Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian empirical likelihood for ridge and lasso regressions

Adel Bedoui [*], Nicole A. Lazar

Department of Statistics, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Article history:
Received 12 October 2018
Received in revised form 8 January 2020
Accepted 9 January 2020
Available online 20 January 2020

Keywords:
Bayesian inference
Empirical likelihood
Ridge regression
Lasso regression
MCMC

ABSTRACT

Ridge and lasso regression models, which are also known as regularization methods, are widely used methods in machine learning and inverse problems that introduce additional information to solve ill-posed problems and/or perform feature selection. The ridge and lasso estimates for linear regression parameters can be interpreted as Bayesian posterior estimates when the regression parameters have Normal and independent Laplace (i.e., double-exponential) priors, respectively. A significant challenge in regularization problems is that these approaches assume that data are normally distributed, which makes them not robust to model misspecification. A Bayesian approach for ridge and lasso models based on empirical likelihood is proposed. This method is semiparametric because it combines a nonparametric model and a parametric model. Hence, problems with model misspecification are avoided. Under the Bayesian empirical likelihood approach, the resulting posterior distribution lacks a closed form and has a nonconvex support, which makes the implementation of traditional Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling and Metropolis–Hastings very challenging. To solve the nonconvex optimization and nonconvergence problems, the tailored Metropolis–Hastings approach is implemented. The asymptotic Bayesian credible intervals are derived.

1. Introduction

A hallmark of many modern data problems is the large number of potential predictors often with a small (or smaller) number of observations. As a result, regularization methods, which introduce additional information to solve these ill-posed problems or perform feature selection, have become increasingly popular in machine learning and statistical modeling more generally. Two well-known members of this family of approaches are ridge regression (Tikhonov and Nikolayevich, 1943) and the lasso (Tibshirani, 1996), although others such as the elastic net (Zou and Hastie, 2005) have recently been developed. Whereas ridge regression introduces an l_2 norm penalty to constrain the solution of the usual ordinary least squares (OLS) procedure, lasso uses an l_1 norm; this has the added effect of performing variable selection and parameter estimation simultaneously. The different regularization methods differ in the shape of the constraint space — diamond for lasso, circle for ridge, and yet other shapes corresponding to combinations of l_1 and l_2 penalties, for instance. Furthermore, the ridge and lasso regression approaches have an intuitive Bayesian interpretation stemming from the use of a Normal or Laplace (double exponential) prior on the regression parameters, respectively. We exploit this connection in the current work, which explores the use of Bayesian empirical likelihood in the regularization paradigm. Empirical likelihood (EL) is a nonparametric method first introduced by Owen (1988, 1990), although it can be considered as an extension of calibration estimation in survey sampling (Hartley and Rao, 1968; Deville and Sarndal, 1992). It is

[*] Correspondence to: P.O. Box 170164, Boston, MA, USA.
E-mail address: bedoui.adel1@gmail.com (A. Bedoui).

https://doi.org/10.1016/j.csda.2020.106917

an estimation method inspired by maximum likelihood but without assuming a parametric model for the data. Hence, we avert the potential problem of model misspecification. One of the advantages of the EL approach is its flexibility to incorporate constraints and prior information (see for instance Kuk and Mak, 1989; Chen and Qin, 1993; Owen, 2001). Qin and Lawless (1994) extend the original work of Owen by linking moment conditions and developing methods of combining information about parameters. In some settings, Owen (1988, 1990, 2001) showed that EL inherits properties of a parametric model. For instance, the limiting distribution of the likelihood ratio test based on EL for a univariate mean is χ^2, in parallel with the classical Wilks result (1938) for parametric likelihood ratio tests. One can obtain data-determined confidence intervals through the Wilks statistics, which does not require the estimation of variance (Owen, 2001). Due to its robustness, and the fact that it inherits many of the desirable properties of a parametric approach, EL has enjoyed a strong line of development beyond the initial rather simple settings in which it was first employed. It has, for example, been extended to linear models, correlation models, ANOVA and variance modeling (Owen, 1991, 2001); generalized linear models (Kolaczyk, 1994); Bayesian settings (Lazar, 2003); weighted empirical likelihood (Wu, 2004); exponentially tilted empirical likelihood (Schennach, 2007); covariance estimation (Chaudhuri et al., 2007), generalized linear models incorporating population level information (Chaudhuri et al., 2008); penalized high-dimensional empirical likelihood (Leng and Tang, 2010); and penalized empirical likelihood and growing dimensional general estimating equations (Leng and Tang, 2012). The exponentially tilted empirical likelihood (ETEL) is another nonparametric method that exhibits the same $O(n^{-1})$ bias and the same $O(n^{-2})$ variance as EL. In addition, the superiority of the ETEL under misspecification has been studied thoroughly by Schennach (2007) and by Chib et al. (2018). Grendar and Judge (2009) showed that the estimates of EL are consistent under the misspecification of the data model, and that EL possesses an exclusive property of Bayesian consistency. In this article, we rely on EL because of the properties aforementioned and the consistency of its estimators under the Bayesian framework (see Supplementary material).

EL under the Bayesian framework has captured the attention of many researchers since the idea was first introduced by Lazar (2003). Lazar (2003) discusses the validity of using EL as an alternative to the likelihood function by exploring the characteristics of Bayesian inference with the profile EL ratio in place of the data density. She provides simulation via Monte Carlo and further discussion to assess the validity and the appropriateness of the resulting posterior by using the method proposed by Monahan and Boos (1992). Grendar and Judge (2009) show that Bayesian empirical likelihood (BEL) and Bayesian maximum a posteriori (MAP) estimators are consistent under misspecification of the model. They also demonstrate that the point estimators obtained by empirical likelihood and Bayesian MAP are asymptotically equivalent. Rao and Wu (2010) apply Bayesian empirical likelihood to survey sampling; Chaudhuri and Ghosh (2011) to small area estimation; Yang and He (2012) to quantile regression; Mengersen et al. (2013) to approximate Bayesian computation; and Chib et al. (2018) to handle moment condition models, where they use the exponentially tilted empirical likelihood framework.

In this paper we suggest a Bayesian empirical likelihood approach to the regularization problem. As is standard for BEL, we replace the usual regression model with the profile EL to derive the ridge and lasso regressions, avoiding concerns about model misspecification because no distributional assumption is made. We use informative Normal and Laplace priors on the regression parameters to derive the BEL ridge and lasso regression versions, respectively. The Bayesian EL framework facilitates construction of credible intervals for the model parameters.

The outline of the remaining sections is as follows. In Section 2, we introduce the Bayesian linear model based on EL. BEL for ridge regression and lasso regression are provided in Sections 3 and 4, respectively. A simulation study is presented in Section 5. In Section 6, we derive the asymptotic Bayesian credible regions of the coefficients. Section 7 covers the estimation of the shrinkage parameter. Section 8 contains some additional remarks.

2. Bayesian empirical likelihood for a linear model

We begin with the derivation of the profile EL for a linear model. Assume that we observe a set of n pairs, which are denoted as $(\boldsymbol{z_1}, y_1), \ldots, (\boldsymbol{z_n}, y_n)$. If the relationship between $\boldsymbol{z_i}$ and y_i is linear, then this association can be explained by the following model:

$$y_i = \theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2} + \cdots + \theta_p z_{ip} + \epsilon_i, \tag{1}$$

where $\boldsymbol{z_i} = \left[z_{i1}, \ldots, z_{ip}\right]^T$ and y_i are the predictor and response variables, respectively; θ_0 is the unknown intercept; θ_j is the unknown slope for explanatory variable z_{ij}; and ϵ_i is the error for data pair $(\boldsymbol{z_i}, y_i)$. In standard parametric models, we would assume that the errors are independent and normally distributed with mean 0 and constant variance; however, in EL we relax the distribution assumption. We rewrite model (1) as

$$y_i = \boldsymbol{x_i}^T \boldsymbol{\theta} + \epsilon_i,$$

where $\boldsymbol{x_i} = [1, \boldsymbol{z_i}]^T$ and $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_p]^T$. In the linear model, our objective is to estimate the coefficients by minimizing

$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i}^T \boldsymbol{\theta}\right)^2, \tag{2}$$

such that $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)=0$, where $\hat{y}_i=\boldsymbol{x}_{\boldsymbol{i}}^T\hat{\boldsymbol{\theta}}$. Let $X=(1,\boldsymbol{x_1},\ldots,\boldsymbol{x_p})$ be our design matrix. Using matrix notation and assuming that X^TX is invertible, the value that minimizes (2) is

$$\hat{\boldsymbol{\theta}}^{OLS}=\left(X^TX\right)^{-1}X^T\boldsymbol{y},$$

where the OLS estimators do not depend on the distributional assumption on the errors. Thus, estimation of the regression parameters can also be approached via estimating equations, with the formulation

$$E\left\{X^T\left(\boldsymbol{y}-X\hat{\boldsymbol{\theta}}^{OLS}\right)\right\}=\mathbf{0}.$$

Now, we can define the profile empirical likelihood ratio for $\boldsymbol{\theta}$ as follows:

$$\max_{w_i}\left\{\prod_{i=1}^{n}nw_i\middle|\ w_i\geq 0,\ \sum_{i=1}^{n}w_i=1,\ \sum_{i=1}^{n}w_i\boldsymbol{x_i}(y_i-\boldsymbol{x}_{\boldsymbol{i}}^T\boldsymbol{\theta})=\mathbf{0}\right\}, \tag{3}$$

where $\boldsymbol{w}=(w_1,\ldots,w_n)^T$ is the vector of weights on the complete set of data points. Eq. (3) describes a function on the n-dimensional simplex:

$$\boldsymbol{w}=\{w_1,\ldots,w_n|\ w_i\geq 0,\ \sum_{i=1}^{n}w_i=1\}\in\Delta_{n-1}.$$

This is an example of the general approach from Qin and Lawless (1994) with specific estimating equations for multiple regression. To maximize equation (3), we apply the Lagrange multipliers by setting the partial derivative of

$$G=\sum_{i=1}^{n}\log nw_i-n\boldsymbol{\lambda}^T\sum_{i=1}^{n}w_i\boldsymbol{x_i}(y_i-\boldsymbol{x}_{\boldsymbol{i}}^T\boldsymbol{\theta})-\gamma(1-\sum_{i=1}^{n}w_i),$$

with respect to w_i equal to 0. This in turn leads to the profile-EL function, which is given by $L_{EL}(\boldsymbol{\theta})=\exp\{l_{EL}(\boldsymbol{\theta})\}$, where

$$l_{EL}(\boldsymbol{\theta})=-n\log(n)-\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i-\boldsymbol{x}_{\boldsymbol{i}}^T\boldsymbol{\theta})\right\}, \tag{4}$$

and $\boldsymbol{\lambda}=\boldsymbol{\lambda}(\boldsymbol{\theta})$ solves

$$\sum_{i=1}^{n}\frac{\boldsymbol{x_i}(y_i-\boldsymbol{x}_{\boldsymbol{i}}^T\boldsymbol{\theta})}{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i-\boldsymbol{x}_{\boldsymbol{i}}^T\boldsymbol{\theta})}=\mathbf{0}. \tag{5}$$

To estimate the coefficients in the penalized linear regression (PLR), we minimize the objective function, which is the sum of the squares of the residuals and the penalty term. PLR is closely connected to Bayesian linear regression and its estimates can be interpreted as Bayes posterior estimates under specific priors for the $\boldsymbol{\theta}$ parameters. That is, the ridge and lasso regressions have a close connection to the Bayesian linear model when the regression parameters have independent Normal and Laplace priors, respectively. The BEL scheme is as follows: let $X=(\boldsymbol{x_1},\ldots,\boldsymbol{x_n})$ be independent multivariate random variables with an unknown distribution $F_{\boldsymbol{\theta}}\in\mathcal{F}_{\boldsymbol{\theta}}$ that depends on a parameter $\boldsymbol{\theta}=\left(\theta_1,\ldots,\theta_p\right)\in\Theta\subseteq\mathbb{R}^p$. $\mathcal{F}_{\boldsymbol{\theta}}$ is a family of distributions described by $\boldsymbol{\theta}$. We assume that both the predictor and response variable are standardized so that the intercept is zero. By placing a prior distribution $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$, the posterior empirical likelihood density is

$$\pi(\boldsymbol{\theta}|\ X,\boldsymbol{y})=\frac{L_{EL}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta}L_{EL}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}\ \propto\ L_{EL}(\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{6}$$

We combine $L_{EL}(\boldsymbol{\theta})$ with a specified prior $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ via Bayes theorem to obtain the empirical likelihood posterior

$$\pi(\boldsymbol{\theta}|X,\boldsymbol{y})\propto\exp\left[\log\{\pi(\boldsymbol{\theta})\}-\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}(y_i-\boldsymbol{x}_{\boldsymbol{i}}^T\boldsymbol{\theta})\right\}\right]. \tag{7}$$

The posterior empirical likelihood of $\boldsymbol{\theta}$ does not have a closed form, which makes the implementation of the Gibbs sampler impossible. One can consider implementing the Metropolis–Hastings (MH) algorithm, as it is suitable for generating samples from a distribution that lacks an analytic form. However, the implementation of MH is challenging, and it fails to achieve convergence due to the nature of the posterior density support, which complicates the process of finding an efficient proposal density for the MH algorithm. The surface of the posterior EL is not smooth and contains many local optima. Often, in our experience, the chain becomes trapped in a region and never reaches the global optimum.

To observe this, we consider 100 independent and identically distributed bivariate observations, which are denoted $\boldsymbol{x_i}=(x_{i1},x_{i2})$ for $i=1,\ldots,100$; we assume that $y_i=\theta_1x_{i1}+\theta_2x_{i2}+e_i$, where $\theta_1=2$, $\theta_2=5$, and e_i is the error term, which follows a standard normal distribution. Fig. 1 depicts the perspective plot of $\log\left(\pi(\boldsymbol{\theta}|X,\boldsymbol{y})\right)$ for various values of θ_1 and θ_2. The support is nonconvex where its surface is rigid. That is, if we start from values that are far from the global optimum, the chain becomes trapped near a local optimum. Therefore, we are required to tune the Metropolis–Hastings

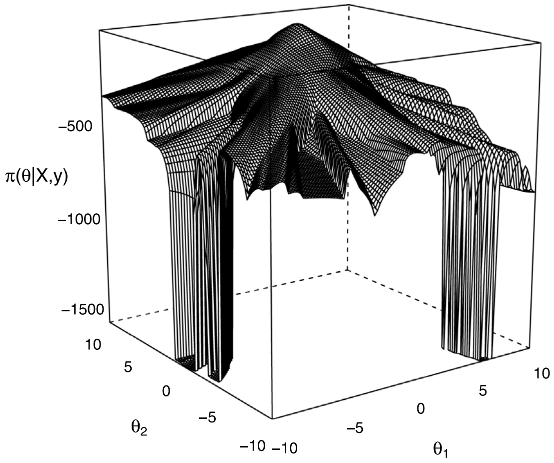

Fig. 1. Perspective plot of $\log(\pi(\boldsymbol{\theta}|, X, \boldsymbol{y}))$ for various values of θ_1 and θ_2.

to find a satisfactory proposal density with an appropriate variance that allows us to reach all states frequently and provides a high acceptance rate. To solve the nonconvex optimization problem, one can use either the Hamiltonian Monte Carlo algorithm (HMC) (Neal, 2011) or the tailored Metropolis–Hastings algorithm introduced by Chib and Greenberg (1995). In HMC, distances between successively generated points are large. Thus, fewer iterations are required to obtain a representative sample. HMC uses the gradient of the density and the Hamiltonian system to sample successive states for the Metropolis–Hastings algorithm with a high jump and a large acceptance probability. This reduces the correlation between successive sampled states, thereby enabling faster convergence. However, it might require an extensive hyper-parameter tuning, especially in a high dimension problem. We instead use the tailored Metropolis–Hastings algorithm, which does not require the hyper-parameter tuning. This approach uses the t location-scale distribution as the proposal density with location parameter equal to the mode of the log EL function and dispersion matrix the inverse of the negative Hessian matrix of the log EL function at the mode.

Finally, we must find $\boldsymbol{\lambda}$, which is the vector of the Lagrange multipliers. In the Bayesian scheme and for each iteration of the MCMC, we obtain the value of $\boldsymbol{\lambda}$, which is the root of Eq. (5). As discussed by Qin and Lawless (1994), the existence and uniqueness of $\boldsymbol{\lambda}$ are guaranteed if the following conditions are satisfied:

1. The vector $\mathbf{0} \in R^{p+1}$ is within the convex hull of $\left\{\boldsymbol{x_i}(y_i - \boldsymbol{x_i^T\theta}),\ i = 1, \dots, n\right\}$.
2. The matrix $\sum_{i=1}^{n} \frac{\boldsymbol{A_i A_i^T}}{[1+\boldsymbol{\lambda}^T \boldsymbol{A_i}]^2}$ is positive definite where $\boldsymbol{A_i} = \boldsymbol{x_i}(y_i - \boldsymbol{x_i^T\theta})$.

$\boldsymbol{\lambda}$ is obtained by widely available numerical approaches. We use the modified Newton–Raphson approach that was introduced by Owen (2001) (see Supplementary Material for details).

3. Bayesian empirical likelihood for ridge regression

Ridge regression (Tikhonov and Nikolayevich, 1943), which is also known as the method of linear regularization, penalizes the size of the regression coefficients by imposing an l_2 penalty. That is, it minimizes the penalized residual sum of squares,

$$\min_{\boldsymbol{\theta}} \left(\frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \alpha\|\boldsymbol{\theta}\|_2^2\right), \tag{8}$$

where

$\alpha \geq 0$,
$\boldsymbol{\theta}$ is a $p \times 1$ vector,
$\boldsymbol{y}$ is a $n \times 1$ vector,
X is a $n \times p$ matrix.

α is a complexity parameter that controls the amount of shrinkage. The larger the value of α, the greater the amount of shrinkage (Hastie et al., 2009). We assume that

$$\sum_{i=1}^{n} x_{ij} = 0,\ \sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij}^2 = 1 \text{ for } j = 1, \dots, p.$$

The term $\alpha\|\boldsymbol{\theta}\|_2^2$ is referred to as the ridge penalty. The solution to the ridge regression problem is given by

$$\hat{\boldsymbol{\theta}}^{\text{ridge}} = \left(X^TX + \alpha I\right)^{-1} X^T\boldsymbol{y}. \tag{9}$$

From the form of the penalty term in (8), one can see that the ridge regression parameters have independent and identical Normal priors. The shrinkage parameter, α, is introduced into the model in the form of a hyperparameter. Encouraged by this connection, one can consider a semiparametric Bayesian model where the $\boldsymbol{\theta}$ follows a Normal prior distribution. Chib et al. (2018) put a prior that follows a student-t distribution. We place a Normal prior of the form

$$\begin{aligned} \pi(\boldsymbol{\theta}|\psi) &= \prod_{j=1}^{p}\sqrt{\frac{\psi}{2\pi}}e^{-\dfrac{\psi}{2}\theta_j^2}, \\ \psi &\sim IG(a,\ b/\alpha), \end{aligned} \tag{10}$$

where IG denotes the inverse gamma distribution with shape parameter a and scale parameter b/α. Note that the shrinkage parameter α plays the role of the prior precision. For example, a small (large) value of α leads to a wider (more concentrated) prior. By replacing the likelihood function with the profile EL ratio in the Bayesian setting, we obtain the following hierarchical representation of the full model:

$$\begin{aligned} L_{EL}(\boldsymbol{\theta}) &\sim \exp\left[-\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i - \boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right\}\right], \\ \boldsymbol{\theta}|\psi &\sim N(\mathbf{0},\ \psi I_{p\times p}), \\ \psi &\sim IG(a,\ b/\alpha), \\ a,\ ,b,\ \alpha &> 0. \end{aligned} \tag{11}$$

The full conditional distribution of $\boldsymbol{\theta}$ and ψ is given by

$$\pi\left(\boldsymbol{\theta},\psi|X,\boldsymbol{y},\alpha\right) \propto \exp\left[-\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right\}\right]\left(\frac{1}{\psi}\right)^{p/2+a+1}\exp\left\{-\frac{1}{\psi}\left(\frac{b}{\alpha}+\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}\right)\right\}.$$

The full conditional distribution for ψ is an inverse-gamma distribution with shape parameter $p/2+a$ and scale parameter $\dfrac{b}{\alpha}+\dfrac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}$. The full conditional distribution for $\boldsymbol{\theta}$ does not have a closed form:

$$\pi\left(\boldsymbol{\theta}|\psi,\alpha,X,\boldsymbol{y}\right) \propto \exp\left[-\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i}^T\boldsymbol{\theta}\right)\right\}-\frac{1}{2\psi}\boldsymbol{\theta}^T\boldsymbol{\theta}\right]. \tag{12}$$

We use a building block of the tailored Metropolis–Hastings and the Gibbs sampler to sample $\boldsymbol{\theta}$ and ψ, respectively. For the tailored Metropolis–Hastings algorithm, we follow similar approach as in Chib et al. (2018). That is, we let $Q(\boldsymbol{\theta}|X,\boldsymbol{y},\ldots)$ be the proposal density in MH such that it follows a student-t distribution whose location parameter is the mode of the log EL function for the linear model and whose dispersion matrix is the inverse of the negative Hessian matrix of the log EL function evaluated at the mode. Then, starting from some initial value $\boldsymbol{\theta}^{(0)}$, we get a sample of draws from the posterior empirical likelihood for $\boldsymbol{\theta}$ by repeating the following steps for $t = 1,\ldots,T$:

1. Sample $\boldsymbol{\theta}^p$ from $Q(\boldsymbol{\theta}|X,\boldsymbol{y})$ and solve the Lagrange multiplier, $\boldsymbol{\lambda}(\boldsymbol{\theta}^p)$, in Eq. (5).
2. Calculate the acceptance probability:

$$\alpha\left(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\theta}^p|\psi,\alpha,X,\boldsymbol{y}\right) = \min\left\{1,\ \frac{\pi(\boldsymbol{\theta}^p|\psi,\alpha,X,\boldsymbol{y})}{\pi(\boldsymbol{\theta}^{(t-1)}|\psi,\alpha,X,\boldsymbol{y})}\frac{Q(\boldsymbol{\theta}^{(t-1)}|X,\boldsymbol{y})}{Q(\boldsymbol{\theta}^p|X,\boldsymbol{y})}\right\}.$$

3. Set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^p$ with probability $\alpha\left(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\theta}^p|\psi,\alpha,X,\boldsymbol{y}\right)$. Otherwise, set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{(t-1)}$. Go to step 1.

Fig. 2 compares posterior mean estimates for the BEL for ridge with posterior mean estimates for Bayesian ridge regression for the diabetes data that were provided by Efron et al. (2004). The data are scaled, corresponding to 442 diabetes patients, and describe the relationships between 10 baseline variables and a quantitative measure of disease progression one year after baseline. The variables include age, sex, body mass index, average blood pressure, and six blood measurements. To sample $\boldsymbol{\theta}$, we use the MCMC technique aforementioned with a burn-in of 2000 and 10,000 iterations. The figure depicts the computational path of the BEL and Bayesian ridge estimates over a grid of values between 0 and 2000. The methods provide similar results and their paths are smooth. As we increase the shrinkage value, the estimates shrink to approximately zero.

4. Bayesian empirical likelihood for lasso regression

The least absolute shrinkage and selection operator, introduced by Tibshirani (1996), is a regression method that involves penalizing the absolute size of the regression coefficient. It performs both variable selection and regularization.

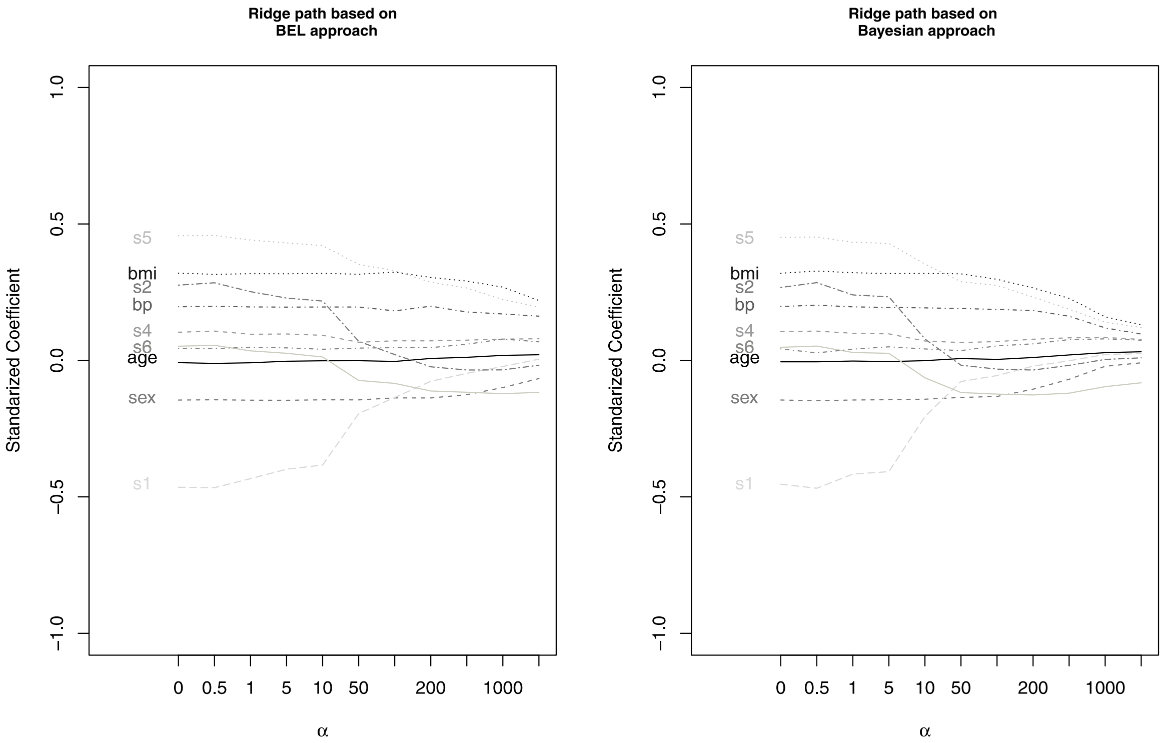

Fig. 2. Bayesian empirical likelihood for ridge (left) and Bayesian ridge (right) trace plots for estimates of the diabetes data regression parameters.

The lasso solves the following regularized optimization problem:

$$\min_{\boldsymbol{\theta}} \left(\frac{1}{2} \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \alpha \|\boldsymbol{\theta}\|_1 \right), \tag{13}$$

where

$\alpha \geq 0$,

$\boldsymbol{\theta}$ is a $p \times 1$ vector,

$\boldsymbol{y}$ is a $n \times 1$ vector,

X is a $n \times p$ matrix,

by using the l_1 penalty. α is a complexity parameter that controls the amount of shrinkage. We assume that

$$\sum_{i=1}^{n} x_{ij} = 0, \ \sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij}^2 = 1 \text{ for } j = 1, \ldots, p.$$

The l_1 penalty has the effect of forcing some of the coefficient estimates to be equal to zero when the tuning parameter α is sufficiently large (James et al., 2013). That is, this penalty term leads to feature/model selection. Fan and Li (2001) show that the lasso is the only model that produces a sparse solution among l_q penalized estimators ($q \geq 1$). The lasso penalty contains an absolute value; thus, the objective function in Eq. (13) is not differentiable. Therefore, in general, the lasso solution lacks a closed form. This requires the implementation of an optimization algorithm to find the minimizing solution. Like ridge regression, the lasso is closely connected to the Bayesian linear model. Tibshirani (1996) suggests that the lasso estimates can be interpreted as posterior mode estimates. That is, using a hierarchical model, one can place an independent identical double-exponential prior, which is also known as a Laplace distribution, on the parameters of the model. Several authors suggest using the Laplace distribution as a prior (Figueiredo, 2003; Bae and Mallick, 2004; Yuan and Lin, 2005). Motivated by this, Park and Casella (2008) consider a fully Bayesian analysis using a conditional double-exponential prior, in which they assumed that errors are independent and identically distributed and follow the Laplace distribution. We consider a conditional prior specification of the form

$$\pi(\boldsymbol{\theta}|\sigma^2, \ \alpha) \ = \prod_{j=1}^{p} \frac{\alpha}{2\sqrt{\sigma^2}} \exp\left(-\alpha|\theta_j|/\sqrt{\sigma^2}\right). \tag{14}$$

The Laplace distribution is sharply peaked at its mean, where a high scale value yields a probability density that is near zero. Another notable feature is that the Laplace distribution assigns a higher density around its mean compared to the Normal density. One can introduce the prior in (14) as a scale mixture of Normals with an exponential mixing density (Andrews and Mallows, 1974), however, it does not allow for a separate mixing variable for each θ_j. By replacing the likelihood function by the profile empirical likelihood ratio for a linear model, our hierarchical representation of the full model becomes

$$
\begin{aligned}
L_{EL}(\boldsymbol{\theta}) &\sim \exp\left[-\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i^T\theta}\right)\right\}\right],\\
\boldsymbol{\theta}|\sigma &\sim \prod_{j=1}^{p}\frac{\alpha}{2\sigma}\exp\left(-\alpha|\theta_j|/\sigma\right),\\
\sigma|\alpha &\sim \pi(\sigma)d\sigma,\\
\sigma &> 0.
\end{aligned}
\tag{15}
$$

We choose $\pi(\sigma) = \text{IG}(a, b)$. One can also impose a noninformative prior $\pi(\sigma) = 1/\sigma$ on σ. Conditioning on σ guarantees the unimodality of the full posterior distribution (Park and Casella, 2008). That is, a large precision ($\frac{1}{\sigma}$) enables sampling from a probability density that is near zero (see Supplementary material). That is, a large penalty forces the estimates to shrink toward zero. The full empirical posterior distribution is:

$$
\begin{aligned}
\pi(\boldsymbol{\theta},\sigma|X,\boldsymbol{y},\alpha) \propto \exp&\left[-\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i^T\theta}\right)\right\}\right],\\
&\left(\frac{\alpha}{\sigma}\right)^p\exp\left(-\frac{\alpha}{\sigma}\sum_{j=1}^{p}|\theta_j|\right)(\sigma)^{-a-1}\exp\left(-\frac{b}{\sigma}\right).
\end{aligned}
\tag{16}
$$

Eq. (16) gives rise to the following sampling scheme:

1. Sample $\boldsymbol{\theta}$ from

$$
\pi(\boldsymbol{\theta}|\sigma) \propto \exp\left[-\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x_i}\left(y_i-\boldsymbol{x_i^T\theta}\right)\right\}-\frac{\alpha}{\sigma}\sum_{j=1}^{p}|\theta_j|\right).
$$

   This is a nonstandard distribution. We use the tailored Metropolis–Hastings described in Section 3.
2. Sample σ from the inverse-gamma distribution with shape parameter $p+a$ and scale parameter $b+\alpha\sum_{j=1}^{p}|\theta_j|$. a and b are fixed.

Fig. 3 compares posterior mean estimates for the BEL for lasso with posterior mean estimates for Bayesian lasso for the diabetes data that are used in Section 3. To sample $\boldsymbol{\theta}$, we use an MCMC scheme with burn-in of 2000 and 10,000 iterations. The figure shows the paths of the estimates as their respective shrinkage parameters change over a grid of values between 0 and 2000. Both methods provide almost identical results and their paths are smooth. In addition, the results can be used as a feature selection method. One easily concludes from the BEL lasso paths that several predictor variables, namely, sex, age, s1, s2, s4, and s6, have faster decrease rates and are less influential on the disease progress compared to other predictors. That is, the quantitative measure of disease progression one year after baseline can be strongly predicted by the body mass index, average blood pressure, s3, and s5. Using BEL lasso (which uses a Laplace prior), the weakest variables are identified by the more rapid decline of their parameter estimates to zero compared to the BEL ridge (which uses a Normal prior). In the lasso we use the l_1 norm constraint, this makes it easier for the coefficient to be exactly 0 and hence the lasso prior can be used for variable selection. However, this is computationally expensive. On the other hand, the ridge prior is easy to implement and faster to compute but the coefficients will never equal to zero. Therefore, it cannot be implemented as a method for variable selection.

5. Simulations

To investigate the performance of the methods derived in Sections 3 and 4, we simulate datasets from the following model:

$$
y_i = 3x_{i1} + x_{i2} + 0.2x_{i3} - 0.5x_{i4} + \epsilon_i,\ \text{for } i = 1,\ldots,n_j \text{ and } j = 1, 2, 3,
\tag{17}
$$

where $x_{i1} \sim \text{N}\left(2,\ \frac{1}{2}\right)$, $x_{i2} \sim \text{unif}\,(-0.1,\ 0.1)$, $x_{i3} \sim \text{G}\,(2,\ 198)$, $x_{i4} \sim \text{N}\,(5,\ 1)$ and ϵ is a normal distribution with mean 0.1 and standard deviation 2. N, unif, and G are the normal, uniform, and gamma distributions, respectively. The errors

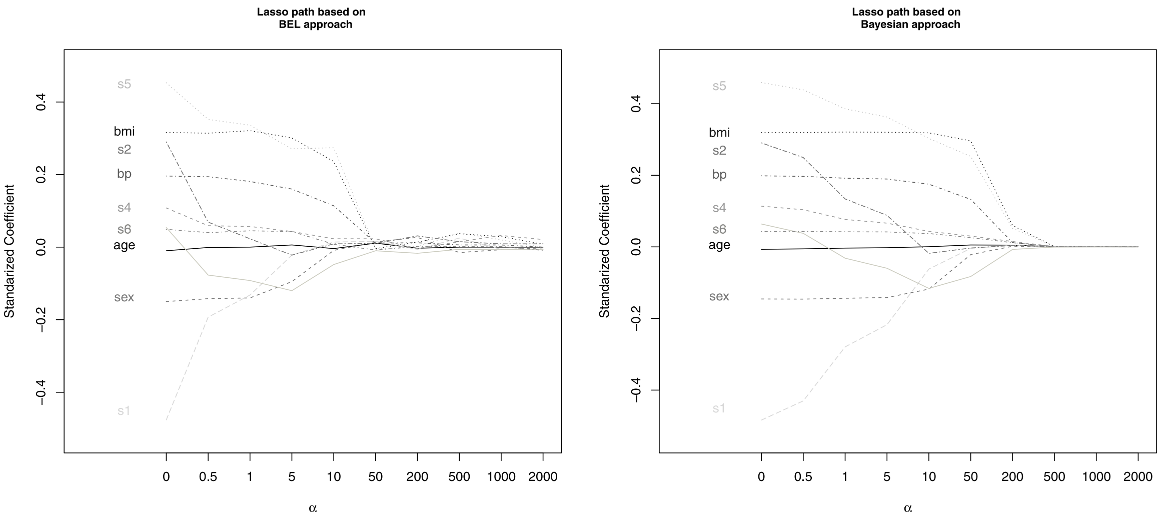

Fig. 3. Bayesian empirical likelihood for lasso (left) and Bayesian lasso (right) trace plots for estimates of the diabetes data regression parameters.

violate the linearity assumption that they should sum up to zero. To evaluate the estimation of the parameters in Eq. (17), we investigate their properties. In particular, the bias

$$\text{Bias}(\hat{\boldsymbol{\theta}}) = E\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right),$$

which is the average difference between the estimator and the truth $\boldsymbol{\theta} = (3, 1, 0.2, -0.5)^T$. Moreover, we investigate the mean square error (MSE) to evaluate the estimator precision,

$$\text{MSE}(\hat{\boldsymbol{\theta}}) = E\left(\left\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\}^2\right).$$

Applying the identity that $var(x) = E(x^2) - E(x)^2$, where $x = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$, the equation for MSE becomes

$$\text{MSE}(\hat{\boldsymbol{\theta}}) = \text{Bias}(\hat{\boldsymbol{\theta}})^2 + \text{var}(\hat{\boldsymbol{\theta}}).$$

We compare the estimator for $\boldsymbol{\theta}$ based on our approach (BEL) to the Bayesian approach (Bayesian) and the frequentist approach (Freq.). We use three different sample sizes: $n_1 = 20$, $n_2 = 40$, and $n_3 = 100$. That is, for each sample size and for each approach:

1. Calculate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_4)^T$.
2. Save $(\hat{\theta}_i - \theta_i)$ and $\left(\hat{\theta}_i - \theta_i\right)^2$ for $i = 1, \ldots, 4$.
3. Repeat step 1–2 $S(= 500)$ times.
4. Compute the means of $(\hat{\theta}_i - \theta_i)$ and $\left(\hat{\theta}_i - \theta_i\right)^2$ for $i = 1, \ldots, 4$, over the S replicates.

For the shrinkage coefficient, α, we use a range of [0, 3]. Precisely, we use the following values: $0,\ 0.5,\ 1,\ 2,\ 3$. For the frequentist approach, we use the Least Angle Regression (LARS) algorithm implemented by Hastie and Efron (2013), which is a stylized version of the stage wise procedure that uses a simple mathematical formula to accelerate the computations. LARS is described in detail in Section 3 in Efron et al. (2004). For the Bayesian ridge and lasso approaches, we use the normal prior and the Laplace prior, respectively, with likelihood function distributed according to a Gaussian distribution.

Tables 1 and 2 show the mean square errors and biases of the estimators based on the Bayesian empirical likelihood, the pure Bayesian approach, and the Frequentist method for the three different sample sizes and fitted using the six different shrinkage values. It is evident that the BEL and the Bayesian methods outperform the Frequentist approach. Moreover, investigating the MSEs of the estimators, the BEL method provides estimators with better precisions. Recall that one of the characteristics of the penalized regression is *bias–variance trade-off*, which means that it is possible for a biased estimator to be more precise than an unbiased estimator. That is, the penalized model introduces a little bias in to the estimate for $\boldsymbol{\theta}$ that leads to a substantial decrease in variance. In most cases, the Bayesian estimator has a small bias (large MSE) whereas the BEL estimator has a large bias (small MSE). In addition, the bias of the BEL estimator, in most cases, is close to the bias of the Bayesian estimator.

Table 1
Mean square error measures for BEL, Bayesian, and Frequentist methods using three different sample sizes and evaluated at different shrinkage values.

	Coefficient:		$\theta_1 = 3$			$\theta_2 = 1$			$\theta_3 = 0.2$			$\theta_4 = -0.5$		
Sample size	Method:	Shrinkage	BEL	Bayesian	Freq.	BEL	Bayesian	Freq.	BEL	Bayesian	Freq.	BEL	Bayesian	Freq.
$n_1 = 20$	*Lasso*:													
		$\alpha_1 = 0$	0.215	0.216	0.216	0.242	0.242	0.252	0.254	0.232	0.242	0.268	0.273	0.273
		$\alpha_2 = 0.5$	0.215	0.216	0.527	0.228	0.241	0.429	0.224	0.232	0.254	0.264	0.273	0.282
		$\alpha_3 = 1$	0.264	0.216	1.307	0.221	0.239	0.729	0.156	0.232	0.242	0.199	0.271	0.219
		$\alpha_4 = 1.5$	0.518	0.216	2.549	0.230	0.236	0.935	0.118	0.231	0.240	0.170	0.269	0.248
		$\alpha_5 = 2$	1.268	0.215	4.268	0.356	0.231	0.992	0.091	0.230	0.240	0.175	0.266	0.250
		$\alpha_6 = 3$	4.290	0.215	7.988	0.643	0.218	1.000	0.070	0.228	0.240	0.224	0.258	0.250
	Ridge:													
		$\alpha_1 = 0$	0.883	0.242	0.216	0.204	0.232	0.232	0.175	0.216	0.242	0.200	0.273	0.273
		$\alpha_2 = 0.5$	0.998	0.224	0.352	0.214	0.217	0.220	0.176	0.207	0.183	0.198	0.256	0.216
		$\alpha_3 = 1$	1.129	0.211	0.661	0.223	0.207	0.207	0.176	0.217	0.157	0.207	0.243	0.193
		$\alpha_4 = 1.5$	1.259	0.199	1.018	0.235	0.200	0.228	0.172	0.239	0.140	0.208	0.232	0.183
		$\alpha_5 = 2$	1.402	0.190	1.380	0.244	0.195	0.253	0.175	0.271	0.127	0.210	0.222	0.178
		$\alpha_6 = 3$	1.642	0.176	2.058	0.270	0.192	0.305	0.171	0.361	0.109	0.213	0.208	0.175
$n_2 = 40$	*Lasso*:													
		$\alpha_1 = 0$	0.118	0.123	0.123	0.085	0.085	0.085	0.094	0.095	0.095	0.113	0.118	0.118
		$\alpha_2 = 0.5$	0.114	0.123	0.359	0.084	0.085	0.311	0.085	0.095	0.043	0.114	0.118	0.177
		$\alpha_3 = 1$	0.120	0.123	1.107	0.085	0.085	0.787	0.066	0.095	0.039	0.098	0.118	0.244
		$\alpha_4 = 1.5$	0.171	0.122	2.327	0.107	0.085	0.972	0.039	0.095	0.040	0.089	0.118	0.250
		$\alpha_5 = 2$	0.346	0.122	4.067	0.167	0.085	1.000	0.038	0.094	0.040	0.091	0.117	0.250
		$\alpha_6 = 3$	2.083	0.122	8.176	0.401	0.084	1.000	0.036	0.093	0.040	0.140	0.116	0.250
	Ridge:													
		$\alpha_1 = 0$	0.426	0.123	0.095	0.092	0.085	0.085	0.067	0.095	0.123	0.087	0.118	0.118
		$\alpha_2 = 0.5$	0.487	0.116	0.070	0.098	0.081	0.070	0.066	0.091	0.216	0.082	0.113	0.082
		$\alpha_3 = 1$	0.551	0.115	0.058	0.101	0.077	0.087	0.067	0.088	0.483	0.087	0.107	0.072
		$\alpha_4 = 1.5$	0.624	0.116	0.052	0.109	0.075	0.117	0.065	0.085	0.816	0.089	0.103	0.072
		$\alpha_5 = 2$	0.704	0.119	0.048	0.116	0.072	0.150	0.065	0.083	1.167	0.092	0.099	0.075
		$\alpha_6 = 3$	0.838	0.136	0.043	0.132	0.069	0.218	0.065	0.078	1.845	0.092	0.092	0.085
$n_3 = 100$	*Lasso*:													
		$\alpha_1 = 0$	0.038	0.049	0.049	0.051	0.051	0.051	0.049	0.038	0.038	0.050	0.051	0.051
		$\alpha_2 = 0.5$	0.037	0.049	0.037	0.051	0.061	0.327	0.047	0.038	0.296	0.050	0.051	0.196
		$\alpha_3 = 1$	0.039	0.049	0.040	0.050	0.061	0.853	0.042	0.038	1.055	0.050	0.051	0.248
		$\alpha_4 = 1.5$	0.049	0.049	0.040	0.065	0.068	0.989	0.037	0.038	2.293	0.053	0.051	0.250
		$\alpha_5 = 2$	0.070	0.049	0.040	0.084	0.121	1.000	0.032	0.038	4.044	0.057	0.051	0.250
		$\alpha_6 = 3$	0.167	0.049	0.040	0.151	0.152	1.000	0.027	0.037	8.453	0.072	0.051	0.250
	Ridge:													
		$\alpha_1 = 0$	0.145	0.038	0.038	0.051	0.051	0.051	0.043	0.049	0.049	0.044	0.050	0.051
		$\alpha_2 = 0.5$	0.165	0.037	0.158	0.055	0.061	0.063	0.044	0.048	0.040	0.043	0.050	0.047
		$\alpha_3 = 1$	0.187	0.037	0.434	0.068	0.071	0.093	0.043	0.048	0.035	0.044	0.050	0.051
		$\alpha_4 = 1.5$	0.212	0.037	0.771	0.070	0.073	0.130	0.042	0.048	0.032	0.045	0.049	0.058
		$\alpha_5 = 2$	0.237	0.038	1.126	0.072	0.073	0.167	0.043	0.047	0.030	0.044	0.049	0.065
		$\alpha_6 = 3$	0.288	0.042	1.812	0.079	0.073	0.240	0.043	0.046	0.027	0.046	0.048	0.081

6. Bayesian credible intervals

Lazar (2003) proves that under standard regularity conditions and as $n \to \infty$, the posterior distribution $\boldsymbol{\theta}(F)$, which is the functional of interest, converges to the Normal distribution (the proof is presented in Supplementary Material). Based on this result, the asymptotic distribution of $\boldsymbol{\theta}$ is easily obtained under the Bayesian EL ridge model and Bayesian EL lasso. In the ridge case, the posterior EL distribution of $\boldsymbol{\theta}$ converges to the Normal distribution with mean $\boldsymbol{m}_{n1}$ and covariance J_{n1}, where

$$\begin{aligned} J_{n1} &= J(\hat{\boldsymbol{\theta}}_n) + \psi I_{p\times p}, \\ \boldsymbol{m}_{n1} &= J_{n1}^{-1} J(\hat{\boldsymbol{\theta}}_n)\hat{\boldsymbol{\theta}}_n. \end{aligned}$$

$\hat{\boldsymbol{\theta}}_n$ is the profile maximum likelihood estimate of $\boldsymbol{\theta}$ and

$$J(\hat{\boldsymbol{\theta}}_n) = \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \sum_{i=1}^{n} \log\left\{1 + \lambda^T \boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T \boldsymbol{\theta})\right\} \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} .$$

Table 2
Bias measures for BEL, Bayesian, and Frequentist methods using three different sample sizes and evaluated at different shrinkage values.

	Coefficient:		$\theta_1 = 3$			$\theta_2 = 1$			$\theta_3 = 0.2$			$\theta_4 = -0.5$		
Sample size	Method:	Shrinkage	BEL	Bayesian	Freq.	BEL	Bayesian	Freq.	BEL	Bayesian	Freq.	BEL	Bayesian	Freq.
$n_1 = 20$	*Lasso*:													
		$\alpha_1 = 0$	0.016	0.011	0.011	0.024	0.013	0.013	0.037	0.0397	0.039	−0.006	−0.020	−0.020
		$\alpha_2 = 0.5$	−0.030	0.010	−0.541	−0.014	0.012	−0.507	0.025	0.039	−0.148	0.014	−0.020	0.309
		$\alpha_3 = 1$	−0.195	0.007	−1.021	−0.135	0.010	−0.809	−0.016	0.038	−0.197	0.090	−0.018	0.452
		$\alpha_4 = 1.5$	−0.470	0.003	−1.508	−0.290	0.005	−0.959	−0.051	0.036	−0.197	0.187	−0.014	0.498
		$\alpha_5 = 2$	−0.888	−0.003	−2.003	−0.466	−0.001	−0.995	−0.078	0.034	−0.199	0.264	−0.010	0.500
		$\alpha_6 = 3$	−1.891	−0.021	−2.810	−0.760	−0.017	−1.000	−0.146	0.027	−0.200	0.397	0.003	0.500
	Ridge:													
		$\alpha_1 = 0$	−0.893	0.011	0.011	−0.266	0.013	0.013	−0.040	0.039	0.039	0.170	−0.020	−0.020
		$\alpha_2 = 0.5$	−0.958	−0.081	−0.400	−0.289	−0.017	−0.123	−0.039	0.032	0.007	0.186	−0.002	0.062
		$\alpha_3 = 1$	−1.021	−0.167	−0.697	−0.308	−0.045	−0.219	−0.054	0.026	−0.016	0.198	0.014	0.120
		$\alpha_4 = 1.5$	−1.082	−0.246	−0.927	−0.333	−0.070	−0.293	−0.047	0.020	−0.033	0.219	0.031	0.163
		$\alpha_5 = 2$	−1.146	−0.320	−1.112	−0.349	−0.093	−0.353	−0.059	0.015	−0.047	0.224	0.046	0.197
		$\alpha_6 = 3$	−1.246	−0.456	−1.393	−0.387	−0.137	−0.445	−0.074	0.004	−0.069	0.247	0.073	0.248
$n_2 = 40$	*Lasso*:													
		$\alpha_1 = 0$	0.046	0.043	0.043	0.051	0.056	0.056	0.045	0.037	0.037	−0.018	−0.025	−0.025
		$\alpha_2 = 0.5$	0.020	0.043	−0.486	0.029	0.056	−0.475	0.034	0.037	−0.175	−0.003	−0.025	0.382
		$\alpha_3 = 1$	−0.070	0.042	−0.987	−0.047	0.056	−0.863	0.002	0.037	−0.196	0.052	−0.024	0.492
		$\alpha_4 = 1.5$	−0.206	0.041	−1.480	−0.156	0.055	−0.985	−0.036	0.036	−0.200	0.121	−0.023	0.500
		$\alpha_5 = 2$	−0.414	0.040	−1.983	−0.277	0.053	−1.000	−0.070	0.036	−0.200	0.183	−0.022	0.500
		$\alpha_6 = 3$	−1.200	0.036	−2.853	−0.564	0.050	−1.000	−0.132	0.033	−0.200	0.315	−0.019	0.500
	Ridge:													
		$\alpha_1 = 0$	−0.607	0.043	0.043	−0.157	0.057	0.056	−0.026	0.037	0.037	0.088	−0.024	−0.025
		$\alpha_2 = 0.5$	−0.656	0.001	−0.338	−0.178	0.042	−0.077	−0.022	0.034	0.005	0.098	−0.018	0.040
		$\alpha_3 = 1$	−0.701	−0.040	−0.628	−0.191	0.027	−0.178	−0.034	0.030	−0.020	0.104	−0.011	0.089
		$\alpha_4 = 1.5$	−0.751	−0.079	−0.859	−0.209	0.014	−0.258	−0.030	0.027	−0.039	0.113	−0.004	0.129
		$\alpha_5 = 2$	−0.800	−0.117	−1.048	−0.223	0.001	−0.322	−0.041	0.024	−0.054	0.122	0.003	0.161
		$\alpha_6 = 3$	−0.879	−0.191	−1.337	−0.254	−0.025	−0.422	−0.052	0.018	−0.078	0.135	0.015	0.210
$n_3 = 100$	*Lasso*:													
		$\alpha_1 = 0$	0.012	0.014	0.014	−0.014	−0.015	−0.015	0.007	0.010	0.010	0.027	0.026	0.026
		$\alpha_2 = 0.5$	0.004	0.014	−0.499	−0.026	−0.015	−0.522	0.002	0.010	−0.185	0.035	0.026	0.424
		$\alpha_3 = 1$	−0.031	0.013	−0.999	−0.060	−0.015	−0.911	−0.011	0.010	−0.200	0.061	0.026	0.498
		$\alpha_4 = 1.5$	−0.086	0.013	−1.495	−0.113	−0.015	−0.993	−0.028	0.010	−0.200	0.102	0.026	0.500
		$\alpha_5 = 2$	−0.155	0.013	−1.996	−0.177	−0.015	−1.000	−0.044	0.010	−0.200	0.139	0.026	0.500
		$\alpha_6 = 3$	−0.323	0.013	−2.904	−0.314	−0.016	−1.000	−0.075	0.010	−0.200	0.210	0.027	0.500
	Ridge:													
		$\alpha_1 = 0$	−0.341	0.014	0.014	−0.135	−0.015	−0.015	−0.008	0.010	0.010	0.077	0.026	0.026
		$\alpha_2 = 0.5$	−0.369	−0.002	−0.350	−0.146	−0.020	−0.136	−0.008	0.009	−0.009	0.081	0.028	0.086
		$\alpha_3 = 1$	−0.397	−0.018	−0.634	−0.156	−0.026	−0.229	−0.013	0.009	−0.026	0.088	0.031	0.131
		$\alpha_4 = 1.5$	−0.426	−0.033	−0.861	−0.162	−0.030	−0.303	−0.012	0.008	−0.040	0.092	0.033	0.168
		$\alpha_5 = 2$	−0.455	−0.048	−1.048	−0.172	−0.035	−0.364	−0.016	0.007	−0.052	0.096	0.036	0.197
		$\alpha_6 = 3$	−0.507	−0.078	−1.337	−0.189	−0.045	−0.459	−0.022	0.006	−0.071	0.103	0.041	0.243

In the lasso case, the posterior EL distribution of $\boldsymbol{\theta}$ converges to the Normal distribution with mean m_{n2} and covariance J_{n2}, where

$$\begin{aligned} J_{n2} &= J(\hat{\boldsymbol{\theta}}_n) + \sigma^2 D_\tau, \\ \boldsymbol{m}_{n2} &= J_{n2}^{-1} J(\hat{\boldsymbol{\theta}}_n)\hat{\boldsymbol{\theta}}_n. \end{aligned}$$

The Normal distribution of the posterior EL of $\boldsymbol{\theta}$ leads to the following lemma:

Lemma 1. *Under standard regularity conditions, $-2\log(\pi(\boldsymbol{\theta}|X, \boldsymbol{y}))$ converges in distribution to χ^2_p as $n \to \infty$.*

Proof of Lemma 1. It is well known (Cochran, 1934) that if $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$ is a vector of order p and Σ is positive definite, then

$$(\boldsymbol{x} - \boldsymbol{\mu})^T\, \Sigma^{-1}\, (\boldsymbol{x} - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{} \chi^2_p.$$

We set $\triangledown \log \pi(\boldsymbol{\theta}) = \mathbf{0}$ and $\triangledown \log \pi(X, \boldsymbol{y}|\boldsymbol{\theta}) = \mathbf{0}$. Under standard regularity conditions and as $n \to \infty$, Lazar (2003) shows that $\boldsymbol{\theta} \sim N(\boldsymbol{m}_n, J_n)$, where $\boldsymbol{m_n}$ and J_n are as defined above. These regularity conditions are introduced to ensure that terms of higher order than quadratic may be ignored and that the sum of the terms from the likelihood will dominate the term from the prior (Bernardo and Smith, 1994). Empirical likelihood inherits the low-order cumulants properties (Mykland, 1994). Therefore, $-2\log(\pi(\boldsymbol{\theta}|X,\boldsymbol{y})) \propto (\boldsymbol{\theta}-\boldsymbol{m}_n)^T J_n (\boldsymbol{\theta}-\boldsymbol{m}_n)$. Now, it suffices to show that $J_n = A^{-1} + J(\hat{\boldsymbol{\theta}}_n)$ is a positive-definite matrix. A^{-1} is positive definite because, by assumption, A is positive definite. A in the ridge case and in the lasso case is $\frac{\sigma^2}{\alpha} I_{p\times p}$ and $\sigma^2 D_\tau$, respectively. We compute the second derivative of the negative logarithm of $\pi(X,\boldsymbol{y}|\boldsymbol{\theta})$

$$\frac{\partial}{\partial\boldsymbol{\theta}^T}\left[\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x}_i(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta})\right\}\right] = -\sum_{i=1}^{n}\frac{\boldsymbol{\lambda}^T\boldsymbol{x}_i\boldsymbol{x}_i^T}{1+\boldsymbol{\lambda}^T\boldsymbol{x}_i(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta})},$$

$$\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\left[-\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x}_i(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta})\right\}\right] = \sum_{i=1}^{n}\frac{(\boldsymbol{x}_i\boldsymbol{x}_i^T)^T\boldsymbol{\lambda}\boldsymbol{\lambda}^T\boldsymbol{x}_i\boldsymbol{x}_i^T}{\left\{1+\boldsymbol{\lambda}^T\boldsymbol{x}_i(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta})\right\}^2} > 0,$$

because the denominator is positive and the numerator has a quadratic form. Therefore, J_n is positive definite because the sum of two positive-definite matrices is positive definite. Thus, $-2\log(\pi(\boldsymbol{\theta}|X,\boldsymbol{y})) \xrightarrow[n\to\infty]{D} \chi_p^2$. □

For $0<\alpha<1$, the property that is presented above provides an asymptotic justification for tests that reject the value of $\boldsymbol{\theta}$ at level α, when $-2\log(\boldsymbol{\theta}|X,\boldsymbol{y}) > \chi_p^{2,1-\alpha}$. The unrejected values of $\boldsymbol{\theta}$ form 100(1-α)% Bayesian empirical credible regions.

7. Estimation of the shrinkage parameter

In the ridge and lasso models, the parameter α controls the trade-off between goodness of fit and model complexity. The selection of α is crucial because a specific value corresponds to a fitted model and it controls the shape of the prior distribution. Moreover, a small value of α produces a better fit in the sense of the residual sum of square. To select the optimal value for the penalty term, one can use empirical approaches such as the Akaike information criterion (AIC) (Akaike, 1974), Bayes information criterion (BIC) (Schwarz, 1978), cross-validation (CV) (Geisser, 1993), and generalized cross-validation (GCV) (Craven and Wahba, 1978). The most frequently used method is K-fold cross-validation. In this section, we consider the lasso model and we treat the penalty term, α, as a random term (unknown variable) by placing a distribution on it (hyperprior).

7.1. Prior for the lasso parameter

An alternative way to selecting α is to give it a diffuse prior. The resulting conditional distribution for α does not involve the data at all, which makes it too sensitive to the choices of its parameters (hyper-hyperparameters). We consider three priors: the Gamma distribution, the Uniform distribution, and the Beta distribution.

Similar to Park and Casella (2008), we place a Gamma prior with shape r and rate d on α^2. In this case, factoring equation (16) by this prior leads to the following posterior conditional distribution:

$$\pi(\alpha|\tau_1^2,\ldots,\tau_p^2) \propto (\alpha^2)^{p+r-1}\exp\left\{-\alpha^2\left(d+\frac{1}{2}\sum_{j=1}^{p}\tau_j^2\right)\right\},$$

which is a Gamma distribution with shape $p+r$ and rate $d+\frac{1}{2}\sum_{i=1}^{p}\tau_j^2$. An alternative approach is to consider a class of Uniform priors on α^2 of the form:

$$\pi(\alpha^2) = \frac{1}{\eta_2-\eta_1};\ \alpha^2>0,\ 0\le\eta_1<\eta_2.$$

When this prior is used in Eq. (16), the full conditional distribution of α^2 is a truncated Gamma distribution

$$\pi(\alpha^2|\tau_1^2,\ldots,\tau_p^2) \propto G(p,\ \frac{1}{2}\sum_{j=1}^{p}\tau_j^2)I_{(\eta_1\le\alpha^2\le\eta_2)}.$$

Another selection is to use a more flexible distribution. Los Campos et al. (2009) placed a Beta distribution with parameters ν_1 and ν_2 on $\tilde{\alpha} = \frac{\alpha}{u}$, where $u>0$ is an upper bound on α. That is,

$$\pi(\alpha) = \text{Beta}\left(\tilde{\alpha}(\alpha)|\nu_1,\nu_2\right)\left|\frac{\partial\tilde{\alpha}(\alpha)}{\partial}\right| \propto \text{Beta}\left(\frac{\alpha}{u}|\nu_1,\nu_2\right).$$

If we know that the shrinkage parameter is between 0 and 1, we can use a Beta distribution without the constraint. When the constrained Beta distribution is used in (16), the full conditional distribution of α (not of α^2) is

$$\pi(\alpha|\tau_1^2,\ldots,\tau_p^2) \propto (\alpha)^{2p+\nu_1-1}(u-\alpha)^{\nu_2-1}\exp\left(-\frac{\alpha^2}{2}\sum_{j=1}^{p}\tau_j^2\right).$$

Table 3
Posterior mean estimates of the shrinkage parameter for the Bayesian lasso based on EL.

Prior	Gamma	Uniform	Beta
α^{lasso}	0.2265	0.2177	0.2324

Table 4
Posterior mean estimates for the BEL lasso coefficients; using different priors.

Prior	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6
Gamma	−0.012	−0.083	0.287	0.079	−0.032	−0.005	−0.104	0.026	0.370	0.054
Uniform	−0.012	−0.085	0.288	0.080	−0.030	−0.007	−0.108	0.025	0.370	0.054
Beta	−0.011	−0.082	0.287	0.079	−0.027	−0.007	−0.106	0.024	0.367	0.054

Table 5
The 95% highest posterior density intervals and the 95% equal-tailed credible regions of the posterior distribution of the coefficients in the lasso model using the diabetes data, under Gamma, Beta, Uniform distributions as prior on the penalty term.

Variables	Gamma prior 95% HPD (95% credible regions)	Beta prior 95% HPD (95% credible regions)	Uniform prior 95% HPD (95% credible regions)
AGE	[−0.092, 0.066] ([−0.089, 0.066])	[−0.092, 0.067] ([−0.089, 0.067])	[−0.097, 0.069] ([−0.102, 0.062])
SEX	[−0.183, 0.013] ([−0.180, 0.010])	[−0.178, 0.015] ([−0.176, 0.012])	[−0.177, 0.009] ([−0.172, 0.009])
BMI	[0.167, 0.406] ([0.173, 0.407])	[0.170, 0.410] ([0.175, 0.409])	[0.169, 0.406] ([0.170, 0.400])
BP	[−0.024, 0.196] ([−0.017, 0.197])	[−0.02, 0.194] ([−0.023, 0.192])	[−0.026, 0.194] ([−0.027, 0.188])
S1	[−0.166, 0.088] ([−0.159, 0.091])	[−0.161, 0.090] ([−0.162, 0.085])	[−0.160, 0.086] ([−0.158, 0.083])
S2	[−0.123, 0.106] ([−0.122, 0.105])	[−0.118, 0.104] ([−0.118, 0.101])	[−0.129, 0.097] ([−0.134, 0.089])
S3	[−0.223, 0.015] ([−0.222, 0.008])	[−0.231, 0.012] ([−0.225, 0.011])	[−0.231, 0.016] ([−0.228, 0.014])
S4	[−0.087, 0.160] ([−0.081, 0.163])	[−0.088, 0.149] ([−0.086, 0.147])	[−0.085, 0.151] ([−0.085, 0.148])
S5	[0.235, 0.504] ([0.233, 0.496])	[0.233, 0.493] ([0.237, 0.492])	[0.235, 0.504] ([0.240, 0.502])
S6	[−0.029, 0.151] ([−0.029, 0.147])	[−0.029, 0.149] ([−0.027, 0.146])	[−0.034, 0.151] ([−0.029, 0.152])

To select an optimal α^2, we divide the data into two datasets: training and validation. We apply K-fold cross-validation on the training data and retrieve a value, which is denoted as $\alpha^2_{training}$, that results in a small prediction error. After that, we place a prior on the shrinkage parameter and choose its hyper-hyperparameters such that the posterior mean estimate is close to $\alpha^2_{training}$.

7.2. Example

We use the diabetes data from Efron et al. (2004), which were presented in Section 3. We split the data into 50% training and 50% testing. Then, we apply 5-fold cross-validation over a grid of α values on the training set. The lasso and ridge parameters that minimize the average cross-validation mean square are 0.25 and 0.32, respectively. We run an MCMC with burn-in of 1000 and 5000 iterations on the testing dataset such that the prior mean for lasso is approximately equal to 0.25. Table 3 lists the posterior mean estimates of the shrinkage parameter in the lasso model under various priors. Table 4 lists the posterior mean estimates of the BEL lasso coefficients using different priors. The value of the estimates are pretty similar and the choice of prior for this α parameter is not that crucial. Figs. 4 depicts the trace plots for each predictor, where the shrinkage coefficient follows the Gamma distribution. The points in the trace plots are concentrated around their centers with reasonable fluctuations, which demonstrate that the chains have good mixing. Trace plots for BEL lasso estimates, where the shrinkage parameter follows the beta and the uniform distributions, are depicted in Figs. S1 and S2 in Supplementary Material. Table 5 lists the 95% highest posterior density intervals and the 95% equal-tailed credible regions for BEL lasso. The lower and upper limits for each predictor are quite similar under various priors and they do agree on the importance of the variables. For instance, both 95% HPD and 95% equal-tail probability intervals for variables

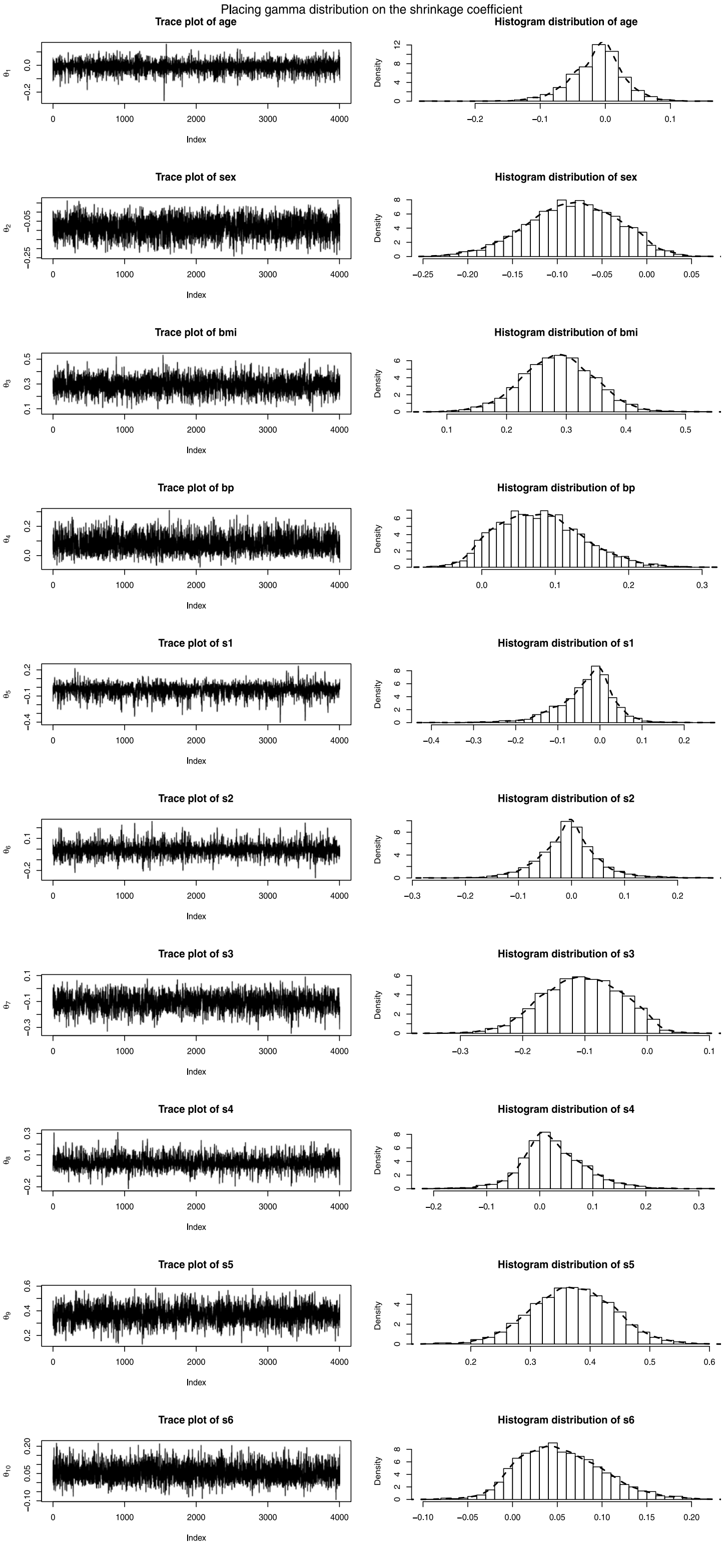

Fig. 4. Trace plot (left) and histogram along with the kernel density (right) of the posterior mean estimates for the shrinkage parameter under Gamma prior in the BEL lasso; using 5000 iterations with 1000 burn-in.

BMI and S5 do not include the value 0. This implies the importance of BMI and S5 on predicting the disease progression one year after the baseline, which aligns with the results shown in Fig. 3. Placing different priors on α results in different coefficient estimates. We use the training data to learn about the shrinkage parameter, such that the mean of the prior equals to $\alpha^{training}$. It is clear from the results provided in Table 5 that the posterior estimates of θ_j's are sensitive to the proposed prior distribution of α. For instance, we place a gamma distribution on α^2 with shape and rate equal to r and d, respectively. The values of r and d are selected such that $\frac{r}{d} = E\left(\left(\alpha^{training}\right)^2\right)$. Note that different values of r and d may satisfy the latter condition. Hence, one has to be careful on how much information is placed on the prior. For example in the gamma distribution if we fix the rate at a certain value, a small (large) value of the shape leads to a wider (more concentrated) prior.

8. Discussion

In this article we developed a new Bayesian approach for ridge and lasso regressions based on empirical likelihood. The method is semiparametric because it combines a nonparametric model and a parametric model. We derived the profile EL for a linear model and employed it as the likelihood piece in the Bayesian setting. To obtain the desired BEL ridge and lasso models then simply involved altering the prior distributions on the regression coefficients. Although EL inherits many properties from the likelihood function without assuming the distributional form of the data (in the regression case, this means that there is no need to assume normality of the errors) in this Bayesian framework, the resulting posterior EL distribution lacks a closed form. Implementation of even standard MCMC methods is very challenging due to the features of the support of the posterior EL function. To overcome the problem of nonconvergence of this nonconvex optimization, we used the tailored Metropolis–Hastings algorithm of Chib and Greenberg (1995).

The role of regularization in regression is to stabilize inference when the model is otherwise unidentifiable. A particularly important case in modern settings occurs when the sample size (n) is smaller than the number of predictor (p) variables. In this situation, the curse of dimensionality is acute. More critically for inference purposes there is insufficient information and too few degrees of freedom to estimate the full model. Penalized regressions such as the lasso were introduced to overcome the sparsity problem and to deal with data for which $p \gg n$. Notably, however, our approach is not applicable when $p \gg n$. The reason is that the profile EL ratio uses the ordinary regression estimating equations, and the penalization for ridge or lasso is incorporated through the prior distribution on the parameters of the regression model. Hence, the noninvertability of X^TX is not avoided in this formulation of the problem, and $w_i = 0$ for $i = 1, \ldots, n$.

A way around this is to modify the estimating equations used in the Bayesian empirical likelihood directly, as for ridge regression. That is, instead of using the estimating equations $X^TX\boldsymbol{\theta} - X^Ty$ one can set $\tilde{\Sigma} = X^TX + cI$ where c is a small value. Then, $\tilde{\Sigma}$ is used in the estimating equations instead of X^TX. This approach is similar to elastic net, a regularization method that linearly combines the ridge and lasso penalties, minimizing the least-squares criterion subject to l_1 and l_2 penalties. In the Bayesian approach, one can place the following prior distribution on $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}) \propto \exp\left\{-\alpha_1\|\boldsymbol{\theta_1}\| - \alpha_2\|\boldsymbol{\theta_2}\|_2^2\right\},$$

where α_1 and α_2 are the shrinkage parameters for the lasso and ridge methods, respectively. Alternatively, one could modify the ridge approach by adding a value c to the diagonal of X^TX and placing a Laplace prior on θ, where c can be selected via cross-validation. We will explore these modifications in future work.

Acknowledgment

Nicole A. Lazar was supported in part by the National Science Foundation under grant NSF IIS-1607919.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2020.106917.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19, 716–723.

Andrews, D.F., Mallows, C.I., 1974. Scale mixtures of normal distributions. J. R. Stat. Soc. Ser. B Stat. Methodol. 36, 99–102.

Bae, K., Mallick, B.K., 2004. Gene selection using a two-level hierarchical Bayesian model. Bioinformatics 20, 3424–3430.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. John Wiley & Sons, Chichester, New York, USA.

Chaudhuri, S., Drton, M., Richardson, T.S., 2007. Estimation of a covariance matrix with zeros. Biometrika 94, 199–216.

Chaudhuri, S., Ghosh, M., 2011. Empirical likelihood for small area estimation. Biometrika 98, 473–480.

Chaudhuri, S., Handcock, M.S., Rendall, M.S., 2008. Generalized linear models incorporating population level information: An empirical-likelihood-based approachs. J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 311–328.

Chen, J., Qin, J., 1993. Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. Biometrika 80, 107–116.

Chib, S., Greenberg, E., 1995. Understanding the metropolis-hastings algorithm. Amer. Statist. 49, 327–335.

Chib, S., Shin, M., Simoni, A., 2018. Bayesian estimation and comparison of moment condition models. J. Amer. Statist. Assoc. 113, 1–13.

Cochran, W.G., 1934. The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. Math. Proc. Camb. Phil. Soc. 30, 178–191.

Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numer. Math. 31, 377–403.

Deville, J.C., Sarndal, C.E., 1992. Calibration estimators in survey sampling. J. Amer. Statist. Assoc. 87, 376–382.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32, 407–451.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Figueiredo, M.A.T., 2003. Adaptive sparseness for supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 25, 1150–1159.

Geisser, S., 1993. Predictive Inference: An Introduction. In: Monographs on Statistics and Apllied probability, vol. 55, Chapman and Hill, New York.

Grendar, M., Judge, G., 2009. Asymptotic equivalence of empirical likelihood and Bayesian MAP. Ann. Statist. 37, 2445–2457.

Hartley, H.O., Rao, J.N.K., 1968. A new estimation theory for sample surveys. Biometrika 55, 547–557.

Hastie, T., Efron, B., 2013. LARS: Least angle regression, lasso and forward stagewise. URL: https://CRAN.R-project.org/package=lars. R package version 1.2.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer Series in Statistics, Springer, New York.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: With Applications in R, first ed. In: Springer Texts in Statistics, Springer New York.

Kolaczyk, E.D., 1994. Empirical likelihood for generalized linear models. Statist. Sinica 4, 199–218.

Kuk, A.Y.C., Mak, T.K., 1989. Median estimation in the presence of auxiliary information. J. R. Stat. Soc. Ser. B Stat. Methodol. 51, 261–269.

Lazar, N.A., 2003. Bayesian empirical likelihood. Biometrika 90, 319–326.

Leng, C., Tang, C.Y., 2010. Penalized high-dimensional empirical likelihood. Biometrika 97, 905–920.

Leng, C., Tang, C.Y., 2012. Penalized empirical likelihood and growing dimensional general estimating equations. Biometrika 99, 703–716.

Los Campos, G., Naya, H., Gianola, D., Grossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, M.J., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genet. Soc. Am. 182, 375–385.

Mengersen, K.L., Pudlo, P., Robert, C.P., 2013. Bayesian computation via empirical likelihood. Proc. Natl. Acad. Sci. USA 110, 1321–1326.

Monahan, J.F., Boos, D.D., 1992. Proper likelihoods for Bayesian analysis. Biometrika 79, 271–278.

Mykland, P.A., 1994. Bartlett type identities for martingales. Ann. Statist. 22, 21–38.

Neal, R., 2011. Handbook of Markov Chain Monte Carlo, Chapter 5: MCMC using Hamiltonian Dynamics. In: Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Chapman & Hall/CRC.

Owen, A.B., 1988. Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75, 237–249.

Owen, A.B., 1990. Empirical likelihood ratio confidence regions. Ann. Statist. 18, 90–120.

Owen, A.B., 1991. Empirical likelihood for linear models. Ann. Statist. 19, 1725–1747.

Owen, A.B., 2001. Empirical Likelihood. Chapman & Hall/CRC, Boca Raton.

Park, T., Casella, G., 2008. The Bayesian lasso. J. Amer. Statist. Assoc. 103, 681–686.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. Ann. Statist. 22, 300–325.

Rao, J.N.K., Wu, C., 2010. Bayesian pseudo-empirical-likelihood intervals for complex surveys. J. R. Stat. Soc. Ser. B Stat. Methodol. 72, 533–544.

Schennach, S.M., 2007. Point estimation with exponentially tilted empirical likelihood. Ann. Statist. 35, 634–672.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.

Tikhonov, A.N., Nikolayevich, N., 1943. On the stability of inverse problems. Dokl. Akad. Nauk SSSR 39, 195–198.

Wu, C., 2004. Weighted empirical likelihood inference. Statist. Probab. Lett. 66, 67–79.

Yang, Y., He, X., 2012. Bayesian empirical likelihood for quantile regression. Ann. Statist. 40, 1102–1131.

Yuan, M., Lin, Y., 2005. Efficient empirical Bayes variable selection and estimation in linear models. J. Amer. Statist. Assoc. 100, 1215–1225.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 301–320.