EUSIPCO 2013 1569746675

BAYESIAN MODEL SELECTION AND PARAMETER ESTIMATION IN PENALIZED REGRESSION MODEL USING SMC SAMPLERS

Thi Le Thu Nguyen[1], *François Septier*[1], *Gareth W. Peters*[2], *Yves Delignon*[1]

[1] Institut Mines-Télécom / Télécom Lille1 / LAGIS UMR CNRS 8219, France
[2] Department of Statistical Science, University College of London, UK

ABSTRACT

Penalized regression methods have received a great deal of attention in recent years, mostly through frequentist models using ℓ_1-regularization. However, all existing works assume that the design matrix, that links the explanatory variables to the observed response, is known a priori. Unfortunately, this is often not the case and thus solving this challenging problem is of considerable interest. In this paper, we look at a fully Bayesian formulation of this problem. This paper proposes the use of Sequential Monte Carlo samplers for joint model selection and parameter estimation. Furthermore, a new class of priors based on α-stable family distribution is proposed as non-convex penalty for regularization of the regression coefficients. The performance of the proposed methodology is demonstrated in two different settings.

Index Terms— Model Selection, Bayesian Inference, Regularization, SMC sampler

1. INTRODUCTION

Sparse regression analysis initially studied in the context of penalized least squares or likelihood has gained increasing popularity since the seminal paper on the LASSO [1]. Since this work, many approaches under both frequentist and Bayesian have been proposed to extend these sparsity inducing regression frameworks. In a frequentist setting the most common choice is the ℓ_1-regularization for the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ known as LASSO, i.e. a penalty term $\gamma \sum_{i=1}^{p} |\beta_i|$. Under a Bayesian modelling paradigm, in which the regression coefficients are treated as a random vector, one may recover the LASSO estimates from the maximum a posteriori (MAP) point estimator of the coefficients via a choice of prior on the coefficients given by the multivariate Laplace distribution, $p(\boldsymbol{\beta}) \propto exp(-\gamma \sum_{i=1}^{p} |\beta_i|)$. Unlike the garrotte or the ridge penalties, the Laplace prior will produce truly sparse solutions as γ increases. Finally, we note that from a Bayesian perspective the use of MAP estimates is not exploiting the full posterior information, see [2] and [3] who explore full posterior distribution using a Laplace prior via Markov chain Monte Carlo (MCMC).

A limitation in this approach is the use of identical penalization on each regression coefficient. This can lead to unacceptable bias in the resulting estimates [4]. Indeed, the classical ℓ_1-regularization can lead to an over-shrinkage of large regression coefficients even in the presence of many zeros. This has resulted in sparsity-inducing non-convex penalties that uses different penalty coefficients on each regression coefficient, i.e. $\sum_{i=1}^{p} \gamma_i |\beta_i|$ have been proposed, as have grouping regularization constraints, see adaptive and sequential estimation approaches in [5–8]. Alternative non-convex approaches include the bridge regression framework, i.e. $\gamma \sum_{i=1}^{p} |\beta_i|^q$ with $q \in (0, 1)$, which leads to the ℓ_q-regularization problem [9]. Compared to previous non-convex prior, the latter possesses the advantage of not introducing additional variables that need to be tuned.

In this work, we focus on the design of efficient algorithms to fully explore the regression coefficient posterior distribution when a non-convex penalty functions with the same penalty coefficient for each regression term are used. In [9], an MCMC algorithm is proposed for the Bayesian Bridge regression problem. Unfortunately, all existing Bayesian approaches assume that the possibly non-linear basis function(s) required to link the input variables (explonatory variables) to the observed response $\mathbf{y}$ is perfectly known. In this paper, we propose an efficient Bayesian algorithm for the joint model selection of these basis functions as well as the regression coefficients under a non-convex penalized regression model. Finally, we also introduce a new class of priors based on the α-stable family. We contrast their performance w.r.t. regularization against ℓ_q priors.

2. PROBLEM FORMULATION

2.1. Generalized Linear Models, the Exponential Family and Basis Function Regression

We consider a widely utilized class of regression models known as the Generalized Linear Model (GLM) structure [10]. These represent a widely developed class of regression models, see discussions in [11, 12].

The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a

link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted mean value. Hence, the data consists of pairs $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ say, where y_i is the response for the ith case in the dataset and $\mathbf{x}_i = (x_{i,1} \ \ldots \ x_{i,p})^T$ is the corresponding vector of explanatory variables. When specifying a GLM regression model one must consider three aspects the distribution for the response, the link function and the mean/variance relationships in terms of the covariates.

A GLM model postulates that given $\mathbf{x}_i$, y_i has some probability distribution with mean μ_i. We consider a general basis function regression structure in which we need to perform model selection to assess the most suitable class of basis functions and we will jointly perform regularization of the regression coefficients on the basis funtions to remove bases (transformed covariates) which are not explanatory of the variation in the response in a given model structure.

Consider for the k-th basis function model a function of the GLM regression mean given by

$$g(\mu_i) = \mathbf{\Phi}_k \left(\mathbf{x}_i\right)^T \boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j \Phi_k^j \left(x_{i,j}\right) \tag{1}$$

$(= \eta_i$, say) for some coefficient vector $\boldsymbol{\beta} = (\beta_1 \ \ldots \ \beta_p)^T$, a basis function regression design matrix $\mathbf{\Phi}_k$ where each element contains the basis function applied to the covariate corresponding to the i-th column $\Phi_k^j(x_{i,j})$ in which $\Phi_k^j : \mathbb{R} \mapsto \mathbb{R}$ and the first column containing 1 corresponding to the intercept term. In specifying the function of the mean $g(\cdot)$ (known as the link function) we must ensure it is selected to be strictly monotonic and differentiable. Then one can say that after application of the link function to the basis function linear model η_i, forms the *linear predictor* for y_i.

In addition, we will focus on distributional models for the response Y's selected from the the *exponential family*. This family of models contains many standard distributions allowing for continous response distributions as well as discrete response distributions such as the normal, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson, Wishart, Inverse Wishart and many others. We illustrate in Section 4 two choices from this GLM : Normal regression model with identity link and Poisson regression model with log link.

2.2. Bayesian Regularization and Priors

We consider the exponential power distribution used in the Bayesian bridge regression (see [9]) and then we introduce the symetric α-stable distribution as an alternative family of regularizing priors.

Prior Class 1: Exponential Power distribution : Bridge
The exponential power distribution (EP) with zero-mean is defined as :

$$f(x; \gamma, q) = \frac{q}{2\gamma\Gamma(1/q)} e^{(-\left|\frac{x}{\gamma}\right|^q)}. \tag{2}$$

Prior Class 2: α-Stable distribution
We propose to study the use of the symetric α-Stable distribution as a new class of prior distributions for the regression coefficients. The α-stable distribution with characteristic exponent $0 < \alpha = q < 2$, dispersion parameter $\gamma > 0$, location parameter δ and skewness parameter $\beta \in [-1; 1]$, is only defined through its characteristic function :

$$\log \phi(t) = \begin{cases} i\delta t - \gamma^q |t|^q \left[1 - i\beta \text{sign}(t) \tan\left(\frac{q\pi}{2}\right)\right] & q \neq 1 \\ i\delta t - \gamma |t| \left[1 + i\beta \text{sign}(t) \frac{2}{\pi} \log |t|\right] & q = 1 \end{cases}$$

In this paper, we are particularly interested as a regularization prior with the symmetric α-stable ($\mathcal{S}\alpha\mathcal{S}$) distribution ($\delta = 0$, $\beta = 0$).

Comparison
Here we present plots of the negative log densities (i.e. penalty function) for the α-stable distribution and the exponential power distribution, see Figure 1 for different values of q. For $q = 2$, these two distributions are equivalent to the normal distribution, producing a convex penalty (Ridge regression). For $q < 1$ the penalty function from the exponential power distribution is non-convex whereas the one from the symetric α-stable distribution is non-convex when the characteristic exponent of the distribution is $0 < q < 2$. In particular, for $q = 1$, we can see the greater Kurtosis and heavier tails provided by the stable distribution. As mentioned previously, the relatively light tails of the exponential power distribution prior is unattractive as it tends to shrink large values of the coefficients even when there is clear evidence from the likelihood that they corresponds to large values. This is an important motivation for the class of α-stable priors we introduce in this paper.

2.3. Bayesian Model Selection

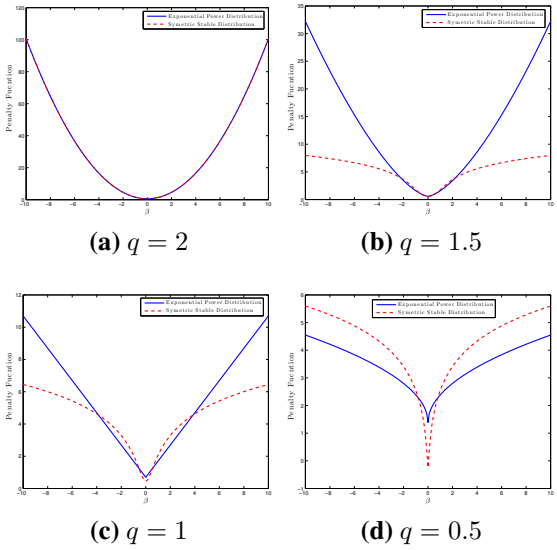
We consider several families of non-nested regression models, each specified by the choice of basis function transforming the covariates. We utilize regularization to remove non-explanatory regressors and model selection for the most suitable choice of basis. To achieve this we perform Bayesian model selection in which we aim to approximate $p(\mathcal{M}_k|\mathbf{y})$, for each of the models $k \in \{1, 2, \ldots, K\}$, which corresponds to the posterior model probability. Using Bayes' theorem,

$$p(\mathcal{M}_k|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_k) p(\mathcal{M}_k) \tag{3}$$

where $p(\mathbf{y}|\mathcal{M}_k)$ denotes the marginal likelihood under model $\mathcal{M}_k$, also known as Bayes evidence, and $p(\mathcal{M}_k)$ corresponds to the model prior. Moreover, we are also interested in estimating the parameters that define each model through the parameter posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_k)$. In the two examples considered in this study, the parameter is defined as

$$\boldsymbol{\theta} = \begin{cases} \{\boldsymbol{\beta}, \sigma_y^2, \gamma\} & \text{Normal model} \\ \{\boldsymbol{\beta}, \gamma\} & \text{Poisson model} \end{cases} \tag{4}$$

In these models, the two distributions of interest, i.e. the conditional parameter posterior, $p(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_k)$, and the associated marginal likelihood $p(\mathbf{y}|\mathcal{M}_k)$ are intractable. Therefore, we



(a) $q = 2$ **(b)** $q = 1.5$

(c) $q = 1$ **(d)** $q = 0.5$

Fig. 1: Comparison of the penalty term induced by the log prior of the regression coefficient to be either the exponential power distribution or the α-sable distribution ($\gamma_{EP} = 2\gamma_{\mathcal{S}\alpha\mathcal{S}} = 1$)

resort to an Importance-Sampling (IS) based Monte Carlo solution to jointly approximate these two quantities. This is a challenge due to the high-dimension of the parameter $\boldsymbol{\theta}$, so classical IS methods will be inefficient and produce high variance estimators. Consequently, we utilize a special class of algorithms known as "Sequential Monte Carlo samplers".

3. PROPOSED BAYESIAN SOLUTION

3.1. Introduction to SMC samplers

Here we describe briefly a special class of SMC algorithms specifically designed to work in settings in which the sequence of target distributions to be sampled from are all defined on the same fixed support, see discussions in [13, 14]. This is different to standard SMC algorithms for state space models (particle filtering) in which the sequence of distributions evolves on a product space, and as a result requires modification to the incremental importance sampling weight expressions.

In short, the SMC sampler generates weighted samples (termed *particles*) from a sequence of arbitrary distributions π_t, for $t = 1, \ldots, T$, where π_T may be of particular interest and referred as the target distribution. Procedurally, this involves mutation (or move), correction (or importance weighting) and selection (or resampling). The final weighted particles at distribution π_T are considered weighted samples from the target distribution π.

In more detail, suppose that at time $t-1$, the distribution π_{t-1} can be approximated empirically using N weighted particles. These particles are first propagated to the next distribution π_t using a mutation kernel $\mathcal{K}_t(\boldsymbol{\theta}_{t-1}; \boldsymbol{\theta}_t)$, and then assigned new weights $w_t = w_{t-1} W_t(\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_t)$, where w_{t-1} is the weight of a particle at time $t-1$ and W_t is the incremental weight given by

$$W_t(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = \frac{\pi_t(\boldsymbol{\theta}_t)\, \mathcal{L}_{t-1}(\boldsymbol{\theta}_t; \boldsymbol{\theta}_{t-1})}{\pi_{t-1}(\boldsymbol{\theta}_{t-1})\, \mathcal{K}_t(\boldsymbol{\theta}_{t-1}; \boldsymbol{\theta}_t)} \tag{5}$$

There is a range of possible things to consider when designing an SMC sampler algorithm, the appropriate sequence of distributions, the choice of mutation kernel and then the optimal choice of backward mutation kernel $\mathcal{L}_{t-1}(\cdot;\cdot)$ (for a given mutation kernel), see discussion on the optimal choices for these components in [13]. In the context of the modelling undertaken in this paper, we will utilize the SMC Sampler algorithm to also perform model selection for the basis function choices as detailed below.

3.2. Proposed SMC sampler

In this paper, we propose to use the SMC sampler on the *artificial* sequence of distributions $\{\pi_t(\boldsymbol{\theta})\}_{t=1}^T$ as follows:

$$\pi_t(\boldsymbol{\theta}) \propto p(y|\boldsymbol{\theta}, \mathcal{M}_k)^{\phi_t} p(\boldsymbol{\theta}|\mathcal{M}_k) \tag{6}$$

where conditioned on a specific model $\mathcal{M}_k$, $p(\boldsymbol{\theta}|\mathcal{M}_k)$ is the prior of the model parameters and ϕ_t is a non-decreasing temperature schedule with $\phi_1 = 0$ and $\phi_T = 1$. We thus sample initially from $\pi_1(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathcal{M}_k)$ directly and introduce the effect of the likelihood gradually in order to obtain at this end ($t = T$) an approximation of the conditional parameter posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_k)$. As shown in [13], the marginal likelihood of interest to make a decision regarding the basis function to use can be approximated with SMC samplers as :

$$Z_T = Z_1 \prod_{t=2}^{T} \frac{Z_t}{Z_{t-1}} \approx \prod_{t=1}^{T} \left(\sum_{i=1}^{N_p} w_t^i \right) \tag{7}$$

where $Z_t = \int p(y|\boldsymbol{\theta}, \mathcal{M}_k)^{\phi_t} p(\boldsymbol{\theta}|\mathcal{M}_k) d\boldsymbol{\theta}$ corresponds to the normalizing constant of the target distribution at iteration t. As a consequence the following procedure is performed :

1. For each model $\mathcal{M}_k$, $k \in 1, \ldots, K_{max}$: approximate the conditional parameter posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_k)$ as well as the marginal likelihood $p(\mathbf{y}|\mathcal{M}_k)$ using Algo. 1.
2. Approximate the model posterior $p(\mathcal{M}_k|\mathbf{y})$ the model posterior, via the approximation of $p(\mathbf{y}|\mathcal{M}_k)$ and model prior $p(\mathcal{M}_k)$ - Eq. (3).

Successive Random Walk Metropolis Hastings within Gibbs proposal kernels is used for the mutation step of the algorithm $\mathcal{K}_t(.;.)$ by randomly partitioning the parameters vector $\boldsymbol{\theta}$ into B blocks.

4. NUMERICAL SIMULATION

In this section, we study the performance of the proposed SMC sampler for joint model selection and parameter estimation in two different settings from GLM. All results have been

Algorithm 1 SMC Sampler Algorithm for Model $\mathcal{M}_k$

1: Initialize particle system from the prior
2: $\left\{\boldsymbol{\theta}_1^i\right\}_{i=1}^{N_p} \sim p(\boldsymbol{\theta}|\mathcal{M}_k)$ and set $\left\{\widetilde{w}_1^i\right\}_{i=1}^{N_p} = 1/N_p$
3: **for** $t = 1, \ldots, T$ **do**
4: Computation of the weights: for each $i = 1, \ldots, N_p$

$$w_t^i = \widetilde{w}_{t-1}^i \frac{\pi_t(\boldsymbol{\theta}_{t-1}^i)}{\pi_{t-1}(\boldsymbol{\theta}_{t-1}^i)} = \widetilde{w}_{t-1}^i \frac{p(y|\boldsymbol{\theta}_{t-1}, \mathcal{M}_k)^{\phi_t}}{p(y|\boldsymbol{\theta}_{t-1}, \mathcal{M}_k)^{\phi_{t-1}}}$$

Normalization of the weights : $\widetilde{w}_t^i = w_t^i \left[\sum_{j=1}^{N_p} w_t^j\right]^{-1}$
5: Selection: if $ESS < N_p/2$ then Resample
6: Mutation: for each $i = 1, \ldots, N_p$: Sample $\theta_t^i \sim \mathcal{K}_t(\boldsymbol{\theta}_{t-1}^i; \cdot)$ where $\mathcal{K}_t(\cdot;\cdot)$ is a $\pi_t(\cdot)$ invariant Markov kernel.
7: **end for**
8: The weighted particle system $\left\{\boldsymbol{\theta}_t^i, \widetilde{w}_t^i|\mathcal{M}_k\right\}_{i=1}^{N_p}$ approximates $\pi_t(\boldsymbol{\theta}) \propto p(y|\boldsymbol{\theta}, \mathcal{M}_k)^{\phi_t} p(\boldsymbol{\theta}|\mathcal{M}_k)$

	Model	Expression Φ_k^i
$\mathcal{M}_1$	Linear	x
$\mathcal{M}_2$	Gaussian	$\exp(-(\rho_i r_i)^2)$
$\mathcal{M}_3$	Inverse quadratic	$\frac{1}{1+(\rho_i r_i)^2}$
$\mathcal{M}_4$	Sigmoidal	$\frac{1}{1+\exp(-r_i/\rho_i)}$
$\mathcal{M}_5$	B-spline	See [16] Order=2
$\mathcal{M}_6$	Mollifier	$\begin{cases} \exp(-\frac{1}{1-(\rho_i r_i)^2}) & \text{if } \lvert\rho_i r_i\rvert < 1 \\ 0 & \text{otherwise} \end{cases}$

Table 1: Description of the different basis functions used in the numerical simulation section - $r_i = ||x - c_i||$ defines the ℓ_1-norm between the input univariate variable and the i-th center of the current basis - ρ_i is a scale factor.

obtained using the approach of Section 3 with the following settings. $N_p = 500$ particles and $T = 50$ iterations have been used to approximate the sequence of distributions. A piecewise linear tempering schedule $\{\phi_t\}$ has been selected. The sequence increased uniformly from 0 to 7/50 for the first 10 iterations, then from 7/50 to 20/50 for the next 20 and finally from 20/50 to 1 for the last 20 iterations. This choice was made to allow an initially slow evolution of the densities and then to allow more complex densities to appear at a faster rate.

The different basis functions considered in this paper are described in Table 1 with equally spaced centers c_i on some chosen bounded support of interest for the univariate input variable $x \in [-4; 12]$. The following priors have been used: $p(\mathcal{M}_k) = 1/K$ and an inverse-gamma prior for both γ and σ_y^2. Finally, in order to validate our proposed algorithm, the results will be compared with the frequentist LASSO implemented using the coordinate descent algorithm [15] and for which the tuning parameter is obtained by a ten-fold cross-validation procedure.

4.1. Normal Regression Model

In this first example, we have generated $n = 40$ observations under model $\mathcal{M}_2$ ($\sigma_y^2 = 2$) with regression coefficients set to zeros except $\beta_0 = 1$, $\beta_3 = \beta_{15} = \beta_{24} = 5$, $\beta_8 = \beta_{20} = -5$ and $\beta_5 = \beta_{17} = 3$. For the basis functions ($\mathcal{M}_2$ to $\mathcal{M}_6$), twelve equally spaced centers c_i with 2 different scale parameters have been used. As shown in Fig. 2, the SMC sampler is able to efficiently predict the unknown function even if only few observations are available. As opposed to frequentist LASSO, the proposed approach can give a confidence interval on the predicted curve which is of great interest in many applications. Table 2 clearly shows the ability of the proposed method to give an accurate estimate of the regression coefficients. We can see a significant gain with proposed SMC sampler. From Fig. 3, we can see the shrinkage effect on the marginal posterior distribution on one true zero coefficient. Finally, from Fig. 4, we study the model choice given by the proposed SMC sampler. The model used to generate the data is selected, thus validating the proposed procedure.

	EP	$\mathcal{S}\alpha\mathcal{S}$	LASSO
$q = 1$	29.13	**28.98**	133.69
$q = 0.8$	29.37	29.06	

Table 2: Median of the mean squared error between true regression coefficients and the estimated ones under the true model $\mathcal{M}3$ based on 50 replications.

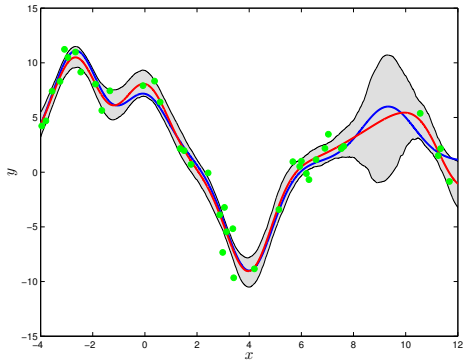
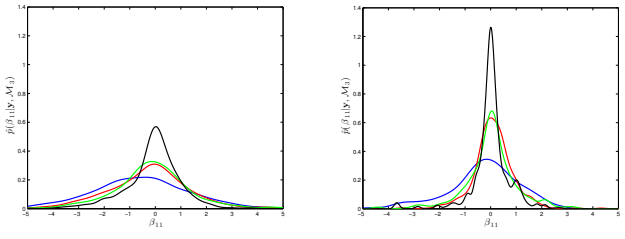
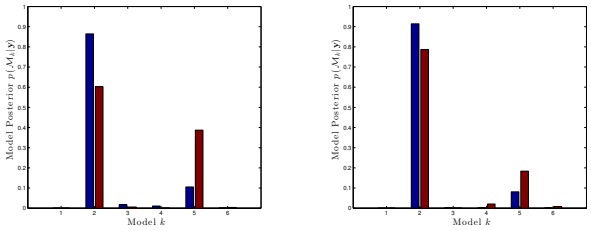


Fig. 2: Regression in Normal regression with EP prior ($q = 1$): true function in blue - observed responses in green filled circles - posterior mean from SMC under model $\mathcal{M}_3$ in red and confidence region in gray (5% to 95% percentiles)



(a) EP **(b)** $\mathcal{S}\alpha\mathcal{S}$

Fig. 3: Comparison of the shrinkage results obtained with the two different priors as q decreases (blue : $q = 1.5$, red : $q = 1$, green : $q = 0.8$, black : $q = 0.5$)



(a) $q = 1$ **(b)** $q = 0.8$

Fig. 4: Comparison of the approximation of the model posterior (blue : $\mathcal{S}\alpha\mathcal{S}$, red : EP)

4.2. Poisson Regression Model

In this second example, a Poisson regression model is considered. $n = 100$ observations have been generated with $\mathcal{M}_3$. For the basis functions, 25 equally spaced centers c_i have been used with the same scale parameter. Fig 5 shows the resulting mean predicted curve (and associated confidence interval) obtained by using the proposed SMC sampler under the true model. As in the previous case, the true curve is always within the confidence region. Table 3 presents the mean squared prediction errors obtained by the proposed approach by using the two different priors as well as the ones obtained by the frequentist LASSO. The SMC sampler with the $\mathcal{S}\alpha\mathcal{S}$ and $q = 1$ slightly outperforms the other ones.

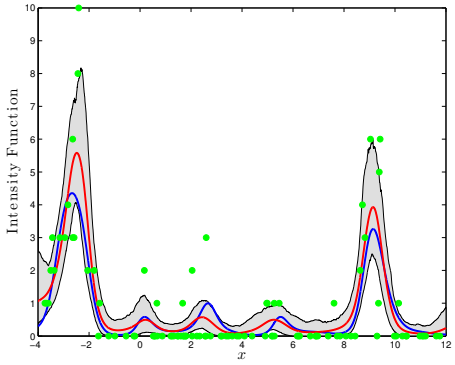


Fig. 5: Regression in Poisson regression with $\mathcal{S}\alpha\mathcal{S}$ prior ($q = 1$): true function in blue - observed count responses in green filled circles - posterior mean from SMC under model $\mathcal{M}_3$ in red and confidence region in gray (5% to 95% percentiles)

5. CONCLUSION

In this paper, we have proposed an efficient algorithm for model selection and parameter estimation in penalized regression models based on SMC samplers. Moreover, we have proposed a new class of priors based on α-stable family distribution that represents an alternative to exponential power distribution commonly used for ℓ_q-regularization. The proposed methodology has shown promising results in two examples from generalized linear models.

	EP		$\mathcal{S}\alpha\mathcal{S}$		LASSO
	$q = 1$	$q = 0.5$	$q = 1$	$q = 0.5$	
$\mathcal{M}_1$	17.4087	17.4370	**17.2961**	17.4117	17.3391
$\mathcal{M}_2$	2.7994	2.4939	**2.3589**	2.5132	3.4125
$\mathcal{M}_3$	2.6987	2.5305	**2.4089**	2.6344	2.5451
$\mathcal{M}_4$	3.0428	3.0087	**2.9484**	3.3708	3.3917
$\mathcal{M}_5$	3.3182	3.6683	**3.2428**	6.4821	4.3435
$\mathcal{M}_6$	3.0162	3.0104	**2.9131**	3.4356	3.1100

Table 3: Median of the mean squared prediction error for the proposed approach using the different priors as well as the LASSO estimate, based on 50 replications.

6. REFERENCES

[1] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. R. Stat. Soc. Series B*, vol. 58, pp. 267–288, 1996.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *J. R. Stat. Soc. Series B*, vol. 73, no. 3, pp. 273–282, 2011.

[3] T. Park and G. Casella, "The Bayesian Lasso," *J. Amer. Statist. Assoc.*, vol. 103, no. 482, pp. 681–686, June 2008.

[4] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, pp. 1348–1360, 2001.

[5] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, pp. 1418–1429, 2006.

[6] A. Lee, F. Caron, A. Doucet, and C. Holmes, "Bayesian Sparsity-Path-Analysis of Genetic Association Signal using Generalized t Priors," *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 2, 2012.

[7] E. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877–905, 2008.

[8] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. ICASSP*, 2008.

[9] N. G. Polson, J. G. Scott, and J. Windle, "The Bayesian Bridge," *arXiv.org*, vol. stat.ME, Sept. 2011.

[10] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *J. R. Stat. Soc. Series A*, pp. 370–384, 1972.

[11] P. McCullagh and J. A. Nelder, *Generalized linear models*, vol. 37, Chapman & Hall/CRC, 1989.

[12] D. G. Denison, C. C. Holmes, B. K. Mallick, and A. F. Smith, *Bayesian methods for nonlinear classification and regression*, vol. 386, Wiley New York:, 2002.

[13] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo samplers," *J. R. Stat. Soc. Series B*, vol. 68, no. 3, pp. 411–436, 2006.

[14] G. Peters, Y. Fan, and S. Sisson, "On sequential Monte Carlo, partial rejection control and approximate Bayesian computation," *Statistics and Computing*, pp. 1–14, 2009.

[15] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, no. 1, Jan. 2010.

[16] E. T. Y. Lee, "A Simplified B-Spline Computation Routine," *Computing*, , no. 29, pp. 365–371, 1982.