

# On homotopy algorithms in statistics

Berwin A Turlach

berwin@maths.uwa.edu.au

School of Mathematics and Statistics (M019)

The University of Western Australia

35 Stirling Highway

Crawley, WA 6009

Australia

# The LASSO

$$\underset{\beta \in \mathbb{R}^p}{\text{minimise}} \quad \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \quad (1a)$$

$$\text{subject to} \quad \|\beta\|_1 \leq t. \quad (1b)$$

where

- $\mathbf{y}$  is an  $n \times 1$  vector of responses,
- $\mathbf{X}$  is the  $n \times p$  design matrix; and
- $\beta$  is the  $p \times 1$  vector of parameters.

Tibshirani (1996)

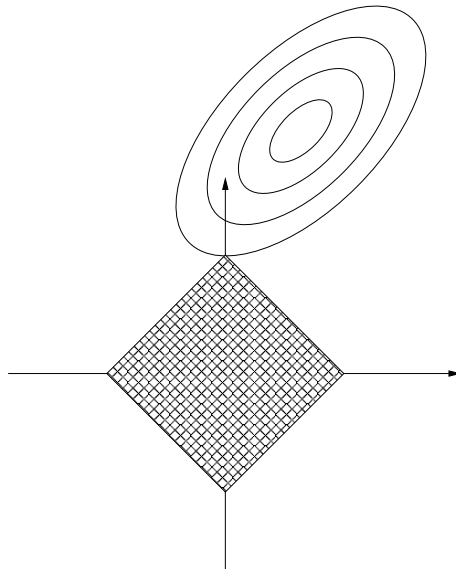
Santosa and Symes (1986)

Further work: Knight and Fu (2000), Osborne *et al.* (2000a,b), Huang (2003), Rosset and Zhu (2004a), Zou *et al.* (2004) ...

Wavelet literature: Chen *et al.* (1999), Sardy *et al.* (2000)

Related work: Fu (1998), Fan and Li (2001), ...

# The LASSO



# Characterisation of solutions

$\beta$  is a solution of (1) if  $\lambda \geq 0$  exists such that

$$\mathbf{X}'\mathbf{r} = \lambda\mathbf{v},$$

where  $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$  and  $\mathbf{v} = (v_1, \dots, v_p)'$  is such that

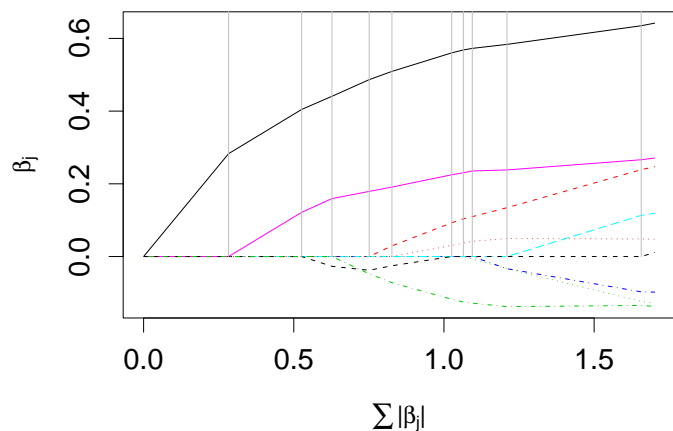
$$v_i = 1 \quad \text{if } \beta_i > 0$$

$$v_i = -1 \quad \text{if } \beta_i < 0$$

$$v_i \in [-1, 1] \quad \text{if } \beta_i = 0$$

Osborne *et al.* (2000a,b)

## The LASSO



Osborne *et al.* (2000a); Efron *et al.* (2004)

## Other homotopy algorithms

Support vector machines:

$$\begin{aligned} &\text{minimise}_{\beta_0, \beta} \quad \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \xi_i \\ &\text{subject to} \quad y_i(\beta_0 + \mathbf{x}_i^T \beta) \geq 1 - \xi_i, \quad i = 1, \dots, n \end{aligned}$$

Written in *Loss + Penalty* form:

$$\text{minimise}_{\beta_0, \beta} \sum_{i=1}^n \left[ 1 - y_i(\beta_0 + \beta^T \mathbf{x}_i) \right]_+ + \frac{\lambda}{2} \|\beta\|^2$$

Hastie *et al.* (2004)

See Lee and Cui (2005) for multiclass SVMs

## Other homotopy algorithms

1-norm vector machines:

$$\begin{aligned} &\text{minimise}_{\beta_0, \beta} \quad \sum_{i=1}^n \left[ 1 - y_i \left( \beta_0 + \sum_{j=1}^p \beta_j h_j(\mathbf{x}_i) \right) \right]_+ \\ &\text{subject to} \quad \|\beta\|_1 = \sum_{j=1}^p |\beta_j| \leq t \end{aligned}$$

Zhu *et al.* (2003)

“Huberised” LASSO:

$$\text{argmin}_{\beta} \sum_{|y_i + \beta^T \mathbf{x}_i| \leq 1} (y_i + \beta^T \mathbf{x}_i)^2 + \sum_{|y_i + \beta^T \mathbf{x}_i| > 1} 2(|y_i + \beta^T \mathbf{x}_i| - 0.5) + \lambda \|\beta\|_1$$

Rosset and Zhu (2004b)

## Multivariate regression

We have  $k$  response variables:

$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nk} \end{pmatrix} = \begin{pmatrix} | & & | \\ \mathbf{y}_1 & \cdots & \mathbf{y}_k \\ | & & | \end{pmatrix} = \begin{pmatrix} - & \mathbf{y}'_{(1)} & - \\ & \vdots & \\ - & \mathbf{y}'_{(n)} & - \end{pmatrix}$$

but one design matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

## Multivariate regression

With

$$\mathbf{B} = \begin{pmatrix} \beta_{11} & \dots & \beta_{1k} \\ \vdots & & \vdots \\ \beta_{p1} & \dots & \beta_{pk} \end{pmatrix} = \begin{pmatrix} | & & | \\ \boldsymbol{\beta}_1 & \dots & \boldsymbol{\beta}_k \\ | & & | \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\beta}'_{(1)} & - \\ \vdots & \\ -\boldsymbol{\beta}'_{(p)} & - \end{pmatrix}$$

our model is

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

Breiman and Friedman (1997)

## Characterisation of solutions

$\mathbf{B}$  is a solution of (2) if  $\lambda \geq 0$  exists such that

$$\mathbf{X}'\mathbf{R} = \lambda\mathbf{V}$$

where  $\mathbf{R} = \mathbf{Y} - \mathbf{XB}$  and  $\mathbf{V}$  has the following form:

- If  $\|\boldsymbol{\beta}_{(l)}\|_\infty = 0$ , then  $\|\mathbf{v}_{(l)}\|_1 \leq 1$ .
- If  $\|\boldsymbol{\beta}_{(l)}\|_\infty > 0$ , then  $\|\mathbf{v}_{(l)}\|_1 = 1$  and, for  $j = 1, \dots, k$ ,
  - $v_{lj} \geq 0$  if  $\beta_{lj} = \|\boldsymbol{\beta}_{(l)}\|_\infty$ ,
  - $v_{lj} \leq 0$  if  $\beta_{lj} = -\|\boldsymbol{\beta}_{(l)}\|_\infty$ ,
  - $v_{lj} = 0$  if  $|\beta_{lj}| \neq \|\boldsymbol{\beta}_{(l)}\|_\infty$ .

The solution of (2), as a function of  $t$ , is piecewise linear with breakpoints at  $0 = t_0 < t_1 < t_2 < \dots$

## Simultaneous variable selection

$$\underset{\mathbf{B} \in \mathbb{R}^{p \times k}}{\text{minimise}} \quad \frac{1}{2} \sum_{j=1}^k (\mathbf{y}_j - \mathbf{X}\boldsymbol{\beta}_j)' (\mathbf{y}_j - \mathbf{X}\boldsymbol{\beta}_j) \quad (2a)$$

$$\text{subject to} \quad \sum_{l=1}^p \|\boldsymbol{\beta}_{(l)}\|_\infty \leq t. \quad (2b)$$

T., Venables and Wright (2005)

## Homotopy algorithm for SVS

Assume we are at point  $t_s$  and we have the following quantities calculated:

- $\boldsymbol{\beta}_j^s$ ,  $j = 1, \dots, k$ , the estimated parameters,
- $\boldsymbol{\mu}_j^s = \mathbf{X}\boldsymbol{\beta}_j^s$ ,  $j = 1, \dots, k$ , the fitted values,
- $\mathbf{r}_j^s = \mathbf{y}_j - \boldsymbol{\mu}_j^s$ ,  $j = 1, \dots, k$ , the residuals,
- $\mathbf{c}_j^s = \mathbf{X}'\mathbf{r}_j^s$ ,  $j = 1, \dots, k$ , the correlations between the residuals and the explanatory variables; and
- $\boldsymbol{\theta}_j^s = \text{sign}(\mathbf{c}_j^s)$ ,  $j = 1, \dots, k$ , where the sign is taken component wise. ( $\text{sign}(0) = 0$ .)

## Homotopy algorithm for SVS

Furthermore, define

- $\sigma \subseteq \{1, \dots, p\}$  is such that  $l \in \sigma$  iff  $\|\beta_{(l)}^s\|_\infty > 0$ .
- The  $p \times |\sigma|$  matrices  $\mathbf{E}_{\sigma,j}$ ,  $j = 1, \dots, k$ , are defined as

$$\mathbf{E}_{\sigma,j} = (\cdots \theta_{lj}^s e_l \cdots)_{l \in \sigma}$$

where  $e_l \in \mathbb{R}^p$  is the  $l^{\text{th}}$  unit vector.

- $\sigma_j \subseteq \sigma$ ,  $j = 1, \dots, k$ , are such that  $l \in \sigma_j$  iff  $c_{lj}^s = 0$  (i.e.  $|\beta_{lj}^s|$  may differ from  $\|\beta_{(l)}^s\|_\infty$ ).
- The  $p \times |\sigma_j|$  matrices  $\mathbf{E}_{\sigma_j}$  are defined as

$$\mathbf{E}_{\sigma_j} = (\cdots e_l \cdots)_{l \in \sigma_j}$$

## Homotopy algorithm for SVS

To determine  $t_{s+1}$  we parameterise the  $\beta_j$ ,  $j = 1, \dots, k$  as follows ( $\tau > 0$ ):

$$\beta_j = \beta_j^s + \tau (\mathbf{E}_{\sigma,j} \Delta + \mathbf{E}_{\sigma_j} \delta_j), \quad j = 1, \dots, k. \quad (4)$$

Straightforward linear algebra yields

$$\delta_j = - \left( \mathbf{X}'_{\sigma_j} \mathbf{X}_{\sigma_j} \right)^{-1} \mathbf{X}'_{\sigma_j} \mathbf{X}_{\sigma,j} \Delta, \quad j = 1, \dots, k.$$

where  $\mathbf{X}_{\sigma_j} = \mathbf{X} \mathbf{E}_{\sigma_j}$  and  $\mathbf{X}_{\sigma,j} = \mathbf{X} \mathbf{E}_{\sigma,j}$ .

## Homotopy algorithm for SVS

Substituting  $\delta_j$ s back into (4) yields:

$$\Delta = \mathbf{A}^{-1} \mathbf{1}$$

where

$$\mathbf{A} = \sum_{j=1}^k \mathbf{X}'_{\sigma,j} (\mathbf{I} - \mathbf{H}_{\sigma_j}) \mathbf{X}_{\sigma,j}$$

and

$$\mathbf{H}_{\sigma_j} = \mathbf{X}_{\sigma_j} \left( \mathbf{X}'_{\sigma_j} \mathbf{X}_{\sigma_j} \right)^{-1} \mathbf{X}'_{\sigma_j}$$

## Homotopy algorithm for SVS

Now,

$$t_{s+1} = t_s + \tau_0$$

where  $\tau_0 > 0$  is the smallest value at which either  $\sigma$  or one of the  $\sigma_j$ s change.

1.  $\sigma$  decreases
2.  $\sigma$  increases
3. One of the  $\sigma_j$  increases
4. One of the  $\sigma_j$  decreases

• The algorithm starts at  $t_0 = 0$  with  $\sigma = \{l_0\}$ , where

$$l_0 = \operatorname{argmax}_{l=1, \dots, p} \|(\mathbf{X}'\mathbf{Y})_{(l)}\|_1$$

and, for  $j = 1, \dots, k$ ,  $\beta_j = \delta_j = \mathbf{0}$  and  $\sigma_j = \emptyset$ .

• The algorithm stops when

•  $\mathbf{X}'\mathbf{R} = \mathbf{0}$ ; or

•  $|\sigma| = p$ ; or

• ...

## References

- Breiman, L. and Friedman, J.H. (1997). Predicting multivariate responses in multiple linear regression (with discussion), *Journal of the Royal Statistical Society, Series B* **59**(1): 3–54.
- Chen, S.S., Donoho, D.L. and Saunders, M.A. (1999). Atomic decomposition by basis pursuit, *SIAM Journal on Scientific Computing* **20**(1): 33–61.  
URL: <http://www-stat.stanford.edu/~donoho/Reports/1995/30401.pdf>
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion), *Annals of Statistics* **32**(2): 407–499.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456): 1348–1360.
- Fu, W.J. (1998). Penalized regression: The Bridge versus the Lasso, *Journal of Computational and Graphical Statistics* **7**(3): 397–416.
- Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2004). The entire regularization path for the support vector machine., *Journal of Machine Learning Research* **5**: 1391–1415.
- Huang, F. (2003). Prediction error property of the lasso estimator and its generalization, *Australian & New Zealand Journal of Statistics* **45**(2): 217–228.
- Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators, *Annals of Statistics* **28**(5): 1356–1378.
- Lee, Y. and Cui, Z. (2005). Characterizing the solution path of multicategory support vector machines, *Technical Report 754*, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210–1247.

2003), Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA.

URL: <http://www-stat.stanford.edu/~hastie/Papers/svm1.ps>

Zou, H., Hastie, T. and Tibshirani, R. (2004). On the "degrees of freedom" of the lasso, *unpublished manuscript*, Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA.

URL: <http://www-stat.stanford.edu/~hastie/Papers/dflasso.pdf>

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000a). A new approach to variable selection in least squares problems, *IMA Journal of Numerical Analysis* **20**(3): 389–403.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000b). On the LASSO and its dual, *Journal of Computational and Graphical Statistics* **9**(2): 319–337.

Rosset, S. and Zhu, J. (2004a). Corrected proof of the result of 'a prediction error property of the Lasso estimator and its generalization' by Huang (2003), *Australian & New Zealand Journal of Statistics* **46**(3): 5005–510.

Rosset, S. and Zhu, J. (2004b). Piecewise linear regularized solution paths, *unpublished manuscript*, Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA.

URL: <http://www-stat.stanford.edu/~saharon/papers/piecewise.ps>

Santosa, F. and Symes, W.W. (1986). Linear inversion of band-limited reflection seismograms, *SIAM Journal on Scientific and Statistical Computing* **7**(4): 1307–1330.

Sardy, S., Bruce, A.G. and Tseng, P. (2000). Block coordinate relaxation methods for nonparametric signal denoising with wavelet dictionaries, *Journal of Computational and Graphical Statistics* **9**(2): 361–379.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B* **58**(1): 267–288.

Turlach, B.A., Venables, W.N. and Wright, S.J. (2005). Simultaneous variable selection, *Technometrics* **47**(3): 349–363.

Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2003). 1-norm support vector machines, *unpublished manuscript (accepted spotlight poster NIPS*