Review of modern logistic regression methods

Enes Makalic Daniel F. Schmidt

Centre for MEGA Epidemiology
The University of Melbourne

23rd Australasian Joint Conference on Artificial Intelligence
2010

Outline

Outline

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

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

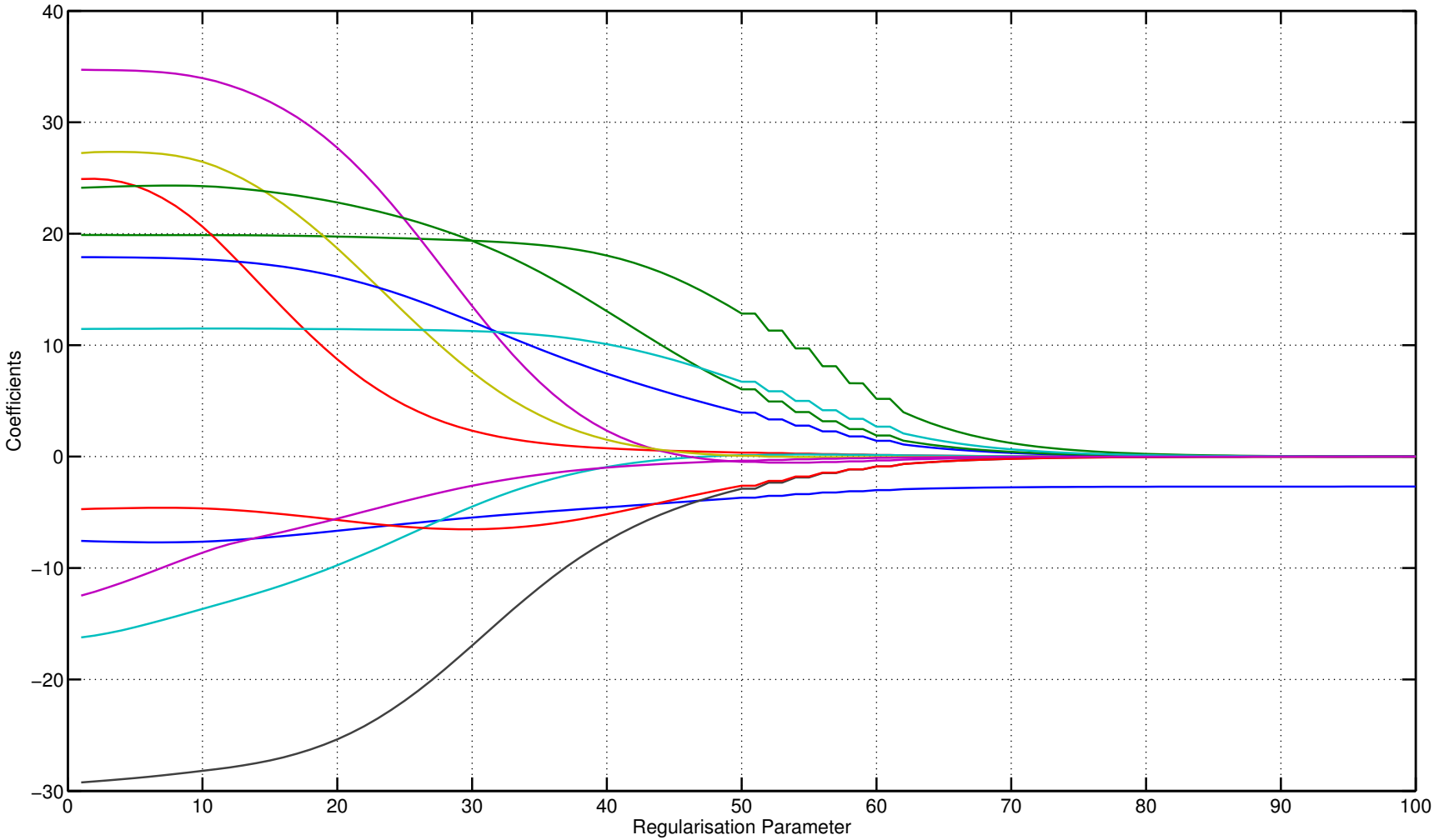
Ridge Regression (1)

- 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)



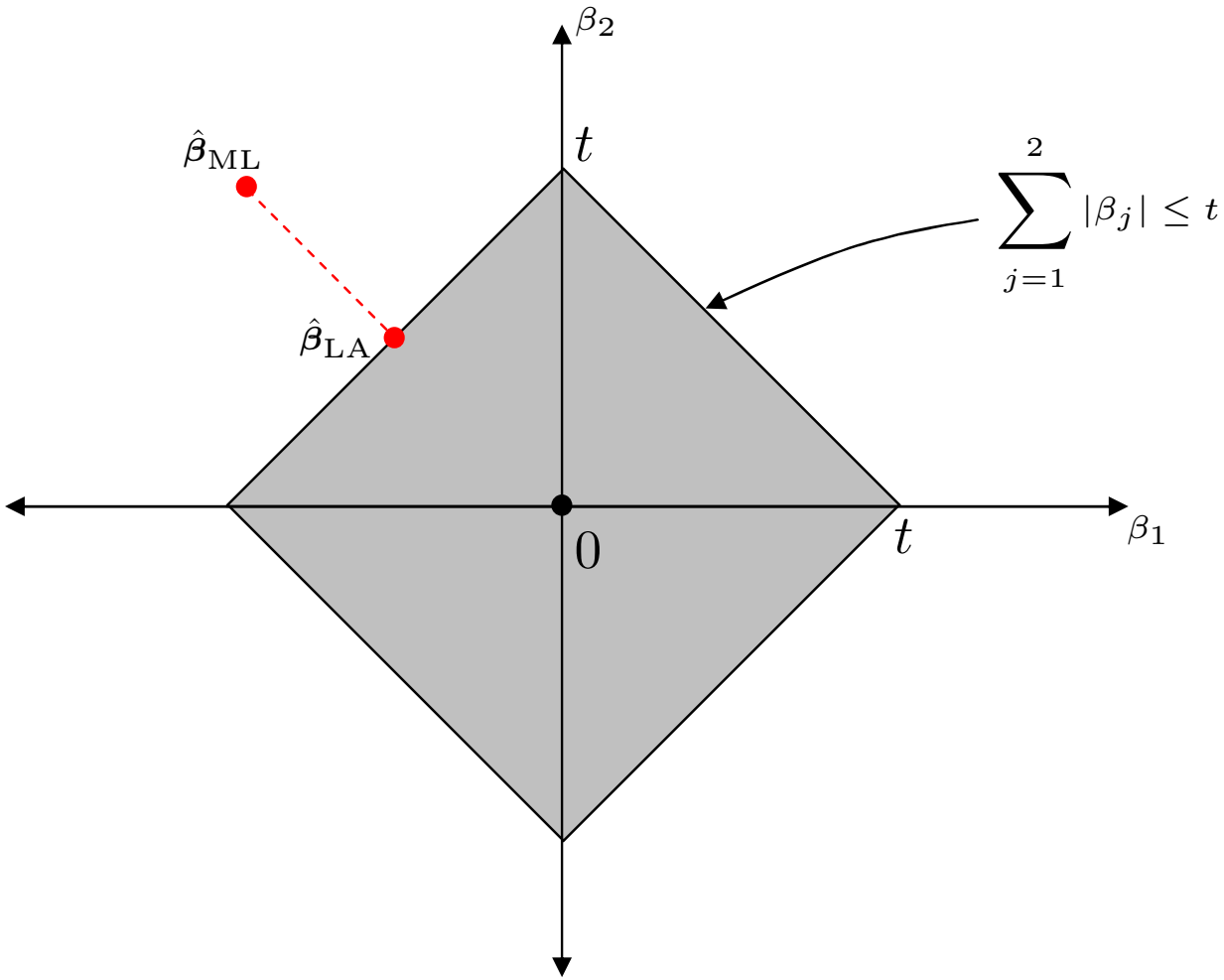
LASSO (1)

- Least absolute shrinkage and selection operator (LASSO)

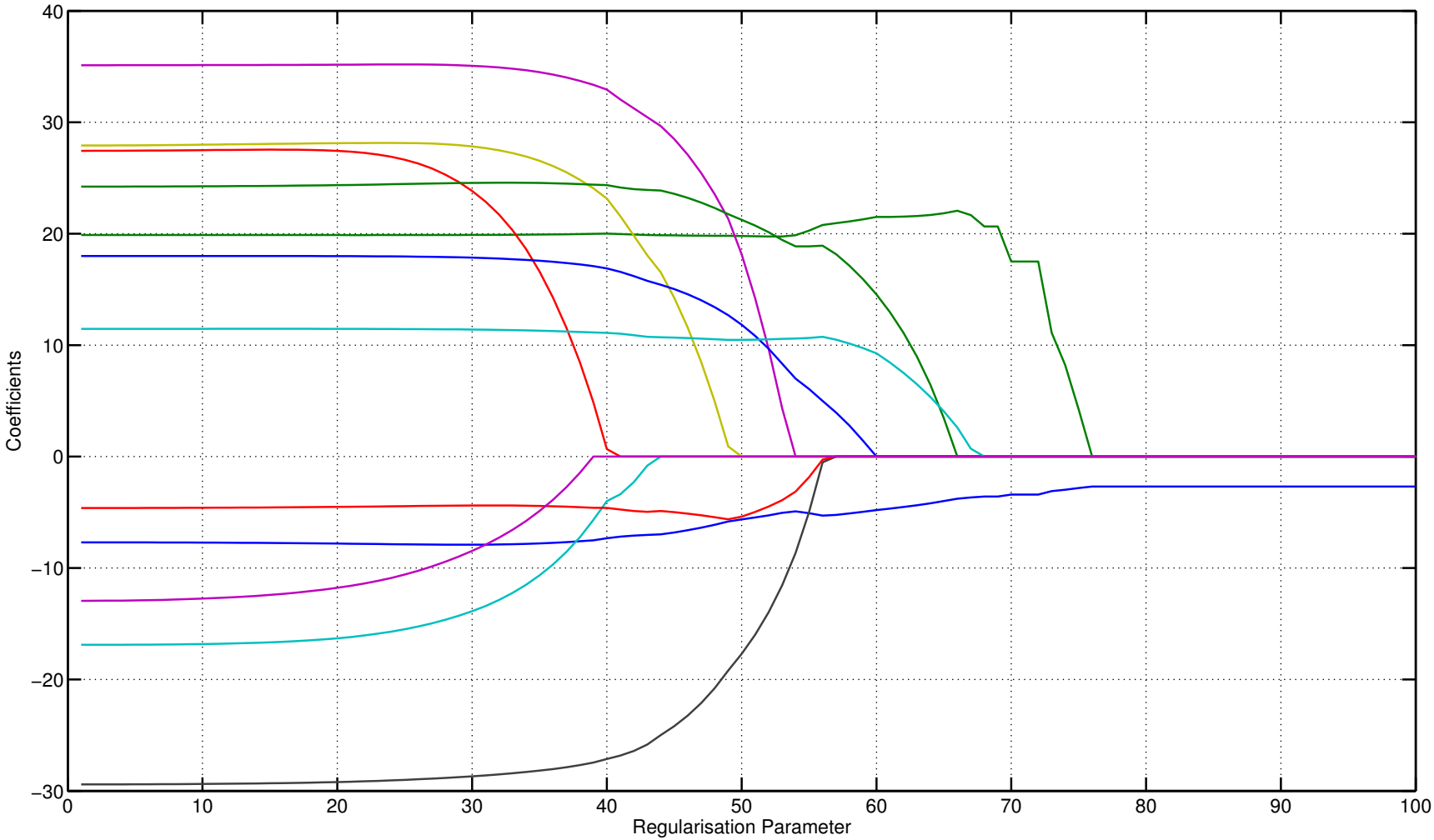
$$\hat{\boldsymbol{\beta}}_{\mathrm{LA}} = \arg\max_{\boldsymbol{\beta}} \{l(\boldsymbol{\beta})\} \quad \text{s.t.} \sum_{j=i}^{p} |\beta_j| \leq t$$

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

LASSO (2)


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

Elastic Net

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

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

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

Non-negative Garrote (1)

- Requires an initial parameter estimate $\boldsymbol{\beta}^*$
  - For example, maximum likelihood, ridge regression, etc.
- Non-negative Garrote (NNG) estimate

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

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

Outline

Simulated Data

- In all simulations ...
  - Sample size $n = \{20, 50, 100\}$.
  - Regressor correlation $\text{corr}(i, j) = 0.5^{|i-j|}$.
- 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: $\boldsymbol{\beta} = (5, 0{\cdot}5, 0{\cdot}5, 0{\cdot}5, 0, 0, 0, 0)'$.
- Test metrics
  - median negative log-likelihood (NLL), mean model size (Size), mean number of false positive / false negative regressors

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

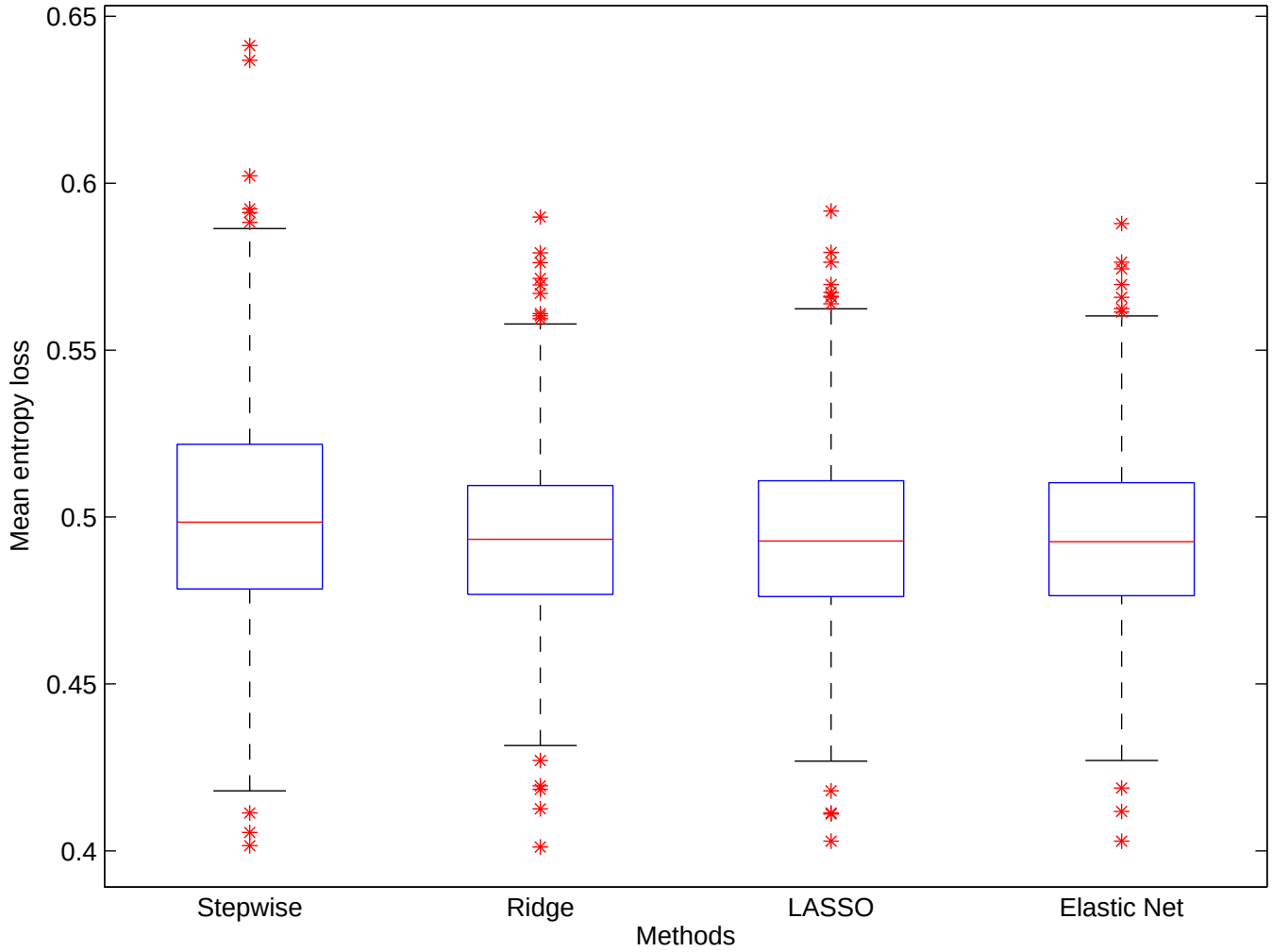
Methods	$n = 20$				$n = 50$			
	NLL	Size	FP	FN	NLL	Size	FP	FN
fwd	19·42	1·17	2·98	0·15	5·57	1·68	2·54	0·22
	(0·78)	(0·04)	(0·03)	(0·02)	(0·03)	(0·04)	(0·02)	(0·02)
gfwd	16·55	0·99	3·09	0·08	5·40	1·63	2·57	0·19
	(0·35)	(0·03)	(0·02)	(0·01)	(0·02)	(0·04)	(0·02)	(0·02)
lasso	16·70	4·08	1·49	1·57	5·45	5·08	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

Real data

- Obtained from the UCI Machine Learning repository
  - Pima Indian diabetes (250/250/268)
  - Wisconsin diagnostic breast cancer (100/100/369)
- Test metric: mean entropy (classification) loss on test data
- Repeated for 1000 iterations

Pima indian diabetes



Wisconsin diagnostic breast cancer