Simultaneous Variable Selection

Berwin A Turlach
berwin@maths.uwa.edu.au
School of Mathematics and Statistics
The University of Western Australia

joint work with William N Venables and Stephen J Wright

Overview

Multivariate regression

The LASSO

Simultaneous variable selection

Homotopy algorithm (Complete solution path)

Example

Concluding remarks

References

Multivariate regression

We have n observations on k response variables:

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

and p regressor variable, i.e. our design matrix is

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} | & & | \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_p \\ | & & | \end{pmatrix}$$

W.l.o.g. columns of $\mathbf{Y}$ and $\mathbf{X}$ are centred and standardised.

Multivariate regression (cont.)

With

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

our model is

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

Mardia *et al*. (1979)
Breiman and Friedman (1997)
Brown *et al*. (1998, 1999, 2002)

LASSO

For multiple linear regression ($k = 1$):

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

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

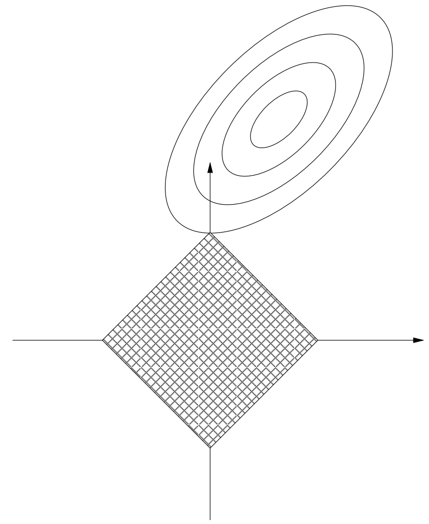
where

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

Santosa and Symes (1986), Tibshirani (1996)

Further work: Knight and Fu (2000), Osborne *et al*. (2000a,b), Huang (2003), Rosset and Zhu (2004), Zou *et al*. (2004), Zou (2006) . . .
Wavelet literature: Chen *et al*. (1999), Sardy *et al*. (2000),. . .
Related work: Fu (1998), Fan and Li (2001), . . .

LASSO (cont.)



Simultaneous variable selection

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

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

$\alpha = \infty$: T., Venables and Wright (2005)
$\alpha = 2$: Bakin (1999), Yuan and Lin (2006)

Other approaches: Fused LASSO (Tibshirani *et al.*, 2005), Elastic Net (Zou and Hastie, 2005)

Characterisation of solution $\alpha = \infty$

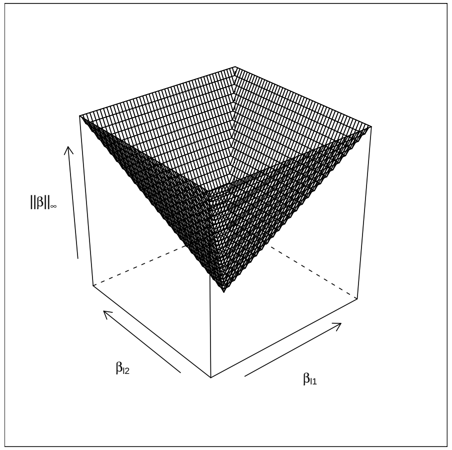
$\mathbf{B}$ is a solution of (2) if $\lambda \geq 0$ exists such that

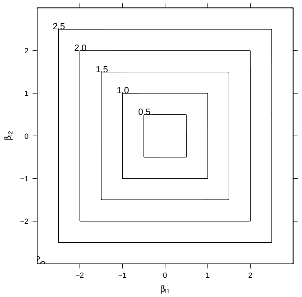
$$\mathbf{X}'\mathbf{R} = \lambda\mathbf{V}$$

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

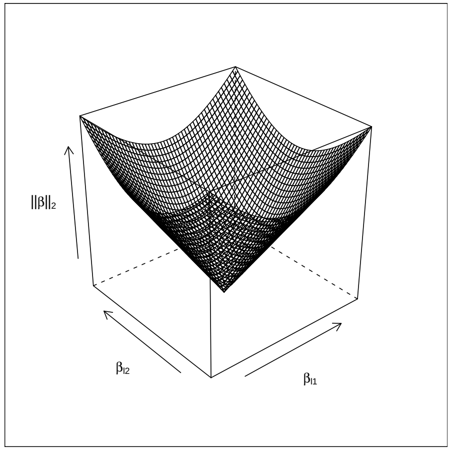
- If $\|\boldsymbol{\beta}_{(l)}\|_\infty = 0$, then $\|\boldsymbol{v}_{(l)}\|_1 \leq 1$.
- If $\|\boldsymbol{\beta}_{(l)}\|_\infty > 0$, then $\|\boldsymbol{v}_{(l)}\|_1 = 1$ and, for $j = 1, \ldots, k$,
  - $v_{lj} \geq 0$ if $\beta_{lj} = \|\boldsymbol{\beta}_{(l)}\|_\infty$,
  - $v_{lj} \leq 0$ if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

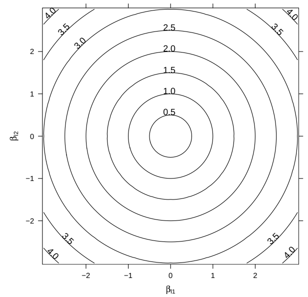
Characterisation of solution $\alpha = \infty$ (cont.)





Characterisation of solution $1 < \alpha < \infty$ (cont.)





Characterisation of solution $1 < \alpha < \infty$

$\mathbf{B}$ 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 $\|\boldsymbol{\beta}_{(l)}\|_\alpha = 0$, then $\|\boldsymbol{v}_{(l)}\|_\gamma \leq 1$.
- If $\|\boldsymbol{\beta}_{(l)}\|_\alpha > 0$, then $\|\boldsymbol{v}_{(l)}\|_\gamma = 1$.

where

$$\frac{1}{\alpha} + \frac{1}{\gamma} = 1$$

Orthonormal Design

For an orthonormal design ($\mathbf{X}'\mathbf{X} = \mathbf{I}$), T., Venables and Wright (2005)

- describe an algorithm that computes the complete solution path.
- show that, if $\boldsymbol{\beta}^0_{(l)}$ denotes the unconstrained solutions ($\mathbf{B}^0 = \mathbf{X}'\mathbf{Y}$), then their approach essentially orders the variables such that

$$\|\boldsymbol{\beta}^0_{(l_1)}\|_1 \geq \|\boldsymbol{\beta}^0_{(l_2)}\|_1 \geq \|\boldsymbol{\beta}^0_{(l_3)}\|_1 \geq \cdots \geq \|\boldsymbol{\beta}^0_{(l_{p-1})}\|_1 \geq \|\boldsymbol{\beta}^0_{(l_p)}\|_1,$$

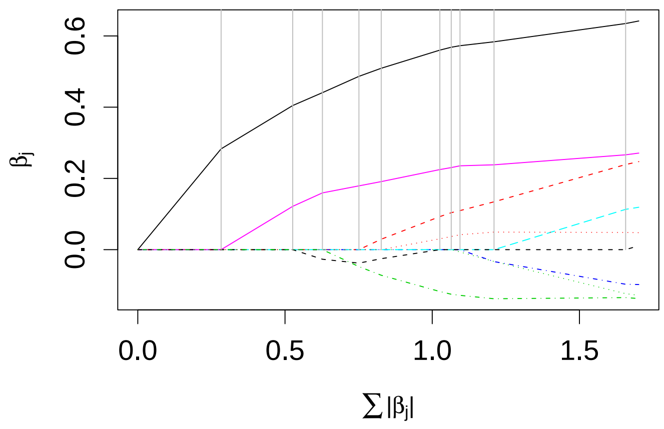
and then selects the variables $\boldsymbol{x}_{l_1}, \boldsymbol{x}_{l_2}, \ldots, \boldsymbol{x}_{l_m}$, where m depends on t, using this ordering.

General Design

For a general design ($\mathbf{X}'\mathbf{X} \neq \mathbf{I}$), T., Venables and Wright (2005)

- develop an interior point algorithm that computes the solution of (2) for a given t.
- their algorithm, by some clever linear algebra, is able to deal efficiently with the $p \gg n$ case.

LASSO



Osborne *et al*. (2000a), Efron *et al*. (2004), T. (2005)

Homotopy algorithm

If $\alpha = \infty$, then the solution of (2), as a function of t, is piecewise linear and continuous with breakpoints at $0 = t_0 < t_1 < t_2 < \dots$.

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

Characterisation of solution $\alpha = \infty$

REMEMBER: If $\alpha = \infty$, then $\mathbf{B}$ is a solution of (2) if $\lambda \geq 0$ exists such that

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

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

- If $\|\boldsymbol{\beta}_{(l)}\|_\infty = 0$, then $\|\boldsymbol{v}_{(l)}\|_1 \leq 1$.
- If $\|\boldsymbol{\beta}_{(l)}\|_\infty > 0$, then $\|\boldsymbol{v}_{(l)}\|_1 = 1$ 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 $|\beta_{lj}| \neq \|\boldsymbol{\beta}_{(l)}\|_\infty$.

Homotopy algorithm (cont.)

Furthermore, define

- $\sigma \subseteq \{1, \ldots, p\}$ is such that $l \in \sigma$ iff $\|\boldsymbol{\beta}_{(l)}\|_\infty > 0$, for $t = t_s + \tau$ and (small) $\tau > 0$.
- $\sigma_j \subseteq \sigma$, $j = 1, \ldots, k$, are such that $l \in \sigma_j$ iff $c_{lj} = 0$ (i.e. $|\beta_{lj}|$ may differ from $\|\boldsymbol{\beta}_{(l)}\|_\infty$), for $t = t_s + \tau$ and (small) $\tau > 0$.

- The $p \times |\sigma|$ matrices $\mathbf{E}_{\sigma,j}$, $j = 1, \ldots, k$, are defined as

$$\mathbf{E}_{\sigma,j} = (\cdots \theta^s_{lj} \boldsymbol{e}_l \cdots)_{l \in \sigma}$$

  where $\boldsymbol{e}_l \in \mathbb{R}^p$ is the l^{th} unit vector.
- The $p \times |\sigma_j|$ matrices $\mathbf{E}_{\sigma_j}$ are defined as

$$\mathbf{E}_{\sigma_j} = (\cdots \boldsymbol{e}_l \cdots)_{l \in \sigma_j}$$

Homotopy algorithm (cont.)

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

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}^s_j + \tau \left(\mathbf{E}_{\sigma,j} \boldsymbol{\Delta} + \mathbf{E}_{\sigma_j} \boldsymbol{\delta}_j, \right) \qquad j = 1, \ldots, k. \qquad (4)$$

Straightforward linear algebra yields

$$\boldsymbol{\delta}_j = -\left(\mathbf{X}'_{\sigma_j} \mathbf{X}_{\sigma_j} \right)^{-1} \mathbf{X}'_{\sigma_j} \mathbf{X}_{\sigma,j} \boldsymbol{\Delta}, \qquad j = 1, \ldots, k.$$

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

Homotopy algorithm (cont.)

Substituting $\boldsymbol{\delta}_j$s back into (4) yields:

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

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 (cont.)

Now,

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

where $\tau_0 > 0$ is the smallest value at which either σ or one of the σ_js change.

1. σ decreases:
   Can only happen if a component of $\boldsymbol{\Delta}$ is negative.
2. σ increases:
   Have to check where a linear function intersects various continuous, convex, piecewise linear functions.
3. One of the σ_j increases:
   Happens if a v_{lj}, which change linearly, becomes zero.
4. One of the σ_j decreases:
   If a v_{lj} is zero, we have to check the rate with which the corresponding β_{lj} changes (linearly) against the rate with which $\|\boldsymbol{\beta}_{(l)}\|_\infty$ changes (linearly).

Homotopy algorithm (cont.)

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

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

  and, for $j = 1, \ldots, k$, $\boldsymbol{\beta}_j = \boldsymbol{\delta}_j = \mathbf{0}$ and $\sigma_j = \emptyset$.
- The algorithm stops when
  - $\mathbf{X}'\mathbf{R} = \mathbf{0}$; or
  - $|\sigma| = p$; or
  - $\ldots$

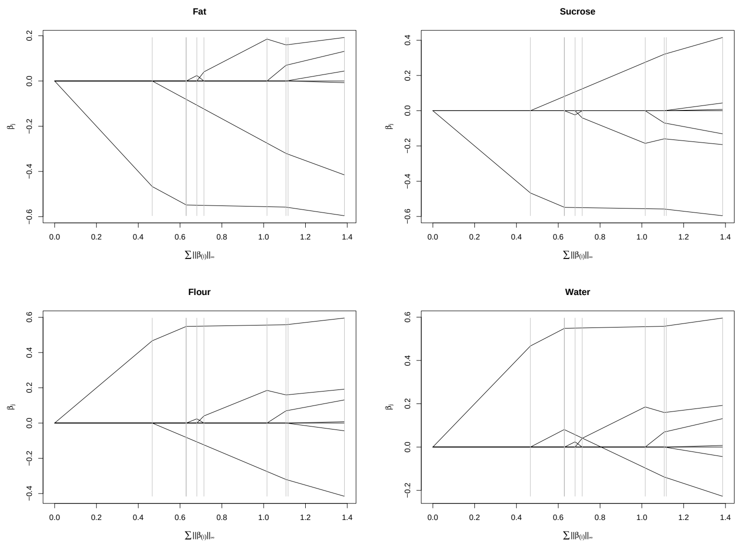
Biscuit dough piece data

The experiment involved varying the composition of biscuit dough pieces.
The calibration data set has

- $k = 4$ response variables; namely the percentage of fat, sucrose, flour and water in the dough,
- $p = 700$ regressor variables; NIR spectral data where the spectral range is 1100–2498nm in steps of 2nm, and
- $n = 40$ observations.

Brown *et al*. (1999, 2001)

Biscuit dough piece data (cont.)



Concluding remarks

Some open questions:

- How should we choose t?
- Would it be more appropriate to use

$$\sum_{l=1}^{p} \|\boldsymbol{\beta}_{(l)}\|_2$$

  as constraint in (2b)? Or any other α norm?
- How to calculate $\mathbf{A}$ efficiently?
- Add an l_2 constraint, à la elastic net (Zou and Hastie, 2005)?
- Add weights

$$\sum_{l=1}^{p} w_i \|\boldsymbol{\beta}_{(l)}\|_1$$

  in the penalty, à la adaptive lasso (Zou, 2006)?
- $\ldots$

References

S. Bakin. *Adaptive Regression and Model Selection in Data Mining Problems*. PhD thesis, Australian National University, Canberra ACT 0200, Australia, 1999

L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression (with discussion). *Journal of the Royal Statistical Society, Series B*, 59(1):3–54, 1997

P. J. Brown, T. Fearn, and M. Vannucci. The choice of variables in multivariate regression: A non-conjugate bayesian decision approach. *Biometrika*, 86(3):635–648, 1999

P. J. Brown, T. Fearn, and M. Vannucci. Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96:398–408, 2001

P. J. Brown, M. Vannucci, and T. Fearn. Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, 60(3):627–641, 1998

References (cont.)

P. J. Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B*, 64(3):519–536, 2002

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *Annals of Statistics*, 32(2):407–499, 2004

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001

W. J. Fu. Penalized regression: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998

F. Huang. Prediction error property of the Lasso estimator and its generalization. *Australian & New Zealand Journal of Statistics*, 45(2):217–228, 2003

References (cont.)

K. Knight and W. Fu. Asymptotics for lasso–type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, Duluth, London, 1979

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

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

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

References (cont.)

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

S. Rosset and J. Zhu. 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):505–510, 2004

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

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, 67(1):91–108, 2005

B. A. Turlach. On algorithms for solving least squares problems under an l_1 penalty or an l_1 constraint. In *2004 Proceedings of the American Statistical Association*, pages 2572–2577, Alexandria, VA, 2005. Statistical Computing Section [CD-ROM], American Statistical Association

References (cont.)

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

M. Yuan and Y. Lin. Model selection and estimation in regression withrouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2006

H. Zou. The adaptive lasso and its oracle propertiesd. *Journal of the American Statistical Association*, 2006. To appear

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005

H. Zou, T. Hastie, and R. Tibshirani. On the “degrees of freedom” of the lasso. Unpublished manuscript, Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA, 2004