

# MPRA

Munich Personal RePEc Archive

## **Regularized Skew-Normal Regression**

Karl Shutes and Chris Adcock

Coventry University, University of Sheffield

24. November 2013

Online at <http://mpra.ub.uni-muenchen.de/52217/>

MPRA Paper No. 52217, posted 14. December 2013 19:39 UTC

# Regularized Skew-Normal Regression

K. Shutes & C.J. Adcock

December 11, 2013

## Abstract

This paper considers the impact of using the regularisation techniques for the analysis of the extended skew-normal distribution. The approach is estimated using a number of techniques and compared to OLS based LASSO and ridge regressions in addition to non- constrained skew-normal regression.

## 1 Introduction & Motivation

The use of regularisation in econometrics is far from widespread, however the use is more commonplace in the statistics literature. This paper addresses the issue raised by Bühlmann [2013] of the lack of non-Gaussian distributions using the regularisation methods. The problem of variable selection is commonly side-stepped with legitimate appeals to theoretical frameworks. Indeed multi-factor models used in finance (for example Chen et al. [1986]) face a variable selection problem, which can be solved using Principal Components or *a priori* judgements. The regularisation approach gives an alternative to these within a standard regression framework.

This paper extends this to consider situations where theory is not prescriptive and into situations where one might be tempted into using hypothesis tests to determine the independent variables in one's analyses. The use of these machine learning techniques is far from a *carte blanche* for mindless data mining. The use and selection of relevant data is still driven by theoretical foundations. However it is informative to ascertain which variables are driving the underlying relationships and thus the problem of variable selection continues to exist. This means that the standard approach of ordinary least squares is not feasible without some form of variable selection.

The literature on the use and abuse of stepwise regression is significant. The situation of 'excessive data' can be dealt with by the regularised regressions, such as the LASSO and elastic net, for example Zou and Hastie [2005] & Zou [2005] where it is possible to have more independent variables than observations, unlike the situation in standard OLS. It is common that approaches such as the Aikake or Schwartz Information criteria are used in the variable selection problem (Akaike [1974] amongst others) albeit less so than stepwise regression techniques. These can further be contrasted with subset regressions, which take the various permutations of individual variables to find the *best* model. These forms of modelling can lead to issues such as inflated  $R^2$ ,  $F$  statistics (as

discussed in, for example, Pope and Webster [1972]) and biases within the estimated parameters.

In the majority of cases, the use of the regularisation techniques are based upon Gaussian distributed errors and Ordinary Least Squares. Though in many cases this is sufficient, there are many cases such as those in finance where normality is not an appropriate assumption. This paper looks to add to the regularisation literature by extending the LASSO (Tibshirani [1996]) to accommodate shrinkage within the higher moments via the use of the extended skew-normal based regression model (Adcock & Shutes [2001] & Shutes [2005]). The method proposed here uses the technique of the LASSO, i.e. the introduction of  $\ell_1$  norms, but in contrast to the literature based on Gaussian regression, a further norm is introduced, that of the skewness parameter. This will imply that in addition to the variable selection made via the standard approach the method also performs a selection of non-normality as the extra parameters control the skewness and kurtosis. It is not necessary to constrain the location of the truncating variable it is only estimated when the skewness parameter is non-zero.

The rest of the paper is organised as follows. A consideration of the extended skew-normal and the LASSO is presented with the relevant estimation and an example to conclude. A standard data set from the machine learning literature, that of diabetes patients is used (see Efron et al. [2004] where it is fully described).

## 2 Literature Review & Definitions

### 2.1 Regularization

Within the econometric literature, regularisation has a limited history. Ridge regression is perhaps the best known example (for example Hoerl & Kennard[1970]), where the problem of multicollinearity is dealt with by the imposition of a constraint on the coefficients of the regressions. This estimator is known to be biased however it is the case that the approach gives estimators with lower standard errors. The penalised function for the estimation is given by:

$$\begin{aligned}
 \beta_R &= \arg \min_{\beta} (Y_i - \beta_0 - X_i \beta^T)^T (Y_i - \beta_0 - X_i \beta^T) \quad \text{s.t.} \quad \beta^T \beta \leq \epsilon \quad (1) \\
 &= \arg \min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \nu \sum \beta_j^2 \\
 &= \arg \min_{\beta} (Y_i - \beta_0 - X_i \beta^T)^T (Y_i - \beta_0 - X_i \beta^T) + \nu \beta^T \beta \\
 &= (X^T X + \nu I)^{-1} X^T y
 \end{aligned}$$

This approach does not perform any form of variable selection as, although it does shrink coefficients, it does not shrink them to 0. The  $\nu$  parameter<sup>1</sup> acts as the shrinkage

---

<sup>1</sup>Traditionally the Lagrangean multiplier is denoted  $\lambda$ , however due to the use of  $\lambda$  as the skewness parameter in the distribution, the Lagrangean is denoted  $\nu$ .

control with  $\nu = 0$  being no shrinkage and therefore ordinary least squares. This can be compared to the Least Absolute Shrinkage & Selection Operator (LASSO). In this case the penalty is based on the  $\ell_1$  norm rather than the  $\ell_2$  norm of the ridge approach. Hence the problem becomes:

$$\begin{aligned} \beta_R &= \arg \min_{\beta} (Y_i - \beta_0 - X_i \beta^T)^T (Y_i - \beta_0 - X_i \beta^T) \quad \text{s.t.} \quad \|\beta\|_1 \leq \epsilon \quad (2) \\ &= \arg \min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \nu \sum \|\beta_j\| \\ &= \arg \min_{\beta} (Y_i - \beta_0 - X_i^T \beta)^T (Y_i - \beta_0 - X_i^T \beta) + \nu \|\beta^T\|_1 \end{aligned}$$

There is a well-known mapping between the multiplier  $\nu$  and the constraint of the sums of the coefficients  $\epsilon$ . In general the constant is not shrunk and remains at  $\bar{y}$ .

The variable selection property is clearly shown graphically when considering two parameter estimates, with the LASSO (black) and ridge (red). The estimator loss functions are shown as ellipses. The point of tangency are the estimates for each technique. The

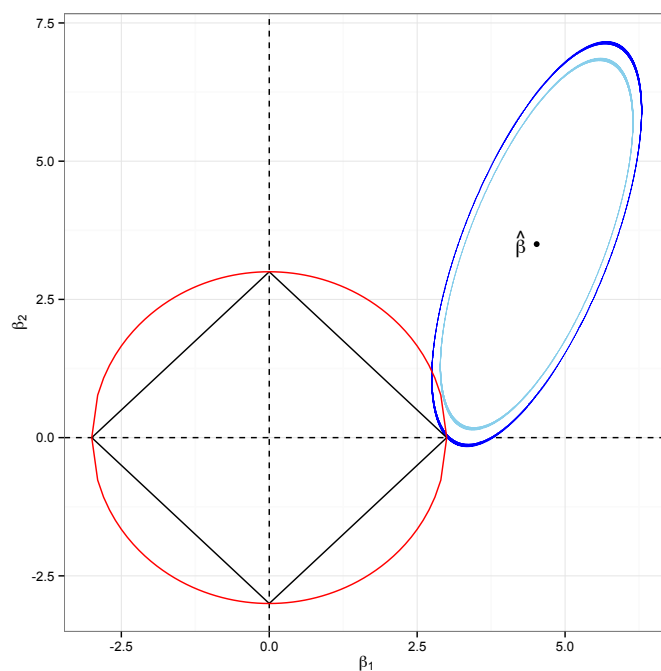


Figure 1: Differences Between LASSO & Ridge Regressions

LASSO shrinks  $\beta_1$  to 0, whereas the ridge regression approaches it. The OLS estimator is given as  $\hat{\beta}$ . The parameter  $\nu$  controls the amount of penalty applied to the parameters for the LASSO. Fu and Knight[2000] show that under certain regularity conditions, the estimates  $\hat{\beta}$  are consistent & that these estimates will have the same limiting distribution as the OLS estimates.

There is a generalisation such that the  $\gamma$ -th norm is used. This is the bridge estimator. There are a number of similarities between the bridge estimator with  $1 < \gamma < 2^2$  however the elastic net approach has non-differentiable corners at the axes (Hastie, Tibshirani, and Friedman[2008]). This therefore implies that the bridge regression, despite first impressions will not select variables unless  $\gamma < 1$  in which case the penalty function is non-concave and the estimates may not be unique, though they may be set at zero. The  $\gamma$ -th norm is defined as:

$$\|\beta\|_\gamma = \left( \sum |\beta_i|^\gamma \right)^{\frac{1}{\gamma}} \quad (3)$$

These estimators, Lasso, bridge and ridge are all forms of Bayesian estimator with priors based on a LaPlace or variants of this based on a log exponential function.

## 2.2 The Skew-Normal Distribution

The skew-normal distribution has become increasingly well used within a number of fields since its initial description by Azzalini [1985]. A particularly attractive feature of the distribution is that it includes the Gaussian as a limiting case. In its simplest form the distribution is described by the following density function:

$$\begin{aligned} h(y) &= 2\phi(y)\Phi(\lambda y) \\ -\infty &< \lambda < \infty \\ -\infty &< y < \infty \end{aligned} \quad (4)$$

with  $\lambda$  controlling the degree of skewness of the distribution. The case  $\lambda=0$  will lead to a standard normal distribution. As  $\lambda$  increases in absolute value, the weighting on  $\Phi$  function increases. This leads to the limiting case being the half or folded normal distribution. Graphically the impact of  $\lambda$  can be seen from the Figure 2.

Azzalini ([1985] & [1986]) proposes that the skew-normal distribution is best thought of as a combination of a symmetric element and a skewing element, which is a truncated normal distribution with mean of 0. This is generalised in Arnold & Beaver [2000] and Adcock & Shutes [2001] where the truncated normal has a mean of  $\tau$ . Thus the density function can be written as:

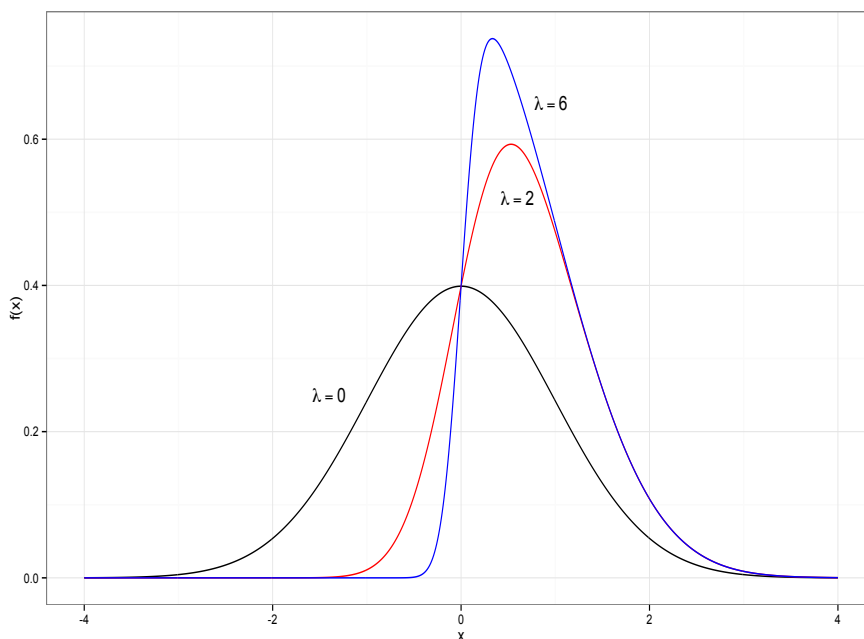
$$f(r) = \frac{1}{\Phi(\tau)} \phi\left(r; \mu + \lambda\tau, \sigma^2 + \lambda^2\right) \Phi\left(\frac{\tau + \frac{\lambda}{\sigma^2}(r - \mu)}{\sqrt{1 + \frac{\lambda^2}{\sigma^2}}}\right) \quad (5)$$

where  $\phi$  and  $\Phi$  are the probability density and cumulative functions of the normal distribution respectively.

---

<sup>2</sup>The limits here are LASSO and ridge regressions.

Figure 2: The Skew-Normal Distribution  $\lambda = 0, 2, 6$



It is possible to use the following parameterization, with  $\gamma$  and  $\omega^2$  being the mean and the variance of the normal part of the distribution respectively:

$$\begin{aligned}\gamma &= \mu + \lambda\tau & (6) \\ \omega^2 &= \sigma^2 + \lambda^2 \\ \psi &= \sqrt{\sigma^2 + \lambda^2} \frac{\lambda}{\sigma} = \omega \frac{\lambda}{\sigma} \\ \frac{\psi^2}{\omega^2} &= \frac{\lambda^2}{\sigma^2}\end{aligned}$$

This parameterisation allows a simpler description of the distribution. This is not a unique transformation. However the definitions used are easily extendable to the multi-variate distribution. The probability density function can be expressed in terms of these parameters as:

$$f_R(r) = \frac{1}{\Phi(\tau)} \phi(r; \gamma, \omega^2) \Phi\left(\tau \sqrt{1 + \frac{\psi^2}{\omega^2}} + \frac{\psi}{\omega^2}(r - \gamma)\right) \quad (7)$$

where  $\phi(x; \mu, \sigma^2)$  is the probability density function of a normally distributed variable with mean  $\mu$  and variance  $\sigma^2$ . This gives an extension to the standard skew-normal distribution.

The application of the LASSO type approach to the skewed family of distributions is limited. Wu et al. [2012] consider the variable selection problem for the skew-normal

family. However they use a fixed but estimated skewness parameter in essence removing the skewness problem in conjunction with a quadratic expansion of the penalised likelihood to give a tractable solution. Their focus is very much on the location and scale parameters rather than the skewness with a view to modelling the variance as an entity as well as the mean i.e. regression style models. The penalised likelihood approach used both in Wu and here is found in Fan and Li [2001]. This allows both the estimation and standard errors to be estimated despite the singularity introduced by the constraint.

### 3 Likelihood Functions

In order to use the LASSO style estimators, it is necessary to consider the relevant likelihood estimators in light of the constraints. We can think of the constrained likelihood as having two elements, the objective and the constraint. Thus we can exploit the first order conditions of the standard skew-normal family to derive the LASSO solution path for various values of the constraint. This is not unlike a co-ordinate descent approach as discussed in Friedman et al. [2007]. Thus the LASSO estimator is broken into  $h(\beta, \lambda) = f(\beta, \lambda) + g(\beta, \lambda, \nu)$  where  $f(\beta)$  is the standard MLE estimator of the skew-normal regression and  $g(\beta, \lambda, \nu)$  the constrained element.

The likelihood function of the extended skew-normal distribution is somewhat non-linear. Using the specification above, the likelihood is given by:

$$\begin{aligned}
 \ell_i(y; \tau, \gamma, \beta, \psi, \omega^2) &= -\ln \Phi(\tau) - \frac{1}{2} \ln \omega^2 - \frac{1}{2} \ln 2\pi - \frac{1}{2\omega^2} (y_i - \beta_0 - \beta x_i - \gamma)^2 \\
 &\quad + \ln \Phi \left( \tau \sqrt{1 + \frac{\psi^2}{\omega^2}} + \frac{\psi}{\omega^2} (y_i - \beta_0 - \beta x_i - \gamma) \right) + \nu_1 (\|\beta\|_1 + \|\psi\|_1) \\
 f(\bullet) &= -\ln \Phi(\tau) - \frac{1}{2} \ln \omega^2 - \frac{1}{2} \ln 2\pi - \frac{1}{2\omega^2} (y_i - \beta_0 - \beta x_i - \gamma)^2 \\
 &\quad + \ln \Phi \left( \tau \sqrt{1 + \frac{\psi^2}{\omega^2}} + \frac{\psi}{\omega^2} (y_i - \beta_0 - \beta x_i - \gamma) \right) \\
 g(\bullet) &= \nu_1 (\|\beta\|_1 + \|\psi\|_1)
 \end{aligned} \tag{8}$$

This is the standard log-likelihood function for the extended skew-normal with the addition of the LASSO penalty for the coefficients and the skewness parameter. Given the formulation of the regression problem, the likelihood of a number of the parameters are

identical to those of the non-penalised regression model. Hence:

$$\frac{\partial \ell}{\partial \tau} = \frac{\partial f}{\partial \tau} = -\zeta_1(\tau) + \zeta_1 \left( \tau \sqrt{1 + \frac{\psi^2}{\omega^2}} + \frac{\psi}{\omega^2}(y - \beta x) \right) \sqrt{1 + \frac{\psi^2}{\omega^2}} \quad (9)$$

$$\frac{\partial \ell}{\partial \gamma} = \frac{\partial f}{\partial \gamma} = \frac{1}{\omega^2}(y - \beta x - \gamma) - \frac{\psi}{\omega^2} \zeta_1 \left( \tau \sqrt{1 + \frac{\psi^2}{\omega^2}} + \frac{\psi}{\omega^2}(y - \beta x) \right) \quad (10)$$

$$\begin{aligned} \frac{\partial \ell}{\partial \omega^2} = \frac{\partial f}{\partial \omega^2} &= -\frac{1}{2\omega^2} + \frac{1}{2\omega^4}(y - \beta x - \gamma)^2 \\ &\quad - \frac{\psi}{\omega^4} \zeta_1 \left( \tau \sqrt{1 + \frac{\psi^2}{\omega^2}} + \frac{\psi}{\omega^2}(y - \beta x) \right) \left( \frac{\tau \psi}{2} \left( 1 + \frac{\psi^2}{\omega^2} \right)^{-1/2} + (y - \beta x + \gamma) \right) \end{aligned} \quad (11)$$

The coefficients where the constraints can potentially bind are given below.

$$\frac{\partial \ell}{\partial \beta} = \frac{\partial f}{\partial \beta} + \frac{\partial g}{\partial \beta} = \frac{x}{\omega^2}(y - \beta x - \gamma) - \frac{\psi}{\omega^2} x \zeta_1 \left( \tau \sqrt{1 + \frac{\psi^2}{\omega^2}} + \frac{\psi}{\omega^2}(y - \beta x) \right) + \text{sgn}(\beta) \nu_1 \quad (12)$$

$$\begin{aligned} \frac{\partial \ell}{\partial \psi} = \frac{\partial f}{\partial \psi} + \frac{\partial g}{\partial \psi} &= \frac{1}{\omega^2} \zeta_1 \left( \tau \sqrt{1 + \frac{\psi^2}{\omega^2}} + \frac{\psi}{\omega^2}(y - \beta x) \right) \\ &\quad \left( y - \beta x - \gamma + \tau \psi \left( 1 + \frac{\psi^2}{\omega^2} \right)^{-1/2} \right) + \text{sgn}(\psi) \nu_1 \end{aligned} \quad (13)$$

## 4 Estimation

It is possible to leverage the co-ordinate descent approach to update the estimates of the relevant coefficients until convergence to the LASSO solution occurs. Assuming uncorrelated predictors, the updating procedure can be based on the product of the residuals and the relevant predictors and the value of the Lagrange multiplier. Thus following Friedman et al. [2007] we can write:

$$\tilde{\beta}_j(\gamma) \leftarrow S \left( \tilde{\beta}_j(\gamma) + \sum x_{ij} \epsilon_i, \gamma \right) \quad j = 1, 2, \dots, p \quad (14)$$

This produces a whole path solution with the different solutions for the problem providing the starting point for the next optimisation thus reducing the issues with convergence<sup>3</sup> and speed. The estimations here are for the Azzalini form of the distribution i.e.  $\tau=0$ .

<sup>3</sup>As noted in Azzalini and Capitanio [1999] the likelihood function of the skew-normal is not convex in its standard form, thus a slight re-formulation not dissimilar to the one presented above is more stable and robust.



## 4.1 Estimation with Maximum Likelihood

Estimation was performed using a maximum likelihood approach with the nuisance parameter,  $\nu$  being based on a grid in the first case and then cross validation being used to optimise the choice of this parameter. Using the non-constrained maximum likelihood estimates as the initial points to aid in convergence, the estimations were performed with a transformation of the parameter  $\nu$  to  $\exp(\nu)$ . This led to more satisfactory convergence of the algorithms and allowed a greater range of the parameter than a simple linear constraint would allow.

## 4.2 Estimation MCMC with LaPlace Priors

Using the approach of Park and Casella [2008], a Markov Chain Monte Carlo approach is proposed. The main estimated parameters,  $\beta$  &  $\lambda$  are all given Laplace prior distributions. The LASSO parameter can be given a diffuse hyperprior based on the gamma distribution or chosen by techniques such as cross-validation. The former approach is taken here. The prior for regression coefficients,  $\beta$  and the LASSO parameter,  $\nu_i$  is based, as suggested by Park & Casella on

$$g(\beta | \sigma^2) = \prod_{j=1}^p \frac{\nu}{2\sqrt{\sigma^2}} \exp^{-\nu|\beta_j|\sqrt{\sigma^2}} \quad (15)$$

$$f(\nu^2 | \xi, \theta) = \frac{\theta^\xi}{\Gamma(\xi)} (\nu^2)^{\xi-1} \exp(-\theta\nu^2) \quad (16)$$

The square ensures a proper posterior distribution. Estimation itself is undertaken using R [2008] and Stan [2013a]<sup>4</sup>. The priors of the regression coefficients are centred at zero and have a variance proportional to  $\nu$ , the langrange multiplier. This gives the variable selection effect. The hyperprior is parameterized as  $\Gamma(1, b)$  where  $b$  is estimated. This is a somewhat hybrid approach. From the estimation underlying posterior generated is demonstrated in Figure 3.

## 5 Data & Results

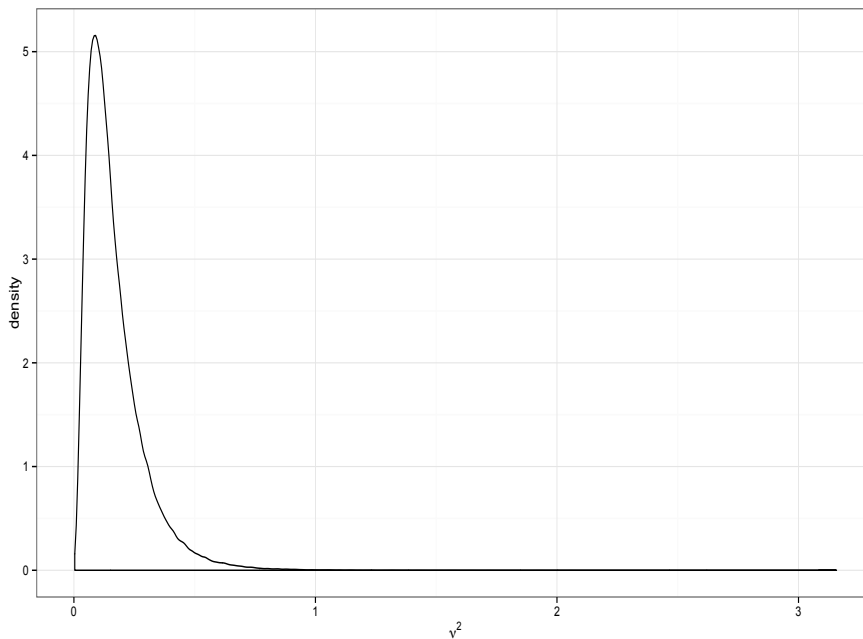
The data used was a standard machine learning example, the diabetes dataset. The summary statistics are presented below. There are 442 observations with the first non-interaction terms were used. Though this is not a  $p \gg n$  situation it serves to demonstrate the technique and places this in the corpus of penalised regression. The estimation was performed using RStan [2013a] & [2013b]. There are a number of convergence (in distribution) issues with shorter chains however using 5 chains with 5000 iterations gave convergence according to the potential scale reduction,  $\hat{R}$  statistic. This statistic measures the average ratio of the variances within chains to the pooled average variance. The estimation converges to  $\hat{R}=1$  supporting convergence in distribution in the chains. The estimation coefficients and associated intervals are given in Table 1.

The MLE approach used a grid of Lagrange multipliers and the coefficients from each of these values are recorded. These are presented as a proportion of the unconstrained

---

<sup>4</sup>Code for replicating the results are available from the corresponding author.

Figure 3: The Posterior Distribution of  $\nu^2$



maximum likelihood estimates<sup>5</sup>. As can be seen the estimates converge to zero as the penalty increases. A number of coefficients were somewhat unstable. It is believed that this is due to the correlations between the variables that makes identification difficult in addition to the relative smoothness of the likelihood functions under specific conditions (examples are given in Azzalini and Capitanio [1999]). The path of the coefficients is given in Figure 4 using a rather course path. These are given as a proportion of the unconstrained estimates (with a sign modification to aid visualisation). This diagram shows the variable selection power of the LASSO, where the variable hits the zero level the variable is not included in the regression.

The skewness parameter has a tendency not to shrink, rather it compensates and becomes more important as the model becomes more parsimonious. It appears to have the impact of dealing with the missing variables' form and the non-normality that this creates. This is demonstrated in Figure 5, with the *leap* in the value occurring where there is the most obvious increase in parsimony.

Using a 10-fold cross validation, the estimates of each of the parameters were plotted to consider the stability of the algorithm. These plots further demonstrate the shrinkage of the coefficients with the increase of the LASSO parameter. We can see that the regression coefficients all shrink towards zero, as does the location parameter, though the skewness parameter,  $\gamma$  increases. This data set therefore trades off the explanatory

<sup>5</sup>Given the the LASSO parameter is re-parameterized as  $\exp^\nu$ , the unconstrained optimum is given as a small step away from the start of the grid search in order to demonstrate the shrinkage across the range.

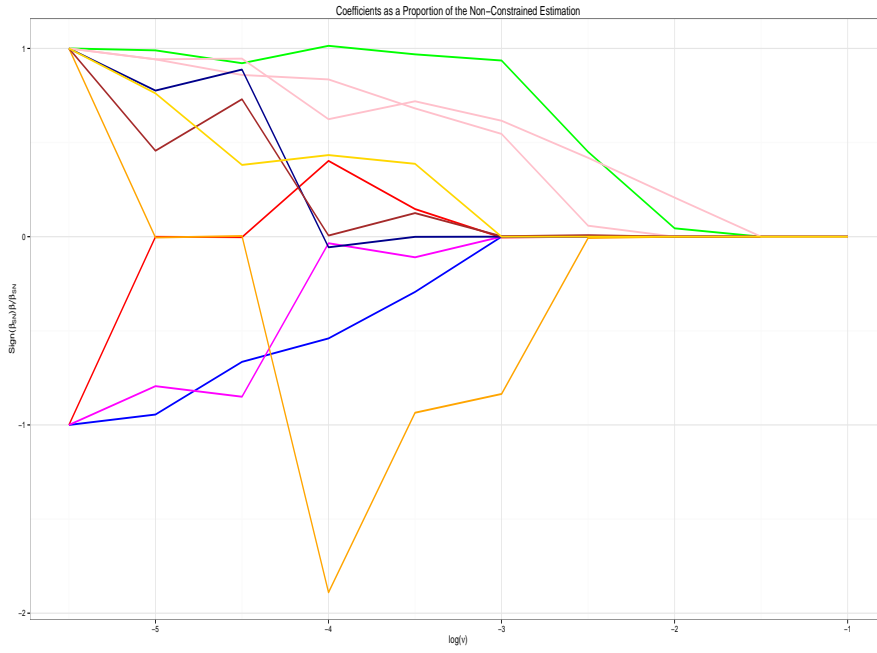


Figure 4: Path of SN Lasso Coefficients by  $\nu$

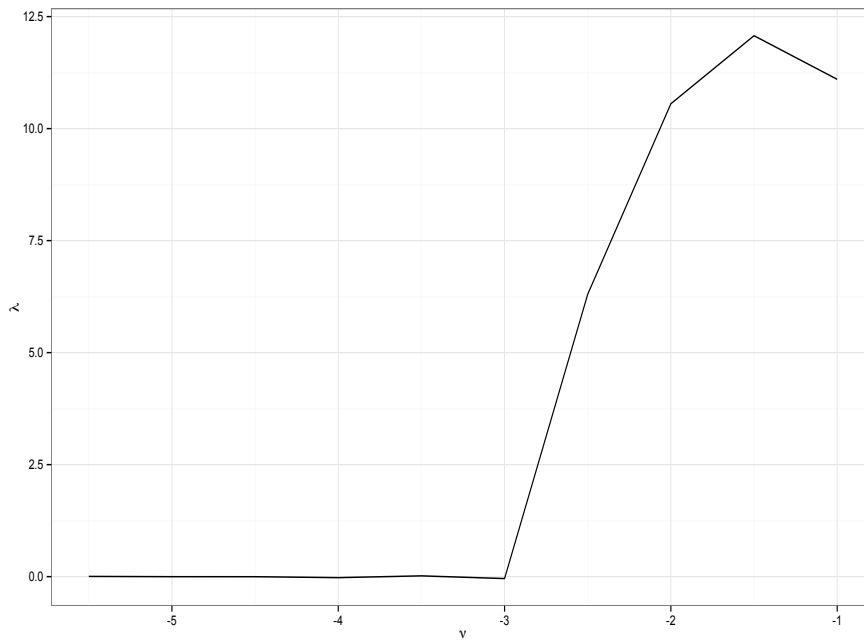


Figure 5: Path of SN Lasso Skewness Coefficients by  $\nu$

power of the regression for increasing the skewness parameter of the distribution.

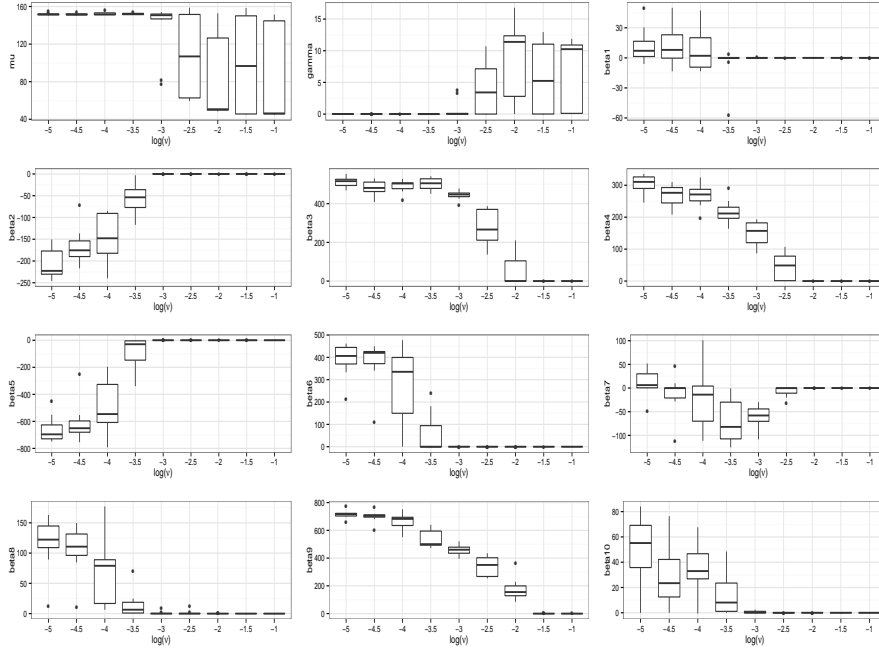


Figure 6: Variation of the Estimates of the Regression Coefficients by  $\nu$

It should be noted that the Bayesian estimates are not point estimates and so there is only limited levels of variable selection- the selection comes through the median being sufficiently close to 0 and the penalty not being as extreme as in the case of the Gaussian LASSO. Following Gelman et al. [2003], the zero point estimate is not considered as a ‘conceptual advantage’. The cross validated (Leave one out) LASSO coefficients are also given in Table 1. These were estimated using `glmnet` (Friedman et al. [2010]) in R. It is noticeable that those variables that the Gaussian LASSO drops are close to zero and certainly within 1.5 standard error of 0 with the exception of  $\beta_8$  which in both the SN and SN-LASSO has a large positive coefficient. OLS and SN-MLE estimates are almost identical due to the low levels of skewness. For a further comparison the OLS based ridge regression is included. The penalty is selected using the approach of Cule and De Iorio [2012] based on cross-validation.

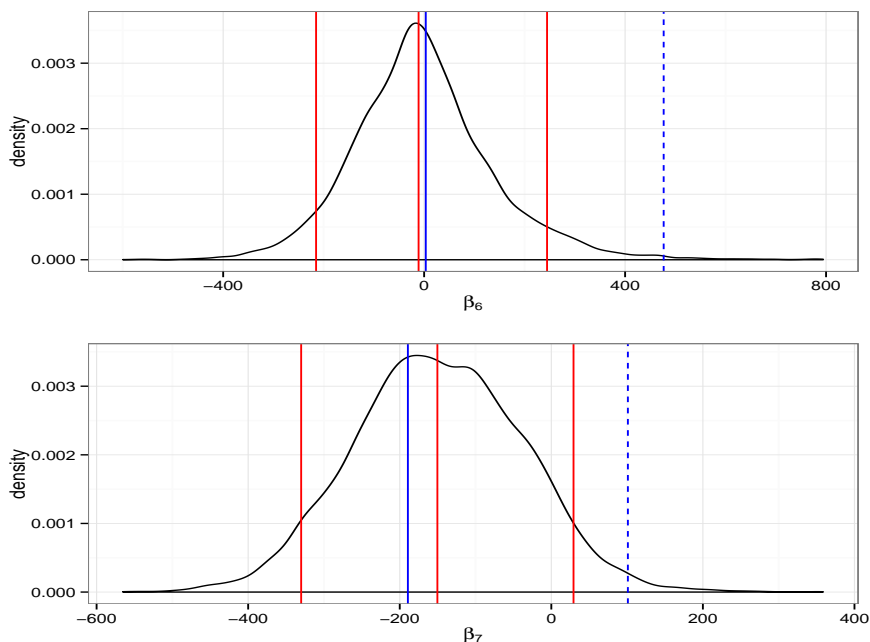
It is clear that in general there is a significant shrinkage in the estimators. Though for  $\beta_6$  and  $\beta_7$  (See Figure 7), this appears to have lead to estimates with the opposite sign; this might be best explained by the relatively high (negative) correlation between the two variables, thus the LASSO does not assign the correct sign as there are known issues with LASSO estimators and highly correlated variables. This is supported by the negative sign on  $\beta_7$  in the LASSO estimation with the Gaussian errors. The MLE point estimates for the Skew Normal regression are included in the Figure as dashed lines. In addition to the fifth and ninety fifth percentiles, the median (red) and mode (blue) are

Table 1: Estimates of the Skew Normal LASSO for Diabetes Data

	Mean	$SE_{mean}$	SD	2.5%	25%	50%	75%	97.5%	$n_{eff}$	$\hat{R}$	SN	SE	CV.LASSO	Ridge	Ridge SE	OLS	$SE_{OLS}$
$\mu$	139.713	0.569	26.289	98.850	117.375	136.296	161.432	188.949	2135	1.002	152.1335	2.544	152.133	152.133	152.133	152.133	2.576
$\beta_1$	-0.207	0.308	28.125	-55.341	-18.665	0.056	18.247	55.503	8313	1.000	-10.012	59.297	-	-4.816	NA	-10.012	59.749
$\beta_2$	-204.564	0.696	61.678	-325.125	-245.895	-204.606	-163.353	-83.008	7845	1.000	-239.819	61.070	-196.053	-228.124	57.599	-239.819	61.222
$\beta_3$	519.982	0.698	64.968	392.485	475.125	519.840	563.844	646.636	8660	1.000	519.840	65.816	522.070	515.391	63.156	519.840	66.534
$\beta_4$	302.573	0.722	64.189	174.912	260.153	302.613	345.071	429.464	7896	1.000	324.390	64.804	296.268	316.125	62.340	324.390	65.422
$\beta_5$	-171.805	3.015	170.602	-545.363	-269.538	-152.920	-53.057	113.150	3203	1.001	-792.184	414.036	-102.047	-206.171	102.045	-792.184	416.684
$\beta_6$	-2.543	2.136	140.929	-260.592	-90.274	-11.062	70.895	309.779	4354	1.001	476.746	337.776	-	13.835	99.620	476.746	339.035
$\beta_7$	-148.884	1.622	111.349	-363.936	-224.283	-150.324	-72.536	64.882	4715	1.001	101.045	209.892	-223.27	-150.203	91.810	101.045	212.533
$\beta_8$	98.629	1.502	116.604	-113.953	18.056	90.506	173.942	343.605	6027	1.001	177.064	159.876	-	115.787	114.508	177.064	161.476
$\beta_9$	517.542	1.482	96.397	335.271	453.122	513.998	580.021	715.689	4229	1.001	751.279	170.958	513.684	518.312	76.632	751.279	171.902
$\beta_{10}$	60.728	0.631	59.268	-52.527	20.130	58.964	100.316	180.243	8826	1.000	67.625	65.334	53.937	75.172	63.061	67.625	65.984
$\nu^2$	58.665	0.692	51.188	12.698	28.214	44.535	71.663	185.667	5469	1.000							
$\nu$	7.665																
$\lambda$	0.356	0.016	0.731	-0.973	-0.215	0.376	0.914	1.700	2147	1.002	0.005	0.101					
$\sigma$	61.377	0.117	6.704	52.316	56.035	59.934	65.563	76.970	3275	1.001	53.476	1.799					
$b$	0.054	0.001	0.062	0.004	0.017	0.035	0.069	0.221	6136	1.000							
$lp$	-2053.50	0.041	2.781	-2059.88	-2055.12	-2053.14	-2051.48	-2049.12	4695	1.000							

also shown. Gelman et al. [2003] suggest that in the case of the LASSO the mode is an useful statistic to use as the coefficient. Though these estimates show little difference from the mean in many cases. The OLS ridge regression also shrinks the coefficients towards 0 however this is not as extreme as that of the LASSO in both the Gaussian and non- Gaussian scenarios.

Figure 7: The Posterior Distribution of Parameters,  $\beta_6$  &  $\beta_7$



Note that  $\nu^2$  is the LASSO tuning parameter with the associated gamma distribution with parameters 1 and  $b$ . The parameter  $b$  supports a posterior not dissimilar to that of Figure 3, though the mode is at 0. This gives a posterior distribution of the LASSO tuning parameter as seen in Figure 8. The 5th and 95th quantiles are marked in red, the median in blue.

It is noticeable that the skewness parameter under the Skew Normal estimation is small and insignificant. The posterior of the  $\lambda$  is seen in Figure 9. The bimodality of the distribution can be traced to the problems associated with the specification of the skewed normal distribution in the *centred* manner as discussed in Azzalini [1985]. This also agrees with the maximum likelihood estimate of the skewness being near zero.

## 6 Conclusions

The skew normal is an example of a well developed class of asymmetric distributions. This paper has shown that it is possible to adapt the estimation of regressions based on this distribution to include a LASSO type penalty. This is seen to shrink the estimates

Figure 8: The Posterior Distribution of the LASSO Parameter

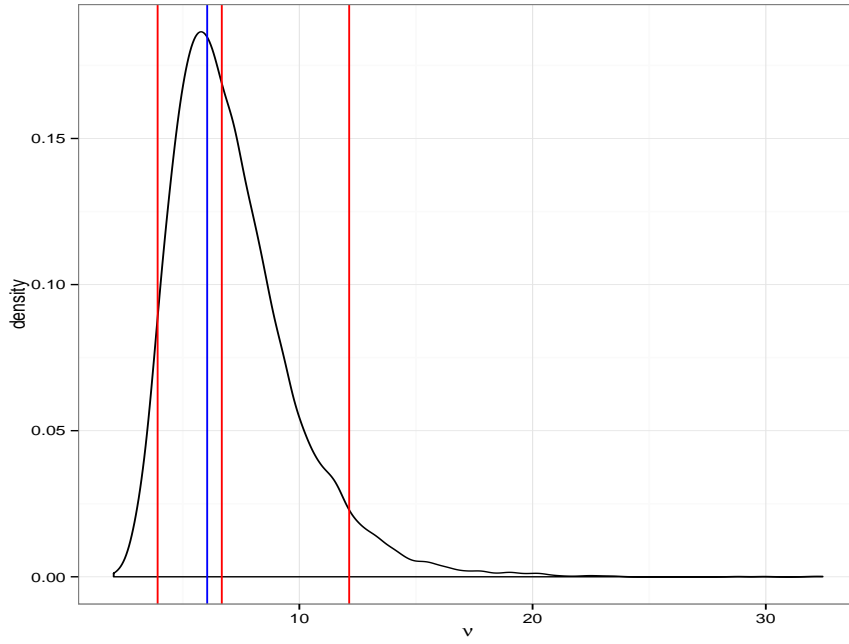
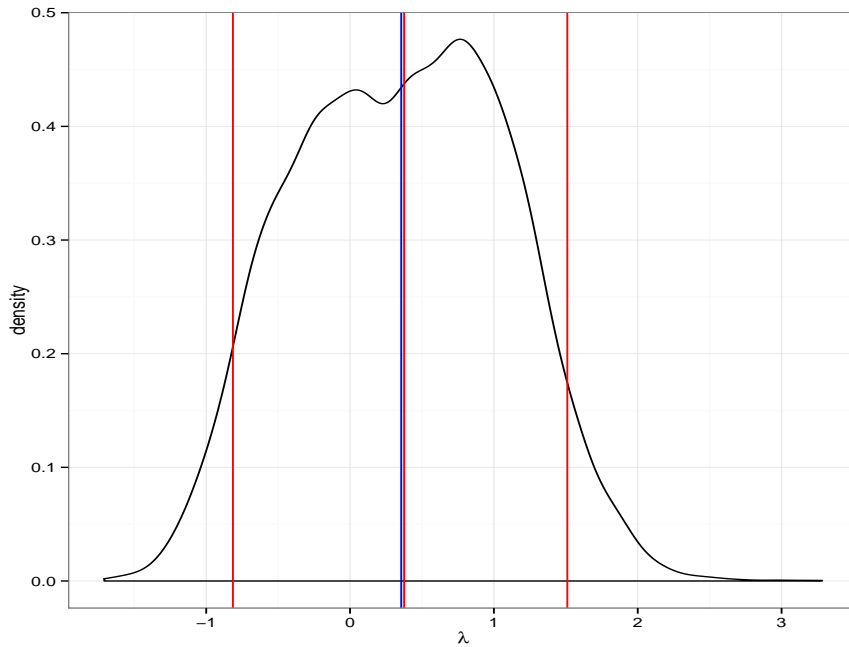


Figure 9: The Posterior Distribution of the Skewness Parameter,  $\lambda$



in most cases and to generate posterior estimates of the parameters of the regressions. These are similar in sign to those of the maximum likelihood with exceptions being potentially driven by high levels of correlation in variables. This behaviour requires further examination.

## References

- C. J. Adock and K. Shutes. Portfolio Selection Based on The Multivariate Skew-Normal Distribution. In A Skulimowski, editor, *Financial Modelling*. Progress and Business Publishers, 2001.
- H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716 – 723, dec 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.
- B. C. Arnold and R. J. Beaver. Hidden Truncation Models. *Sankhya, Series A*, 62 (22-35), 2000.
- A. Azzalini. A Class of Distributions Which Includes The Normal Ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.
- A. Azzalini. Further Results on a Class of Distributions which Includes The Normal Ones. *Statistica*, 46(2):199–208, 1986.
- A. Azzalini and A. Capitanio. Statistical Applications of The Multivariate Skew Normal Distribution. *Journal of The Royal Statistical Society Series B*, 61(3):579–602, 1999.
- P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19 (4):1212–1242, 2013.
- N Chen, R Roll, and S A Ross. Economic Forces and The Stock Market. *Journal of Business*, 59(3):383–403, 1986.
- E. Cule and M. De Iorio. A semi-automatic method to guide the choice of ridge parameter in ridge regression. *ArXiv e-prints*, May 2012.
- B. Efron, R. Tibshirani, I. Johnstone, and T. Hastie. Least angle regression. *The Annals of Statistics*, 32(2):407–499, April 2004. ISSN 0090-5364. doi: 10.1214/009053604000000067. URL <http://projecteuclid.org/Dienst/getRecord?id=euclid.aos/1083178935/>.
- 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.
- J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.



- J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL <http://www.jstatsoft.org/v33/i01/>.
- W. Fu and K. Knight. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.
- A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. CRC press, 3 edition, 2003.
- T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning; Data Mining, Inference & Prediction*. Springer Verlag, 2008.
- A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL <http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634>.
- T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. doi: 10.1198/016214508000000337. URL <http://amstat.tandfonline.com/doi/abs/10.1198/016214508000000337>.
- P. T. Pope and J. T. Webster. The use of an f-statistic in stepwise regression procedures. *Technometrics*, 14(2):327–340, 1972.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.
- K. Shutes. *Non-Normality in Asset Pricing- Extensions and Applications of the Skew-Normal Distribution*. PhD thesis, University of Sheffield, 2005.
- Stan Development Team. Stan: A c++ library for probability and sampling, version 1.3, 2013a. URL <http://mc-stan.org/>.
- Stan Development Team. *Stan Modeling Language User’s Guide and Reference Manual, Version 1.3*, 2013b. URL <http://mc-stan.org/>.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- L.-C. Wu, Z.-Z. Zhang, and D.-K. Xu. Variable selection in joint location and scale models of the skew-normal distribution. *Journal of Statistical Computation and Simulation*, pages 1–13, 2012. doi: 10.1080/00949655.2012.657198. URL <http://www.tandfonline.com/doi/abs/10.1080/00949655.2012.657198>.
- H. Zou. *Some Perspectives of Sparse Statistical Modeling*. PhD thesis, Stanford University, 2005.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. URL <http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x>.

Preliminary Draft