Regularized Skew-Normal Regression

Shutes, Karl and Adcock, Chris

Coventry University, University of Sheffield

24 November 2013

Regularized Skew-Normal Regression

K. Shutes & C.J. Adcock

March 31, 2014

Abstract

This paper considers the impact of using the regularisation techniques for the analysis of the (extended) skew-normal distribution. The models are estimated using Maximum Likelihood and Bayesian estimation techniques and compared to OLS based LASSO and ridge regressions in addition to non- constrained skew-normal regression. The LASSO is seen to shrink the model's coefficients away from the unconstrained estimates and thus select variables in a non- Gaussian environment.

1 Introduction & Motivation

Variable selection is an important issue for many fields. Further it is noticeable that not all data conforms to the standard of normality. This paper addresses the issue raised by Bühlmann [2013] of the lack of non-Gaussian distributions using the regularisation methods. Within the statistics literature there are many applications of penalised regressions. There are other fields such as finance and econometrics where these approaches are less common. 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. 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. The situation of 'excessive data' can be dealt with by the regularised regressions, such as the Least Absolute Shrinkage & Selection Operator (henceforth 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.

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 [2004]). The method proposed here uses the technique of the LASSO, i.e. the introduction of ℓ_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 more fully described). All estimation was performed in R [2008] with package Azzalini [2013] and RStan [2013a] & [2013b][1].

2 Literature Review & Definitions

2.1 Regularization

Within the econometric literature, regularisation has a limited history, though in many other fields it is a well established technique. In circumstances of ill-formed problems, such as multi-collinearity or non-full rank in the independent variable matrix, it is possible to use to use these approaches. 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

[1] Code for replicating the results are available from the corresponding author.

with lower standard errors. The penalised function for the estimation is given by:

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

This approach does not perform any form of variable selection as, although it does shrink coefficients, it does not shrink them to 0. The ν parameter[2] acts as the shrinkage 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 ℓ_1 norm rather than the ℓ_2 norm of the ridge approach. Hence the problem becomes:

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

There is a well-known mapping between the multiplier ν and the constraint of the sums of the coefficients ϵ. 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 LASSO shrinks β_1 to 0, whereas the ridge regression approaches it. The OLS estimator is given as $\hat{\beta}$. The parameter ν 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 γ-th norm is used. This is the bridge estimator. There are a number of similarities between the bridge estimator[3] with $1 < \gamma < 2$ 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 γ-th norm is defined as:

[2]Traditionally the Lagrangean multiplier is denoted λ, however due to the use of λ as the skewness parameter in the distribution, the Lagrangean is denoted ν.

[3]The limits here are LASSO and ridge regressions.

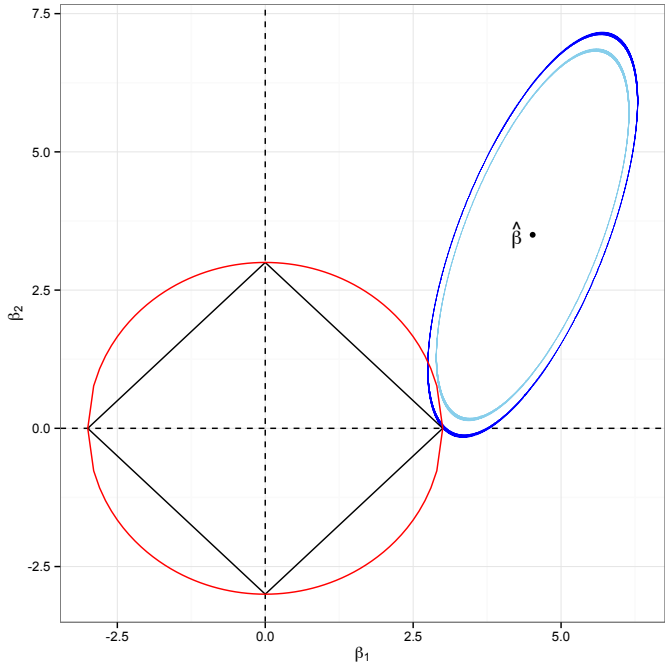


Figure 1: Differences Between LASSO & Ridge Regressions

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

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.

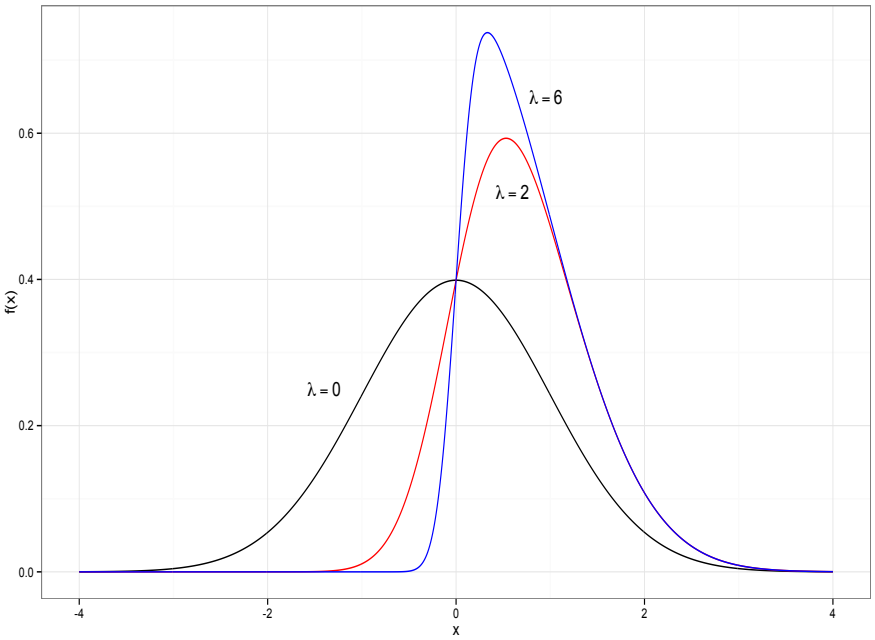
2.2 The Skew-Normal Distribution

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:

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

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

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



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 τ. Thus the density function can be written as:

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

where ϕ and Φ are the probability density and cumulative functions of the normal distribution respectively.

It is possible to use the following parameterization, with γ and ω^2 being the mean and the variance of the normal part of the distribution respectively:

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

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:

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

where $\phi(x;\mu,\sigma^2)$ is the probability density function of a normally distributed variable with mean μ and variance σ^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:

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

This is the standard log-likelihood function for the extended skew-normal with the addition of the LASSO penally 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:

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

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

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

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

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

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

4 Estimation

For Gaussian based estimations. 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. 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[4] and speed.

The estimations here are for the Azzalini form of the distribution i.e. $\tau = 0$.

4.1 Estimation with Maximum Likelihood

Estimation was performed using a maximum likelihood approach with the nuisance parameter, ν 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

[4]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.

estimates as the initial points to aid in convergence, the estimations were performed with a transformation of the parameter ν to $\exp(\nu)$. This leads 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, β & λ 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[5]. The former approach is taken here. The prior for regression coefficients, β and the LASSO parameter, ν_i is based, as suggested by Park & Casella, on

$$g(\beta \mid \sigma^2) = \prod_{j=1}^{p} \frac{\nu}{2\sqrt{\sigma^2}} \exp^{-\nu|\beta_j|\sqrt{\sigma^2}} \tag{14}$$

$$f(\nu^2 \mid \xi, \theta) = \frac{\theta^{\xi}}{\Gamma(\xi)} (\nu^2)^{\xi-1} \exp(-\theta\nu^2) \tag{15}$$

The square ensures a proper posterior distribution. The priors of the regression coefficients are centred at zero and have a variance proportional to ν, the Lagrange multiplier. This gives the variable selection effect. The hyperprior is paramterized as $\Gamma(1, b)$ where b is estimated. This is a somewhat hybrid approach.

5 Data & Results

The data used was a standard machine learning example, the diabetes dataset. These relate the progress of diabetes over a year to the age, weight, BMI and various serum measurements. There are 442 observations with the first non-interaction terms used. The data are standardised to have 0 mean and an unit ℓ_2-norm. Though this is not a $p >> n$ situation it serves to demonstrate the technique and places this approach in the corpus of penalised regression.

There are two elements to the results, the frequentist and the Bayesian analyses. The Maximum Likelihood approach used a grid of Lagrange multipliers and the coefficients from each of these values are recorded. These are presented in Table 2 and graphically in Figure 3 with the coefficients presented as a proportion of the unconstrained maximum likelihood estimates[6]. 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

[5] In the case of estimation of an elasticnet problem, a modified prior including a term in $|| \beta ||_2^2$ is used in addition to that of the L_1 norm.

[6] Given that 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.

(examples are given in Azzalini and Capitanio [1999]). The path of the coefficients is given in Figure 3 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 ability of the LASSO. Table 1 shows the convergence (and implicitly the speed of the convergence to 0). If the mean and median are both close to 0, the coefficient is constrained early in the path and remains at or near 0 for the path. Otherwise there is a substantial range in which the variable appears in the model. It is noticeable that the location and dispersion parameter vary; this is related to the variation in the skewness parameter- in certain cases the skewness increases to fit the model and this has an impact (especially on the dispersion, σ).

	μ	σ	λ	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Min.	98.05	100.00	-519.40	-0.05	-0.00	-0.00	0.00	-0.00	-0.00	-119.40	-12.61	0.00	0.00
1st Qu.	99.34	100.20	147.50	-0.00	-0.00	0.00	0.00	-0.00	0.00	-65.55	0.00	0.00	0.00
Median	99.58	105.00	295.70	-0.00	0.00	90.44	53.36	0.00	0.00	-13.94	0.00	57.30	0.03
Mean	99.54	117.00	278.40	1.67	25.31	61.49	42.20	21.10	21.00	-28.96	12.89	47.95	24.16
3rd Qu.	99.62	144.00	462.60	2.23	78.20	95.41	89.73	71.48	72.36	-0.00	47.33	90.85	67.06
Max.	100.80	144.00	1259.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	109.10

Table 1: Summary of Convergence of Coefficients
100% represents equality to unconstrained MLE

For the Bayesian approach, there are a number of convergence (in distribution) issues with shorter chains however using 10 chains with 7500 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, indeed all are less than 1.05 which is considered a practical level of the measure. The estimation coefficients and associated intervals are given in Table 2.

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 4, 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. There is some variability in the estimation of the regression coefficients at small penalties (i.e. near unconstrained solutions), but this is reduced as the constraint is more strongly enforced. In contrast the location and skewness parameter variability increases in value as the penalty increases. 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, γ increases. This data set therefore trades off the explanatory power of the regression for increasing the skewness parameter of the distribution.

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. [2013], the zero point estimate is not considered as a

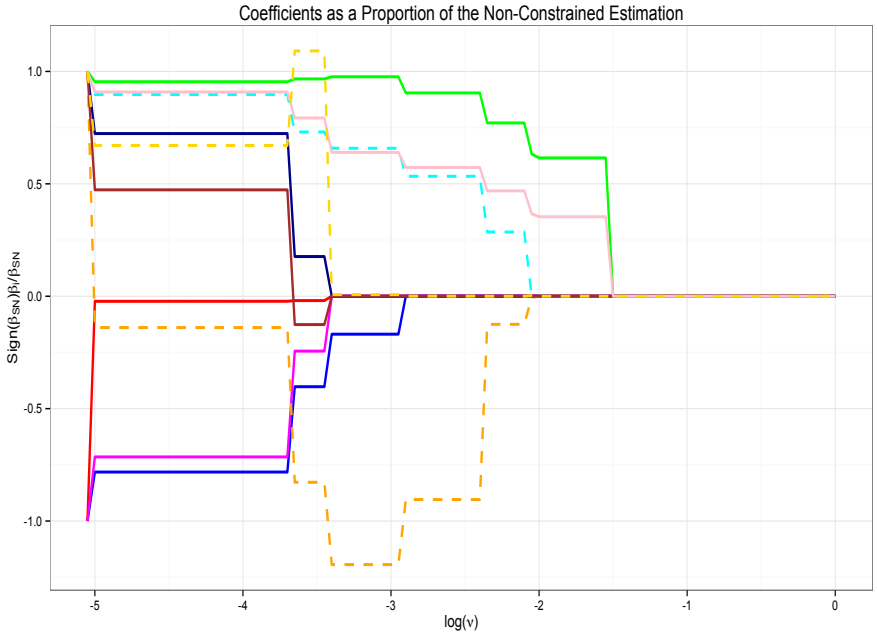


Figure 3: Path of SN Lasso Coefficients by ν

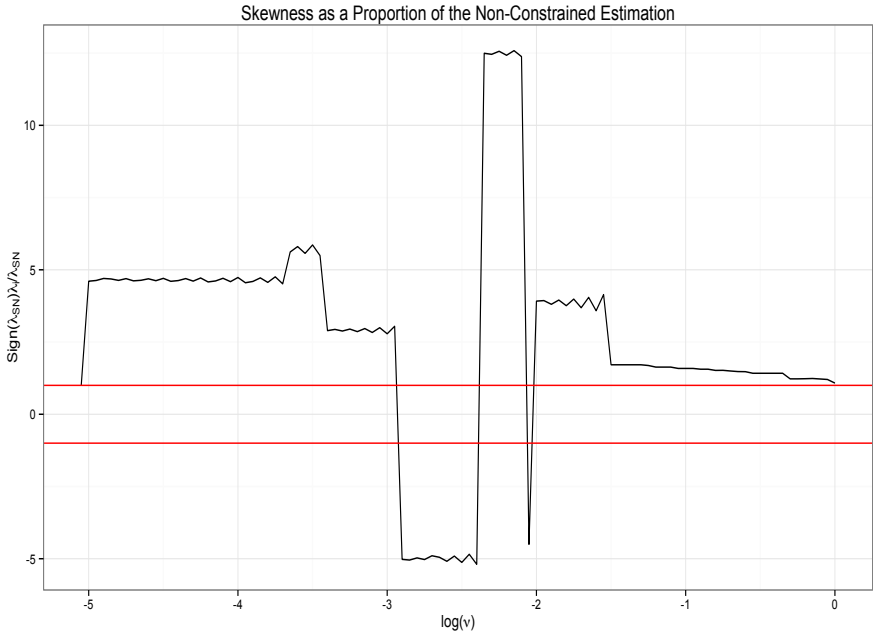


Figure 4: Path of SN Lasso Skewness Coefficients by ν

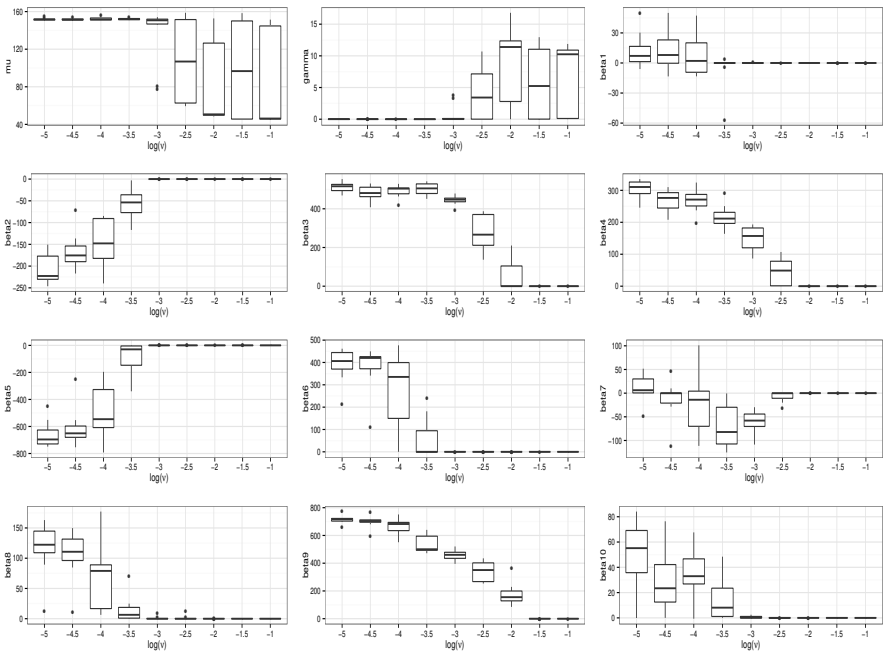


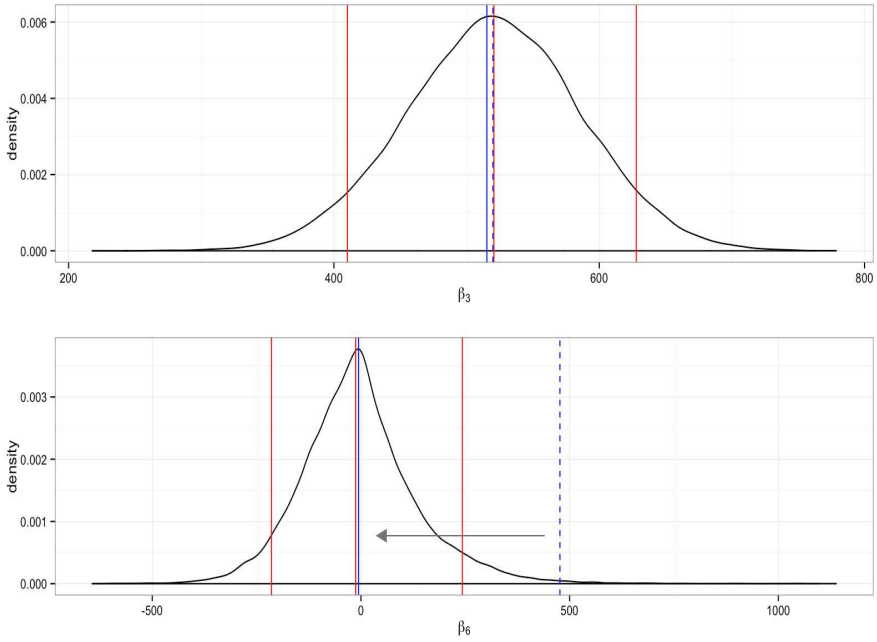
Figure 5: Variation of the Maximum Likelihood Estimates of the Regression Coefficients by ν

'conceptual advantage'.

The (leave one out) cross validated LASSO gaussian (OLS) coefficients are also given in Table 2. 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 β_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, with the variable selection demonstrated by the comparison of the coefficients of β_6 and β_3 as shown in Figure 6. The first is reduced towards zero in comparison with its unconstrained Maximum Likelihood Estimator, whereas that of β_3 remains near the ML estimate under this approach. This bears out the frequentist MLE where the coefficient β_3 remains relatively large over much of the path. The 5th and 95th quantiles are marked in red, the median in blue. For coefficients β_6 and β_7 (See Figure 7), the shrinkage

Figure 6: The Posterior Distribution of Parameters, β_3 & β_6



appears to have led 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 β_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, and it should be noted that the

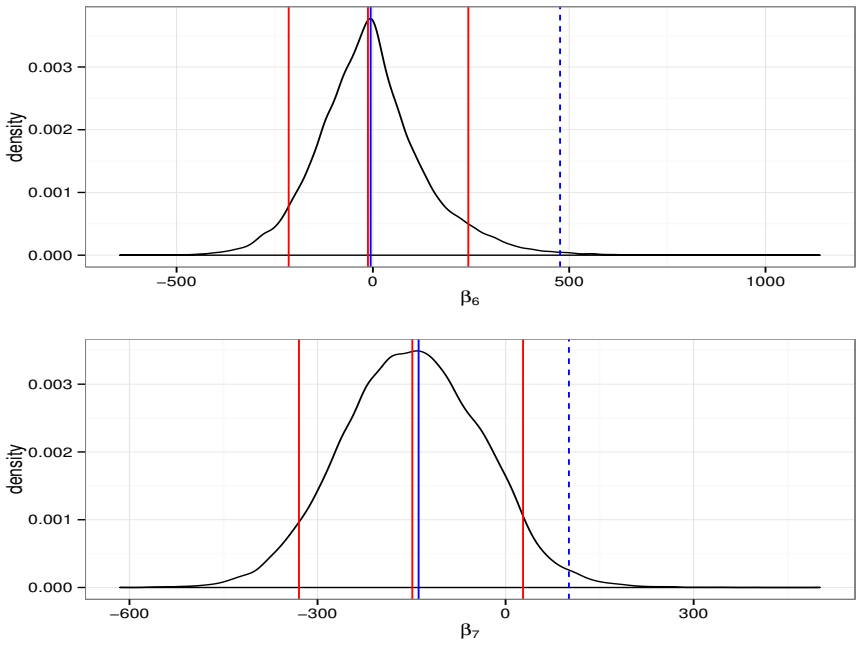
Table 2: Estimates of the Skew Normal LASSO for Diabetes Data

Parameters	Bayesian Estimation Results										SN MLE		LASSO	Ridge		OLS	
	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}
μ	138.804	0.305	26.168	99.076	116.685	134.873	160.292	188.557	7376	1.002	152.1335	2.544	152.133	152.133	NA	152.133	2.576
β_1	-0.439	0.181	27.541	-54.742	-18.563	-0.503	17.496	54.207	23066	1.000	-10.012	59.297	-	-4.816	57.599	-10.012	59.749
β_2	-204.579	0.397	60.746	-325.125	-245.895	-204.606	-163.353	-83.008	23371	1.000	-239.819	61.070	-196.053	-228.124	58.710	-239.819	61.222
β_3	520.142	0.403	65.989	389.406	476.073	520.622	564.471	648.292	26823	1.000	519.840	65.816	522.070	515.391	63.156	519.840	66.534
β_4	302.730	0.406	64.373	176.437	259.475	302.902	346.036	428.455	25090	1.000	324.390	64.804	296.268	316.125	62.340	324.390	65.422
β_5	-169.967	1.717	169.493	-553.173	-266.653	-151.408	-52.354	110.4963	9750	1.001	-792.184	414.036	-102.047	-206.171	102.045	-792.184	416.684
β_6	-4.908	1.316	140.106	-261.343	-90.961	-12.553	67.062	309.328	11338	1.001	476.746	337.776	-	13.835	99.620	476.746	339.035
β_7	-148.517	0.968	110.571	-362.968	-224.255	-148.753	-72.136	63.029	13039	1.000	101.045	209.892	-223.27	-150.203	91.810	101.045	212.533
β_8	99.930	0.902	115.514	-109.083	18.142	91.136	175.254	344.467	16399	1.000	177.064	159.876	-	115.787	114.508	177.064	161.476
β_9	515.742	0.809	96.940	332.851	450.464	513.426	577.528	714.491	14373	1.001	751.279	170.958	513.684	518.312	76.632	751.279	171.902
β_{10}	61.216	0.384	59.344	-49.782	20.012	58.940	100.863	181.131	23945	1.000	67.625	65.334	53.937	75.172	63.061	67.625	65.984
ν^2	58.326	0.412	50.850	12.279	28.207	44.386	71.11	188.332	15247	1.000							
ν	7.637																
λ	0.382	0.008	0.729	-0.954	-0.185	0.414	0.942	1.694	7424	1.002	0.005	0.101					
σ	61.508	0.065	6.744	52.361	56.100	60.051	65.881	76.906	10668	1.001	53.476	1.799					
b	0.055	0.0004	0.063	0.004	0.017	0.036	0.069	0.223	20109	1.000							
lp	-2053.46	0.024	2.726	-2059.69	-2055.07	-2053.12	-2051.49	-2049.15	13097	1.001							

Key:
Bayesian Estimation Results= Estimation of MCMC Skew Normal LASSO
SN MLE= Estimation of Skew Normal by MLE
LASSO= Gaussian based LASSO with penalty parameter estimated using Cross Validation
Ridge= Gaussian based Ridge with penalty parameter estimated using Cross Validation
OLS= Gaussian based regression

standard error on this coefficient is relatively large, thus the coefficient may actually be negative and within the estimate of the posterior. In addition to the fifth and ninety fifth percentiles, the median (red) and mode (blue) are also shown. Gelman et al. [2013] & Hastie et al. [2008] suggest that in the case of the LASSO the posterior mode is an useful measure to use as the coefficient, rather than the posterior mean in the case of the ridge estimator. 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, β_6 & β_7



Note that ν^2 is the LASSO tuning parameter with the associated gamma distribution with parameters 1 and b. This gives a posterior distribution of the LASSO tuning parameter as seen in Figure 8.

It is noticeable that the skewness parameter under the Skew Normal estimation is small and insignificant. The posterior of the λ 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

Figure 8: The Posterior Distribution of the LASSO Parameter

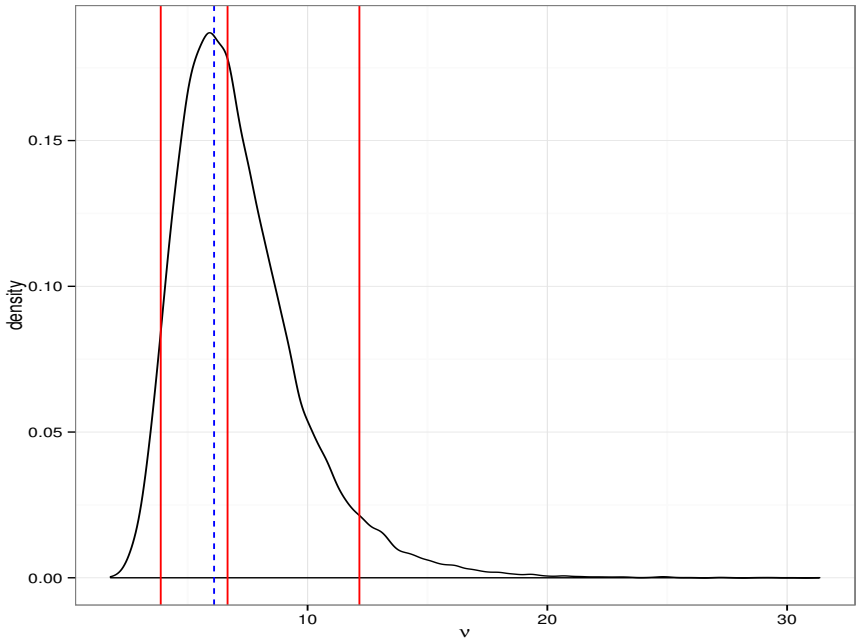
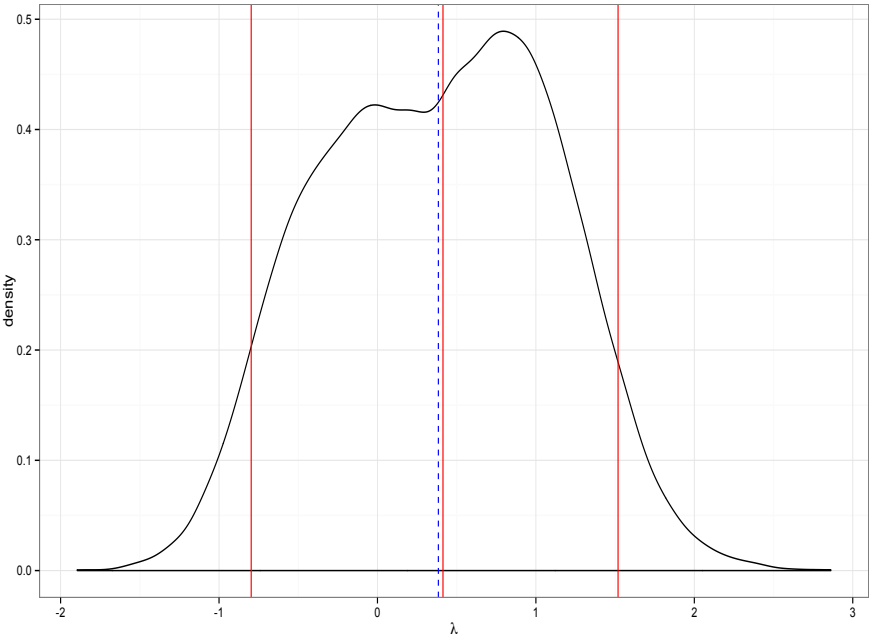


Figure 9: The Posterior Distribution of the Skewness Parameter, λ



this distribution to include a LASSO type penalty. This is seen to shrink the estimates of regression coefficients and thus perform a variable selection role. The shrinkage occurs in both Bayesian and frequentist paradigms. It is possible 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.

Natural extensions from this work include a generalisation from the skew normal distribution to include other, spherically symmetric distributions. These such as the skew Student distribution would increase the application of these approaches to situations where higher moments are critical such as finance.

References

C. J. Adcock and K. Shutes. Portfolio Selection Based on The Multivariate Skew-Normal Distirbution. 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. *R package* `sn`*: The skew-normal and skew-t distributions (version 0.4-18)*. Università di Padova, Italia, 2013. URL `http://azzalini.stat.unipd.it/SN`.

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, 3rd edition, 2013.

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, 2004.

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, Standford 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.