Review of modern logistic regression methods

Enes Makalic   Daniel F. Schmidt

Centre for MEGA Epidemiology
The University of Melbourne

Work in Progress, 2010

Outline

Outline

1. Introduction
   - Problem Description
   - Traditional Logistic Regression
   - Motivation
2. Regularised Logistic Regression
   - Ridge Regression
   - LASSO
   - Elastic Net
   - Non-negative Garrote
   - Summary
3. Simulation
   - Consistency
   - Simulated Data
   - Real data

Problem Description (1)

- We have a binary classification problem
  - Data $\mathbf{y} = (y_1, \ldots, y_n)'$, $y_i = \{0, 1\}$
  - Matrix of p covariate vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, $\mathbf{x}_j \in \mathbb{R}^n$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

- Use a logistic regression model (n samples, p predictors)

Problem Description (2)

- Logistic regression model for explaining data $\mathbf{y}$

$$p(y_i = 1 | \mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{x}_i' \boldsymbol{\beta})}$$

  - $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of logistic regression coefficients
- Task:
  - Estimate parameters
  - Select significant regressors

Problem Description (2)

- Logistic regression model for explaining data $\mathbf{y}$

$$p(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta})}$$

    - $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of logistic regression coefficients
- Task:
    - Estimate parameters
    - Select significant regressors

Problem Description (2)

- Logistic regression model for explaining data $\mathbf{y}$

$$p(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta})}$$

  - $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of logistic regression coefficients
- Task:
  - Estimate parameters
  - Select significant regressors

Parameter Estimation (1)

- Parameter estimation: maximum likelihood (ML)
  - Use β that maximises the (log-)likelihood $l(\boldsymbol{\beta}) : \mathbb{R}^p \to \mathbb{R}$

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\beta}} \{l(\boldsymbol{\beta})\}$$

  - Log-likelihood is convex; single maximum
  - Consistent when $p < n$
  - Solution found with numerical optimization
    - Newton-Raphson, iteratively re-weighted least squares, etc.

Parameter Estimation (2)

- Problems with maximum likelihood
  - Non-zero estimates for all coefficients
  - Over-estimation; $\hat{\beta}_{\mathrm{ML}}$ biased away from 0
  - Estimate not defined if $p > n$
  - Estimate does not exist if the data is (quasi)completely separated
  - Poor small sample performance
  - Correlated predictors are problematic

Parameter Estimation (2)

- Problems with maximum likelihood
  - Non-zero estimates for all coefficients
  - Over-estimation; $\hat{\beta}_{\mathrm{ML}}$ biased away from 0
  - Estimate not defined if $p > n$
  - Estimate does not exist if the data is (quasi)completely separated
  - Poor small sample performance
  - Correlated predictors are problematic

Parameter Estimation (2)

- Problems with maximum likelihood
  - Non-zero estimates for all coefficients
  - Over-estimation; $\hat{\beta}_{\mathrm{ML}}$ biased away from 0
  - Estimate not defined if $p > n$
  - Estimate does not exist if the data is (quasi)completely separated
  - Poor small sample performance
  - Correlated predictors are problematic

Parameter Estimation (2)

- Problems with maximum likelihood
  - Non-zero estimates for all coefficients
  - Over-estimation; $\hat{\beta}_{\mathrm{ML}}$ biased away from 0
  - Estimate not defined if $p > n$
  - Estimate does not exist if the data is (quasi)completely separated
  - Poor small sample performance
  - Correlated predictors are problematic

Parameter Estimation (2)

- Problems with maximum likelihood
  - Non-zero estimates for all coefficients
  - Over-estimation; $\hat{\beta}_{\mathrm{ML}}$ biased away from 0
  - Estimate not defined if $p > n$
  - Estimate does not exist if the data is (quasi)completely separated
  - Poor small sample performance
  - Correlated predictors are problematic

Parameter Estimation (2)

- Problems with maximum likelihood
  - Non-zero estimates for all coefficients
  - Over-estimation; $\hat{\beta}_{\mathrm{ML}}$ biased away from 0
  - Estimate not defined if $p > n$
  - Estimate does not exist if the data is (quasi)completely separated
  - Poor small sample performance
  - Correlated predictors are problematic

Parameter Estimation (2)

- Problems with maximum likelihood
  - Non-zero estimates for all coefficients
  - Over-estimation; $\hat{\beta}_{\mathrm{ML}}$ biased away from 0
  - Estimate not defined if $p > n$
  - Estimate does not exist if the data is (quasi)completely separated
  - Poor small sample performance
  - Correlated predictors are problematic

Regressor selection (1)

- Stepwise regression
  - Forward selection
  - Backward elimination
- All-subset selection
  - Examine all possible 2^p regressor subsets

Regressor selection (2)

- Problems with stepwise regression
  - Multiple hypothesis testing problem
  - Correlated predictors
- Problems with all-subset selection
  - Not feasible for moderate to large p
    - If $p = 10$, need to check $2^p = 1024$ groups of regressors!
  - Stability
    - Small change in the data $\rightarrow$ big changes in parameter estimates

Regressor selection (2)

- Problems with stepwise regression
  - Multiple hypothesis testing problem
  - Correlated predictors
- Problems with all-subset selection
  - Not feasible for moderate to large p
    - If $p = 10$, need to check $2^p = 1024$ groups of regressors!
  - Stability
    - Small change in the data $\rightarrow$ big changes in parameter estimates

Regressor selection (2)

- Problems with stepwise regression
  - Multiple hypothesis testing problem
  - Correlated predictors
- Problems with all-subset selection
  - Not feasible for moderate to large p
    - If $p = 10$, need to check $2^p = 1024$ groups of regressors!
  - Stability
    - Small change in the data $\rightarrow$ big changes in parameter estimates

Regressor selection (2)

- Problems with stepwise regression
  - Multiple hypothesis testing problem
  - Correlated predictors
- Problems with all-subset selection
  - Not feasible for moderate to large p
    - If $p = 10$, need to check $2^p = 1024$ groups of regressors!
  - Stability
    - Small change in the data $\rightarrow$ big changes in parameter estimates

Regressor selection (2)

- Problems with stepwise regression
  - Multiple hypothesis testing problem
  - Correlated predictors
- Problems with all-subset selection
  - Not feasible for moderate to large p
    - If $p = 10$, need to check $2^p = 1024$ groups of regressors!
  - Stability
    - Small change in the data $\rightarrow$ big changes in parameter estimates

Regressor selection (2)

- Problems with stepwise regression
  - Multiple hypothesis testing problem
  - Correlated predictors
- Problems with all-subset selection
  - Not feasible for moderate to large p
    - If $p = 10$, need to check $2^p = 1024$ groups of regressors!
  - Stability
    - Small change in the data $\rightarrow$ big changes in parameter estimates

Motivation

- Many problems with maximum likelihood and stepwise regression
- Ideally, want a method that
  - consistently selects true predictors
  - automatically shrinks parameters
  - selects important variables
  - can be applied when $p >> n$
  - has the Oracle property (asymptotically)

Outline

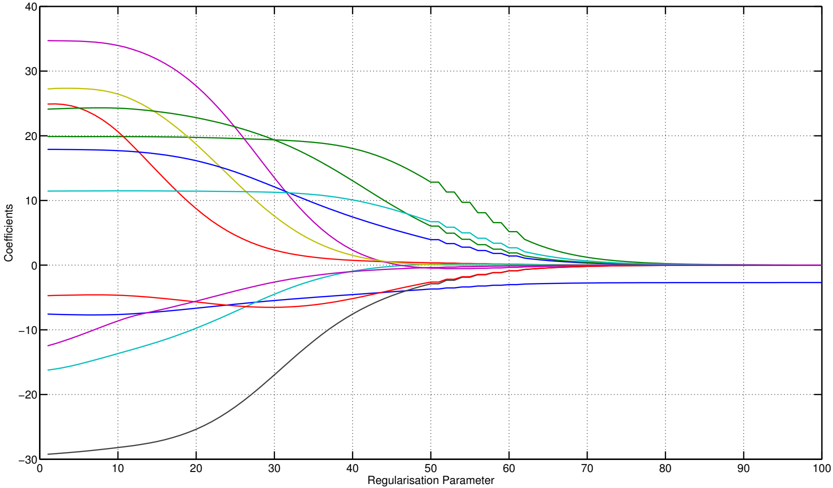
Ridge Regression (1)

- Ridge regression estimate

$$\hat{\boldsymbol{\beta}}_{\mathrm{RR}} = \arg\max_{\boldsymbol{\beta}} \{l(\boldsymbol{\beta})\} \quad \text{s.t. } \sum_{j=i}^{p} \beta_j^2 \leq t$$

- Restricted maximum likelihood estimator
- Introduced originally for linear regression models
  - Dominates least squares in terms of MSE
- Shrinks all parameters to zero ($t = 0$) or includes all predictors ($t \to \infty$)
- Introduces little bias; large reduction in variance

Ridge Regression (2)



Ridge Regression (3)

- Disadvantages
  - Cannot infer sparse models
    - Includes all predictors or none
  - Not suitable for large p, or $p >> n$
  - Must estimate regularisation parameter

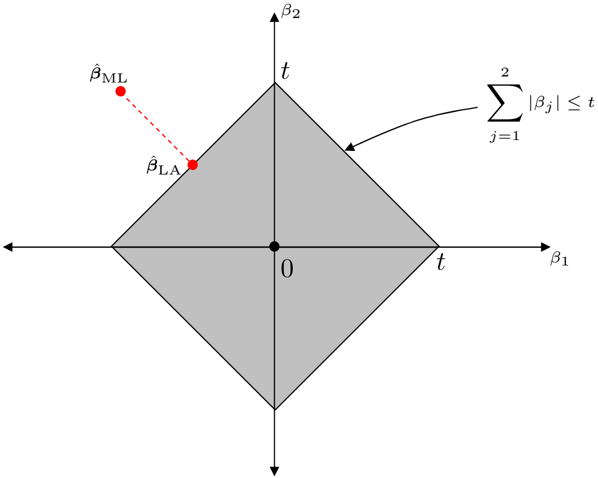
LASSO (1)

- Least absolute shrinkage and selection operator (LASSO)

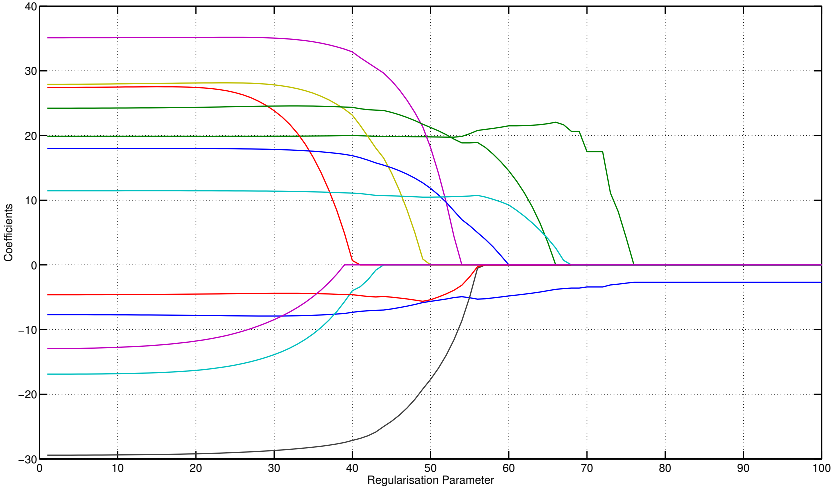
$$\hat{\boldsymbol{\beta}}_{\mathrm{LA}} = \arg\max_{\boldsymbol{\beta}} \{l(\boldsymbol{\beta})\} \quad \text{s.t. } \sum_{j=i}^{p} |\beta_j| \leq t$$

  - Simultaneous parameter shrinkage and variable selection
  - Shrinkage determined by t
  - Consistent if $p >> n$ under certain assumptions

LASSO (2)


β_2
$\hat{\beta}_{\mathrm{ML}}$
t
$\sum_{j=1}^{2} |\beta_j| \leq t$
$\hat{\beta}_{\mathrm{LA}}$
0
t
β_1

LASSO (3)


Coefficients
Regularisation Parameter
40
30
20
10
0
−10
−20
−30
0
10
20
30
40
50
60
70
80
90
100

LASSO (4)

- Disadvantages
  - Correlated predictors are problematic
    - Can lead to inconsistent estimates
  - Produces biased estimates for large coefficients
    - All coefficients penalized equally
    - Solution: data dependent weights (Adaptive LASSO, Iterative LASSO, etc.)
  - Optimal t for prediction gives inconsistent variable selection

Elastic Net

- LASSO can perform poorly if the predictors are correlated
  - Ridge regression performs well
- Solution: combine LASSO and ridge regression penalties

$$\hat{\boldsymbol{\beta}}_{\mathrm{EN}} = \arg\max_{\boldsymbol{\beta}} \{l(\boldsymbol{\beta})\} \quad \text{s.t. } \sum_{j=i}^{p} |\beta_j| \leq t_1, \sum_{j=i}^{p} \beta_j^2 \leq t_2$$

  - Handles correlated predictors
  - Tends to perform better than LASSO when $p > n$

Non-negative Garrote (1)

- Requires an initial parameter estimate β^*
  - For example, maximum likelihood, ridge regression, etc.
- Non-negative Garrote (NNG) estimate

$$\hat{\beta}_{\mathrm{NG}} = \arg\max_{\tilde{\beta}} \left\{ l(\tilde{\beta}_1, \ldots, \tilde{\beta}_p) \right\} \quad \text{s.t. } c_j \geq 0, \sum_{j=1}^{p} c_j \leq t$$

where $\tilde{\beta}_j = c_j \beta_j^*$, $j = 1, \ldots, p$.

Non-negative Garrote (2)

- Properties
  - Consistent in terms of parameter estimation and variable selection (linear regression)
    - Remains true even if β^* is inconsistent (with caveats)
  - Oracle property
    - It performs as well as if the true underlying model were given in advance

But where are the p-values?

- Can use LASSO, NNG, etc. in two phases
  - Phase I: dimensionality reduction
  - Phase II: classical significance testing
- Automatic multiplicity correction
  - Control FWER, FDR

Software

- MATLAB™
  - S:\Staff\EnesAndDaniel\MATLAB Code\Penalized Logistic Regression\
- R and S-PLUS
  - Packages: brdgrun, elasticnet, glars, glasso, glmpath, grplasso, lars, lasso2, relaxo
- STATA
  - G. Ambler, www.homepages.ucl.ac.uk/~ucakgam/stata.html

Outline

1. Introduction
   - Problem Description
   - Traditional Logistic Regression
   - Motivation
2. Regularised Logistic Regression
   - Ridge Regression
   - LASSO
   - Elastic Net
   - Non-negative Garrote
   - Summary
3. Simulation
   - Consistency
   - Simulated Data
   - Real data

A Simple Example (1)

- Path consistency
  - Sample sizes $n = \{20, 50, 100, 200, 500\}$
  - Regressor matrix $\mathbf{X}$ has four regressors $(p = 4)$

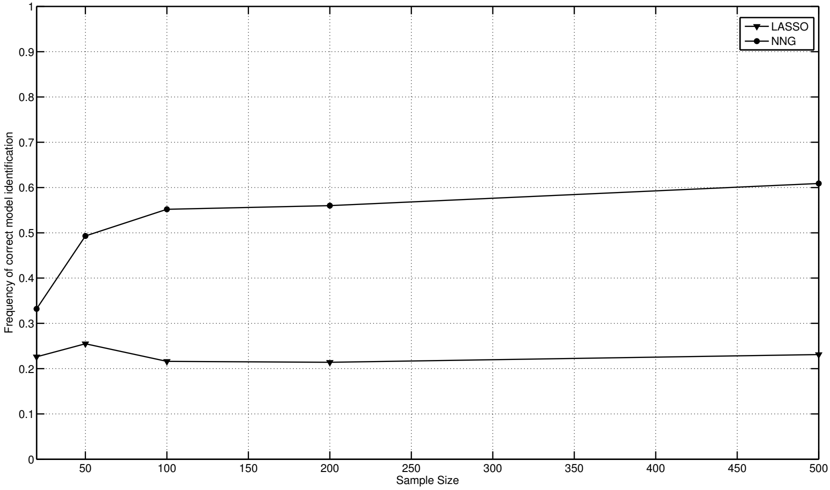
$$\begin{aligned} \mathbf{x}_i &= (1, X_1, X_2, X_3)' \\ X_1, X_2 &\sim N(0,1) \\ X_3 &\sim N(\alpha(X_1 + X_2), 1 - 2\alpha^2) \end{aligned}$$

  where $\alpha \in (0 \cdot 35, 0 \cdot 55)$ and $i = (1, \ldots, n)$.
  - True coefficients

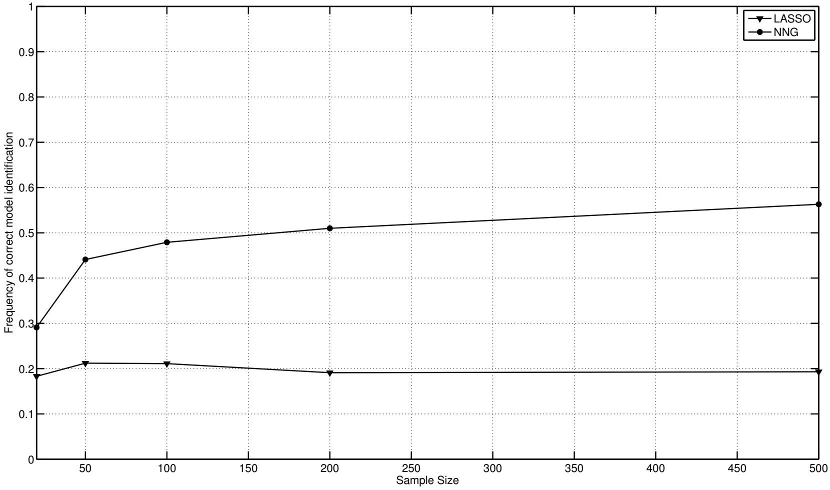
$$\boldsymbol{\beta} = (0, 1, 1, 0)'$$

  - Regularization parameters selected using an independently generated data set
  - NNG uses ridge regression as initial estimates

A Simple Example (2)



A Simple Example (3)



Simulated Data

- In all simulations ...
  - Sample size $n = \{20, 50, 100\}$.
  - Regressor correlation $\mathrm{corr}(i, j) = 0.5^{|i-j|}$.

- Example 1: $\boldsymbol{\beta} = (3, 2, 1{\cdot}5, 0, 0, 0, 0, 0)'$.
- Example 2: $\beta_j = 0.85$ for all j.
- Example 3: $\boldsymbol{\beta} = (5, 0{\cdot}5, 0{\cdot}5, 0{\cdot}5, 0, 0, 0, 0)'$.

Methods	$n = 20$				$n = 50$			
	NLL	Size	FP	FN	NLL	Size	FP	FN
fwd	30·70 (0·80)	1·24 (0·04)	1·96 (0·03)	0·20 (0·02)	7·12 (0·07)	2·79 (0·04)	0·63 (0·03)	0·42 (0·03)
gfwd	26·64 (0·41)	1·18 (0·04)	1·97 (0·03)	0·15 (0·02)	6·74 (0·05)	2·75 (0·04)	0·64 (0·02)	0·38 (0·03)
lasso	21·13 (0·15)	4·53 (0·05)	0·50 (0·02)	2·03 (0·04)	6·50 (0·04)	5·78 (0·04)	0·05 (0·01)	2·83 (0·04)
glasso	22·40 (0·18)	2·89 (0·04)	0·93 (0·03)	0·82 (0·03)	6·44 (0·05)	3·97 (0·04)	0·23 (0·01)	1·20 (0·04)
rr	21·13 (0·14)	8·00 (0·00)	0·00 (0·00)	5·00 (0·00)	6·76 (0·03)	8·00 (0·00)	0·00 (0·00)	5·00 (0·00)
grr	21·86 (0·21)	3·37 (0·05)	0·77 (0·02)	1·14 (0·03)	6·45 (0·03)	4·32 (0·04)	0·17 (0·01)	1·49 (0·04)
enet	20·64 (0·12)	6·10 (0·05)	0·21 (0·01)	3·31 (0·05)	6·50 (0·03)	6·40 (0·04)	0·02 (0·00)	3·42 (0·04)
genet	22·20 (0·22)	3·07 (0·04)	0·85 (0·02)	0·92 (0·03)	6·42 (0·04)	4·06 (0·04)	0·20 (0·01)	1·26 (0·04)
ilasso	22·41 (0·19)	2·90 (0·04)	0·93 (0·02)	0·83 (0·03)	6·42 (0·05)	3·98 (0·04)	0·22 (0·01)	1·20 (0·04)
nng	21·34 (0·25)	3·50 (0·04)	0·66 (0·02)	1·16 (0·03)	6·34 (0·03)	4·35 (0·04)	0·11 (0·01)	1·46 (0·04)

Example 1: median negative log-likelihood (NLL), mean model size (Size), mean number of false positive regressors (FP) and mean number of false negative regressors (FN) included in the selected model. Tests are based on 1000 iterations with standard errors included in parentheses

Methods	$n = 20$				$n = 50$			
	NLL	Size	FP	FN	NLL	Size	FP	FN
fwd	35·10	1·17	6·83	0·00	10·20	4·27	3·73	0·00
	(0·08)	(0·05)	(0·05)	(0·00)	(0·09)	(0·07)	(0·07)	(0·00)
gfwd	34·66	1·11	6·89	0·00	9·19	4·05	3·94	0·00
	(0·03)	(0·04)	(0·04)	(0·00)	(0·08)	(0·07)	(0·07)	(0·00)
lasso	24·83	4·95	3·06	0·00	7·76	6·94	1·06	0·00
	(0·19)	(0·05)	(0·05)	(0·00)	(0·04)	(0·03)	(0·03)	(0·00)
glasso	28·24	3·14	4·86	0·00	8·44	5·66	2·34	0·00
	(0·19)	(0·05)	(0·05)	(0·00)	(0·05)	(0·05)	(0·05)	(0·00)
rr	21·23	8·00	0·00	0·00	7·18	8·00	0·00	0·00
	(0·11)	(0·00)	(0·00)	(0·00)	(0·03)	(0·00)	(0·00)	(0·00)
grr	26·97	3·80	4·20	0·00	8·23	6·10	1·90	0·00
	(0·19)	(0·05)	(0·05)	(0·00)	(0·04)	(0·04)	(0·04)	(0·00)
enet	21·59	7·40	0·60	0·00	7·24	7·88	0·12	0·00
	(0·12)	(0·04)	(0·04)	(0·00)	(0·02)	(0·01)	(0·01)	(0·00)
genet	27·20	3·65	4·35	0·00	8·24	6·06	1·94	0·00
	(0·22)	(0·05)	(0·05)	(0·00)	(0·04)	(0·04)	(0·04)	(0·00)
ilasso	28·23	3·15	4·85	0·00	8·44	5·67	2·33	0·00
	(0·21)	(0·05)	(0·05)	(0·00)	(0·05)	(0·05)	(0·05)	(0·00)
nng	26·49	4·08	3·92	0·00	8·05	6·33	1·67	0·00
	(0·23)	(0·05)	(0·05)	(0·00)	(0·04)	(0·04)	(0·04)	(0·00)

Example 2: median negative log-likelihood (NLL), mean model size (Size), mean number of false positive regressors (FP) and mean number of false negative regressors (FN) included in the selected model. Tests are based on 1000 iterations with standard errors included in parentheses

Methods	n = 20				n = 50			
	NLL	Size	FP	FN	NLL	Size	FP	FN
fwd	19·42	1·17	2·98	0·15	5·57	1·68	2·54	0·22
	(0·78)	(0·04)	(0·03)	(0·02)	(0·03)	(0·04)	(0·02)	(0·02)
gfwd	16·55	0·99	3·09	0·08	5·40	1·63	2·57	0·19
	(0·35)	(0·03)	(0·02)	(0·01)	(0·02)	(0·04)	(0·02)	(0·02)
lasso	16·70	4·08	1·49	1·57	5·45	5·08	0·98	2·06
	(0·14)	(0·05)	(0·03)	(0·03)	(0·03)	(0·05)	(0·02)	(0·04)
glasso	15·63	2·13	2·35	0·48	5·35	2·86	1·90	0·76
	(0·15)	(0·03)	(0·02)	(0·02)	(0·02)	(0·05)	(0·03)	(0·03)
rr	20·35	8·00	0·00	4·00	6·02	8·00	0·00	4·00
	(0·18)	(0·00)	(0·00)	(0·00)	(0·03)	(0·00)	(0·00)	(0·00)
grr	15·74	2·59	2·12	0·71	5·36	3·21	1·76	0·97
	(0·11)	(0·04)	(0·03)	(0·02)	(0·02)	(0·05)	(0·03)	(0·03)
enet	16·70	4·84	1·19	2·03	5·49	5·50	0·83	2·34
	(0·14)	(0·06)	(0·03)	(0·04)	(0·03)	(0·05)	(0·02)	(0·04)
genet	15·64	2·21	2·31	0·52	5·34	2·94	1·86	0·80
	(0·12)	(0·04)	(0·02)	(0·02)	(0·02)	(0·05)	(0·03)	(0·03)
ilasso	15·63	2·13	2·35	0·48	5·35	2·87	1·89	0·76
	(0·15)	(0·04)	(0·02)	(0·02)	(0·02)	(0·05)	(0·03)	(0·03)
nng	15·83	2·76	1·99	0·75	5·30	3·46	1·49	0·95
	(0·11)	(0·04)	(0·03)	(0·03)	(0·02)	(0·05)	(0·03)	(0·03)

Example 3: median negative log-likelihood (NLL), mean model size (Size), mean number of false positive regressors (FP) and mean number of false negative regressors (FN) included in the selected model. Tests are based on 1000 iterations with standard errors included in parentheses

BRCA1 and BRCA2 Pathology Data (1)

- Women with early onset (< 40 years) breast cancer ($n = 426$)
- Predict BRCA1 and/or BRCA2 mutation carriers
  - BRCA1: 29 carriers, 397 non-carriers
  - BRCA2: 16 carriers, 410 non-carriers
- 14 predictors
  - age of diagnosis
  - pathology: pushing margins, mitosis, tubule form, etcetera.
  - family history

BRCA1 (2)

Variables	Methods						
	ML	subset	fwd	rr	lasso	enet	nng
proband age	-0.09	0	0	-0.09	-0.06	-0.06	0
num. mitosis	2.11	2.25	2.25	2.11	1.84	1.80	1.83
nuc. grade I	0.14	0	0	0.14	0	0	0
tubule form	-1.78	0	0	-1.78	0	0	0
synctial	-0.47	0	0	-0.47	0	0	0
pushing	-2.67	0	0	-2.67	0	0	0
circum	0.85	0	0	0.85	0.33	0.27	0
lyphocytic infiltrate	0.65	0	0	0.65	0	0	0
growth	2.54	2.37	2.37	2.54	1.68	1.61	1.91
necrosis	1.03	0	0	1.03	0	0	0
deg1 < 60	1.69	1.05	1.05	1.69	0.72	0.66	0.56
deg1 ≥ 60	0.46	0	0	0.47	0	0	0
deg2 nBreast	-0.01	0	0	-0.01	0	0	0
nOvary	1.14	0	0	1.14	0	0	0

BRCA2 (3)

Variables	Methods						
	ML	subset	fwd	rr	lasso	enet	nng
proband age	-0.09	0	0	-0.09	0	0	0
num. mitosis	-0.90	0	0	-0.95	0	0	0
nuc. grade I	1.11	0	0	1.12	0	0	0
tubule form	0.04	0	0	0.04	0	0	0
synctial	0.17	0	0	0.17	0	0	0
pushing	1.21	0	0	1.21	0	0	0
circum	0.21	0	0	0.21	0	0	0
lyphocytic infiltrate	-0.55	0	0	-0.54	0	0	0
growth	0.99	0	0	0.99	0	0	0
necrosis	-1.41	0	0	-1.40	0	0	0
deg1 < 60	0.31	0	0	0.31	0	0	0
deg1 ≥ 60	2.40	2.17	2.17	2.40	0	0	1.26
deg2 nBreast	0.13	0	0	0.13	0	0	0
nOvary	-427	0	0	-8.56	0	0	0

Predicting Methylation (4)

Variables	Methods						
	ML	subset	fwd	rr	lasso	enet	nng
proband age	-0.06	0	0	-0.06	-0.05	-0.05	0
num. mitosis	2.13	1.84	1.84	2.13	1.18	1.18	1.71
nuc. grade I	-0.51	0	0	-0.51	0	0	0
tubule form	-1.27	0	0	-1.27	0	0	0
synctial	0.02	0	0	0.02	0	0	0
pushing	-1.29	0	0	-1.29	0	0	0
circum	1.25	1.24	1.24	1.25	0.64	0.64	0.75
lyphocytic infiltrate	1.01	0	0	1.01	0	0	0
growth	0.96	0	0	0.96	0	0	0
necrosis	-0.59	0	0	-0.59	0	0	0
deg1 < 60	0.42	0	0	0.42	0	0	0
deg1 ≥ 60	-0.48	0	0	-0.48	0	0	0
deg2 nBreast	-0.04	0	0	-0.04	0	0	0
nOvary	1.65	0	0	1.64	0	0	0