Regularized Extended Skew-Normal Regression

Karl Shutes and Chris Adcock

Coventry University, University of Sheffield

24 November 2013

Regularized Extended Skew-Normal Regression

K. Shutes & C.J. Adcock

October 29, 2016

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 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. A number of approaches such as stepwise regression or best subset regression are widely used with metrics such Aikake's Information Criteria (Akaike [1974]) or Mallows C_p employed as the decision criterion. There are well documented problems with these approaches. The use of regularized regressions mitigate these problems; the coefficients are shrunk towards zero, which creates a selection process. In the majority of cases, the use of the regularization techniques is based upon a linear model and Ordinary Least Squares. That is, it is assumed implicitly that the residuals in the model are normally or at least symmetrically distributed. There are, however, applications for which the residuals in a model are neither normally nor symmetrically distributed. This paper addresses the issue raised by Bühlmann [2013] of the lack of non-Gaussian distributions using regularisation methods. Specifically, this paper adds to the regularization literature by applying the Least Absolute Shrinkage & Selection Operator (henceforth LASSO (Tibshirani [1996]) to accommodate shrinkage when the residuals in a linear model follow the extended skew normal based regression model (Adcock and Shutes [2001], Arnold and Beaver, [**?**]). The procedures described in this paper provide regularization not only for the mean, but also for the parameters that regulate skewness, kurtosis and higher moments.

The paper is organized as follows. Section 2 presents necessary background material and Section 3 summarizes the estimation procedures. Section 4 contains a substantial example based on simulated data. The aim of the example is to demonstrate the skew-normal LASSO in operation. This is followed in Section 5 by three empirical studies based on real data. The first study uses a standard data set from the machine learning literature, namely that of diabetes patients (see Efron et al. [2004] where it is more

fully described). The second demonstrates how to construct a replicating portfolio; that is, a portfolio that designed to replicate the performance of a given index using the ESN-LASSO to minimize the number of constituent names in the replicator. The final empirical study applies the ESN-LASSO to the bicycle hire data set of Fanaee-T and Gama Fanaee-T and Gama [2013]. In general the paper uses standard notation.

2 Background

There are two sub-sections. The first presents a short summary of standard regularization methods. The second presents the skew-normal and extended skew-normal distributions.

2.1 Regularization

Regularization has a substantial history and is widely used in many fields, often for problems which are ill-conditioned.

Ridge regression is perhaps the best known example (see Hoerl & Kennard [1970], for example, for further details), where the problem of multicollinearity may be dealt with by the imposition of a penalty on the coefficients of the regressions. The resulting estimators of the parameters are biased, but have lower estimated standard errors than those obtained from the standard application of OLS.

In the usual notation the penalised objective function to be minimised is:

$$\begin{aligned} \beta_R &= \arg\min_{\beta} \left(Y_i - \beta_0 - X_i\beta^T\right)^T \left(Y_i - \beta_0 - X_i\beta^T\right) + \nu\beta^T\beta \\ &= \left(X^TX + \nu I\right)^{-1} X^Ty \end{aligned}$$

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[1] 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) introduced by Tibshirani [1996]. 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:

$$\beta_L = \arg\min_{\beta} \left\{ (Y - X\beta)^T (Y - X\beta) + \nu \mid\mid \beta \mid\mid_1 \right\} \tag{1}$$

In general the intercept is not shrunk in which case the quadratic form in the above equations is

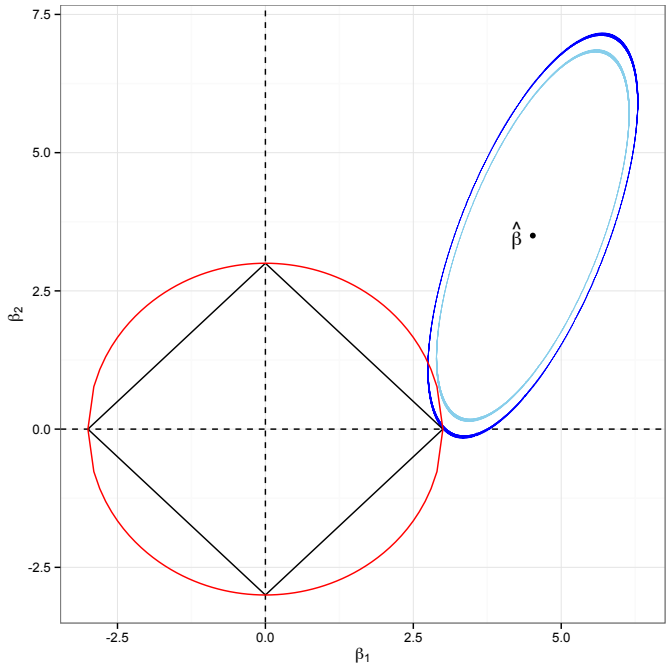
$$(Y - \beta_0 1 - X\beta)^T (Y - \beta_0 1 - X\beta),$$

where 1 denotes a vector of ones.

[1] 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 throughout this paper.

The variable selection property is clearly shown graphically in Figure 1 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

Figure 1: Differences Between LASSO & Ridge Regressions



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 of the coefficients are consistent & that these will have the same limiting distribution as the OLS estimates.

2.2 The Skew Normal Distribution

The skew skew normal distribution [SN henceforth] has become increasingly well used within a number of fields since its initial description by Azzalini [1985] and [1986]. The standard form of the distribution has the probability density function

$$f(y) = 2\phi(y)\,\Phi(\lambda y)\,; -\infty < \lambda < \infty, -\infty < y < \infty, \tag{2}$$

with λ controlling the degree of skewness of the distribution. The case λ=0 will lead to a standard normal distribution.

Azzalini [1985] & [1986] show that the SN distribution may be thought of as the convolution of a normally distributed variable and an independently distributed normal variable which has a mean of zero and which is truncated from below at zero. This

is generalized in Arnold & Beaver [2000] and Adcock & Shutes [2001] where the truncated normal variable has a mean of τ, which may take any real value.Using what is normally referred to as the central parametrization[2], the probability density function of the extended skew normal [ESN] distribution is

$$f(y) = \frac{1}{\omega\Phi(\tau)}\phi\left(\frac{y-\mu}{\omega}\right)\Phi\left\{\tau\sqrt{1+\lambda^2}+\lambda\left(\frac{y-\mu}{\omega}\right)\right\}. \tag{3}$$

where ϕ and Φ are the standard normal probability density and distribution functions respectively and for a linear regression model $\mu = x^T\beta$. Using this notation, the objective function to be minimized for a random sample of size n corresponding to equation (2.1) is

$$-\sum_{i=1}^{n}\log f(y_i) + \nu \mid\mid \beta \mid\mid_1 . \tag{4}$$

Under the ESN distribution, it is also possible to shrink the parameters λ and τ, which control estimates of skewness and other higher moments, using a different shrinkage parameter in each case. The objective function to be minimized is

$$-\sum_{i=1}^{n}\log f(y_i) + \nu_1 \mid\mid \beta \mid\mid_1 + \nu_2 \mid\mid \lambda \mid\mid_! + \nu_3 \mid\mid \tau \mid\mid_1 . \tag{5}$$

Applications of the LASSO in conjunction with the SN or ESN distributions are limited. Wu et al. [2012] consider the variable selection problem for the SN distribution, for which $\tau = 0$. The skewness parameter λ is estimated but is then treated as fixed and omitted from the SN-LASSO; that is, Wu et al minimize the objective function at (4) using a fixed value of $\hat{\lambda}$. Their 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.

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

$$\begin{aligned}\ell_i(y;\tau,\lambda,\omega.\beta) &= -\log(\sqrt{2\pi}) - \log(\omega) - \frac{1}{2}z_i^2 \\ &\quad + \log\left(\Phi\left(\tau\sqrt{1+\lambda^2}+\lambda z_i\right)\right) \\ &\quad - \log\left(\Phi(\tau)\right) + \nu\left(\mid\mid \beta \mid\mid_1 + \mid\mid \lambda \mid\mid_1 + \mid\mid \tau \mid\mid_1\right) \\ \text{where } z_i &= \frac{y_i - \beta x_i}{\omega}\end{aligned} \tag{6}$$

[2]A different parameterization is given in Adcock and Shutes [2001]. This form of ESN distribution is not considered in this paper.

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.

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[3] and speed. The approach taken here is to use direct estimation of the likelihood function for the distributions where τ is unconstrained (the extended skew normal) and where it is constrained to $\tau = 0$, the skew normal. Each estimator used the previous estimate as the starting point of the algorithm to increase the speed of the estimation. Where λ is small, it is sometimes difficult to estimate τ stably. This is reflected in a number of the results where the skewness based estimates are rather volatile and not always non-increasing as one would hope with the LASSO type estimators.

4.1 Estimation with Maximum Likelihood

The procedure uses maximum likelihood optimisation with the penalty parameter based upon a grid. The value of the penalty was selected using a cross-validation procedure. Initially the unconstrained maximum likelihood estimation was performed. The results of this are used as the first estimates for the penalised optimisation. Each optimisation is used as the next starting point for the following procedure. This speeds up the estimation.

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 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.

The estimation of ν used a 10-fold cross-validation over an identical grid of ν parameter values. The mean squared cross validation errors are calculated off the hold-out sample of this, with the ν selected by the min+1S.E. rule of thumb being used as a fixed parameter within the final, whole sample estimation. Thus the process involves sampling in order to estimate the nuisance parameter, with that value then being used to select the model using the whole data set. Alternatives such as a BIC minimisation

[3] As noted in Azzalini and Capitanio [1999] the likelihood function of the skew normal is not convex in its standard form.

or cross-validation based on the (negative) log likelihood are also possible as suggested in Städler et al. [2010] , though these were not used.

The results of the estimations are presented graphically using the logarithm of the penalty and the coefficients as a proportion of the unconstrained maximum likelihood value [4].

5 Data & Maximum Likelihood Estimation Results

A number of data sets are included to demonstrate the approach. These are a simulated dataset with the second being the diabetes data set (from Efron et al. [2004]). The simulations are based on 10 variables with 10000 and 1000 observations each. Fifty different sets of data are used to demonstrate the properties of the estimation. These are aggregated using the mean coefficients and their standard error, as well as the median and 25th and 75th quantiles of the estimates in order to demonstrate the properties of the estimators.

The diabetes data relate the progress of diabetes over a year to the age, weight, BMI, sex, blood pressure and six serum measurements. There are 442 observations. 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.

The bicycle dataset considers the bike rental in Washington DC and inevitably has a degree of skewness generated by the non-zero nature of the data, though this is reduced somewhat by using logarithms in the model. The path for all the variables in the estimation is as one might expect for a LASSO type estimator, though the skewness parameter remains strong throughout.

Further empirical studies using financial data and the rent of bicycles are also considered. These data both exhibit skewness and kurtosis and so the extended skew normal distribution is an attractive tool for modelling these data. The financial application looks to replicate two Chinese indices using other, global indices, which may be more liquid or accessible. The non-convexity of the likelihood gives rise to some local issues in the solution to the LASSO path in these data sets, however these are seen to be only localised and away from the cross validated solutions.

All estimations are performed using RR Core Team [2016] and the packages bbmle Bolker and Team [2016], glmnet Friedman et al. [2010] and sn Azzalini [2015].

[4] This may be substituted for the maximum value in some cases, e.g. where the Maximum Likelihood estimator is not available.

5.1 Simulated Data

Fifty simulated data sets of 1000, 10000 and a smaller sample of 100 observations were created with specific seeding points to ensure reproducibility. These contained 10 independent variables. Two data generating processes were used identical for all of the simulations ($\beta_1 < 0$, $\beta_2 < 0$, $\beta_3 > 0$ and $\beta_5 > 0$ are all non-zero; other coefficients are equal to zero), the second had coefficients of ± 1 for β_1, β_2, β_3 and β_5). The estimates are reported as a proportion of the full Maximum Likelihood estimators. In each case, the inter-quartile range and median are plotted in the first graph and the mean of the estimators is plotted in the second. These demonstrate the important variables in the data generating process clearly.

Those variables that are included in the data generating process are stable around the MLE coefficients (qv. Figures 2, 4a and 5a), whereas those omitted from the data generating process are restricted and converge to zero (Figure 2, 4a & 5a) . These have a wider dispersion than the variables included in the data generating process.

5.1.1 Non-Unit Coefficients

Results for both simulation lengths are similar in substance, though the dispersion is higher in the smaller data sets. In both cases the skewness parameters (λ and τ) converge to zero as the penalty increases even though the actual value is not zero (Figures 3, 4b, 5b). This is in part due to the non-linearities associated with the likelihood function. The instability that this creates gives a median value of zero. The model is penalising the asymmetry and removing it from the regression in these cases.

Figure 2: Spread of LASSO Regression Coefficients (β) of Variables by ν (N=10000)

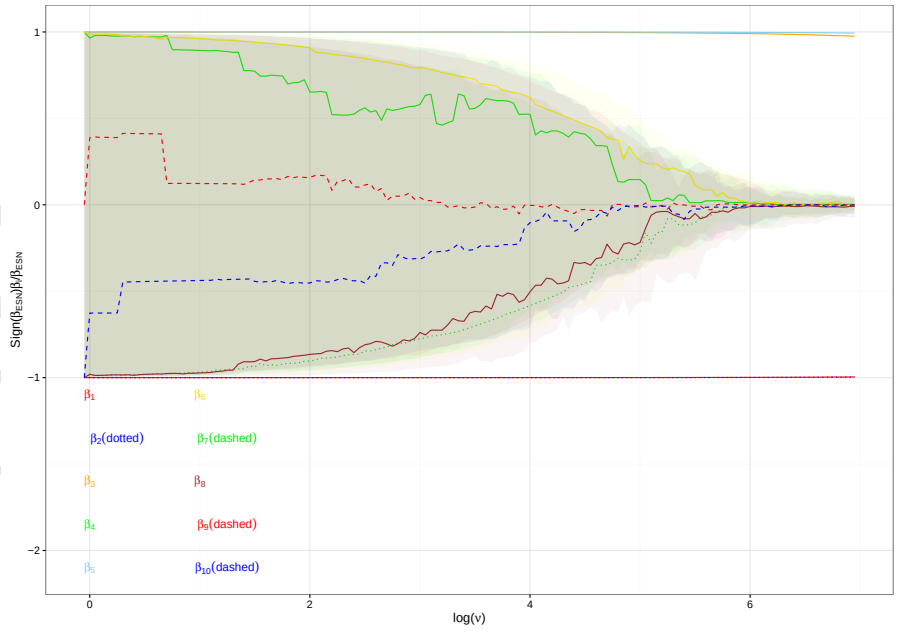
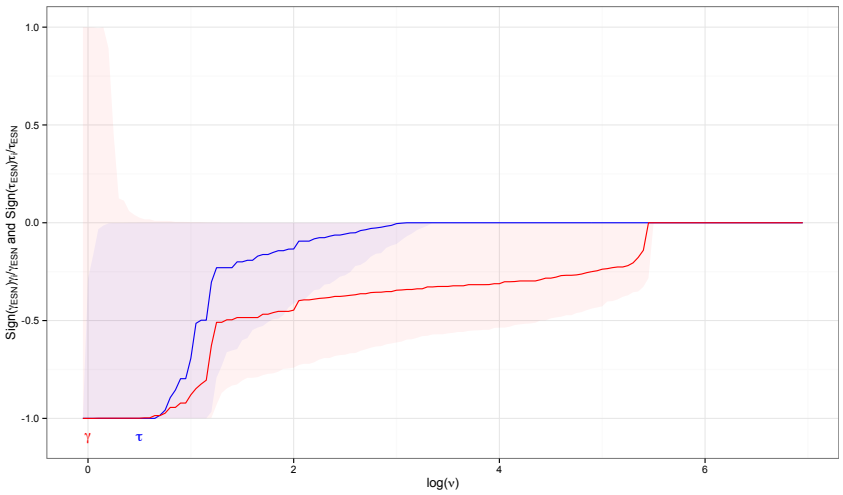


Figure 3: Spread of Skewness Parameter Estimates of Simulation Data (N=10000)

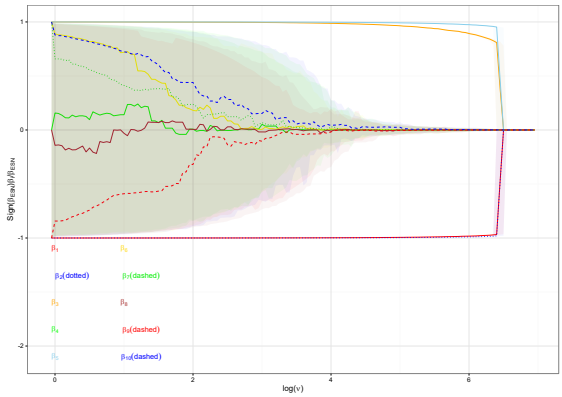


5.1.2 Unit Coefficients

Using a simulation in the same vein as other, but using unit coefficients, a similar result is observed, though the skewness and τ parameters are erratic and sometimes have the 'wrong' sign. This is due to the local instability of the maximum likelihood estimates for these parameters. The regression coefficients however are not affected by these and converge in a more expected manner. The results of these runs are presented in Figures 6a - 7c. Convergence takes place relatively quickly when n is small, with the penalty becomning more binding relatively soon in the estimations.

Figure 4: Paths of LASSO Coefficients for the Skew Family of Distributions for the Simulated Data

(a) Spread of LASSO Regression Coefficients (β) of Variables by ν (N=1000)



(b) Spread of Skewness Parameter Estimates of Simulation Data (N=1000)

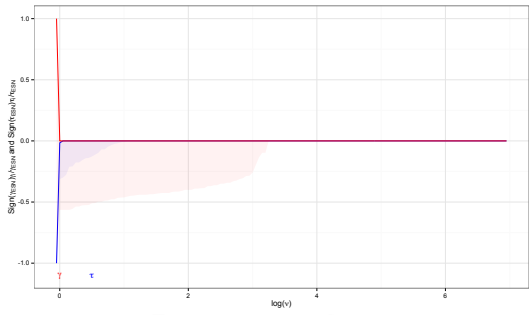
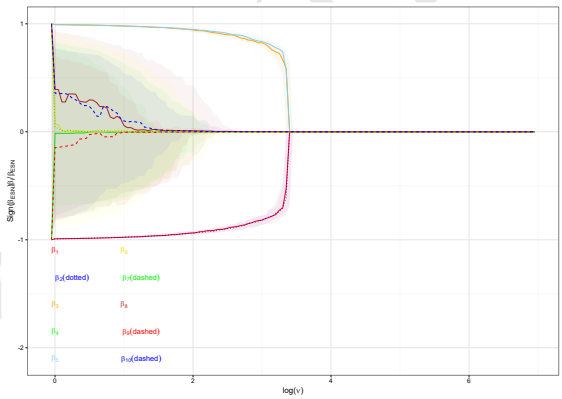


Figure 5: Paths of LASSO Coefficients for the Skew Family of Distributions for the Simulated Data

(a) Spread of LASSO Regression Coefficients (β) of Variables by ν (N=100)



(b) Spread of Skewness Parameter Estimates of Simulation Data (N=100)

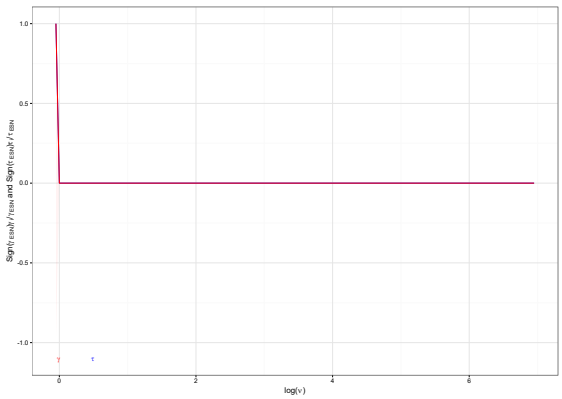


Figure 6: Regression Parameter Estimates for Simulations with Unit Coefficients

(a) Simulation Data (N=10000) (b) Simulation Data (N=1000) (c) Simulation Data (N=100)

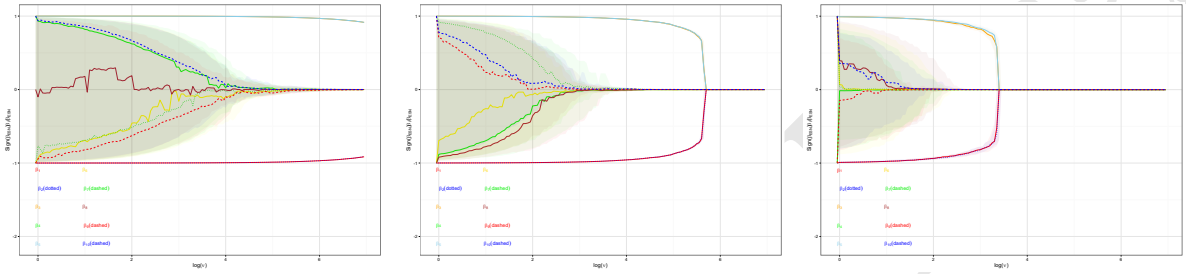
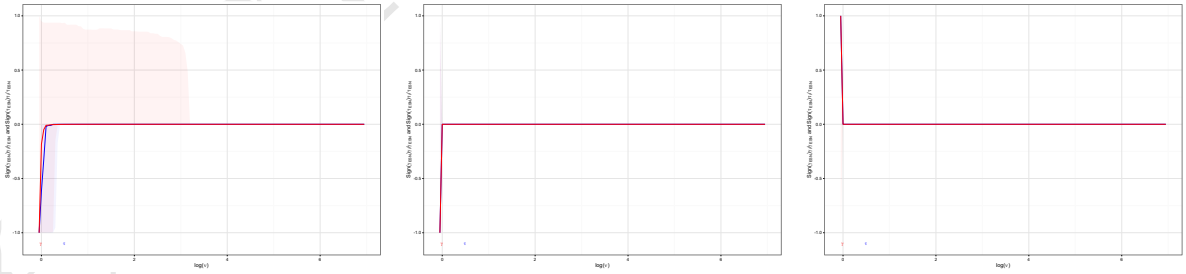


Figure 7: Skewness Parameter Estimates for Simulations with Unit Coefficients

(a) Simulation Data (N=10000) (b) Simulation Data (N=1000) (c) Simulation Data (N=100)



5.2 Diabetes Data Results

The results are presented with skew normal ($\tau = 0$) and extended skew normal ($\tau \neq 0$), Gaussian LASSO and Ridge regressions in Table 1. The Maximum Likelihood approach used a grid of Lagrange multipliers and the coefficients from each of these values are recorded. These are presented graphically in Figure 8 with the coefficients presented as a proportion of the unconstrained maximum likelihood estimates[5]. As can be seen the estimates converge to zero as the penalty increases. The original skew normal distribution is known to have issues with stability in estimation; this is less problematic for the extended skew normal distribution. This is due to the relative smoothness of the likelihood functions under specific conditions (examples are given in Azzalini and Capitanio [1999]).

The path of the regression coefficients are given in Figure 8 using a grid-based path. These are given as a proportion of the unconstrained estimates (with a sign modification to aid visualisation). These diagrams show the variable selection ability of the LASSOs.

The LASSO parameter, ν is selected using the 10-fold cross validation. Using the rule of thumb that one should maximise the cross validated parameter within a standard error (Breiman et al. [1984]) of the MSE of the minimum, the optimal value of $\ln(\nu)$ is -3.6 for the extended skew normal LASSO as is shown in 9. The relevant ν parameters are shown in Figure 8 as the vertical dashed line. These results demonstrate that there is variable selection under the extended skew normal LASSOs.

The selection implies that the variables 2, 3, 4, 7, 9 and 10 are to be included in the extended skew normal model model with the other coefficients being less than 1% of their standard MLE estimate as is the case for the Gaussian LASSO.

The parameters associated with the skewness, λ and τ, are estimated from the likelihood function. These are presented below in Figure 10. The skewness parameter under the extended formulation of the skew normal demonstrates direct convergence.

The OLS ridge regression shrinks the coefficients towards 0 however this is not as extreme as that of the LASSO in both the Gaussian and non- Gaussian scenarios. The (leave one out) cross validated LASSO Gaussian coefficients are also given in Table 1. These were estimated using `glmnet` (Friedman et al. [2010]). The penalty for the ridge regression is selected using the approach of Cule and De Iorio [2012] based on cross-validation. There is more shrinkage under the skew normal approaches to the LASSO. Thus the skew normal creates a more parsimonious regression but the skewness parameters are non-zero. There is therefore a trade-off between a more parsimonious regression and a parsimonious distribution. The skew parameters are acting to counteract the variable not included.

[5]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.

Figure 8: Path of Extended Skew Normal LASSO Regression Coefficients (β) by ν

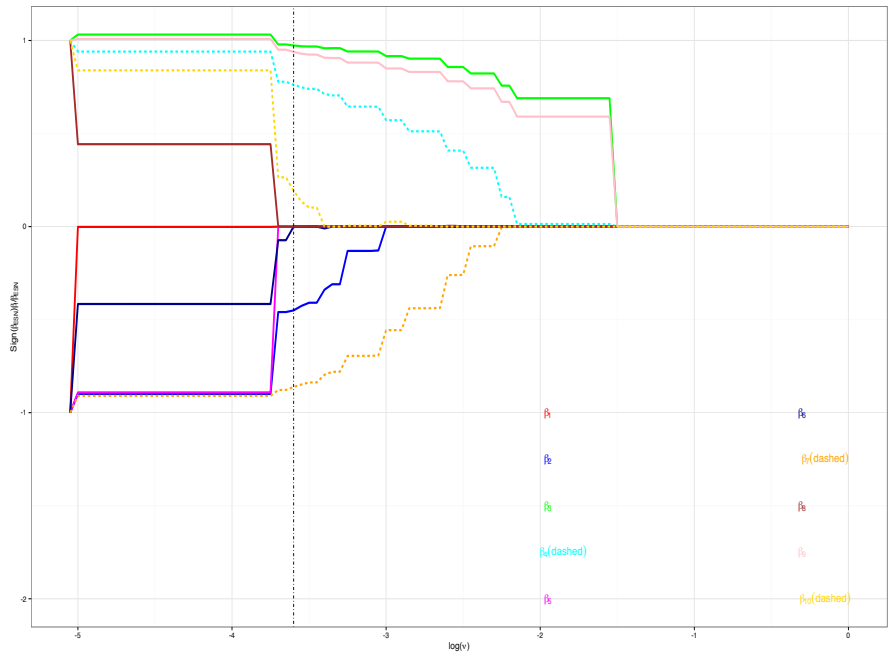


Figure 9: Cross Validation Results for the Selection of ν, the LASSO parameter for the Extended Skew Normal

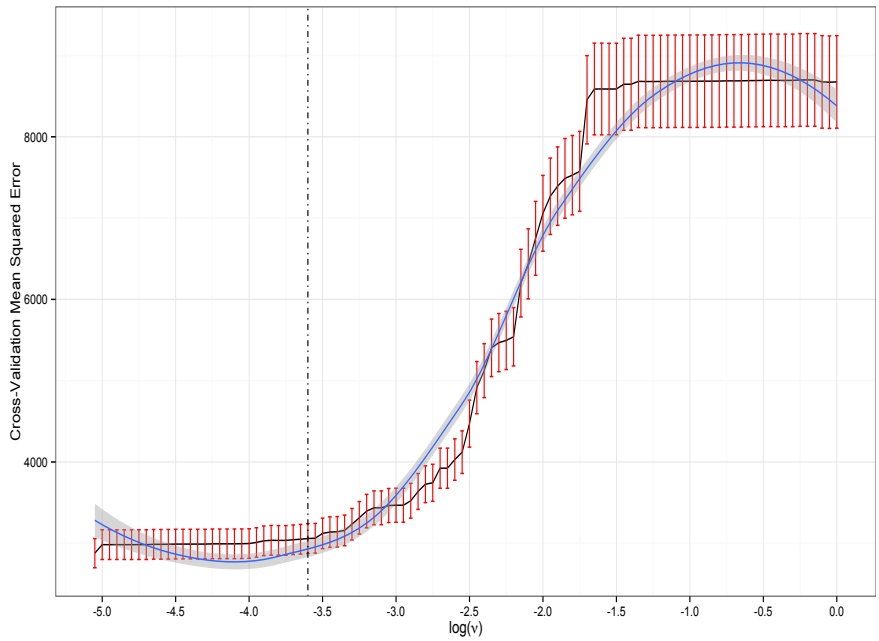


Figure 10: Path of Skewness Parameters λ & τ for the Extended Skew Normal LASSO

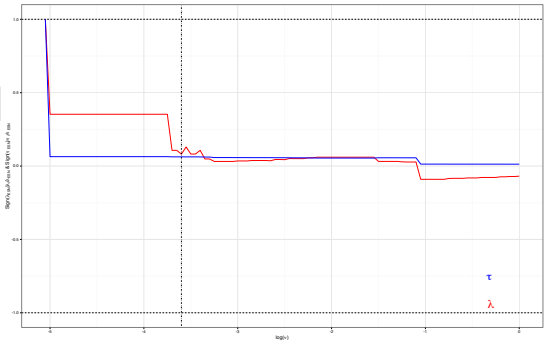


Table 1: Estimates of the Extended Skew Normal LASSO for Diabetes Data

	ESN LASSO	ESN MLE		LASSO	Ridge		OLS	
	Coef	ESN	SE	CV.LASSO	Ridge	Ridge SE	OLS	OLS SE
μ	152.719	152.138	2.553	152.133	152.133	NA	152.133	2.576
β_1	-	-6.580	59.923	-	-4.816	57.599	-10.012	59.749
β_2	-105.654	-237.086	60.687	-196.053	-228.124	59.923	-239.819	61.222
β_3	514.916	529.915	65.955	522.070	515.391	63.156	519.840	66.534
β_4	244.548	323.484	64.849	296.268	316.125	62.340	324.390	65.422
β_5	-	-64.026	415.98	-102.047	-206.171	102.045	-792.184	416.684
β_6	-	-121.526	338.50	-	13.835	99.620	476.746	339.035
β_7	-170.463	-208.798	209.892	-223.27	-150.203	91.810	-208.80	211.720
β_8	-	118.206	160.11	-	115.787	114.508	177.064	161.476
β_9	458.722	463.841	171.51	513.684	518.312	76.632	751.279	171.902
β_{10}	13.586	75.179	65.409	53.937	75.172	63.061	67.625	65.984
λ	-9.627	-3.807	0.000					
σ	55.237	53.680	1.8192					
τ	2.710	10.133	0.000					
lp	-2434.91	-2387.62						

Key:
ESN LASSO= Estimation of Extended Skew Normal LASSO
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

6 Bicyle Hire

This data was acquired from Capital Bikeshare system, Washington DC. This is based on hourly data with the aggregation being created by Fanaee-T and Gama [2013]. This data examines the determinants of bicycle hire based on season, holiday/ work day and weather. There are over 17000 observations per variable. The weather variables include wind speed, humidity, temperature (normalised to 41 degrees) and a weather situation, a general weather variable that describes the weather eg mist, clouds etc. The seasonality was adapted, rather than use Spring, Summer etc.. The seasons were termed as Quarters. In the analysis these entered as dummy variables, with Q4 (September to early December) being the base. The weather situation variable was also recoded to reflect whether it was clear, misty and cloudy (this was taken as the base), there was light snow or rain or heavy rain or snow.

	Jan	Feb	Mar	Apr	May	June	Jul	Aug	Sep	Oct	Nov	Dec
Q 1	1429	1341	949	0	0	0	0	0	0	0	0	523
Q 2	0	0	524	1437	1488	960	0	0	0	0	0	0
Q 3	0	0	0	0	0	480	1488	1475	1053	0	0	0
Q 4	0	0	0	0	0	0	0	0	384	1451	1437	960

Table 2: Classification of Months By Quarter

The analysis allows us to consider the important drivers of bicycle hire. As one might expect, there is a considerable inflation at low levels of hiring with even the early hours seeing some rental(as one finds with international trade statistics). The data is logged to smooth way a degree of the inflation and also to minimise corner solution issues with the estimation[6].

Time of Day	Median(Number of Rentals)
[0,6]	16.00
(6,12]	204.00
(12,18]	281.00
(18,24]	155.00

Table 3: Rentals by 6 hour Period

Using the same approach of constrained optimisation, with the logarithm of the number of hires as the dependent variable the paths of the coefficients of the independent variables were mapped. These are shown in Figures 12a-12c. The standard cross validation techniques were used to ascertain the optimal shrinkage parameter. Interestingly the ESN produced a more parsimonious model than the standard Gaussian model. Indeed the ESN has the identical CV constraint for minimum and minimum +1 standard

[6]The skew normal family of distributions can have problems in the presence of extreme skewness where the distribution is close to a truncated distribution.

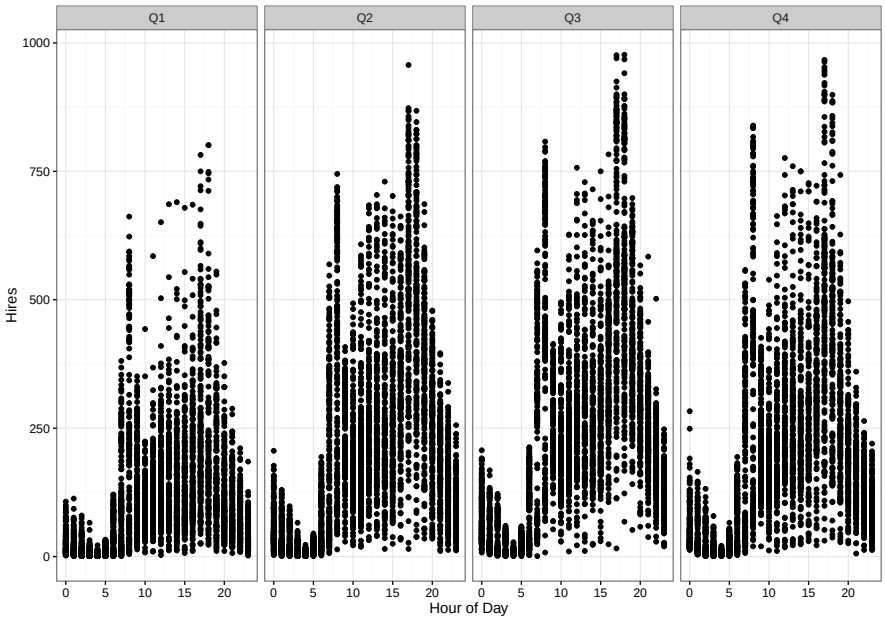


Figure 11: Bicycle Rentals By Hour of the Day

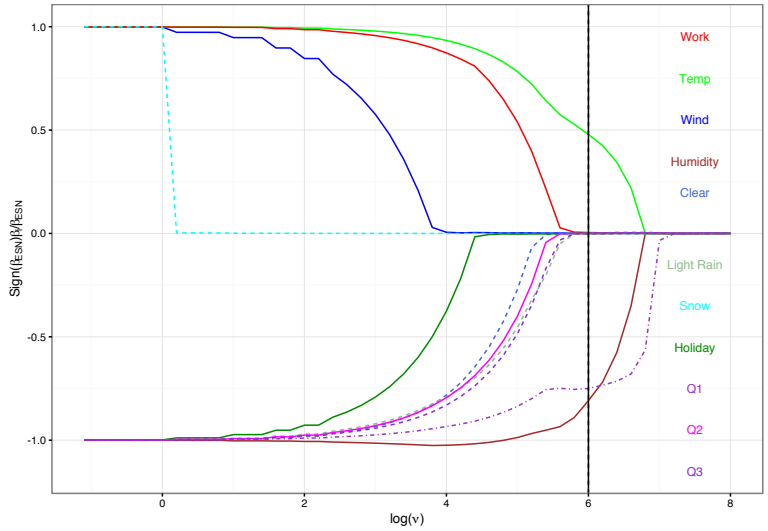
error values, whereas the Gaussian model has distinct and less constrained optima.

Both the ESN and Gaussian LASSOs converge to the same intercept, however the ESN selects on the Temperature, Humidity and the Quarter 1 dummy as important in the 10 fold CV as can be seen in Table 4. It is also worthy of note that the skewness parameter (γ) is not constrained to be zero in this case, with the other parameter (τ) also staying important in the results for a substantial part of the range of the LASSO constraint. Qualitatively the coefficients in the regressions are similar, though there are differences in the magnitudes as one would expect and the least constrained coefficients have different signs for the working day variable. Both models keep the same variables, temperature and humidity and Q1 until last.

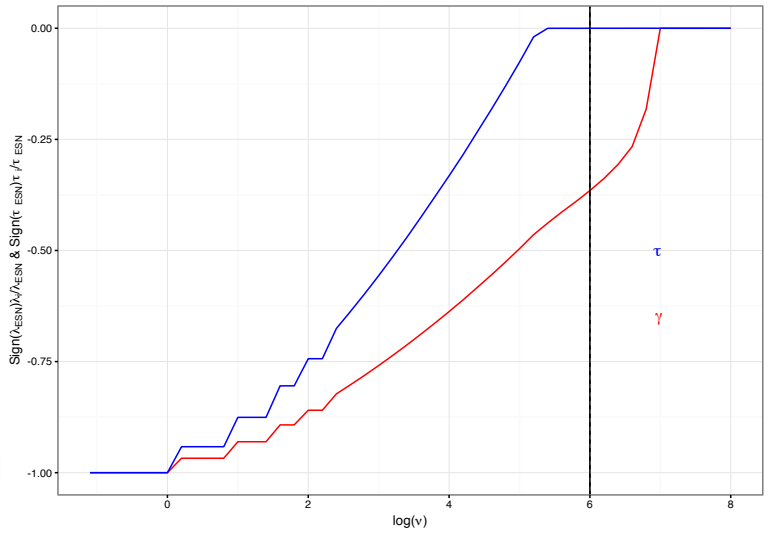
As with the previous analyses, the skewness parameters are included in the constraints and this may explain the differences in the estimations and fits of the model. Thus the parsimony of the ESN is driven in part at least by the extra parameter in the optimisation.

Figure 12: Regularised Path for Bicycle Hire Data

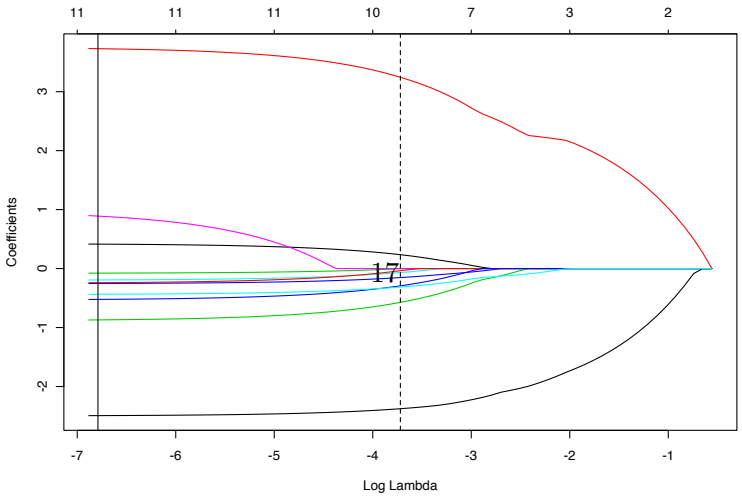
(a) Path of Index Parameters for the Extended Skew Normal LASSO



(b) Path of Skewness Parameters λ & τ for the Extended Skew Normal LASSO



(c) Path of Gaussian LASSO Parameters for the Bicycle Hire Data



	OLS CV Min	ESN CV Min	OLS CV 1SE	ESN CV 1SE
(Intercept)	4.87362	6.10894	4.77815	6.10894
windspeed	0.41454	.	0.23569	.
temp	3.72884	1.18051	3.24406	1.18051
workingday	-0.07714	.	-0.00312	.
clear	-0.25263		-0.15279	
lightrain	-0.19318		-0.06773	
snow	0.88982	.	.	.
hum	-2.49376	-1.04791	-2.37679	-1.04791
holiday	-0.23950	.	-0.03591	.
Q3	-0.87053	.	-0.57338	.
Q2	-0.52293	.	-0.29302	.
Q1	-0.43596	-0.29146	-0.31395	-0.29146
λ	.	-2.984	.	-2.984
σ	.	1.898	.	1.898
τ	.	0.0001	.	0.0001
l_p	.	-30131.74	.	-30131.74

Table 4: Results of 10- fold Cross Validation for LASSO constraint Selection

6.1 Financial Data

In a number of cases, financial data such as stock returns are seen to be non-normal. Thus the extended skew normal distribution allows the characterisation of both of the potentially useful higher moments whilst nesting the normal distribution as a special case. The example here uses the LASSO to identify the important relationships between a number of indices. The Shanghai Stock Exchange Index (SSE) and Shenzhen Index (SZSE) are two of the exchanges in China; neither are completely open to foreign investors with restrictions being placed on trading in the assets that constitute the indices Shanghai Stock Exchange [2015]. Though those restrictions might not bind in many cases, these restrictions might lead to requirement of a replicating portfolio such that the return on the index might be replicated by other more tradable indices. Using the LASSO will give the most effective replication- reducing the number of indices invested in. The indices used as the constituents of the replicating portfolio are the ASX 200, Dow Jones, CAC 40, FTSE 100, Dax 30, Hang Seng, NASDAQ, KLCI, Nikkei and TAIEX indices.

Following the previous method, an OLS and the extended skew normal regression are used as comparisons. Cross-validation was used for the choice of the ν. The approach implicitly ignores any time series issues. The cross validation is the standard sampling rather than the forecast evaluation approach with a rolling origin. This allows the demonstration of the LASSO rather than the data's use for replication.

For the Shanghai index, using OLS and extended skew normal approaches the Hang Seng is highly significant with the Dax and NASDAQ also being statistically significant. For the Shenzhen only the Hang Seng is statistically significant. Using 10 fold cross validation, the LASSO for the extended skew normal was estimated in addition to that of the Gaussian equivalent. The paths are broadly similar in trajectory with the Hang Seng again clearly being the most important index in explaining the Shanghai index. The skewness and τ parameters are somewhat volatile. This is due to the interaction that exists between them in dealing with the estimation of the likelihood function. Using the critierion that a variable is dropped when it is less than 1% of the unregularised coefficient, the extended skew normal are the CAC, DAX, Hang Seng, Nikkei and TAIEX,though the CAC and DAX are only marginal in the regression[7]. The Gaussian equivalent run though a similar 10 fold cross-validation gives the DAX, Hang Seng, KLCI and TAIEX as important variables. It is interesting that the KLCI is included in the Gaussian and not the skew normal LASSO. The KLCI is marginally removed from the asymmetric LASSO. The Gaussian model produces a slightly simpler model. The paths are given in Figure 13a and 13b. This can be compared with the OLS based LASSO from Figure 13c using `glmnet` from Friedman et al. [2010], which uses the coefficients rather than the proportion of the unconstrained coefficient. These are a simple transformation from one to the other, though the proportions approach is sometimes simpler to view when the coefficients are widely dispersed.

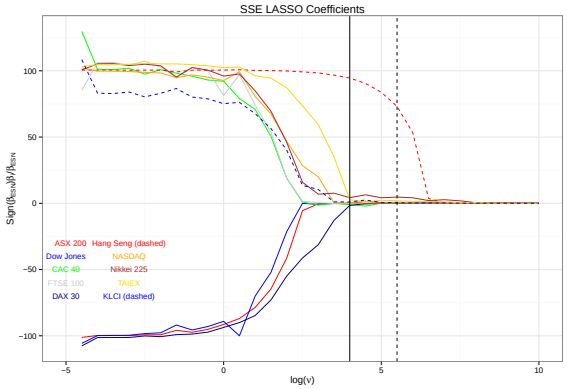
Shenzhen is a smaller market than Shanghai. The OLS and extended skew normal

[7]Increasing the cut-off to 2.5% removes all the non-Asian indices.

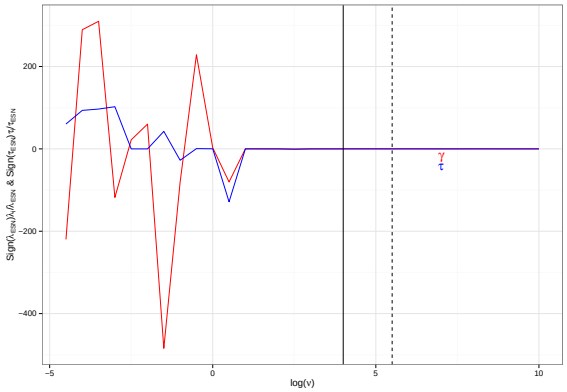
regressions are run and again the Hang Seng is significant. However the DAX and NASDAQ are also significant with the extended skew normal and OLS. As before the ESN and Gaussian LASSOs are estimated. The Gaussian LASSO selects the DAX, Hang Seng and KLCI, whereas the ESN LASSO selects the FTSE, the Hang Seng and the TAIEX. The paths are given in Figures 14a- 14c. One can see that there are parallels between the two LASSOs; the two LASSOs select the Hang Seng (as one would expect), an European index and an Asian index. Again there is some instability in the estimates of the various regression coefficients, though these are often almost equal and opposite, suggesting that these instabilities are caused by local optima.

Figure 13: Regularised Path for the Shanghai Index

(a) Path of Index Parameters for the Extended Skew Normal LASSO



(b) Path of Skewness Parameters λ & τ for the Shanghai Index



(c) Path of Gaussian LASSO Parameters for the Shanghai Index

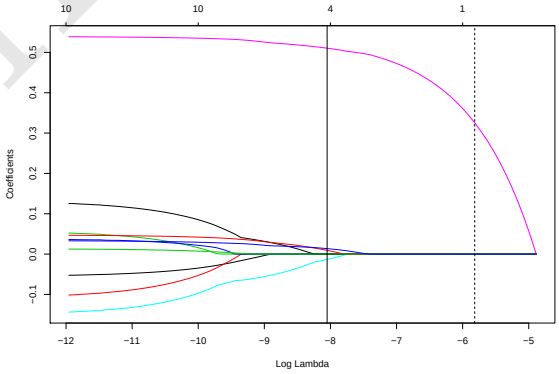
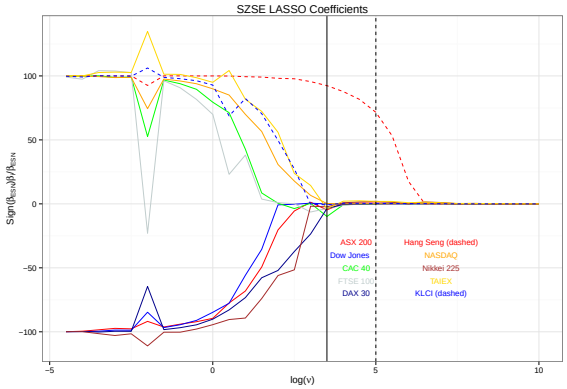
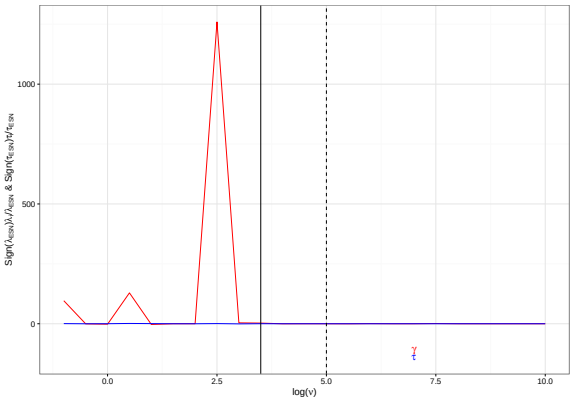


Figure 14: Regularised Path for the Shenzhen Index

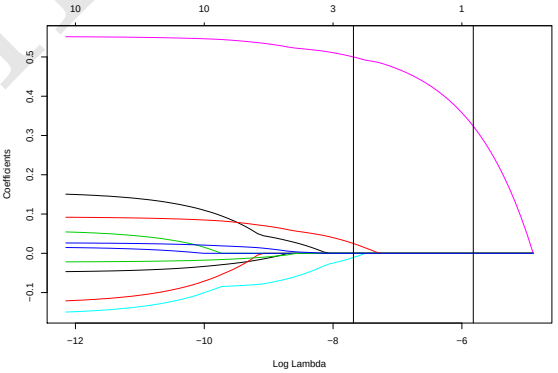
(a) Path of Index Parameters for the Extended Skew Normal LASSO



(b) Path of Skewness Parameters λ & τ for the Shenzhen Index



(c) Path of Gaussian LASSO Parameters for the Shenzhen Index



7 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 of regression coefficients and thus perform a variable selection role. There are issues with instability in certain situations, though other formulations of the various distributions might minimise these problems.

This therefore allows the analysis of data using a non- Gaussian toolbox and thus address the issue raised by Bühlmann [2013]. 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. Further the extension of the LASSO to its generalisation of the elastic net is also possible as is the Bayesian estimation using double exponential priors on the regularised coefficients.

The skew normal family of LASSOs will trade off the distribution complexity with the regression complexity relative to the Gaussian distribution. The skewness parameters act in the same manner fundamentally as the regression coefficients with the approach constraining them towards 0 as the penalty increases. Thus the Gaussian and the skewed variants will converge if the skewness parameters are driven towards 0 relatively soon in the process.

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. *The R package* `sn`*: The Skew-Normal and Skew-t distributions (version 1.3-0)*. Università di Padova, Italia, 2015. URL `http://azzalini.stat.unipd.it/SN`.

Table 5: Estimates for Multifactor Models for Shanghai Index

	SSE		SSE		SSE	SSE
	OLS	SE	ESN	SE	LASSO	ESN LASSO
μ	-0.0001	0.0003	0.0005	NA	-0.0001	-0.001
ASX 200	-0.0555	0.0464	-0.0555	0.0464	-	-
Dow Jones	-0.1111	0.0819	-0.1118	0.0819	-	-
CAC 40	0.0614	0.0729	0.0612	0.0729	-	-0.0006
FTSE 100	0.0386	0.0695	0.0364	0.0695	-	-
DAX 30	-0.1548	0.0708	-0.1532	0.0708	-0.0134	-0.0025
Hang Seng	0.5393	0.0359	0.5367	0.03588	0.5105	0.5073
NASDAQ	0.1333	0.0661	0.1341	0.06611	-	-
KLCI	0.0474	0.0655	0.0575	0.0655	0.0093	-
Nikkei 225	0.0133	0.0301	0.0127	0.0301	-	0.0005
TAIEX	0.0335	0.0403	0.0318	0.0403	0.0135	0.0009
λ			-0.0905	NA		0.0000
σ			0.0116	0.0002		0.0116
τ			0.3627	NA		-0.0001
lp			3681.0020			3646.1410

Table showing the regression coefficients of the various indices in replicating the Shanghai Index for OLS, the Extended Skew Normal and associated LASSO approaches.

Table 6: Estimates for Multifactor Models for Shenzhen Index

	SZSE		SZSE		SZSE	SZSE
	OLS	SE	ESN	SE	LASSO	ESN LASSO
μ	0.0003	0.0004	0.0000	NA	0.0003	0.0003
ASX 200	-0.0485819	0.0592407	-0.0486	0.0592	-	-
Dow Jones	-0.1297524	0.1044750	-0.1299	0.1044	-	-
CAC 40	0.0611437	0.0930687	0.0614	0.0930	-	-
FTSE 100	0.0178131	0.0886362	0.0177	0.0886	-	-0.0006
DAX 30	-0.1584424	0.0903323	-0.1586	0.0903	-0.0106	-
Hang Seng	0.5520044	0.0457921	0.5520	0.04576	0.5000	0.5094
NASDAQ	0.1575918	0.0843702	0.1578	0.0843	-	-
KLCI	0.0927831	0.0835566	0.0931	0.0835	0.0249	-
Nikkei 225	-0.0222821	0.0383614	-0.0223	0.0383	-	–
TAIEX	0.0271307	0.0513720	0.0270	0.05133	-	-0.0010
λ			0.0205	NA		0.0002
σ			0.0148	0.00029		0.0148
τ			0.0330	NA		0.0001
lp			3386.2805			3363.4005

Table showing the regression coefficients of the various indices in replicating the Shenzhen Index for OLS, the Extended Skew Normal and associated LASSO approaches.

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.

Ben Bolker and R Development Core Team. *bbmle: Tools for General Maximum Likelihood Estimation*, 2016. URL `https://CRAN.R-project.org/package=bbmle`. R package version 1.0.18.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. Chapman & Hall, New York, 1984. ISBN 0-412-04841-8. URL `http://www.crcpress.com/catalog/C4841.htm`.

P. Bühlmann. Statistical Significance in High-Dimensional Linear Models. *Bernoulli*, 19 (4):1212–1242, 2013.

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.

H. Fanaee-T and J. Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013. ISSN 2192-6352. doi: 10.1007/s13748-013-0040-3. URL `http://dx.doi.org/10.1007/s13748-013-0040-3`.

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. 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`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

Shanghai Stock Exchange. Restriction on Proportion of Shareholding. Technical report, Shanghai Stock Exchange, 2015. URL `http://english.sse.com.cn/investors/shhkconnect/rules/restriction/`.

Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. 1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

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`.