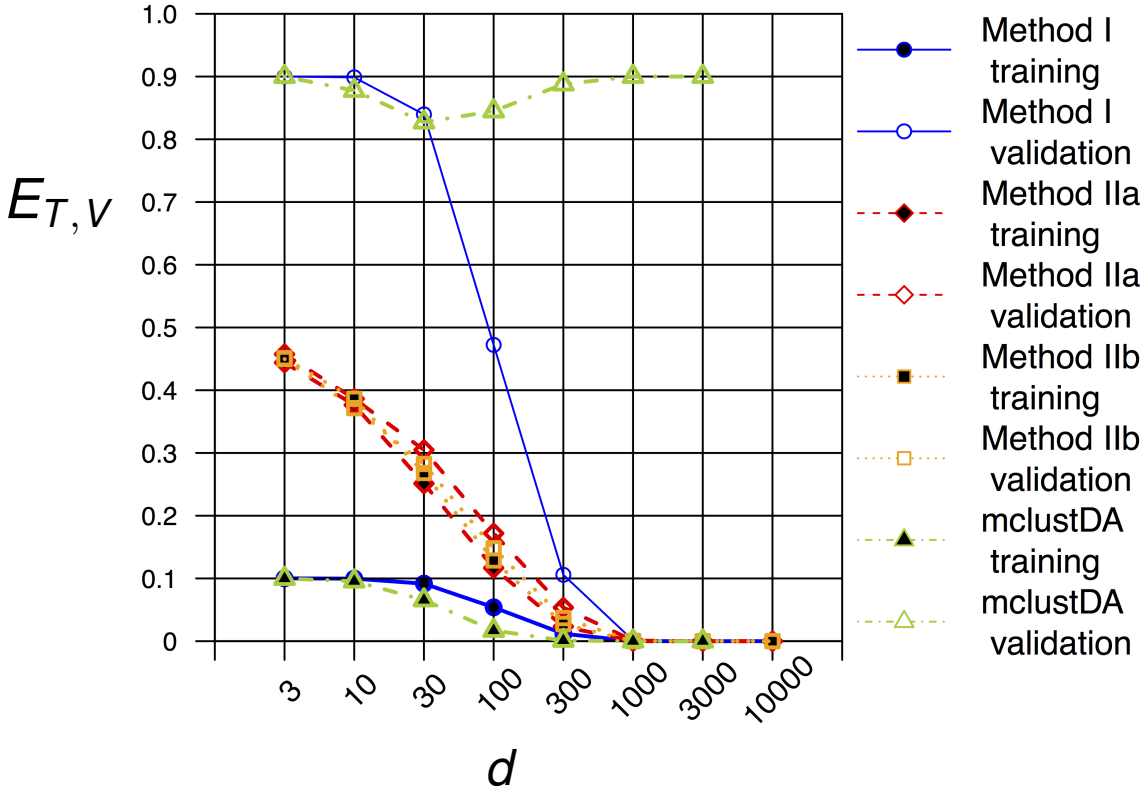
mclustDA struggles since class centres are identical



Error curves (100 training/100 validation), averaged over 100 data sets

	n	f_1	f_2	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	α_1	α_2
Left:	100	0.5	0.5	$(-1, -\frac{1}{2}, -\frac{1}{3}, \ldots, -\frac{1}{d})$	$(1, \frac{1}{2}, \ldots, \frac{1}{d})$	1	1
Right:	100	0.5	0.5	$(-1, -\frac{1}{\sqrt{2}}, \ldots, -\frac{1}{\sqrt{d}})$	$(1, \frac{1}{\sqrt{2}}, \cdots, \frac{1}{\sqrt{d}})$	1	1

left: $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|$ finite, right: diverges as $d \to \infty$



Error curves (100 training/100 validation), averaged over 100 data sets

	n	f_1	f_2	μ_1	μ_2	α_1	α_2
T	100	0.1	0.9	$(0,\ldots,0)$	$(0,\ldots,0)$	0.24	0.28
V	100	0.9	0.1	$(0,\ldots,0)$	$(0,\ldots,0)$	0.24	0.28

mclustDA and method I both struggle when training and validation sets differ in imbalance of class membership

1 Discriminant analysis in high dimensional spaces

2 Bayesian multi-class outcome prediction

3 Application to synthetic data

4 Application to cancer data

5 Summary

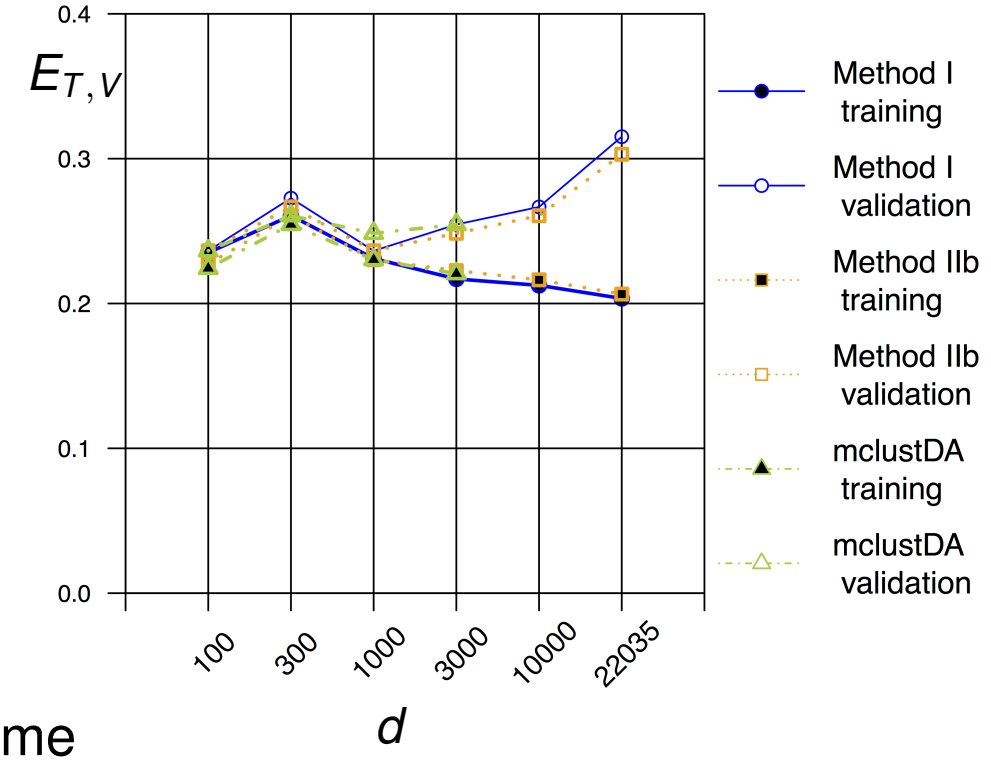
Triple-negative breast cancer

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

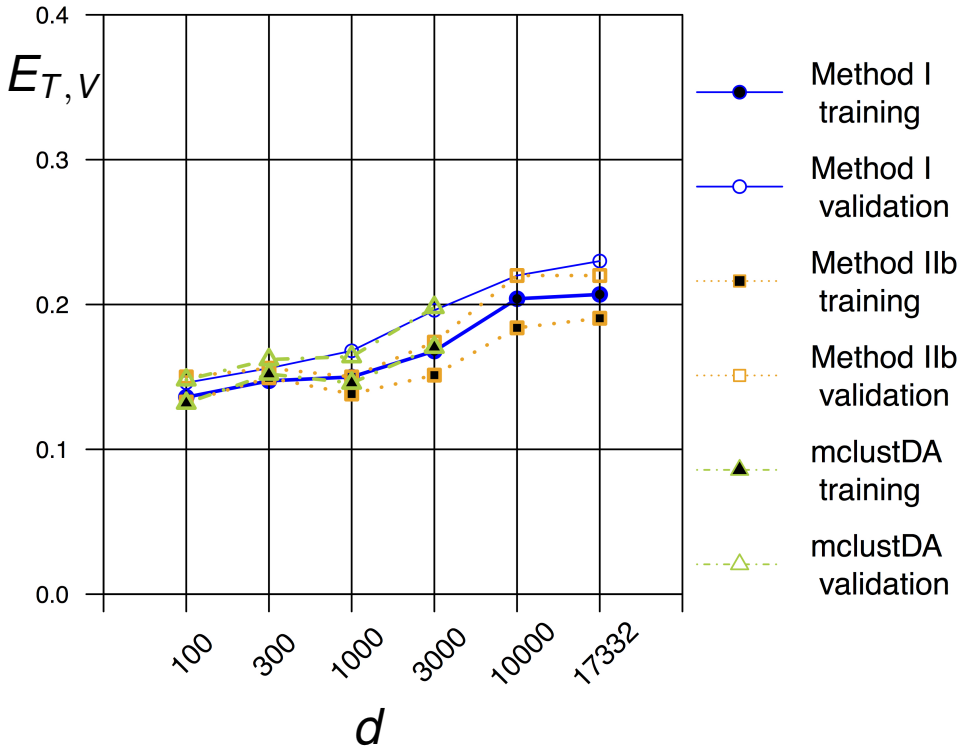
TCGA Breast cancer data

prediction of receptor status

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

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

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

1 Discriminant analysis in high dimensional spaces

2 Bayesian multi-class outcome prediction

3 Application to synthetic data

4 Application to cancer data

5 Summary

Summary

- By *solving integrals analytically*, full Bayesian discriminant analysis *fast* and feasible for *arbitrary covariate dimension d*.
- High *d*: full Bayesian method will increasingly use intra-class variability as opposed to class signatures.
- Synth data, small *d*: performance similar to *mclustDA*.
  Synth data, large *d*: lower E_V and less overfitting than *mclustDA*.
- Version II robust against mismatch in class sizes between training and validation sets.
- predict survival from gene expression in TNBC: X
  predict HER2 and ER status from gene expr in BC: all methods ✓

Future work:

(i) more complicated covariance matrices
(ii) Gaussian mixture models
(iii) $c \to \infty$, predict real-valued outcomes