Building Gene Expression Profiles via Multivariate Analyses

Manuela Zucknick and Sylvia Richardson

manuela.zucknick@imperial.ac.uk

Centre for Biostatistics

Imperial College London

Outline

- Introduction:
  - Usefulness of gene expression arrays for risk analysis in multifactorial diseases such as cancer
  - Statistical issues related to gene expression data
- Stability: How stable is a gene expression profile when the training data are varied (i.e. resampled)?
- Resampling study: Assess prediction accuracy and stability for variety of classification methods and several data sets
- Alternative: Bayesian variable selection

Introduction

Gene expression microarrays allow to measure the simultaneous mRNA expression of thousands of genes.

Studies have different **objectives:**

1. Gene expression association studies: relate gene expression changes to biological outcomes by comparing samples under different conditions
2. Gene prediction: building molecular profiles based on gene expression which can characterise different phenotypes (e.g. clinical outcomes), usually binary
3. Gene classification: finding new subgroups or entities (e.g. subtypes of tumours, typology) in an unsupervised manner

Here we are concerned with Aim 2 and the building of a set of parsimonious models for binary phenotypes.

Specificity of the context of application

- In genomic applications, there are typically many more covariates than samples: large p (thousands of genes), small n (50 to 100 samples) paradigm $\longrightarrow$ Multi-collinearity of genes implies there is no unique best solution, but many alternative models have similar explanatory power

Specificity of the context of application

- In genomic applications, there are typically many more covariates than samples: large p (thousands of genes), small n (50 to 100 samples) paradigm $\longrightarrow$ Multi-collinearity of genes implies there is no unique best solution, but many alternative models have similar explanatory power
- Complex dependence structure between genes linked to underlying biological pathways and networks.

Specificity of the context of application

- In genomic applications, there are typically many more covariates than samples: large p (thousands of genes), small n (50 to 100 samples) paradigm $\longrightarrow$ Multi-collinearity of genes implies there is no unique best solution, but many alternative models have similar explanatory power
- Complex dependence structure between genes linked to underlying biological pathways and networks.
- Sparseness: Out of the thousands of genes usually only a few are expected to be related to the response.

Specificity of the context of application

- In genomic applications, there are typically many more covariates than samples: large p (thousands of genes), small n (50 to 100 samples) paradigm $\longrightarrow$ Multi-collinearity of genes implies there is no unique best solution, but many alternative models have similar explanatory power
- Complex dependence structure between genes linked to underlying biological pathways and networks.
- Sparseness: Out of the thousands of genes usually only a few are expected to be related to the response.
- Need to estimate uncertainty in molecular profiles related to the role of each gene (probabilities, standard errors, ...)

Competing goals

- Make good predictions
- Figure out genes which play an important role

Competing goals

- Make good predictions
- Figure out genes which play an important role

What makes a good molecular profiling method?

- Prediction accuracy: low generalisation error (robustness)
- Interpretability
  - Parsimony: small number of genes that can be followed up in biological experiments
  - Interpretable model: explicit modelling of relationship between genes and response (no “black box”)
  - Stability: little variation in resulting profile when training data are varied (resampled)

Types of approaches

- Methods based on a sequence of univariate steps
- Multivariate regression models with regularisation/shrinkage:
  - penalised likelihood methods (ridge, lasso, elastic net,...)
  - fully Bayesian approach with mixture priors on coefficients $\longrightarrow$ Bayesian variable selection
- Multivariate machine learning approaches: e.g. support vector machines, tree-based methods (random forests), bagging and boosting, neural nets,...

Prior expectations	Accuracy	Parsimony	Structure explicit?	Stability
Univariate	**low?**	**medium**	**medium**	?
Regression with shrinkage	**high?**	**high-low**	**high**	?
Machine learning	**high?**	**low-high**	**low** (initially)	?

How to assess the stability of a profiling method?

- Using resampling methods (bootstrap, repeated training/validation set splits):
  generate m training data sets from original data set

  $\longrightarrow$ generate m profiles and assess degree of overlap

How to assess the stability of a profiling method?

- Using resampling methods (bootstrap, repeated training/validation set splits): generate m training data sets from original data set

  $\longrightarrow$ generate m profiles and assess degree of overlap

- Between any two of m profiles:
- 2×2-contingency table

  $\longrightarrow$ size of intersection $O_{11} = \#(Z^1 \cap Z^2)$

		Z^1		
		1	0	
Z^2	1	O_{11}	O_{10}	$O_{1.}$
	0	O_{01}	O_{00}	$O_{0.}$
		$O_{.1}$	$O_{.0}$	p

- Need to relate O_{11} to profile sizes $O_{1.}$ and $O_{.1}$ to make comparisons between methods possible if their profile sizes are different

Approaches used in microarray literature (e.g.)

- Average size of intersection between any two of all m profiles (Ein-Dor et al. 2005): $\frac{1}{\binom{m}{2}} \sum O_{11}$

Approaches used in microarray literature (e.g.)

- Average size of intersection between any two of all m profiles (Ein-Dor et al. 2005): $\frac{1}{\binom{m}{2}} \sum O_{11}$
- Proportion of variables included in $> 50\%$ of all m profiles (Michiels et al. 2005): generalisation of intersection

Approaches used in microarray literature (e.g.)

- Average size of intersection between any two of all m profiles (Ein-Dor et al. 2005): $\frac{1}{\binom{m}{2}} \sum O_{11}$
- Proportion of variables included in $> 50\%$ of all m profiles (Michiels et al. 2005): generalisation of intersection
- Ratio observed intersection to expected intersection (Blangiardo and Richardson 2007) assuming independence between sets

  - either assuming fixed marginal values ($O_{11} \sim$ Hypergeometric$(O_{1.}, O_{.1}, p)$): $r_e = O_{11}/(O_{1.}O_{.1}/p)$

  - or multinomial distribution for (O_{11}, O_{10}, O_{01}) with fixed p

Approaches used in microarray literature (e.g.)

- Average size of intersection between any two of all m profiles (Ein-Dor et al. 2005): $\frac{1}{\binom{m}{2}} \sum O_{11}$
- Proportion of variables included in $> 50\%$ of all m profiles (Michiels et al. 2005): generalisation of intersection
- Ratio observed intersection to expected intersection (Blangiardo and Richardson 2007) assuming independence between sets

  - either assuming fixed marginal values ($O_{11} \sim$ Hypergeometric$(O_{1.}, O_{.1}, p)$): $r_e = O_{11}/(O_{1.}O_{.1}/p)$

  - or multinomial distribution for (O_{11}, O_{10}, O_{01}) with fixed p
- Number of profiles in which a gene is included $\longrightarrow$ average over all genes, that are selected at least once (Díaz-Uriarte and Alvarez de Andrés. 2006)

Relate O_{11} to profile sizes $O_{1.}$ and $O_{.1}$

1. **Ratio of observed intersection to expected intersection** under assumption of independence between sets:

$$r_e = \frac{O_{11}}{O_{1.}O_{.1}/p}$$

Problems:

- Assumes fixed margins $\longrightarrow$ only true for univariate filtering methods where profile size is fixed in advance.
- In resampling setup, profiles are not independent because of the overlap in training data sets (even if there is no link to the response).

  $\longrightarrow$ Denominator underestimates size of expected random intersection.
- If profile sizes are small, small denominator leads to unreliable estimates of ratio.

2. **Similarity measures** (e.g. environmental sciences) for binary data, which are asymmetric (do not include number of negative matches O_{00}):

   e.g. Jaccard (1901), Dice (1945), Ochiai (1957)

   - Jaccard index:

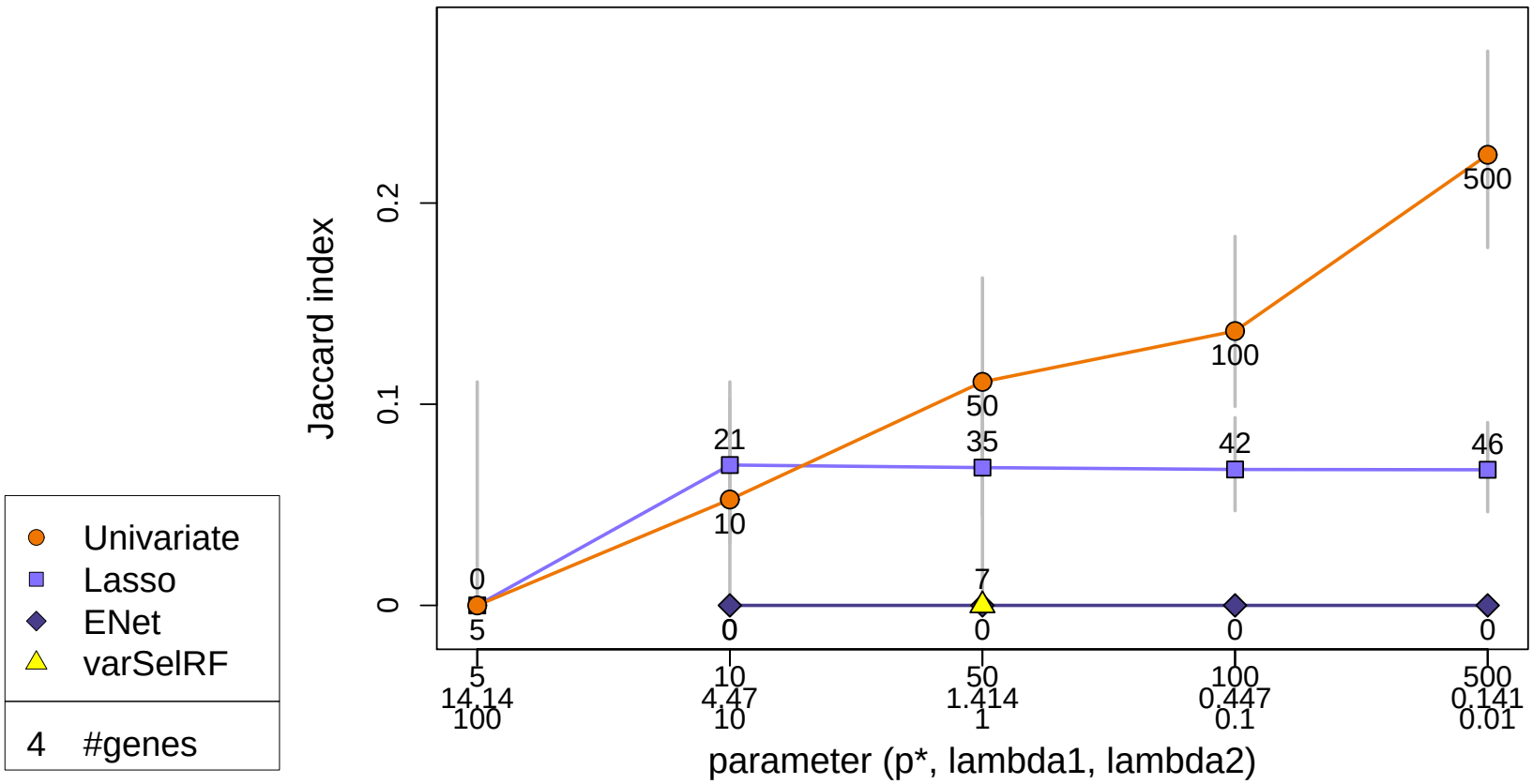
$$r_j = \frac{\#(Z^1 \cap Z^2)}{\#(Z^1 \cup Z^2)} = \frac{O_{11}}{O_{1.} + O_{.1} - O_{11}}$$

   - Advantage: fulfills all criteria of similarity measures, in particular $r_j \in [0, 1]$
   - Does not take random overlap into account. Same problem as before: expected random overlap is larger than under independence of training sets. $\longrightarrow$ Simulations

Simulations: Jaccard index

Ovarian cancer data with randomised reponse variable

Jaccard index median (inter-quartile range) of all $\binom{m}{2}$ pairs of profiles



Comments

- Hard to adjust stability measures for expected size of intersection under null in resampling setup

  $\longrightarrow$ Measures for different methods only comparable if profiles are of same size

- Often high correlations between genes expected due to their joint involvement in biological processes and co-regulation etc:

  Replacing a gene X_1 in one by a near-perfectly correlated gene X_2 in another profile does not necessarily mean decreased stability

  $\longrightarrow$ Use of "fuzzy intersections" accounting for correlations?

Study setup

Resampling scheme: Multiple random validation (Michiels *et al.* 2005):

- *Divide data randomly into training and validation data (ratio 2:1).*
- *Repeat 50 times (*$i = 1, \ldots, 50$*)*
- *For each* i*: Fit model using training data and find molecular profile across a range of parameter values*
- *Compare prediction performance on validation data and make-up ("which genes selected") of resulting* 50 *profiles.*

Data

	p	n	Response (binary)	Clinical covariates	Independent validation data
Breast cancer (van't Veer 2002)	4770	97	survival (dichotomised)	yes	yes (Vijver 2002)
Ovarian cancer (Schwartz 2002)	7129	104	histology	no	yes (Lu 2004)
Leukaemia (Golub 1999)	7129	72	tumour type	no	no
Prostate cancer (Singh 2002)	12625	102	tumour vs. normal	no	no
Acute myeloid leukaemia (Valk 2004)	22283	273	normal vs. abnormal karyotype	no	no

Classification methods (logistic regression/decision trees)

- **Univariate:**

  Select genes by estimated effects $\frac{|\hat{\beta}_j|}{\text{s.e.}(\hat{\beta}_j)}$ from univariate logistic regressions.

  Use selected genes for classification:

  – Nearest-centroid classification

  – Diagonal linear discriminant analysis

Classification methods (logistic regression/decision trees)

- **Univariate:**

  Select genes by estimated effects $\frac{|\hat{\beta}_j|}{\text{s.e.}(\hat{\beta}_j)}$ from univariate logistic regressions.

  Use selected genes for classification:

  - Nearest-centroid classification
  - Diagonal linear discriminant analysis

- **Multivariate regression with penalty term:** Maximise penalised log-likelihood
  - Lasso regression (Tibshirani 1996): $\lambda||\beta||_1 = \lambda \sum_{g=1}^{p} |\beta_g|$
  - Ridge regression (Hoerl and Kennard 1970): $\lambda||\beta||_2 = \lambda \sum_{g=1}^{p} \beta_g^2$
  - Elastic net (Zou and Hastie 2005): $\lambda_1||\beta||_1 + \lambda_2||\beta||_2$

Classification methods (logistic regression/decision trees)

- **Univariate:**

  Select genes by estimated effects $\frac{|\hat{\beta}_j|}{\text{s.e.}(\hat{\beta}_j)}$ from univariate logistic regressions.

  Use selected genes for classification:

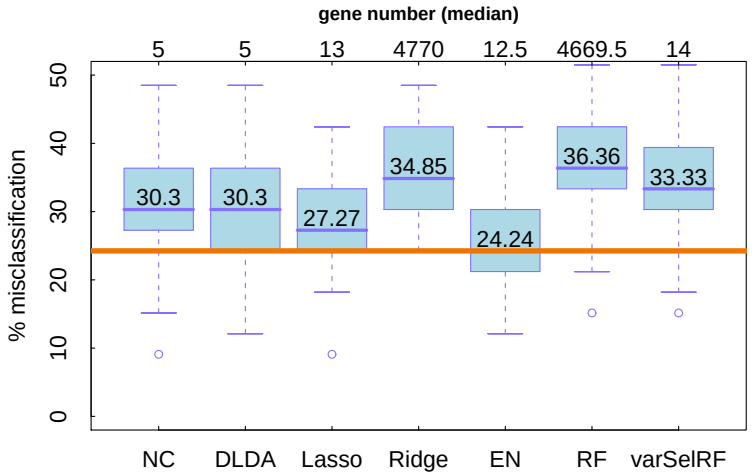
  - Nearest-centroid classification
  - Diagonal linear discriminant analysis

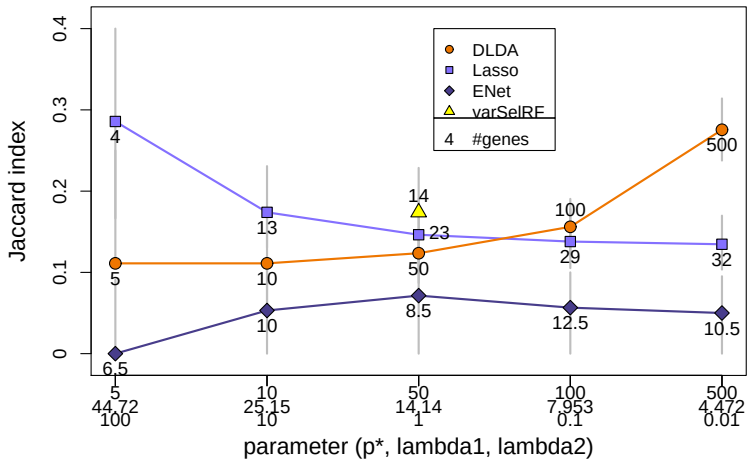
- **Multivariate regression with penalty term:** Maximise penalised log-likelihood
  - Lasso regression (Tibshirani 1996): $\lambda||\beta||_1 = \lambda \sum_{g=1}^{p} |\beta_g|$
  - Ridge regression (Hoerl and Kennard 1970): $\lambda||\beta||_2 = \lambda \sum_{g=1}^{p} \beta_g^2$
  - Elastic net (Zou and Hastie 2005): $\lambda_1||\beta||_1 + \lambda_2||\beta||_2$

- **Multivariate machine learning:**
  - Random forests (Breiman 2001)
  - Random forests with variable selection varSelRF (Díaz-Uriarte 2006)

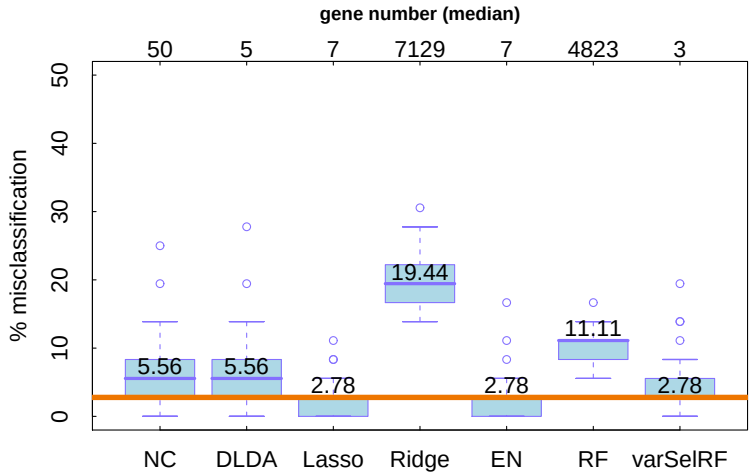
Results: Validation errors and Jaccard index

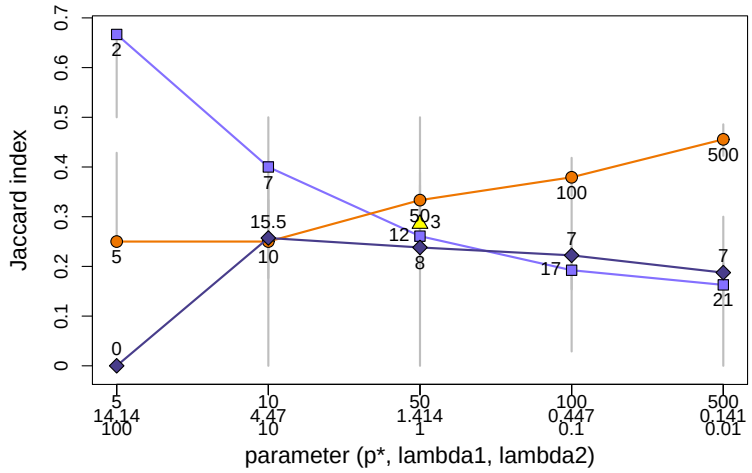
Breast cancer





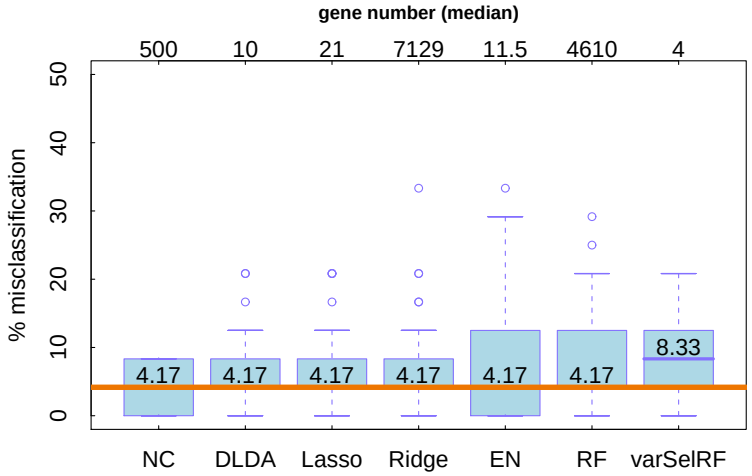
Ovarian cancer

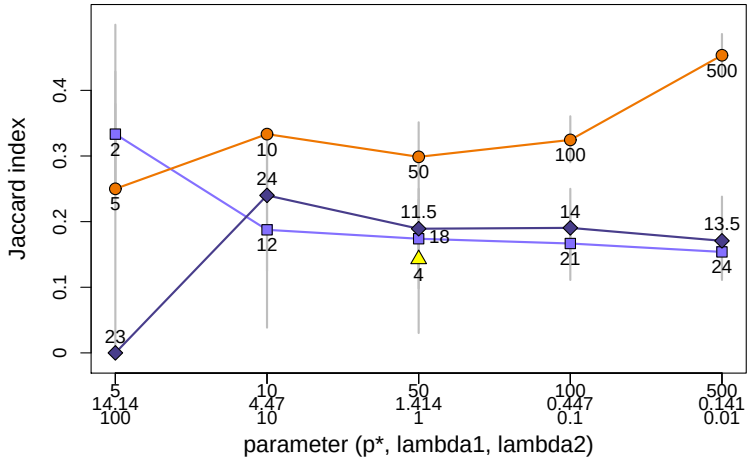




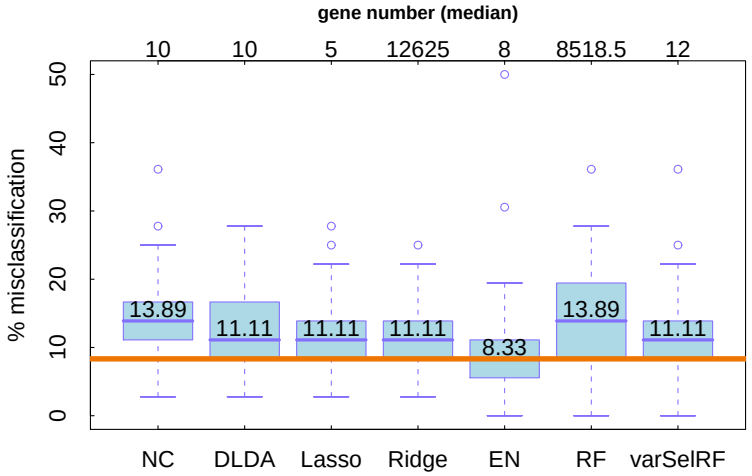
Results: Validation errors and Jaccard index

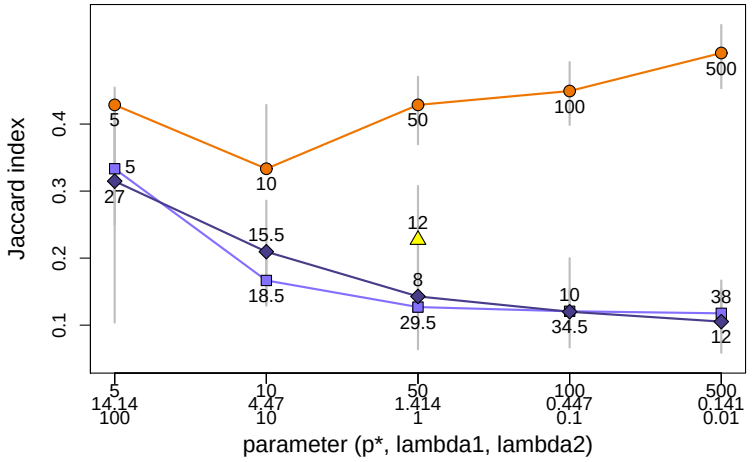
Leukaemia





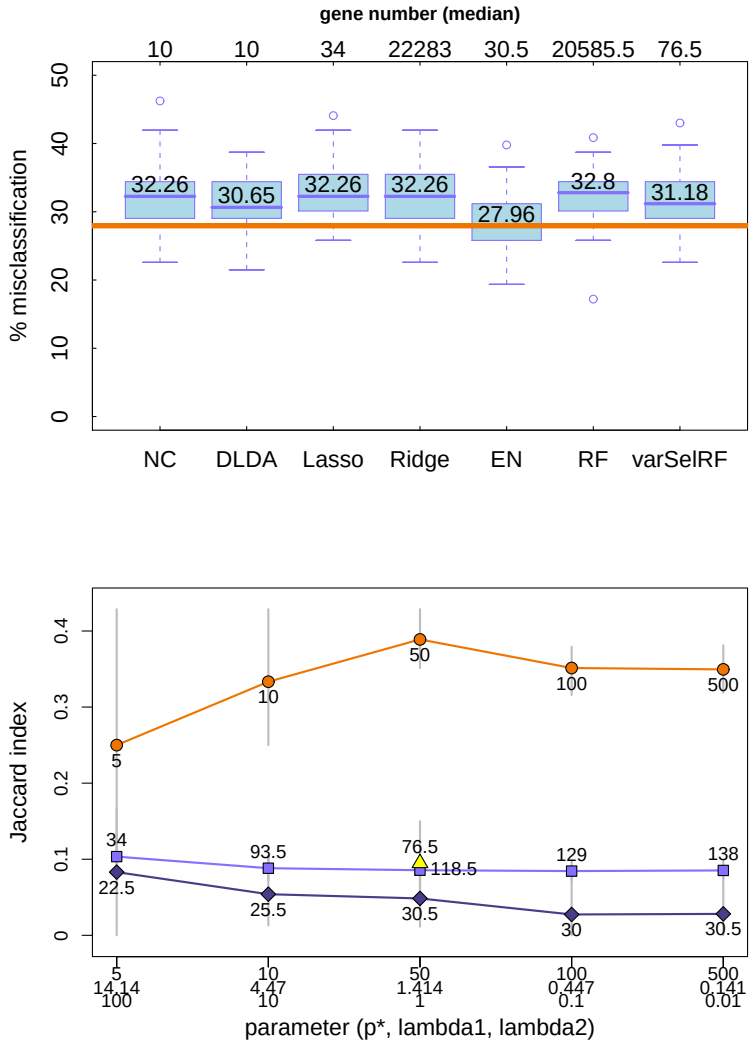
Prostate cancer





Results: Validation errors and Jaccard index

Acute myeloid leukaemia



Results: Validation on independent data

Ovarian cancer data (Lu *et al.* 2004): % misclassifications in new data when using genes found for Schwartz *et al.* (2002) in logistic regression

	Complete profiles (median over all profiles)		Genes in $> 50\%$ of all profiles	
	Error	median # Genes	Error	median # Genes
Naïve classifier (all into most frequent class)	38.1%		38.1%	
Univariate	21.4%	5	21.4%	3
Lasso	16.7%	7	14.3%	5
Elastic Net	14.3%	7	14.3%	5
VarSelRF	21.4%	3	38.1%	2

Note: The same five genes appear in $> 50\%$ of profiles for all multivariate methods (lasso, elastic net, varSelRF): S100P, ABP1, ANX4, CYP2C18, SPINK1

Summary

- Sparsity-inducing methods perform better in terms of prediction than those keeping all/most genes
- Elastic net has overall best prediction accuracy (both in resampling setup and in independent data set), but is usually slightly less stable than lasso
- Caution when comparing Jaccard indices: model sizes differ
- Multivariate methods most stable when profiles are smallest, univariate method most stable when they are largest
  But: simulation study $\longrightarrow$ Jaccard indices inflated for very large profiles
- Predictive performance of univariate methods only slightly worse than best multivariate methods

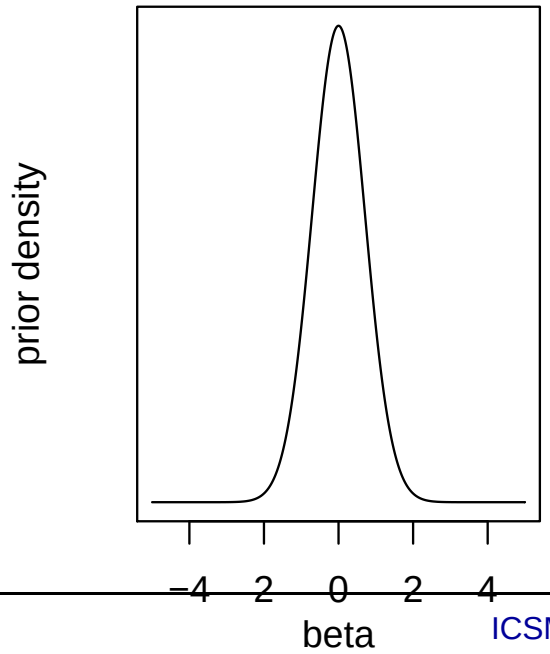
Alternative: Bayesian variable selection

Bayesian interpretation for lasso/ridge

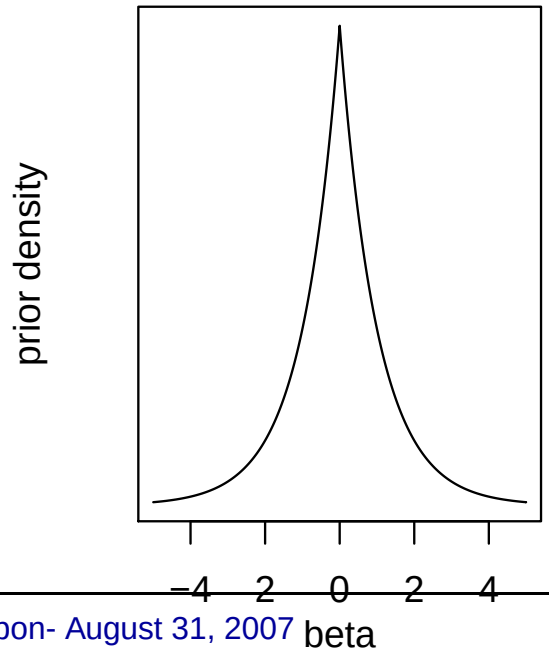
Regression model with shrinkage prior on regression coefficient vector β:

- Ridge: MAP (*maximum a posteriori*) estimator with Gaussian prior
  $p(\beta|\tau) = N(0, \tau I_p)$, where $\tau = 1/(2\lambda)$
- Lasso: MAP estimator with Laplace (double exponential) prior
  $p(\beta|\tau) = \text{Laplace}(0, \tau I_p)$, where $\tau = 2/\lambda^2$.

Ridge (q = 2): Gaussian prior



Lasso (q = 1): Laplace prior



Alternative: Bayesian variable selection

BVS model with indicator variable $\gamma_i = \begin{cases} 1 & \text{variable i is included} \\ 0 & \text{variable i is excluded} \end{cases}$

Shape of prior to encourage parsimony:

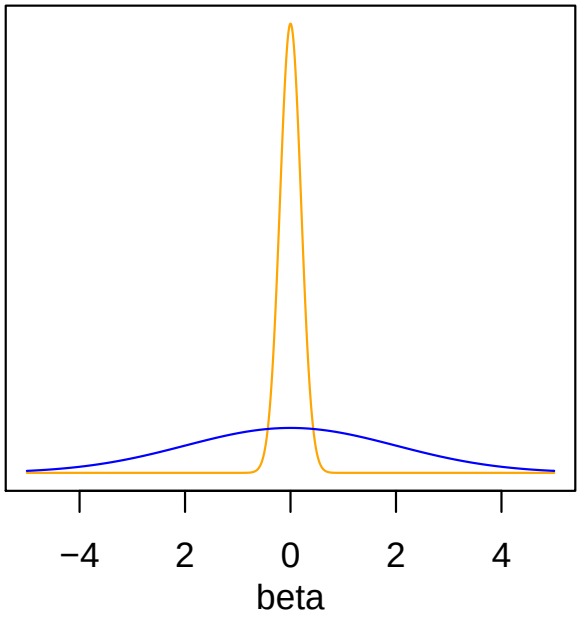
- spike in zero (variable exclusion),
- heavy tails (variable inclusion).

Alternative: Bayesian variable selection

BVS model with indicator variable $\gamma_i = \begin{cases} 1 & \text{variable i is included} \\ 0 & \text{variable i is excluded} \end{cases}$

Shape of prior to encourage parsimony:

- spike in zero (variable exclusion),
- heavy tails (variable inclusion).



E.g. Normal mixture prior

(George and McCulloch 1997):

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0, \sigma^2 v_{0\gamma_i}) + \gamma_i N(0, \sigma^2 v_{1\gamma_i})$$

Here: Similar Bayesian variable selection model in context of logistic regression (Holmes and Held 2006), conditioning the model on the components where $\gamma_i = 1$

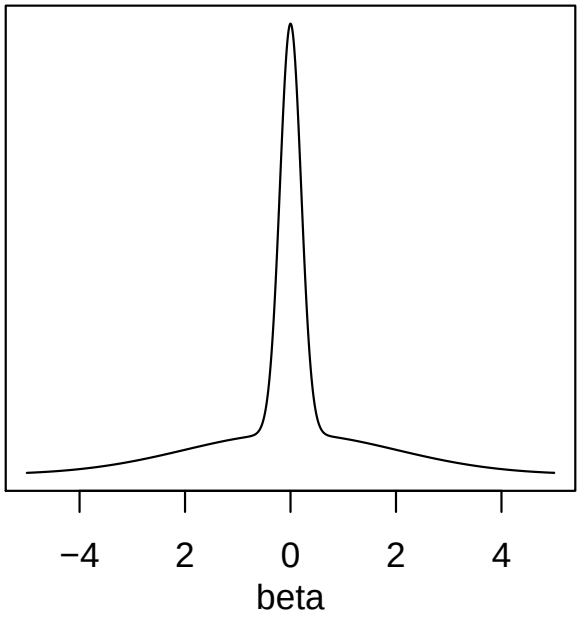
Alternative: Bayesian variable selection

BVS model with indicator variable $\gamma_i = \begin{cases} 1 & \text{variable i is included} \\ 0 & \text{variable i is excluded} \end{cases}$

Shape of prior to encourage parsimony:

- spike in zero (variable exclusion),
- heavy tails (variable inclusion).

Normal mixture prior



E.g. Normal mixture prior

(George and McCulloch 1997):

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0, \sigma^2 v_{0\gamma_i}) + \gamma_i N(0, \sigma^2 v_{1\gamma_i})$$

Here: Similar Bayesian variable selection model in context of logistic regression (Holmes and Held 2006), conditioning the model on the components where $\gamma_i = 1$

Alternative: Bayesian variable selection

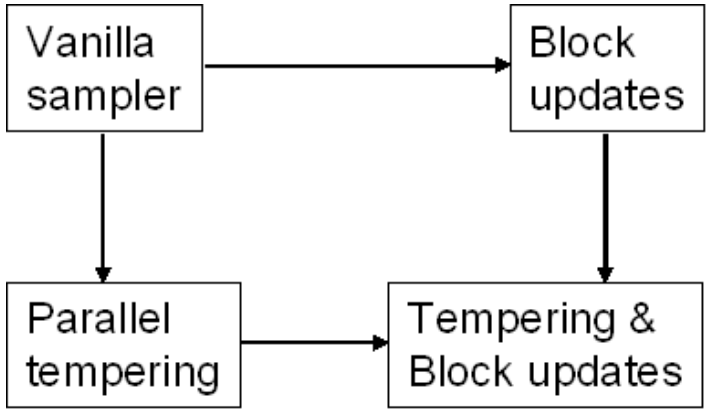
- Advantage: Fully probabilistic framework $\longrightarrow$ posterior probabilities for gene variables to be selected. No resampling necessary to assess uncertainty in selecting individual genes or models.
- But: Validation data still needed to estimate generalisation error (different aim)
- Markov chain Monte Carlo (MCMC) is used as a stochastic search algorithm to find models with high posterior probability.
- Difficulties: Large scale of applications renders standard MCMC algorithms impractical (full Gibbs sampling too time-consuming, and fast single-variable addition/deletion algorithms mixing too slowly).

Alternative: Bayesian variable selection

There are many ways to improve on standard MCMC and to make Bayesian variable selection (BVS) practicable.

For example:

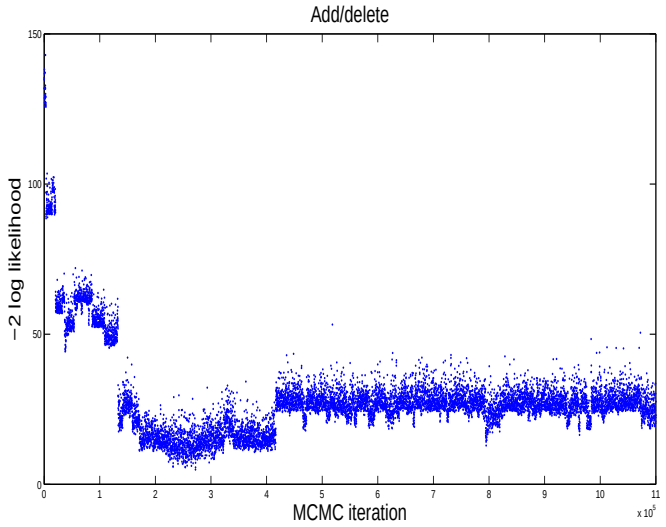
- Block sampling: Employ the dependence structure among covariates to find variables to update together in blocks - to construct Markov chains which can move quickly around the vast model space.
- Metropolis-coupled MCMC (parallel tempering): Run parallel chains at higher temperatures T to improve mixing and "borrow" better mixing by proposing swaps between chains.

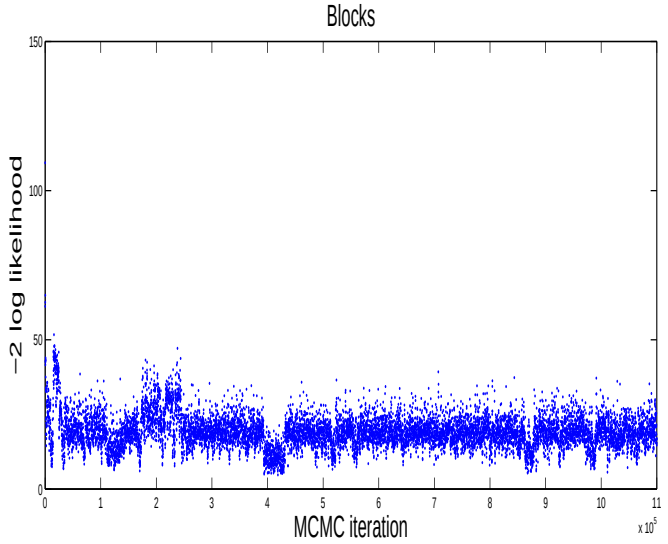


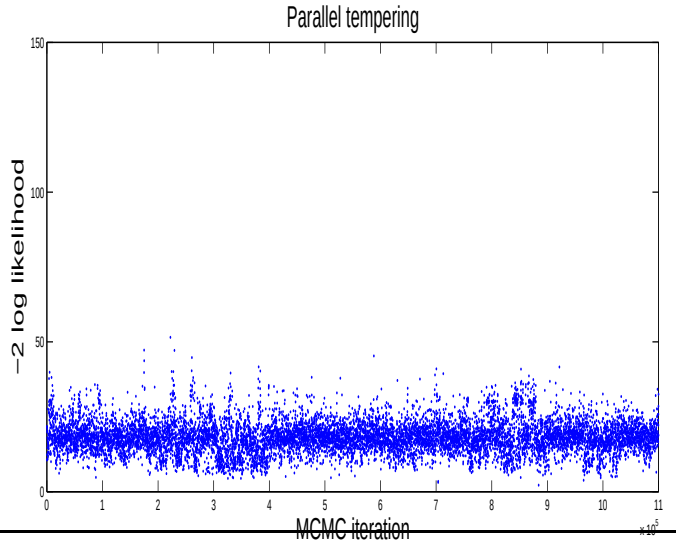
Alternative: Bayesian variable selection

Results for ovarian cancer data: Same 5 genes from resampling study found.

Deviance trace shows convergence rate and when chain gets stuck in local optima.

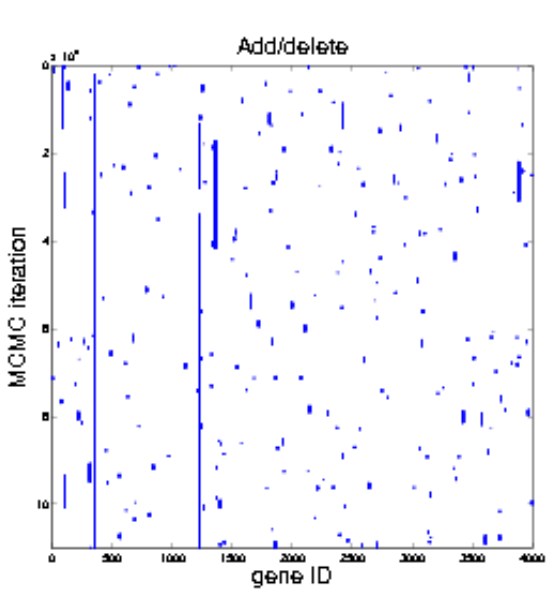


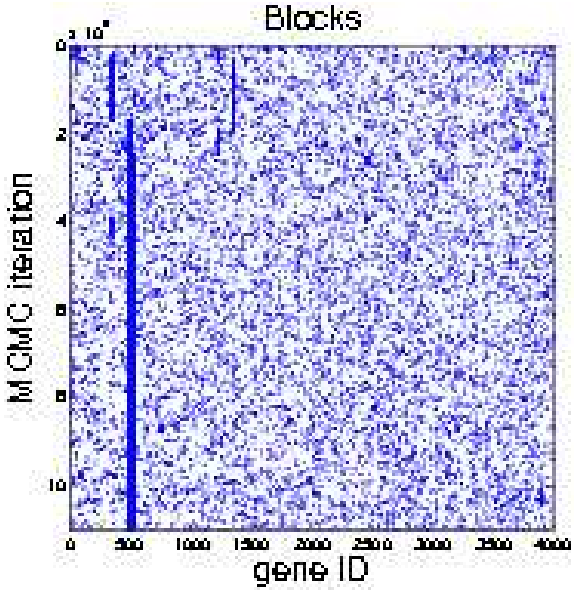


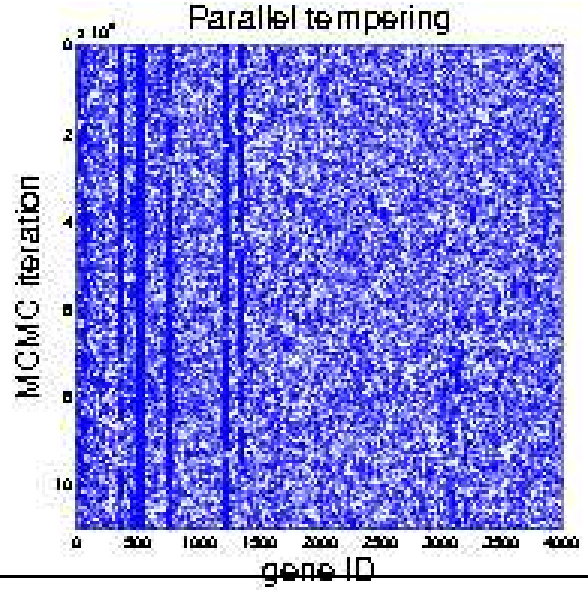


Alternative: Bayesian variable selection

Trace plots for γ: Mixing improves dramatically when using blocks/parallel chains.







Summary

- Prediction accuracy is often not the only goal when estimating molecular profiles
  $\longrightarrow$ interpretability is important
- $\longrightarrow$ Regression framework: shrinkage methods that induce sparsity (elastic net, lasso) do very well
- The uncertainty associated with genes being selected into a profile needs to be estimated:
  - either resampling framework (frequencies for gene inclusion)
  - or Bayesian modelling (posterior probabilities for gene inclusion)
- Open problem: How best to assess the stability of molecular profiles?

Key references

- Díaz-Uriarte and Alvarez de Andrés. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7:3 (2006)
- Hoerl and Kennard. Ridge Regression: biased estimation for nonorthogonal problems. Technometrics, 12:55-67 (1970)
- Holmes and Held. Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Analysis, 1:145-168 (2006)
- Michiels *et al.* Prediction of cancer outcome with microarrays: a multiple random validation study. The Lancet, 488-492 (2005)
- Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 58:267-288 (1996)
- Zou and Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, 67(2):301-320 (2005)

Acknowledgements

Molecular Therapeutics Section, Department of Oncology, Imperial College London

Hani Gabra

Euan Stronach

Institute for Mathematical Sciences, Imperial College London

Leonardo Bottolo

Oxford Centre for Gene Function, Department of Statistics, University of Oxford

Chris Holmes

Financial support

wellcome trust