Robust Lasso Regression with Student-t Residuals

Daniel F. Schmidt and Enes Makalic

The University of Melbourne
Centre for Epidemiology and Biostatistics
Carlton VIC 3053, Australia
{dschmidt,emakalic}@unimelb.edu.au

Abstract. The lasso, introduced by Robert Tibshirani in 1996, has become one of the most popular techniques for estimating Gaussian linear regression models. An important reason for this popularity is that the lasso can simultaneously estimate all regression parameters as well as select important variables, yielding accurate regression models that are highly interpretable. This paper derives an efficient procedure for fitting robust linear regression models with the lasso in the case where the residuals are distributed according to a Student-t distribution. In contrast to Gaussian lasso regression, the proposed Student-t lasso regression procedure can be applied to data sets which contain large outlying observations. We demonstrate the utility of our Student-t lasso regression by analysing the Boston housing data set.

Keywords: lasso, robust regression, expectation-maximisation algorithm

1 Introduction

Despite their apparent simplicity, linear regressions remain an important tool in statistics, machine learning and signal processing. Given a vector of features $\mathbf{x}_i \in \mathbb{R}^p$, a linear regression models the corresponding target $y_i \in \mathbb{R}$ by

$$y_i = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\beta_0 \in \mathbb{R}$ is an intercept, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression parameters relating the features to the target, and ε_i is an unobserved, random disturbance. It is common to model the disturbance using a normal distribution with a mean of zero and an unknown variance, and much of the theory of linear regression is based on this assumption.

In practice, it is usually the case that one has observed the features $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and targets $\mathbf{y} = (y_1, \ldots, y_n)'$, but does not know the values of the regression coefficients $\boldsymbol{\beta}$, which must instead be estimated from the data. The principle of least-squares can be used to provide unbiased estimates of $\boldsymbol{\beta}$; however the resulting coefficient estimates are never exactly zero, even if a feature is unrelated to the targets. A considerable body of research exists surrounding

the problem of variable selection; a popular estimator for linear models that can perform variable selection as well as provide estimates for non-zero coefficients is the lasso [10]. The lasso solves the following penalized least-squares problem:

$$\{\hat{\boldsymbol{\beta}}(\gamma), \hat{\beta}_0(\gamma)\} = \arg\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \beta_0 \mathbf{1}_n||_2^2 + \gamma \sum_{j=1}^{p} |\beta_j| \right\}, \tag{1}$$

where $\mathbf{1}_n$ is a column vector of n ones and $\gamma \geq 0$ is a regularisation parameter that controls the sparsity of the solution, i.e., the number of coefficients estimated to be exactly zero. Despite its popularity, the sum-of-squared residuals used by the regular lasso implicitly assumes that the errors (approximately) follow a normal distribution. If the data contains large outlying values, either because certain observations are anomalous, or because of gross errors in the recorded data, the estimates produced by the regular lasso can degrade dramatically. To counter this problem, a number of researchers have combined robust regression techniques, such as quantile regression and Huber estimation, with lasso-style estimators [5, 3]. In this paper we instead examine an alternative approach to handling large deviations by treating the errors ε_i as arising from the Student-t distribution which possesses tails heavier than the Gaussian distribution.

1.1 Regression with t-distributed errors

A random variable $y_i \sim t_\nu(\mu, \sigma^2)$ is said to follow a Student-t distribution with degrees-of-freedom ν if its probability density is given by

$$p_\nu(y_i \,|\, \mu, \sigma^2) = \left(\frac{1}{\pi\nu\sigma^2}\right)^{\frac{1}{2}} \left(\frac{\Gamma([\nu+1]/2)}{\Gamma(\nu/2)}\right) \left(1 + \frac{(y_i - \mu)^2}{\nu\sigma^2}\right)^{-\left(\frac{\nu+1}{2}\right)},$$

where μ is a location parameter and σ^2 is a scale parameter. The t-distribution can model data with large outlying values much more effectively than the usual Gaussian distribution, with the value of ν determining the heaviness of the tails of the distribution. For $\nu > 1$, the mean of the distribution is given by μ, and for $\nu > 2$ the variance of the distribution is given by $\sigma^2\nu/(\nu - 2)$; important special cases of ν are $\nu = 1$, which yields the Cauchy distribution, and $\nu = \infty$ which reduces to the usual Gaussian distribution. Fitting a regression model with t-distributed errors using maximum likelihood can be efficiently performed using the expectation-maximisation (EM) algorithm [4].

Surprisingly, there has been little work examining the t-distribution within a lasso framework. Some work has discussed the use of t-distributions in conjunction with the Bayesian lasso (see, for example, [12]), but these are sampling based approaches that do not yield sparse point estimates. A recent paper [2] proposed to estimate graphical models using a multivariate t-likelihood with a lasso-type penalty. Their proposal used the expectation-maximisation (EM) algorithm in conjunction with a coordinate-wise descent algorithm to find the lasso estimates. In contrast, in this paper we use a Bayesian interpretation of the lasso estimates

as the posterior mode of a linear regression model with a double-exponential prior distribution over the coefficients $\boldsymbol{\beta}$, which we call the t-lasso. This allows us to efficiently solve for the lasso estimates using the expectation-maximisation algorithm [8]. We express the t-lasso by the following Bayesian hierarchy:

$$y_i \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2, \mathbf{x}_i \sim t_\nu(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0, \sigma^2), \qquad i = 1, \ldots, n, \tag{2}$$

$$\beta_j \,|\, \sigma^2, \tau \sim \mathrm{La}(0, 2^{-1/2}\,\sigma\tau), \qquad j = 1, \ldots, p, \tag{3}$$

where $\mathrm{La}(a, b)$ denotes the Laplace distribution with mean a and scale b. The t-lasso is defined as the posterior mode of the hierarchy (2)–(3), which can be found by solving

$$\{\hat{\boldsymbol{\beta}}(\tau), \beta_0(\tau)\} = \arg\min_{\boldsymbol{\beta}, \beta_0} \left\{ -\log p_\nu(\mathbf{y} \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2, \mathbf{X}) + \left(\frac{2}{\tau^2\sigma^2}\right)^{\frac{1}{2}} \sum_{j=1}^{p} |\beta_j| \right\}, \tag{4}$$

where

$$p_\nu(\mathbf{y} \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2, \mathbf{X}) = \prod_{i=1}^{n} p_\nu(y_i \,|\, \mathbf{x}_i'\boldsymbol{\beta} + \beta_0, \sigma^2)$$

is the likelihood function, and hyperparameter τ plays the role of the regularisation parameter γ in the case of the regular lasso, and controls the degree of sparsity of the estimates $\hat{\boldsymbol{\beta}}(\tau)$. In contrast to posterior mean or median estimates obtained by sampling from the posterior distribution of $\boldsymbol{\beta}$, maximisation of the posterior distribution can produce sparse estimates. A crucial difference between our t-lasso and the one proposed by Finegold and Drton [2] is that we explicitly condition our lasso penalty on the noise scale parameter σ^2; in the case of the standard Gaussian lasso regression failing to condition the penalty on σ^2 when the objective function is a penalised likelihood, rather than simply a penalised sum of squared residuals, leads to potential problems involving multiple minima in the objective function [7].

2 Finding *t*-lasso estimates using the EM algorithm

In our Bayesian lasso hierarchy, the residuals and the regression parameters are modelled using the Student-t and double exponential distributions respectively. Both of these distributions can be represented as exchangable Gaussian variance mixture distributions [11] by introducing appropriate latent variables, which allows us to use expectation-maximisation to efficiently find the posterior mode [8]. Using the scale-mixture representations of the t and Laplace distributions from [4, 7] we can rewrite the hierarchy (2)–(3) as

$$y_i \,|\, u_i \sim N(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0, \sigma^2/u_i^2), \quad u_i^2 \sim \chi^2_\nu/\nu, \tag{5}$$

$$\beta_j \,|\, \lambda_j^2 \sim N(0, \lambda_j^2\tau^2\sigma^2), \qquad \lambda_j^2 \sim \mathrm{Exp}(1), \tag{6}$$

where χ^2_ν denotes a chi-squared random variate with k degrees of freedom, $\mathrm{Exp}(1)$ denotes a standard exponential distribution and $u_1^2, \ldots, u_n^2$ and $\lambda_1^2, \ldots, \lambda_p^2$ are latent variables.

We note that conditional on the latent variables u_i^2 and λ_j^2, the posterior distribution of $\boldsymbol{\beta}$ and β_0 using the hierarchy (5)–(6) is Gaussian, for which maximisation is straightforward. Further, conditional on u_i^2 and λ_j^2, the distribution of the data and the prior distributions of the regression coefficients are conjugate. Maximisation of this conditional posterior distribution with respect to $\boldsymbol{\beta}$ is therefore equivalent to maximising the likelihood of an appropriately augmented data set. The augmented targets are $\tilde{\mathbf{y}} = (\mathbf{y}', \mathbf{0}_p')'$, and the augmented $\mathbf{X}$ matrix is given by

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} & \mathbf{1}_n \\ \mathbf{I}_p & \mathbf{0}_p \end{pmatrix},$$

where $\mathbf{0}_p$ denotes a column vector of p zeros and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. This formulation is also convenient as it combines $\boldsymbol{\beta}$ and β_0 into a single parameter vector $\tilde{\boldsymbol{\beta}}$. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)$ denote the parameters of this alternative likelihood. The equivalent "complete data" likelihood, conditional on the latent variables u_i^2 and λ_j^2, up to terms independent of $\tilde{\boldsymbol{\theta}}$ is

$$\left(\frac{n+p}{2}\right)\log(\tilde{\sigma}^2) + \left(\frac{1}{2\tilde{\sigma}^2}\right)\left[\sum_{i=1}^{n} u_i^2(y_i - \tilde{\mathbf{x}}_i\tilde{\boldsymbol{\beta}})^2 + \left(\frac{1}{\tau^2}\right)\sum_{j=1}^{p}\frac{\tilde{\beta}_j^2}{\lambda_j^2}\right]. \qquad (7)$$

Conditional on the latent variables, maximisation of (7) with respect to $\boldsymbol{\beta}$ is a straightforward weighted least squares problem. However, the values of the latent variables are unknown, so the expectation-maximisation algorithm replaces them with their conditional expectations. All the latent variables are conditionally independent, given $\tilde{\boldsymbol{\beta}}$, $\tilde{\beta}_0$ and $\tilde{\sigma}^2$, and the conditional expectations are given by [4, 7]

$$\mathbb{E}\left[u_i^2 \,|\, \tilde{\boldsymbol{\theta}}\right] = w_i(\tilde{\boldsymbol{\theta}}) = \frac{\nu+1}{\nu + (\tilde{y}_i - \tilde{\mathbf{x}}_i'\tilde{\boldsymbol{\beta}})^2/\tilde{\sigma}^2}, \quad i = 1, \ldots, n,$$

$$\left(\frac{1}{\tau^2}\right)\mathbb{E}\left[\lambda_j^{-2} \,|\, \tilde{\boldsymbol{\theta}}\right] = w_{n+j}(\tilde{\boldsymbol{\theta}}) = \left(\frac{2\tilde{\sigma}^2}{\tau^2\tilde{\beta}_j^2}\right)^{\frac{1}{2}}. \qquad j = 1, \ldots, p.$$

Let $\tilde{\boldsymbol{\theta}}_{(t)} = (\tilde{\boldsymbol{\beta}}_{(t)}, \tilde{\sigma}^2_{(t)})$ denote the parameters at iteration t, with some suitable starting values chosen for $t = 1$. The expectation-maximisation algorithm for solving (4) involves iterating the following steps:

$$\mathbf{W}_{(t)} \leftarrow \mathrm{diag}(w_1(\tilde{\boldsymbol{\theta}}_{(t)}), \ldots, w_{n+p}(\tilde{\boldsymbol{\theta}}_{(t)})), \qquad (8)$$

$$\tilde{\boldsymbol{\beta}}_{(t+1)} \leftarrow \left(\tilde{\mathbf{X}}'\mathbf{W}_{(t)}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\mathbf{W}_{(t)}\tilde{\mathbf{y}}, \qquad (9)$$

$$\tilde{\sigma}^2_{(t+1)} \leftarrow \left(\frac{1}{n+p}\right)(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}_{(t+1)})'\mathbf{W}_{(t)}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}_{(t+1)}), \qquad (10)$$

until convergence is achieved.

Once the algorithm (8)–(10) has converged to some values of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$, the lasso estimates for a given regularisation parameter τ are given by

$$\hat{\beta}_j(\tau) = \tilde{\beta}_j, \ j = 1, \ldots, p, \quad \hat{\beta}_0(\tau) = \tilde{\beta}_{p+1}.$$

A quirk of finding the lasso by minimising a negative log-posterior of the form (4) is that the value of the scale parameter $\tilde{\sigma}^2$ produced by the expectation-maximisation algorithm does not maximise the likelihood for the estimates $\hat{\boldsymbol{\beta}}(\tau)$ and $\hat{\beta}_0(\tau)$; instead it maximises the product of the t-likelihood (2) and Laplace prior distributions (3). The regular lasso itself does not provide estimates for σ^2 even in the case of Gaussian noise. To facilitate comparisons between models based on their likelihoods, we do not use $\tilde{\sigma}^2$ as our estimate of the t-distribution scale parameter; instead, we use

$$\hat{\sigma}^2(\tau) = \arg\max_{\sigma^2} \left\{ p_\nu(\mathbf{y} \mid \sigma^2; \hat{\boldsymbol{\beta}}(\tau), \hat{\beta}_0(\tau), \mathbf{X}) \right\},$$

which maximises the likelihood when the regression coefficients are fixed at the t-lasso estimates $\hat{\boldsymbol{\beta}}(\tau)$ and $\hat{\beta}_0(\tau)$.

2.1 Generating regularisation paths

The EM algorithm discussed in Section 2 gives lasso solutions for the regression coefficients $\hat{\boldsymbol{\beta}}(\tau)$ for a single value of the regularisation parameter τ. By varying τ, the algorithm given by equations (8)–(10) can be used to produce a complete regularisation path for a given data set, from the model in which all the coefficients are zero up to (approximately) the maximum likelihood estimates. In the case of Gaussian lasso regression (1), the minimum and maximum values of the regularisation parameter γ can be obtained in closed form [6].

In the case of our t-lasso regression formulation (4), exact estimates of the minimum and maximum values of the regularisation parameter τ are not easily obtained, as the t-likelihood in the objective function is no longer a simple linear function of the squared residuals. We define the quantity $\tau_{\min}$ as the largest value of τ that leads to the "all zeros" solution $\hat{\boldsymbol{\beta}}(\tau) = \mathbf{0}_p$. For the solution of (4) to be the "all-zeros" solution, the gradient of the objective function evaluated at $\boldsymbol{\beta} = \mathbf{0}_p$ must equal $\mathbf{0}_p$; that is,

$$\mathbf{g} + \left(\frac{2}{\tau^2 \sigma^2} \right)^{\frac{1}{2}} \mathbf{v} = \mathbf{0}_p \tag{11}$$

where $v_j \in [-1, 1]$ are sub-derivatives of the absolute value function,

$$\mathbf{g} = -\left(\frac{\nu + 1}{\nu \hat{\sigma}^2} \right) \sum_{i=1}^{n} \left[\left(1 + \frac{(y_i - \hat{\beta}_0)^2}{\nu \hat{\sigma}^2} \right)^{-1} (y_i - \hat{\beta}_0) \mathbf{x}_i' \right]$$

is the gradient of the t-regression negative log-likelihood with respect to $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta} = \mathbf{0}_p$, and $\hat{\beta}_0$ and $\hat{\sigma}^2$ are the maximum likelihood estimates of β_0 and σ^2, respectively, when $\boldsymbol{\beta} = \mathbf{0}_p$. Solving (11) for τ yields the approximate value

$$\tau_{\min}^2 \approx \frac{2}{\hat{\sigma}^2 \, ||\mathbf{g}||_\infty^2},$$

where $||\mathbf{g}||_\infty$ denotes the maximum absolute value of the entries of the vector $\mathbf{g}$.

Due to the formulation of the objective function, there is no finite value of τ such that $\hat{\boldsymbol{\beta}}(\tau) = \hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, and in the case that $p \geq n$, the maximum likelihood estimates do not exist. If $p < n$, one can take some sufficiently large value of τ, say $\tau_{\max}$, such that there is little discrepancy between $\hat{\boldsymbol{\beta}}(\tau_{\max})$ and the maximum likelihood estimates. Our implementation uses the heuristic choice

$$\tau_{\max}^2 \approx \frac{2K\,||\hat{\boldsymbol{\beta}}_{\mathrm{ML}}||_1^2}{p^2 \hat{\sigma}_{\mathrm{ML}}^2}, \tag{12}$$

where $K > 1$ is a large positive constant, which in our implementation is $K = 100$. The value (12) is obtained by interpreting the penalty function as a negative log-prior distribution for the coefficients $\boldsymbol{\beta}$ and maximising with respect to τ with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ (see Appendix A); it has the advantage of ensuring that the maximum value adapts to the magnitude of the regression coefficients that would be obtained using maximum likelihood.

Given the minimum and maximum values of the τ, the regularisation path is computed over a logarithmically spaced grid of n_τ values of τ between $\tau_{\min}$ and $\tau_{\max}$; in our implementation, the default choice is $n_\tau = 100$. Values of the negative log-likelihood for $\hat{\boldsymbol{\beta}}(\tau)$ are automatically produced by our implementation and can be used to guide selection of the regularisation parameter τ and the degrees-of-freedom parameter ν.

2.2 Selecting τ and the degrees-of-freedom ν

In the regular lasso (1), selection of the regularisation parameters is usually performed by minimising either a cross-validation prediction error, or an information criterion such as Akaike's information criterion (AIC) [1] or the Bayesian information criterion (BIC) [9]. The t-lasso also requires a choice of the t-distribution degrees-of-freedom parameter ν, and this can also be selected by cross-validation or an information criteria approach. Generally, the values of τ under consideration are determined by the grid used to produce the regularisation path (see Section 2.1). While ν is also a continuous parameter, the nature of the parameter means it is common to examine only a small number of discrete candidates.

To select τ and ν using information criteria we minimise a function of the form

$$\{\hat{\tau}, \hat{\nu}\} = \underset{\tau \in T, \nu \in N}{\arg\min} \left\{ -\log p_\nu(\mathbf{y} \,|\, \hat{\boldsymbol{\beta}}(\tau), \hat{\beta}_0(\tau), \sigma^2(\tau), \mathbf{X}) + \alpha_n ||\hat{\boldsymbol{\beta}}(\tau)||_0 \right\}$$

where T is the grid of n_τ values of τ, N is the set of candidate ν values, $||\boldsymbol{\beta}||_0$ is the number of non-zero elements of the vector $\boldsymbol{\beta}$, and α_n is a suitable penalty term; standard choices are $\alpha_n = 1$ for AIC and $\alpha_n = (1/2)\log n$ for BIC.

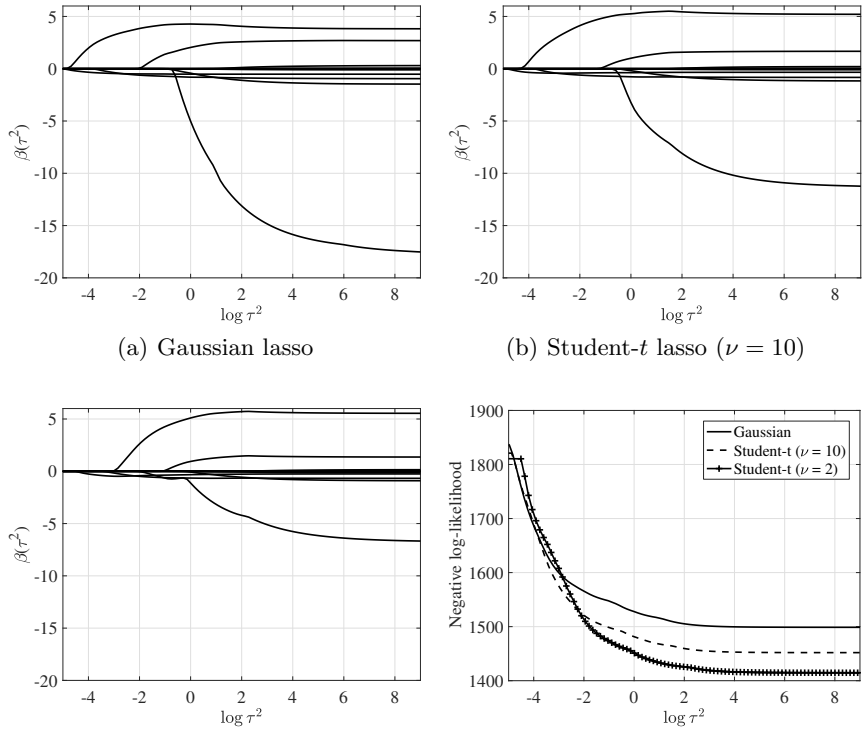
Cross-validation works by dividing the data into disjoint training and testing sets, fitting models to the training data and then calculating the prediction errors of the fitted models on the reserved test data. The value of τ which minimises the cross-validation prediction error is then chosen as optimal. The usual prediction

error metric used in cross-validation is the mean squared-prediction error. This measure is highly sensitive to the presence of large outlying observations in the testing data, which is exactly the situation we expect when we are using t-regression. In our implementation of t-lasso regression, we choose to minimise the cross-validation estimates of the negative log-likelihood on the testing data. This choice of prediction error explicitly takes into account the heavy tails of the t-distribution and is resistant to large outlying observations when ν is small. For large ν, the negative log-likelihood of the t-distribution is essentially equivalent to the usual mean-squared prediction error. A further advantage of using negative log-likelihood prediction scores is that cross-validation can then also be used to select the degrees-of-freedom ν in addition to the regularisation parameter τ.

3 Real data example

We demonstrate the utility of our proposed t-lasso by analysing the Boston housing data set. This data set contains 506 observations ($n = 506$), 13 covariates ($p = 13$) and the target variable is the median value of owner-occupied homes in suburbs of Boston (measured in \$1,000s). Some variables in the data set are strongly positively or negatively correlated; for example, the correlation between the index of accessibility to radial highways and full-value property-tax rate is 0.91 while the correlation between nitric oxides concentration and the weighted distances to five Boston employment centres is -0.77. To estimate the association between the target variable and the 13 covariates, we used Gaussian lasso linear regression and lasso regression with Student-t residuals with degrees-of-freedom $\nu \in \{1, 2, 10, 10000\}$. This set of ν values includes a wide spectrum of distributions ranging from distributions that have light tails (essentially Gaussian for $\nu = 10,000$) to distributions with very heavy tails (Cauchy distribution for $\nu = 1$). An efficient software implementation of our t-lasso, and the scripts required to recreate the analyses in this example are available from the MATLAB Central File Exchange (ID: 63037).

Full regularisation paths for the Gaussian and Student-t regression models where ($\nu = 2$) and ($\nu = 10$) are shown in Figure 1 (a)–(c). It is clear that the Gaussian regularisation path (Figure 1(a)) is different from both of the Student-t regression paths for the Boston housing data. In all three regularisation paths, the variable with the largest regression coefficient is nitric oxides concentration (`nox`). In the Gaussian lasso analysis, the regression coefficient for `nox` starts at approximately -17 and is shrunk to zero when the penalty parameter $\log \tau^2 \approx -1/2$. In case of the t-lasso for ($\nu = 2$) and ($\nu = 10$), the `nox` regression coefficients starts at approximately -6.7 and -11.2, and are shrunk to zero at $\log \tau^2 \approx 0$ and $\log \tau^2 \approx -1/2$, respectively. The variable with the second largest regression coefficient is the average number of rooms per dwelling (`rm`). Interestingly, for the Gaussian lasso path, the regression coefficient for `rm` starts at around 3.8, and shrinks to zero at $\log \tau^2 \approx -4.5$, while for both t-lasso paths, the `rm` coefficient is larger, starting at approximately 5.5 and 5.2 for $\nu = 2$



(a) Gaussian lasso

(b) Student-t lasso ($\nu = 10$)

(c) Student-t lasso ($\nu = 2$)

(d) Negative log-likelihood for Gaussian and Student-t lasso models

Fig. 1. Analysis of the Boston housing data using lasso linear regression with Gaussian and Student-t residuals

and $\nu = 10$, respectively. When $\nu = 2$, this variable is shrunk more aggressively to zero than when $\nu = 10$.

The negative log-likelihoods for all models in the Gaussian and Student-t regularisation paths are shown in Figure 1(d). Based solely on the negative log-likelihood, it appears that some form of robust regression, such as the proposed t-lasso regression, is necessary for this data set. The negative log-likelihoods for the two t-lasso regression models are generally smaller when compared to the negative log-likelihoods of the Gaussian lasso models. To examine this further, we used cross-validation to select one model from each lasso regularisation path as discussed in Section 2.2. In the case of Gaussian lasso regression, cross-validation selected a model which included 11 variables with the following two variables omitted: (1) proportion of non-retail business acres per town, and (2) proportion of owner-occupied units built prior to 1940. For Student-t lasso regression, we used cross-validation to select the important variables as well as the estimate of the degrees-of-freedom ν. The model selected by cross-validation for the t-lasso

had degrees of freedom $\hat{\nu} = 2$ and included all 13 covariates. The sum of the absolute values of the coefficients was 26.4 and 15.8 for the best Gaussian and $\nu = 2$ model, respectively, which show that the t-lasso has estimated a model with smaller coefficients on average than the Gaussian lasso.

We also conducted a small simulation study to evaluate the predictive performance of the t-lasso and the regular Gaussian lasso on the Boston housing data. We randomly reserved half of the data for training and half for testing, and used cross-validation to select the best Gaussian lasso model, and the best Student-t model, with $\nu \in \{1, 2, 10, 10000\}$ from the training data only. The performance of the selected models was then evaluated by computing the negative log-likelihood of the testing data for both the best t-lasso model and the best Gaussian lasso model. This procedure was repeated one hundred times. In 96 out of the 100 tests, the t-lasso selected $\nu = 2$. The mean negative log-likelihoods attained on the test data were 723.8 and 773.0 for the t-lasso and regular Gaussian lasso respectively, clearly demonstrating the utility of lasso regression with Student-t residuals.

4 Appendix A

To find an appropriate maximum value $\tau_{\max}$ of τ for producing a regularisation path we use the following heuristic procedure: let $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$ denote the maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2. The negative log-prior probability of the maximum likelihood estimates, under the Laplace prior (3), is given by

$$p \log(\tau) + \left(\frac{\sqrt{2}}{\tau \hat{\sigma}_{\mathrm{ML}}} \right) ||\hat{\boldsymbol{\beta}}_{\mathrm{ML}}||_1 + \text{const},$$

where const denotes terms that do not depend on either τ or $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$. The value of τ that maximises the prior probability for $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is given by

$$\tilde{\tau} = \frac{\sqrt{2}\, ||\hat{\boldsymbol{\beta}}_{\mathrm{ML}}||_1}{p \hat{\sigma}_{\mathrm{ML}}}.$$

We then choose $\tau_{\max} = c\tilde{\tau}$, where $c > 1$ is a constant that controls the distance of the maximum likelihood estimates to the final point on the regularisation path.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (December 1974)
2. Finegold, M., Drton, M.: Robust graphical modelling with t-distributions. In: 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009) (2009)
3. Lambert-Lacroix, S.: Robust regression through the Hubers criterion and adaptive lasso penalty. Electronic Journal of Statistics 5, 1015–1053 (2011)

4. Lange, K.L., Little, R.J.A., Taylor, J.M.G.: Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84(408), 881–896 (1989)
5. Li, Y., Zhu, J.: l_1-norm quantile regression. Journal of Computational and Graphical Statistics 17(1), 1–23 (2008)
6. Osborne, M.R., Presnell, B., Turlach, B.A.: On the LASSO and its dual. Journal of Computational and Graphical Statistics 9(2), 319–337 (2000)
7. Park, T., Casella, G.: The Bayesian lasso. Journal of the American Statistical Association 103(482), 681–686 (June 2008)
8. Polson, N.G., Scott, J.G.: Data augmentation for non-Gaussian regression models using variance-mean mixtures. Biometrika 100(2), 459–471 (2013)
9. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6(2), 461–464 (1978)
10. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society (Series B) 58(1), 267–288 (1996)
11. West, M.: On scale mixtures of normal distributions. Biometrika 74(3), 646–648 (1987)
12. Yi, N., Xu, S.: Bayesian LASSO for quantitative trait loci mapping. Genetics 179(2), 1045–1055 (2008)