LASSO for Clustered Data

Rosanna Overholser* Ronghui Xu†

Abstract
The LASSO was introduced by Tibshirani for the purposes of estimation and variable selection in linear regression. Most work on the LASSO has included the assumption of independent observations. Several papers have recently extended the LASSO to linear mixed models for clustered data. We will examine through simulations a further extension of the LASSO to general linear mixed models for clustered data that contain within-cluster correlation. Regression splines for correlated data can be formulated as a general linear mixed model so the problem of knot selection for splines is equivalent to variable selection of fixed effects. We can therefore use the LASSO to simultaneously select knots and estimate variance parameters. We apply our methods to functional MRI time courses from several subjects.

Key Words: LASSO, general linear mixed model, regression splines, correlated data, knot selection, fMRI.

1. Introduction

1.1 Background on Lasso

The LASSO (least absolute shrinkage and selection operator) was introduced by Tibsharini (1996) for the purposes of estimation and variable selection in linear regression. In the usual linear regression model $y = X\beta + \epsilon$ the parameter β is estimated by minimizing the residual sum of squares and a penalty term:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Model selection is performed by the choice of the parameter λ in the penalty term: as λ increases from zero, the number of covariates in the model may change, since $\hat{\beta}_j$'s are allowed to be set to exactly zero. A discussion of the LASSO in comparison with other shrinkage techniques is presented by Hastie, Tibshirani and Friedman (2009). Knight and Fu (2000) studied the asymptotics of LASSO estimates in the context of linear regression with iid errors. There is some cost to using a penalty: Zou (2006) showed that lasso can lead to inconsisent estimates of β in some situations and suggested an 'adaptive' LASSO as a means of bias reduction.

While much of the theory of the LASSO is derived under the assumption of a parametric model, the LASSO has been applied to non-parametric problems: the problem of knot selection in regression splines by Osbourne, Presnell and Turlach (1998) is one example and some theory is presented in Bḧlmann and van de Geer (2011).

Limited results have been obtained for correlated data. Bunea and Gupta (2010) studied the asymptotics of LASSO estimates in the presence of correlated errors and a large number

*Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, MC 0112, La Jolla, CA 92093-0112

†Department of Mathematics, Department of Family and Preventative Medicine, University of California, San Diego, 9500 Gilman Drive, MC 0112, La Jolla, CA 92093-0112

of covariates. In particular, they study the following estimate of β:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ (y - X\beta)^T \hat{R}^{-1} (y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where the errors ϵ are assumed to be $N(0, R)$ and $\hat{R}$ is assumed to be consistent estimator of R. Wang, Li and Tsai (2007) study the problem of joint selection of covariates and type of autogressive process via the LASSO for linear regression with autoregressive errors.

Mixed effects models are a common device for exploring correlated data: between-cluster correlations, such as those formed by repeated measures on the several subjects, are modeled by random effects. Two applications of LASSO for selection of both random and fixed effects have recently been proposed. Ibrahim et al. (2010) used one penalty to select fixed effects and another penalty on the cholesky decomposition of the random effects covariance matrix to select random effects. Their method extends to generalized linear mixed models. Bondell, Krishna and Ghosh (2010) purposed methods similar to that of Ibrahim et al (2010). In both works, the variance components of the model are estimated using REML and the number of covariates is less than the number of datapoints. Most recently, Schelldorfer, Bühlmann and van de Geer (2011) suggested a LASSO for linear mixed models but allowed the number of covariates to be much larger than the number of observations. Their LASSO is appropriate for situations where it is known which covariates will be used as random effects, such as the random intercept model.

In this paper, we will consider clustered data with both between and within-cluster correlation. While a parametric model will be assumed for the within-cluster correlation, we will not use a parametric mean structure. Instead, a regression spline to estimate the mean; the LASSO will be used to select the knots. Unlike previous work on mixed effect models, we will examine the situation where the number of observations per cluster is much larger than the number of clusters.

2. Method of Estimation

2.1 Notation

For clusters $i = 1 \dots m$, let $y_i = [y_i(t_{i1}), \dots, y_i(t_{in_i})]^T$ be a vector of data at n_i measurement times. Assume that each $y_i(t)$ follows the model

$$y_i(t) = \mu(t) + d_i(t) + \epsilon_i(t) \tag{1}$$

where $\mu(t)$ is a smooth function of t, $d_i(t)$ is a cluster specific deviation from $\mu(t)$ and $\epsilon_i = [\epsilon_i(t_{i1}), \dots, \epsilon_i(t_{in_i})]^T$ is distributed $N(0, \sigma^2 R_i)$ for some correlation matrix R_i.

For simplicity, we consider the case where $n_i = n$, $t_{ij} = (1, \dots, n)^T$ and $R_i = R_1(\rho)$, for all $i = 1, \dots, m$, $j = 1, \dots, n$, where R_1 is known up to some $r \times 1$ vector of parameters ρ. In addition, we assume $d_i(t_{ij}) = b_i$ where b_i are independently drawn from $N(0, D_1)$ for $i = 1, \dots, m$.

2.2 Criterion

We estimate the function $\mu(t)$ over the interval $[1, n]$ by a linear combination of basis functions. We chose the cubic truncated power basisfor ease of presentation; in practice, B-splines might be preferred for their computational stablity. Denote the basis function by $\{1, t, t^2, t^3, (t - \tau_1)_+^3, \dots, (t - \tau_K)_+^3\}$ where $(x)_+$ is x if $x > 0$ and is 0 otherwise. The points $\tau_1, \dots \tau_K$ are called knots; the selection of the number and position of knots are

essential to obtaining a good estimate of $\mu(t)$: we start with a large K and evenly space the knots in the interval $[1, n]$: from this initial set of basis functions we simultaneously estimate μ and reduce the number of knots used by placing an L_1 penalty on the elements of β that correspond to basis functions with knots. Our model then takes the form $y = X\beta + Zb + \epsilon$ where

$$X_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & t_{i1}^3 & (t_{i1}-\tau_1)_+^3 & \dots & (t_{i1}-\tau_K)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & t_{in_i} & t_{in_i}^2 & t_{in_i}^3 & (t_{in_i}-\tau_1)_+^3 & \dots & (t_{in_i}-\tau_K)_+^3 \end{bmatrix},$$

$X = (X_1, \dots, X_m)^T$, $\beta = (\beta_0, \dots, \beta_{K+3})^T$, $Z = I_m \otimes \mathbf{1_n}$, $b = (b_1, \dots, b_m)^T$, $\epsilon = (\epsilon_1^T, \dots, \epsilon_m^T)^T$and $y = (y_1^T, \dots, y_m^T)^T$. To make the distinction between unpenalized and penalized elements clear, we partition β as $\beta^T = (\beta_u^T, \beta_p^T)^T$ and $X = (X_u, X_p)$ so that $X\beta = X_u\beta_u + X_p\beta_p$. Let $D = \text{diag}(D_1, \dots, D_1)$, $R(\rho) = \text{diag}(R_1(\rho), \dots, R_1(\rho))$ and $V = R + ZDZ^T$.

Under the model in (1), the marginal log likelihood of y is (up to a constant)

$$l(y|\mu, V) = -\frac{1}{2}\log|V| - \frac{1}{2}(y-\mu)^T V^{-1}(y-\mu). \tag{2}$$

Assuming that a β exists so that the linear combination of basis functions, $X\beta$, is close to μ, we replace μ in (2) by $X\beta$ and choose $\hat{\mu} = X\hat{\beta}$ by

$$(\hat{\beta}, \hat{V}) = \underset{\beta, V}{\text{argmin}} \left\{ \frac{1}{2}\log|V| + \frac{1}{2}(y-X\beta)^T V^{-1}(y-X\beta) + \lambda \sum_{k=1}^{K} |\beta_{pk}| \right\}. \tag{3}$$

Following Osbourne, Presnell and Turlach (1998), we penalize β to prevent overfitting of the data. Once the estimates of μ and V are obtained, predictions of the random effects b can be by maximizing the 'joint' likelihood of μ and b with μ replaced by $X\beta$. This maximization results in

$$\hat{b} = \hat{D}Z^T\hat{V}^{-1}(y - X\hat{\beta}).$$

We choose λ to minimize BIC, as in Schelldorfer, Bühlmann and van de Geer (2011).

2.3 Computational Aspects

The criterion may be non-convex in β and the parameters in V. Following Wang, Li and Tsai (2007), we iterate between

$$(\hat{\beta}|V) = \underset{\beta}{\text{argmin}} \left\{ (y-X\beta)^T V^{-1}(y-X\beta) + \lambda \sum_{k=1}^{K} |\beta_{k+3}| \right\} \tag{4}$$

and

$$(\hat{V}|\beta) = \underset{V}{\text{argmin}} \left\{ \frac{1}{2}\log|V| + \frac{1}{2}(y-X\beta)^T V^{-1}(y-X\beta) \right\} \tag{5}$$

until convergence. In each step, $\hat{\beta}$ may be obtained from (4) using any of the standard methods for LASSO in linear regression models (LARS, cyclic or greedy coordinate descent, homotopy), The covariance parameters may be obtained from (5) by numerical optimatization. An alternative to iteration between (4) and (5) would be an EM algorithm as in Ibrahim et al. (2010) and Bondell, Krishna and Ghosh (2010).

3. Simulations

For each of 100 simulations, we generated $m = 10$ curves, each of $n_i = 50$ points in $[0, 1]$. The generating model had overall mean $\mu(t) = a \sin(2\pi t)$, within curve errors ϵ_i from $N(0, \mathrm{R}_1(0.4))$ and curve specific deviations b_i independently drawn from $N(0, 25)$ where a was chosen to give a signal to noise ratio of 7 for the overall mean curve and $R_1(0.4)$ is the autocorrelation matrix of a first order autoregressive process with parameter 0.4.

We fit a regression spline model ref with 49 knots evenly spaced in the interval $[0, 1]$. We parameterized the covariance parameters so that unconstrained optimization might be used and performed the iterate in section with the R packages LARS and nonop for the basis coefficent and covariance parameter estimation, respectively. The criterion used to determine convergence was sum of absolute values of differences in parameters less than 10^{-4}.

The estimation of μ from three randomly selected simulation runs with $m = n = 10$ are shown in Figure 1. Table 1 shows the mean (sd) of the covariance parameters from the 100 simulation runs. Note that increasing the number of clusters from 10 to 50 improved estimation of three parameters while increasing the number of observations per cluster from 50 to 100 had less of an effect.

Figure 1: The true mean, μ, in black and estimated mean, $X\hat{\beta}$, in green from three randomly selected simulation runs with $m = n = 10$. When λ was chosen to minimize BIC, the number of knots was reduced from 49 to 18, 10, and 6, from left to right.

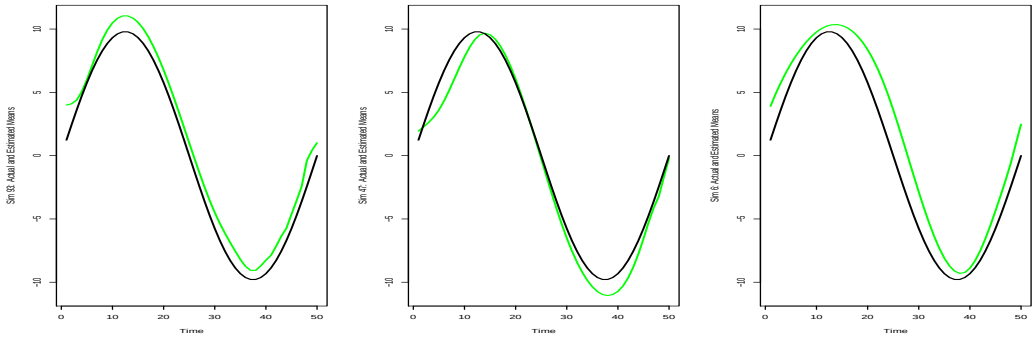


Table 1: The effect of increasing either the cluster size ($n = 50$ to 100) or number of clusters ($m = 10$ to 50) on the mean (sd) of the estimates of random effect variance σ_b^2, the error variance σ_ϵ^2 and the correlation ρ.

(m, n)	$\sigma_b^2 = 25$	$\sigma_\epsilon^2 = 1$	$\rho = 0.4$
$(10, 50)$	22.0 (3.1)	1.27 (0.08)	0.43(0.02)
$(10, 100)$	21.6 (3.2)	1.12 (0.04)	0.44 (0.01)
$(50, 50)$	24.4 (3.3)	1.00 (0.03)	0.39 (0.01)

4. fMRI Data

Functional MRI is a popular method of estimating brain activity by measuring blood flow to the brain. The Center for Functional MRI at UCSD performed a study to examine the effect of caffeine on the blood oxygenation level dependent (BOLD) signal from fMRI sessions (Rack-Gomer, Liau, and Liu, 2009). The study had 11 subjects, but 2 were dropped due to head movement during the scans. A block design was used to observe fingertapping: after an initial period of 20 seconds, the subjects were told to alternate fingertapping (30 seconds) and not fingertapping (30 seconds) for five cycles. The BOLD signal was measured every 2 seconds and the first 4 seconds were dropped from each scan, giving a total of 156 time points for the duration of each scan. These 156 points of 2 second intervals will be referred to as the 'time' variable. The block design was performed twice for each subject, once for a 'pre-caffience' session and again after ingested 200 mg of caffeine (the 'post-caffeine' session). During fingertapping periods, the voxels in the motor-cortex region of the brain become 'activated' - more oxygen was sent to this part of the brain. Following standard techinques in the field, the voxels that were significantly 'activated' in the motor cortex for each subject were selected.

In the analysis of the pre and post caffeine scans over all subjects for the fingertapping sessions, Rack-Gomer, Liau, and Liu (2009) compared four measures. They found a significant difference in 1) time to reach 50% of peak response and 2) time to fall to 50% of peak response but not a significant difference in 3) the full width-half maximum (difference between the previous two times) or 4) the maximum amplitude of the BOLD response.

For each session type and subject, we averaged over the BOLD signal of activated voxels at each timepoint. This average will be referred to as the 'signal' and is shown in Figure 2. We used regression splines with a cubic truncated basis to model the difference between pre and post-caffeine signals. We assumed each time series had AR(1) errors, as that was the structure used in selecting significantly activated voxels in Rack-Gomer, Liau, and Liu, (2009). The result is shown in Figure 3, where shading denotes periods of fingertapping. Note that the smallest and largest signal differences between pre- and post-sessions occur directly after the start and end, respectively, of each of the five fingertapping periods. This suggests that post-caffeine signals increased and decreased, on average, more rapidly then pre-caffeine signals for each fingertapping block.

5. Discussion

In this paper, we have combined several methods to obtain a non-parametric estimate of a mean from correlated data. The correlation is counted as either between-cluster, and handled through random effects, or within-cluster, and parametrically modeled. The resulting model can easily be fit using existing software.

In future work, we consider various methods of obtained a confidence interval for the estimate of $\mu(t)$; among them are the sandwich estimator (Fan and Li, 2001 and 2004) and bootstrap (Chatterjee and Lahiri, 2011). Other topics of interest are the choice of smoothing parameter (Hall and Park, 2010), other penalties (Fan and Li, 2001 or Zou, 2006), and the use of REML to estimate covariance parameters.

REFERENCES

Bühlmann, P. and van de Geer, S. (2011). "Statistics for high dimensional data: methods, theory, and applications". Springer.

Bondell, H. D., Krishna, A. Gosh, S. K. (2010). "Joint variable selection of fixed and random effects in linear mixed-effects models". *Biometrics* 66, 10691077.

Figure 2: Time series of average activated signals from the motor cortex of each of 9 subjects, for pre and post-caeine sessions.

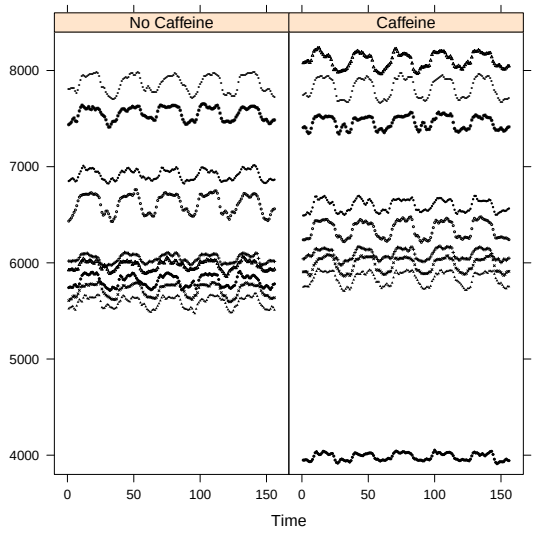
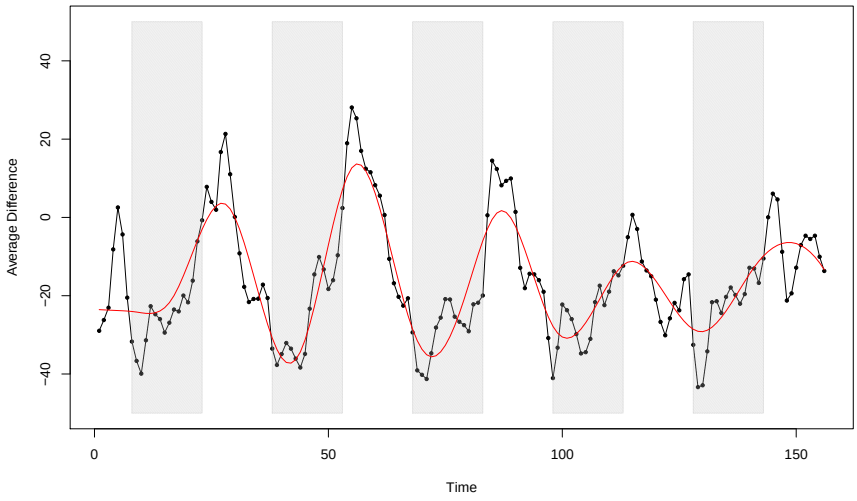


Figure 3: Estimated (red) and observed (black) difference in average activated signal between pre and post-caeine sessions and observed : the shaded regions denote fingertapping sections.



Bunea, F. and Gupta, S. (2010). "A study of the asymptotic proerties of Lasso for correlated data". Technical Report, Florida State University.

Chatterjee A. and Lahiri, S. (2011). "Bootstrapping lasso estimators". *Journal of the American Statistical Association* 106, 608-625.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). "Least angle regression". *Annals of Statistics* 32, 407-451.

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

Fan, J. and Peng, H. (2004). "Nonconcave penalized likelihood with a diverging number of parameters". *Annals of Statistics* 32, 928-961.

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

Hall, P., Lee, E. and Park, B. (2009). "Bootstrap-based penalty choice for lasso, achieving oracle performance".

Statistica Sinica 19, 449-471.

Hastie, T. Tibsharani, R. and Friedman, J. (2001). "The Elements of Statistical Learning: Data Mining, Inference and Prediction". Springer Verlag.

Ibrahim, J. G., Zhu, H., Garcia, R. I. Guo, R. (2010). "Fixed and random effects selection in mixed effects models". *Biometrics*, no. doi: 10.1111/j.1541-0420.2010.01463.x.

Osborne, M.R., Presnell, B., and Turlach, B.A. (1998), "Knot Selection for Regression Splines via the Lasso", in *Dimension Reduction, Computational Complexity, and Information*, ed. S. Weisberg, vol. 30 of Computing Science and Statistics, Fairfax Station, VA: Interface Foundation of North America, pp.44-49

Rack-Gomer, A. L., Liau, J., Liu, T. T. (2009). "Caffeine reduces resting-state BOLD functional connectivity in the motor cortex". *NeuroImage* 46(1), 56-63.

Schelldorfer, J., Bühlmann, P. van de Geer, S. (2011). "Estimation for High-Dimensional Linear Mixed-Effects Models Using l1-Penalization". *Scandinavian Journal of Statistics* 38, 197214.

Wang, H., Li, G. and Tsai, C. (2007). "Regression coefficient and autoregressive order shrinkage and selection via the lasso". *Journal of the Royal Statistical Society, Series B* 69, 63-78.

Zou, H. (2006). "The adaptive Lasso and its oracle properties". *Journal of the American Statistical Association* 101, 1418-1429.