




Tuning Parameter Selection for the Adaptive Lasso Using ERIC

Francis K. C. Hui, David I. Warton & Scott D. Foster


To cite this article: Francis K. C. Hui, David I. Warton & Scott D. Foster (2015) Tuning Parameter Selection for the Adaptive Lasso Using ERIC, Journal of the American Statistical Association, 110:509, 262-269, DOI: [10.1080/01621459.2014.951444](https://doi.org/10.1080/01621459.2014.951444)

To link to this article: <http://dx.doi.org/10.1080/01621459.2014.951444>

 View supplementary material 

 Accepted author version posted online: 15 Aug 2014.

 Submit your article to this journal 

 Article views: 563

 View Crossmark data 

 Citing articles: 1 View citing articles 

Tuning Parameter Selection for the Adaptive Lasso Using ERIC

Francis K. C. HUI, David I. WARTON, and Scott D. FOSTER

The adaptive Lasso is a commonly applied penalty for variable selection in regression modeling. Like all penalties though, its performance depends critically on the choice of the tuning parameter. One method for choosing the tuning parameter is via information criteria, such as those based on AIC and BIC. However, these criteria were developed for use with unpenalized maximum likelihood estimators, and it is not clear that they take into account the effects of penalization. In this article, we propose the extended regularized information criterion (ERIC) for choosing the tuning parameter in adaptive Lasso regression. ERIC extends the BIC to account for the effect of applying the adaptive Lasso on the bias-variance tradeoff. This leads to a criterion whose penalty for model complexity is itself a function of the tuning parameter. We show the tuning parameter chosen by ERIC is selection consistent when the number of variables grows with sample size, and that this consistency holds in a wider range of contexts compared to using BIC to choose the tuning parameter. Simulation show that ERIC can significantly outperform BIC and other information criteria proposed (for choosing the tuning parameter) in selecting the true model. For ultra high-dimensional data ($p > n$), we consider a two-stage approach combining sure independence screening with adaptive Lasso regression using ERIC, which is selection consistent and performs strongly in simulation. Supplementary materials for this article are available online.

KEY WORDS: BIC; Consistency; High-dimensional data; Information criteria; Penalized likelihood; Regularization parameter; Variable selection.

1. INTRODUCTION

Penalized likelihood functions are a powerful approach to variable selection in regression modeling. One commonly applied penalty is the adaptive Lasso, based on a weighted sum of the absolute value of the coefficients (Zou 2006). Like all penalties though, its performance depends critically on the choice of tuning parameter. This is because only a certain range of the tuning parameter values leads to selection consistency, that is, asymptotically identifies the true model. In this article, we focus on using information criteria for tuning parameter selection in adaptive Lasso regression. Other methods of choosing the tuning parameter include cross-validation (Zou 2006) and selection stability (Sun, Wang, and Fang 2013), for instance.

The two most common used criteria for selecting the tuning parameter are those based on AIC (Akaike 1974) and BIC (Schwarz 1978). However, both criteria were designed for regression models with unpenalized maximum likelihood estimators, and so their use in the penalized likelihood setting somewhat lacks motivation (Bühlmann and van de Geer 2011). It is not clear that either criterion account for the prior information introduced by the adaptive Lasso penalty, and the effect of this penalization on the bias-variance tradeoff.

Recently, several modifications to BIC have been proposed, with the aim been to construct a criterion for selecting the tuning parameter which is selection consistent in high-dimensional

settings (e.g., Wang, Li, and Leng 2009; Wang and Zhu 2011; Fan and Tang 2012). These BIC-type criteria were formulated by establishing the conditions necessary to achieve consistency, that is, considering changes in the criterion for the cases of overfitting (tuning parameter too small) or underfitting (tuning parameter too large), rather than by directly considering the effects of penalization on the bias-variance tradeoff. To this end, Konishi, Ando, and Imoto (2004) formulated a BIC-type criterion for choosing smoothing parameters in radial basis function network models, by directly applying Laplace's approximation on the joint distribution of the likelihood and the basis functions which assumed Gaussian priors (similar to ridge regression; Hoerl and Kennard 1970). Kawano (2012) adopted the same approach to derive a criterion for choosing the tuning parameter in bridge regression (Frank and Friedman 1993), although the resulting criterion had a rather complicated form and there was no consideration of selection consistency.

In this article, we propose the extended regularized information criterion (ERIC) for selecting the tuning parameter in adaptive Lasso regression. ERIC is an extension of BIC to account for the effect of applying the adaptive Lasso, and in particular the effect of the Laplace prior on coefficients not shrunk to zero. This leads to a criterion whose penalty for model complexity is itself a function of the tuning parameter. For Generalized Linear Models (GLMs; McCullagh and Nelder 1989) with the adaptive Lasso, we show the tuning parameter chosen by ERIC is selection consistent when the number of variables grows at a smaller rate than sample size, and that this consistency holds in a wider range of contexts compared to using BIC to choose the tuning parameter. We use ERIC in conjunction with a hybrid estimator, whereby the final coefficient estimates are based on applying unpenalized maximum likelihood to the selected covariates (Efron et al. 2004). Simulations show that ERIC can

Francis K. C. Hui (E-mail: fhui28@gmail.com) and David I. Warton (E-mail: david.warton@unsw.edu.au), School of Mathematics and Statistics, The University of New South Wales 2052, Sydney, Australia and CSIRO Computational Informatics, Australia. Scott D. Foster, CSIRO Computational Informatics, Australia, and CSIRO Wealth from Oceans Flagship, Hobart 7001, Australia (E-mail: scott.foster@csiro.au). F.K.C.H. is supported by a Research Excellence Award from the University of New South Wales and a CSIRO PhD Scholarship. D.I.W. is supported by Australian Research Council Discovery Projects and Future Fellow funding schemes (project number DP130102131 and FT120100501). S.D.F. was supported in part by the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Research Program (NERP). Thanks to Berwin Turlach for useful discussions and Eric Foster for inspiration.

significantly outperform BIC and some recently proposed modifications in selecting the true model.

We conclude the introduction by reviewing two methods of variable selection which have been recently developed for ultra high-dimensional settings, that is, $p > n$, and which are outside of the penalized likelihood framework. First, the extended BIC (EBIC, Chen and Chen 2008) is a modification of BIC which imposes a prior distribution to account for the increasing complexity of the model space. Although EBIC is used in conjunction with Lasso or SCAD (Fan and Li 2001) penalties for instance, it only uses penalized regression for the purpose of fitting a sequence of models quickly. That is, EBIC is constructed from the unpenalized likelihood at each of the subset models on the regularization path, and it was not originally designed for tuning parameter selection. It has been proven though that EBIC is selection consistent in high-dimensional linear models (Chen and Chen 2008; Luo and Chen 2013) and GLMs (Chen and Chen 2012; Luo and Chen 2011). Also, Kim, Kwon, and Choi (2012) considered generalized information criterion, with EBIC as a special case, and established selection consistency under quite general conditions. Second, sure independence screening (SIS; Fan and Lv 2008; Fan et al. 2010b) is based on using marginal statistics for each covariate, for example, marginal likelihood ratio values, to screen out those which are uninformative. For ultra high-dimensional data, we show that a two-stage approach combining SIS and adaptive Lasso regression using ERIC is selection consistent and performs strongly in simulation.

2. ADAPTIVE LASSO REGRESSION

We apply the adaptive Lasso to generalized linear models. Let $\{(x_i, y_i); i = 1, \dots, n\}$ be a sample of independent and identically distributed observations, where y is a univariate response and x a p -dimensional vector of covariates. The number of covariates, p , is allowed to grow with sample size. The conditional density of y_i given x_i is assumed to come from the exponential family of distributions,

$$\begin{aligned} f(y_i | x_i, \beta, \phi) &= \exp \left(\frac{1}{\phi} (y_i \theta_i - b(\theta_i)) + c(y_i, \phi) \right) \\ &= \exp \left(\frac{1}{\phi} (y_i x_i' \beta - b(x_i' \beta)) + c(y_i, \phi) \right), \end{aligned}$$

for suitably chosen functions $b(\cdot)$ and $c(\cdot, \cdot)$, where $\beta = (\beta_1, \dots, \beta_p)$ denotes the regression coefficients, $\theta = x' \beta$ is the canonical parameter, and $E(y | x) = b'(\theta) = \mu$. We use the canonical link function, $g(\mu) = \theta$, with the nuisance parameter ϕ either known in advance or requiring estimation. Finally, we assume the covariates have been standardized to have mean zero and unit variance.

Let $\tilde{\beta}$ be the maximum likelihood estimate (MLE) of β . Note that $\tilde{\beta}$ is, under general regularity conditions, well-defined provided $p < n$. The penalized log-likelihood function for the adaptive Lasso is given by

$$\ell_p(\beta) = \ell(y | \beta) - \lambda \sum_{j=1}^p \tilde{w}_j |\beta_j|, \quad (1)$$

where $\ell(y | \beta) = \sum_{i=1}^n \log f(y_i | x_i, \beta, \phi)$, and $\tilde{w}_j = 1/|\tilde{\beta}_j|^\gamma$ with $\gamma > 0$ the power parameter. The penalized estimates, de-

noted as $\hat{\beta}_\lambda$, are obtained by maximizing (1). In a Bayesian context, the adaptive Lasso can be regarded as imposing independent Laplace priors on the coefficients (Park and Casella 2008),

$$\rho(\beta_j | \lambda) = \frac{\lambda \tilde{w}_j}{2\phi} e^{-\lambda \tilde{w}_j |\beta_j| / \phi} \quad \text{for } j = 1, \dots, p, \quad (2)$$

such that maximizing (1) is equivalent to using the posterior mode as a point estimate. Note the scaling by $1/\phi$ in (2) is necessary to cancel out the $1/\phi$ present in the exponential family—this ensures the nuisance parameter does not enter into the calculation of $\hat{\beta}_\lambda$. Besides, in penalized likelihood analysis, one usually estimates ϕ separately after the penalized coefficients have been estimated. For example, in linear models with unknown $\phi \equiv \sigma^2$, the variance is estimated afterward as $\hat{\sigma}_\lambda^2 = (1/n) \|y - x' \hat{\beta}_\lambda\|^2$ (Wang, Li, and Tsai 2007).

The adaptive Lasso introduces prior information which purposefully biases the MLEs, such that some of the unpenalized estimates are shifted to zero. Given an appropriate choice of λ , only the truly zero coefficients are asymptotically shifted to zero, leading to selection consistency. The penalization also reduces the variance of the mean response—for any fixed sample size, larger values of λ lead to greater bias and less variability in predicted values.

3. THE EXTENDED REGULARIZED INFORMATION CRITERION

One approach for choosing the tuning parameter, λ , is via information criteria, with the most commonly used based on BIC (Wang, Li, and Tsai 2007; Zhang, Li, and Tsai 2010),

$$\text{BIC}(\lambda) = -2\ell(y | \hat{\beta}_\lambda) + |\alpha_\lambda| \log(n),$$

where $\alpha_\lambda = \{j : \hat{\beta}_{\lambda,j} \neq 0\}$ is the active set excluding the intercept, and $|\alpha_\lambda|$ is its size. It has been shown that $\text{BIC}(\lambda)$ is selection consistent in the case of fixed p (Zhang, Li, and Tsai 2010). Furthermore, in Section 4 we show $\text{BIC}(\lambda)$ is consistent provided $p/n^\kappa \rightarrow 0$ with $\kappa < 1/2$. As with all information criteria, $\text{BIC}(\lambda)$ can be interpreted in terms of a bias-variance tradeoff—the first term measures bias or goodness of fit, and the second term measures variance of the predicted values (Hastie, Tibshirani, and Friedman 2009). As more covariates are removed from the model, the goodness of fit term increases as we move away from the MLEs of the full model, while the variance penalty decreases due to reduced model complexity.

The BIC was designed for model selection with unpenalized maximum likelihood estimators (Schwarz 1978). It is based on an asymptotic approximation of the model likelihood, and is derived by applying Laplace's approximation and assuming the uninformative priors on all coefficients, $\rho(\beta_j) = O(1)$; $j = 1, \dots, p$. Since $\text{BIC}(\lambda)$ directly adapts the form of BIC to the penalized likelihood setting for choosing λ , it is not clear that the criterion takes into account the effect of applying the adaptive Lasso on the bias-variance tradeoff. This is most clearly seen if we consider an interval $[\lambda_a, \lambda_b]$, with $\lambda_a < \lambda_b$, for which $|\alpha_\lambda|$ remains unchanged. As we increase λ within this interval, the bias of the adaptive Lasso estimates not shrunk to zero is increased, and this is captured by a corresponding increase in $-2\ell(y | \hat{\beta}_\lambda)$. At the same time however, the

variance of these adaptive Lasso estimates (and hence the variance of the predicted response) decreases, yet this is not accounted for since $|\alpha_\lambda| \log(n)$ remains unchanged. This decrease in variance of the nonzero estimates is a result of the increased information from the Laplace prior, as captured by the normalization by constant in Equation (2), $\lambda \tilde{w}_j / (2\phi)$. This suggests that the variance penalty in $\text{BIC}(\lambda)$ should be modified accordingly to $|\alpha_\lambda| \log(n) - |\alpha_\lambda| \log(\lambda/\phi) = |\alpha_\lambda| \log(n\phi/\lambda)$. Therefore, we propose the following ERIC for tuning parameter selection in adaptive Lasso GLMs,

$$\text{ERIC}_v(\lambda) = -2\ell(y|\hat{\beta}_\lambda) + 2v|\alpha_\lambda| \log(n\phi/\lambda). \quad (3)$$

From (3), $\text{ERIC}_v(\lambda)$ adjusts the penalty for model complexity in $\text{BIC}(\lambda)$ by accounting for the effect of prior information introduced by the adaptive Lasso. Coefficients shrunk to zero contribute toward the bias term but not the variance term in (3). On the other hand, for any fixed sample size, coefficients not shrunk to zero have less variance than if they were estimated with no penalty, that is, their variance is less than the variance of the MLEs calculated from the submodel α_λ . Therefore, $\text{ERIC}_v(\lambda)$ accounts for this by reducing the variance penalty to $|\alpha_\lambda| \log(n\phi/\lambda)$. Large values of λ incur greater shrinkage and a smoother (less variable) mean response, so the variance penalty in $\text{ERIC}_v(\lambda)$ decreases to reflect this. The additional parameter, $v > 0$, provides flexibility to control the severity of penalization (similar to EBIC, Chen and Chen 2008). We choose between $v = 0.5$ or 1 , with the former better suited to high-dimensional data.

It is important to emphasize that the way $\text{ERIC}_v(\lambda)$ penalizes for model complexity is fundamentally different to $\text{BIC}(\lambda)$ and its modifications proposed for high-dimensional settings—for a fixed sample size, $\text{BIC}(\lambda)$ penalizes a constant value for every new covariate entered into the model. In contrast, $\text{ERIC}_v(\lambda)$ has a “dynamic” variance penalty which depends on λ itself, meaning it also depends on how complex the model is already.

On the surface then, given the reduced variance penalty, it appears $\text{ERIC}_v(\lambda)$ has a tendency to select larger models compared to $\text{BIC}(\lambda)$. However, it turns the opposite actually occurs—when we consider the relative values of a criteria, which are of more importance in model selection, it turns out that $\text{ERIC}_v(\lambda)$ penalizes *more* severely for overfitting than $\text{BIC}(\lambda)$. This point is detailed at the end of Section 4.

Regarding large sample behavior, the form of $\text{ERIC}_v(\lambda)$ assumes λ grows with sample size, since the effect of penalization on the bias-variance tradeoff is nontrivial in such case. Note that λ must grow with sample size anyway to achieve selection consistency (Zou 2006; Zou and Zhang 2009). On the other hand, if $\lambda = O(1)$ and $v = 0.5$ then $\log(\lambda/\phi) = O(1)$ and $\hat{\beta}_\lambda$ tends to the maximum likelihood estimates for the submodel α_λ , in probability. In other words, if adaptive Lasso penalty is asymptotically negligible, then $\text{ERIC}_v(\lambda)$ reduces to the BIC with unpenalized MLEs.

As is typically done for linear models where $\phi \equiv \sigma^2$ is unknown, we estimate the variance term in $\text{ERIC}_v(\lambda)$ as $\hat{\sigma}_\lambda^2 = (1/n)\|y - x'\hat{\beta}_\lambda\|^2$.

4. SELECTION CONSISTENCY

In this section, we show that $\text{ERIC}_v(\lambda)$ is selection consistent for adaptive Lasso GLMs where p grows at a smaller rate than sample size, that is, $p/n^\kappa \rightarrow 0$ and $\kappa < 1$. This is often referred to literature as a diverging number of parameters setting, after Fan and Peng (2004). We use p_n to reflect this, although for other quantities, for example, α_λ , the subscript is suppressed for clarity of notation. Let $\beta^0 = (\beta_1^0, \dots, \beta_p^0)$ denote the true parameter values, with the true model identified by $\alpha_0 = \{j : \beta_j^0 \neq 0\}$ and $p_0 = |\alpha_0|$. Assume the following regularity conditions are satisfied.

- (A1) The range of tuning parameters considered lies in the interval $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\max}/n \rightarrow 0$.
- (A2) $\lim_{n \rightarrow \infty} \log(p_n)/\log(n) = \kappa$, where $\kappa \in [0, 1)$.
- (A3) For all $i = 1, \dots, n$, $j = 1, \dots, p_n$ in the model matrix x , $|x_{ij}| = O(1)$. Furthermore, for any model α identified in the interval $[\lambda_{\min}, \lambda_{\max}]$, $\Sigma_\alpha = (1/n)x'_\alpha x_\alpha$ has minimum and maximum eigenvalues bounded below and above by c_1 and c_2 , respectively, where $0 < c_1 < c_2 < \infty$.
- (A4) $\lim_{n \rightarrow \infty} \sqrt{n/(p_0 \log(n))} \min_{j \in \alpha_0} \{|\beta_j^0|\} \rightarrow \infty$.
- (A5) (a) $\lambda\sqrt{p_0/n} \rightarrow 0$; (b) $\lambda(n/p_n)^{\gamma/2}/\sqrt{np_n} \rightarrow \infty$.

Condition (A1) ensures the maximum tuning parameter considered does not grow at a faster rate than the sample size. Moreover, for each model α identified in $[\lambda_{\min}, \lambda_{\max}]$, there exists a unique population parameter to which the unpenalized MLEs $\tilde{\beta}_\alpha$ converges to in probability. If $\alpha \supset \alpha_0$, then this population parameter will coincide with β^0 , but if $\alpha \not\supset \alpha_0$ then this will not be the case. Condition (A2) permits the number of covariates to diverge with sample size, with the same rate of divergence as in Wang, Li, and Leng (2009); and Zou and Zhang (2009). Along with a mild regularity condition on the likelihood function (see Supplementary Material A), Condition (A3) ensures the Fisher information matrix exists and is singular for each n . Note that under Conditions (A2) and (A3), the MLEs for the full model are well-defined, $\sqrt{n/p_n}$ -consistent, and thus weights $\tilde{w}_j = 1/|\tilde{\beta}_j|^\gamma$ can be calculated for use in the adaptive Lasso.

The size of the true model, p_0 , is permitted to grow with sample size, provided it satisfies Conditions (A4) and (A5) (and obviously Condition (A2)). Condition (A4) is similar to one seen in Wang, Li, and Leng (2009), and permits the truly nonzero coefficients to tend to zero. In other words, it allows competing underfitted models to get close to true model at a certain rate. When the number of covariates is fixed, Condition (A5a) simplifies to $\lambda/\sqrt{n} \rightarrow 0$ and $\lambda n^{(\gamma-1)/2} \rightarrow \infty$, which is the condition necessary for the adaptive Lasso to achieve consistency in the fixed p setting (Zou 2006). On the other hand, when the number of covariates diverges according to Condition (A2), we have the following result.

Lemma 1. Assume Conditions (A1)–(A5) are satisfied, then the adaptive Lasso estimates $\hat{\beta}_\lambda$ must satisfy:

- Estimation consistency: $\|\hat{\beta}_\lambda - \beta^0\| = O_p(\sqrt{p_n/n})$.
- Selection consistency: $P(\hat{\beta}_{\lambda_{\alpha_0^c}} = 0) \rightarrow 1$.

All proofs may be found in Supplementary Material A. Lemma 1 is a generalization of the consistency result for the

adaptive elastic net (Zou and Zhang 2009) to non-Gaussian responses and GLMs.

To study the asymptotic behavior of $\text{ERIC}_v(\lambda)$, we first define a proxy version of the criterion as follows:

$$\text{ERIC}_v^*(\lambda) = -2\ell(\mathbf{y}|\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) + 2\nu|\alpha_\lambda|\log(n\phi/\lambda),$$

where $\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}$ is the MLE under model α_λ . For linear models, we estimate the variance as $\tilde{\sigma}_{\alpha_\lambda}^2 = (1/n)\|\mathbf{y} - \mathbf{x}'_{\alpha_\lambda}\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}\|^2$ where $\mathbf{x}_{\alpha_\lambda}$ is the model matrix for α_λ . By examining the behavior of proxy $\text{ERIC}_v^*(\lambda)$ for the cases of underfitted and overfitted models, we can infer what happens to $\text{ERIC}_v(\lambda)$ in the large sample limit. We partition the interval $[\lambda_{\min}, \lambda_{\max}]$ into three segments, analogous to Zhang, Li, and Tsai (2010),

$$\begin{aligned} \text{Underfitted Models: } \Lambda_- &= \{\lambda : \alpha_\lambda \not\supseteq \alpha_0\} \\ \text{True Model: } \Lambda_0 &= \{\lambda : \alpha_\lambda = \alpha_0\} \\ \text{Overfitted Models: } \Lambda_+ &= \{\lambda : \alpha_\lambda \supsetneq \alpha_0\}. \end{aligned}$$

The three partitions are closely related to Condition (A5). For $\lambda \in \Lambda_0$, both Conditions (A5a) and (A5b) must be satisfied. For $\lambda \in \Lambda_-$, Condition (A5b) is satisfied but not (A5a), that is, λ is growing too quickly and removing one or more truly nonzero coefficients from the model. For $\lambda \in \Lambda_+$, Condition (A5a) is satisfied but not (A5b), that is, λ is growing too slowly to achieve the correct amount of sparsity. Using proxy $\text{ERIC}_v^*(\lambda)$ and the partitions above, we have the following lemma.

Lemma 2. Assume Conditions (A1)–(A4) are satisfied, and that there exists a $\lambda_0 \in \Lambda_0$ satisfying Condition (A5). Then,

$$\begin{aligned} (a) \quad & P\left(\inf_{\lambda \in \Lambda_-} \min_{\alpha_\lambda \not\supseteq \alpha_0} \text{ERIC}_v^*(\lambda) > \text{ERIC}_v^*(\lambda_0)\right) \rightarrow 1. \\ (b) \quad & \text{If } \gamma \geq 2\kappa/(v(1-\kappa)) - 1 \quad \text{also, then} \\ & P\left(\inf_{\lambda \in \Lambda_+} \min_{\alpha_\lambda \supsetneq \alpha_0} \text{ERIC}_v^*(\lambda) > \text{ERIC}_v^*(\lambda_0)\right) \rightarrow 1. \end{aligned}$$

From parts (a) and (b) of Lemma 2, we cannot asymptotically choose a λ that identifies an overfitted or underfitted model respectively, since we could always select λ_0 to produce a lower value of proxy $\text{ERIC}_v^*(\lambda)$. In Lemma 2b, the condition on γ reduces to a simple inequality for the two cases of $v = 0.5$ ($\gamma \geq (5\kappa - 1)/(1 - \kappa)$) and $v = 1$ ($\gamma \geq (3\kappa - 1)/(1 - \kappa)$). In simulations, choosing γ is straightforward since κ is known. In real applications, we suggest trying several values of γ , for example, $\gamma = 1, 2, 4, 6$, while taking into account the dimensionality of the dataset at hand.

Since λ_0 satisfies Condition (A5), then based on Lemmas 1 and 2 we obtain the following result.

Theorem 1. Let $\hat{\lambda}$ be the tuning parameter chosen by minimizing $\text{ERIC}_v(\lambda)$. Under Conditions (A1)–(A5), then $P(\alpha_{\hat{\lambda}} = \alpha_0) \rightarrow 1$.

If the true model is contained within the set of candidate models, then Theorem 1 guarantees it is selected by $\text{ERIC}_v(\lambda)$. Note that for fixed p , the same property is shared by $\text{BIC}(\lambda)$ (Zhang, Li, and Tsai 2010, Theorem 1B). However, if the number of covariates increases with sample size, then $\text{BIC}(\lambda)$ may be selection inconsistent as formalized below.

Corollary 1. Let $\hat{\lambda}$ be the tuning parameter chosen by minimizing $\text{BIC}(\lambda)$. If $\kappa > 0.5$ in Condition (A2), then $P(\alpha_{\hat{\lambda}} = \alpha_0) \not\rightarrow 1$.

The corollary above extends the result shown in Chen and Chen (2008) to the penalized likelihood setting. It also implies that $\text{ERIC}_v(\lambda)$ is selection consistent for a wider range of settings compared to $\text{BIC}(\lambda)$.

To further understand how $\text{ERIC}_v(\lambda)$ and $\text{BIC}(\lambda)$ penalize differently for models that lie apart on the interval $[\lambda_{\min}, \lambda_{\max}]$, consider the ratio of variance penalties (RVP) between an incorrect model with $\lambda \in \Lambda_+ \cup \Lambda_-$ and a correct model with λ_0 ,

$$\text{RVP}_{\text{BIC}} = \frac{|\alpha_\lambda|}{p_0}; \quad \text{RVP}_{\text{ERIC}_v} = \frac{|\alpha_\lambda|}{p_0} \frac{\log(n/\lambda)}{\log(n/\lambda_0)},$$

where for simplicity we have assumed $\phi = 1$. For $\text{BIC}(\lambda)$, and more generally for all criteria with fixed variance penalties, the RVP depends solely on the sizes of the true and overfitted models. In contrast, the dynamic variance penalty means $\text{RVP}_{\text{ERIC}_v} > \text{RVP}_{\text{BIC}}$ when $\lambda < \lambda_0$, implying that $\text{ERIC}_v(\lambda)$ penalizes more severely than $\text{BIC}(\lambda)$ and its modifications for overfitted models. Put another way, $\text{ERIC}_v(\lambda)$ is more “aggressive” at shrinking coefficients to zero. Conversely, $\text{RVP}_{\text{ERIC}_v} < \text{RVP}_{\text{BIC}}$ for underfitted models, meaning that this aggressive shrinkage comes at the risk of underfitting. In Section 5, we shall see empirically that this aggressive approach of $\text{ERIC}_v(\lambda)$ can produce substantial improvements in variable selection.

4.1 Hybrid Estimation Approach

While $\text{ERIC}_v(\lambda)$ performs strongly in variable selection, it does have a tendency to overshrink correctly selected coefficients. This is not surprising given its dynamic variance penalty makes it a more aggressive criterion in terms of shrinkage and selection, as discussed in the previous section. On the other hand, $\text{BIC}(\lambda)$ and its modifications are more conservative criteria in that, for any active set α_λ , they always choose the smallest possible λ . This is illustrated in Figure 1 where, for datasets whereby both $\text{ERIC}_v(\lambda)$ and $\text{BIC}(\lambda)$ select the true model, almost all tuning parameters chosen by $\text{ERIC}_v(\lambda)$ lie above the median tuning parameter value chosen by $\text{BIC}(\lambda)$. Note, however, that the variability in λ values chosen by $\text{ERIC}_v(\lambda)$ is significantly less compared to $\text{BIC}(\lambda)$. While some shrinkage is useful for improving predictive accuracy (Hoerl and Kennard 1970), over shrinkage of correctly selected covariates can be detrimental to prediction since it introduces an excessive amount of bias.

To resolve this problem, we adopt the hybrid estimation approach of Efron et al. (2004), and use the unpenalized MLEs of the selected covariates as the final estimates. That is, we use $\text{ERIC}_v(\lambda)$ to select the model but not to estimate the coefficients. Several authors have suggested this approach in various applications of penalized likelihood methods (e.g., Meier, Van De Geer, and Bühlmann 2008; Schellendorfer et al. 2014). Since the motivation behind $\text{ERIC}_v(\lambda)$ was to develop an information criterion for consistent variable selection, as opposed to efficient prediction, then it seems sensible to use maximum likelihood estimation to obtain asymptotically unbiased final estimates. The use of a hybrid estimator with $\text{ERIC}_v(\lambda)$ also echoes the underlying aim of the adaptive Lasso, namely to achieve the oracle property, that is, asymptotically unbiased and as efficient as the MLEs from the true model (Zou 2006).

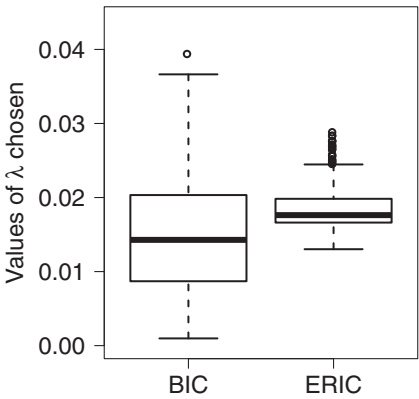


Figure 1. Boxplots of λ values chosen by $\text{BIC}(\lambda)$ and $\text{ERIC}_1(\lambda)$, in datasets where both criteria selected the true model. The boxplots are based on datasets of with $n = 200$ simulated from Model 2 (see Section 5.2).

5. SIMULATION STUDY

We conducted a simulation study to compare the performance of $\text{ERIC}_\nu(\lambda)$, with $\nu = 0.5$ and 1, against several BIC-type criteria used to select λ in adaptive Lasso GLMs. These included $\text{BIC}(\lambda)$, $\text{BIC}_{\text{wll}}(\lambda) = -2\ell(\mathbf{y}|\hat{\boldsymbol{\beta}}_\lambda) + |\alpha_\lambda| \log(n) \log(\log(p_n))$ (Wang, Li, and Leng 2009), $\text{BIC}_{\text{ft}}(\lambda) = -2\ell(\mathbf{y}|\hat{\boldsymbol{\beta}}_\lambda) + |\alpha_\lambda| \log(\log(n)) \log(p_n)$ (Fan and Tang 2012), and $\text{BIC}_{\text{wz}}(\lambda) = -2\ell(\mathbf{y}|\hat{\boldsymbol{\beta}}_\lambda) + |\alpha_\lambda| \log(p_n)$ (Wang and Zhu 2011). We also included 10-fold cross-validation (CV) as a method of selecting λ , as this is the standard choice in many software packages for performing adaptive Lasso regression (e.g., the `glmnet` package; Friedman, Hastie, and Tibshirani 2010).

Note that all information criteria except for $\text{ERIC}_\nu(\lambda)$ have a fixed variance penalty. That is, the same amount is penalized for each covariate included in the model. In other words, for any model α_λ , these criteria will always choose the smallest possible λ attaining that model. Also, note these criteria and CV use the penalized estimator based on maximizing (1), whereas we apply $\text{ERIC}_\nu(\lambda)$ with the hybrid estimator (see Section 4.1).

We generated 1000 datasets for each simulation model, and assessed performance by looking at the proportion of true models selected, the false positive and false negative rates (which are presented in the Supplementary Material B), and the Kullback–Leibler (KL) distance between the selected and

true models based on a validation dataset of $n = 5000$ observations. A smaller KL distance implies better predictive performance.

5.1 Model 1

We used a similar setup to Wang, Li, and Leng (2009), considering a linear model with $p_n = \lfloor 7n^{1/4} \rfloor$ where $\lfloor \cdot \rfloor$ denotes the floor function. Covariates $\{\mathbf{x}_i; i = 1, \dots, n\}$ were generated independently from an AR1 multivariate Gaussian distribution with $\rho = 0.5$. The first $\lfloor p_n/3 \rfloor$ coefficients were generated randomly from a uniform distribution on $[0.5, 1.5]$, and the remaining coefficients set to zero. We considered sample sizes $n = 100, 200, 400, 800, 1600$ and error variances $\sigma^2 = 4, 16$, and used $\gamma = 1$ in the adaptive Lasso weights for both $\nu = 0.5$ and 1.

Both $\text{ERIC}_\nu(\lambda)$ with $\nu = 0.5$ and 1 significantly outperformed the other criteria and 10-fold CV in terms of variable selection and predictive performance, irrespective of n and σ^2 (Table 1). The improvement was driven largely by a reduction in the false positive rates for $\text{ERIC}_\nu(\lambda)$, with little compromise in the false negative rates (see Supplementary Material B). This is consistent with the discussion at the end of Section 4 regarding the tendency for $\text{ERIC}_\nu(\lambda)$ to choose larger values of λ and thus smaller models compared to the BIC-type criteria. Of the four BIC-type criteria tested, $\text{BIC}_{\text{wll}}(\lambda)$ was the best performer, although this was somewhat expected given our simulation design was based on Wang, Li, and Leng (2009), from which the criteria originates. Its performance though was behind $\text{ERIC}_{0.5}(\lambda)$ and $\text{ERIC}_1(\lambda)$ both in variable selection and prediction. Finally, 10-fold CV performed very poorly with regards to selecting the true model, although its predictive performance was similar to the BIC-type criteria. This result was not surprising—in the fixed p setting at least, it has been shown that using cross-validation to choose λ is asymptotically loss efficient rather than consistent (Wang, Li, and Tsai 2007; Zhang, Li, and Tsai 2010).

5.2 Model 2

We considered a logistic regression model with the rate of divergence of p_n and covariates generated in the same manner as Model 1. The first five coefficients were equal to $(-3, 1.5, 0, 0, 2)$. Afterward, every fifth coefficient took

Table 1. Proportion of true models selected and KL distance ($\times 10$, in parentheses) for Model 1

n	$\text{ERIC}_{0.5}(\lambda)$	$\text{ERIC}_1(\lambda)$	$\text{BIC}(\lambda)$	$\text{BIC}_{\text{wll}}(\lambda)$	$\text{BIC}_{\text{ft}}(\lambda)$	$\text{BIC}_{\text{wz}}(\lambda)$	CV
$\sigma^2 = 4$							
100	0.70(0.76)	0.86(0.66)	0.55(1.07)	0.58(1.07)	0.55(1.07)	0.36(1.09)	0.24(1.14)
200	0.92(0.39)	0.99(0.36)	0.78(0.55)	0.83(0.56)	0.78(0.55)	0.55(0.57)	0.32(0.62)
400	0.97(0.25)	1(0.24)	0.85(0.33)	0.90(0.33)	0.86(0.33)	0.66(0.33)	0.36(0.35)
800	0.99(0.15)	1(0.14)	0.90(0.23)	0.94(0.23)	0.90(0.23)	0.82(0.21)	0.34(0.21)
1600	0.99(0.08)	1(0.08)	0.95(0.08)	0.99(0.08)	0.97(0.08)	0.90(0.08)	0.36(0.10)
$\sigma^2 = 16$							
100	0.21(1.34)	0.26(1.31)	0.16(1.47)	0.17(1.49)	0.16(1.47)	0.10(1.45)	0.04(1.51)
200	0.35(0.66)	0.40(0.68)	0.27(0.92)	0.31(0.96)	0.27(0.92)	0.14(0.83)	0.09(0.82)
400	0.72(0.27)	0.86(0.24)	0.54(0.40)	0.61(0.41)	0.55(0.41)	0.34(0.39)	0.15(0.39)
800	0.93(0.19)	0.94(0.21)	0.78(0.28)	0.85(0.31)	0.79(0.28)	0.54(0.25)	0.29(0.25)
1600	0.99(0.09)	0.99(0.09)	0.90(0.17)	0.97(0.19)	0.90(0.18)	0.85(0.14)	0.28(0.14)

Table 2. Proportion of true models selected and KL distance ($\times 10$, in parentheses) for Model 2

n	$\text{ERIC}_{0.5}(\lambda)$	$\text{ERIC}_1(\lambda)$	$\text{BIC}(\lambda)$	$\text{BIC}_{\text{wll}}(\lambda)$	$\text{BIC}_{\text{fit}}(\lambda)$	$\text{BIC}_{\text{wz}}(\lambda)$	CV
100	0.22(0.71)	0.30(0.76)	0.16(0.82)	0.19(0.86)	0.16(0.83)	0.09(0.77)	0.09(0.68)
200	0.28(0.40)	0.59(0.30)	0.20(0.40)	0.28(0.42)	0.22(0.41)	0.08(0.39)	0.06(0.33)
400	0.39(0.23)	0.70(0.13)	0.29(0.21)	0.36(0.22)	0.30(0.21)	0.10(0.19)	0.08(0.15)
800	0.45(0.15)	0.80(0.07)	0.32(0.13)	0.45(0.15)	0.34(0.13)	0.11(0.11)	0.09(0.09)
1600	0.59(0.05)	0.87(0.04)	0.49(0.05)	0.59(0.05)	0.50(0.05)	0.24(0.04)	0.08(0.04)

alternating values of ± 2 while the rest were set equal to zero. An unpenalized intercept of 1 was also included in the true model. We used $\gamma = 1$ for the adaptive Lasso weights.

$\text{ERIC}_\nu(\lambda)$ performed best in this setting, with substantial gains in both the proportion of correct models and KL distance over all four BIC-type criteria as sample size increased (Table 2). As in Model 1, the improvement could be attributed to a reduction in the false positive rates for $\text{ERIC}_\nu(\lambda)$, resulting in less overfitting compared to the BIC-type criteria (see Supplementary Material A). Using $\nu = 1$ generally lead stronger performance compared to $\nu = 0.5$, although the KL distance for $\text{ERIC}_1(\lambda)$ was larger than $\text{ERIC}_{0.5}(\lambda)$ at $n = 100$. This was due to $\text{ERIC}_1(\lambda)$ being overly aggressive in shrinking coefficients to zero, and thus underfitting more compared to $\text{ERIC}_{0.5}(\lambda)$ at the smallest sample size. Both $\text{BIC}_{\text{wz}}(\lambda)$ and 10-fold CV consistently overfitted for the five sample sizes tested (see also the higher false positive rates in Supplementary Material B), although this overfitting did lead to lower KL distance compared to the other methods.

5.3 Model 3

We considered a linear model with $p_n = \lceil 4n^{1/2} \rceil - 5$ and $\sigma^2 = 1$ and 16. The covariates and coefficients were generated in the same manner as Model 2, although the rate of divergence is now greater. An unpenalized intercept of -1 was also included in the model. We used $\gamma = 3$ in the adaptive Lasso weights to ensure selection consistency for both $\nu = 0.5$ and 1. Results were similar to those obtained for Models 1 and 2, with $\text{ERIC}_{0.5}(\lambda)$ and $\text{ERIC}_1(\lambda)$ outperforming the four BIC-type criteria and 10-fold CV in model selection and prediction (Table 3). This was especially the case at $\sigma^2 = 16$, where $\text{ERIC}_\nu(\lambda)$ had significantly higher proportions of true model selected and lower KL distance for $n = 200, 400, 800$. Again, these improvements were driven by a significant reduction in the false positive rate

at the expense of a very small rise in the false negative rate (see Supplementary Material B).

5.4 Additional Simulations

In the results above, one might hypothesize that the improvements in prediction were mostly a result of using the hybrid estimator in conjunction with $\text{ERIC}_\nu(\lambda)$. To test this, we also assessed predictive performance using the hybrid estimator in conjunction with all four BIC-type criteria considered above. Results presented in the Supplementary Material C show that, while the differences in KL distance were smaller when the hybrid estimator was used for all criteria, using it in conjunction with $\text{ERIC}_\nu(\lambda)$ still produced the lowest or equal lowest KL distances in all settings. This suggested that the improvement in prediction made by $\text{ERIC}_\nu(\lambda)$ was a result of both the use of a hybrid estimator, and its improvement in variable selection (due to better choice of tuning parameter).

As pointed out by a reviewer, given $\text{ERIC}_\nu(\lambda)$ tends to choose larger tuning parameters compared to BIC-type criteria, one might hypothesize that its performance may deteriorate when the true model is not (as) sparse. To test this, we performed an additional simulation with modified Models 1 and 2 to include more nonzero coefficients. Specifically, we considered a modified Model 1 where the first $\lfloor 2p_n/3 \rfloor$ coefficients were nonzero, and a modified Model 2 where every *third* coefficient took alternating values of ± 2 . Results are presented in Supplementary Material D, and show that $\text{ERIC}_\nu(\lambda)$ can still outperform BIC-type criteria and 10-fold CV even if the true model is relatively large, that is, over 50% of the coefficients are truly nonzero in the case of modified Model 1. $\text{ERIC}_\nu(\lambda)$ continued to have the lowest false positive rates, but not surprisingly a larger true model also meant it had slightly higher false negative rates at the smaller sample sizes compared to the other methods.

Table 3. Proportion of true models selected and KL distance ($\times 10$, in parentheses) for Model 3

n	$\text{ERIC}_{0.5}(\lambda)$	$\text{ERIC}_1(\lambda)$	$\text{BIC}(\lambda)$	$\text{BIC}_{\text{wll}}(\lambda)$	$\text{BIC}_{\text{fit}}(\lambda)$	$\text{BIC}_{\text{wz}}(\lambda)$	CV
$\sigma^2 = 1$							
100	0.99(0.95)	1(0.95)	0.56(1.11)	0.86(1.03)	0.82(1.06)	0.54(1.29)	0.25(1.70)
200	1(0.58)	1(0.58)	0.74(0.63)	0.92(0.61)	0.89(0.61)	0.65(0.69)	0.14(1.06)
400	1(0.41)	1(0.41)	0.92(0.51)	0.98(0.41)	0.96(0.41)	0.82(0.43)	0.10(0.73)
800	1(0.27)	1(0.27)	0.99(0.27)	0.99(0.27)	0.99(0.27)	0.96(0.27)	0.05(0.47)
1600	1(0.21)	1(0.21)	0.99(0.21)	1(0.21)	1(0.21)	0.98(0.21)	0.05(0.34)
$\sigma^2 = 16$							
100	0.18(1.89)	0.20(2.01)	0.13(1.98)	0.18(1.98)	0.17(1.97)	0.10(2.03)	0.04(2.12)
200	0.60(0.73)	0.64(0.71)	0.35(0.97)	0.47(1.00)	0.43(0.97)	0.27(0.98)	0.08(1.25)
400	0.96(0.43)	0.94(0.43)	0.71(0.52)	0.84(0.52)	0.80(0.52)	0.53(0.54)	0.08(0.81)
800	0.99(0.28)	0.99(0.28)	0.87(0.31)	0.94(0.31)	0.93(0.31)	0.81(0.31)	0.04(0.51)
1600	1(0.21)	1(0.21)	0.98(0.22)	0.99(0.22)	0.99(0.22)	0.92(0.22)	0.02(0.36)

Table 4. Positive selection rates/false discovery rates for the ultra high-dimensional data simulations. Results are presented to three decimal places to better differentiate between the methods

n	$\text{ERIC}_{0.5}(\lambda)$	$\text{ERIC}_1(\lambda)$	$\text{BIC}(\lambda)$	$\text{BIC}_{\text{wll}}(\lambda)$	$\text{BIC}_{\text{ft}}(\lambda)$	$\text{BIC}_{\text{wz}}(\lambda)$
Model 1 (logistic regression)						
80	0.136/0.861	0.131/0.853	0.143/0.870	0.139/0.869	0.141/0.869	0.144/0.870
300	0.385/0.670	0.348/0.405	0.427/0.712	0.380/0.525	0.403/0.623	0.442/0.756
500	0.630/0.224	0.608/0.062	0.635/0.290	0.620/0.156	0.626/0.210	0.645/0.437
Model 2 (linear models)						
80	0.159/0.843	0.147/0.783	0.163/0.861	0.157/0.857	0.159/0.858	0.168/0.867
300	0.502/0.455	0.485/0.145	0.538/0.688	0.500/0.420	0.519/0.557	0.553/0.763
500	0.758/0.044	0.748/0.002	0.778/0.284	0.759/0.065	0.763/0.106	0.787/0.505

Finally, while the theory and simulations above were focused toward cases where p grows with n , we also performed simulations for fixed p settings. Results presented in Supplementary Material E again showed the strong performance $\text{ERIC}_v(\lambda)$, whereas the three BIC-type criteria tailored made for high-dimensional settings performed worse than $\text{BIC}(\lambda)$ when the number of covariates was fixed.

6. ULTRA HIGH-DIMENSIONAL DATA

For settings where $p > n$, often referred in the literature as the ultra high-dimensional setting (Fan and Lv 2010a), adaptive Lasso weights based on the MLEs of the full model cannot be calculated. To overcome this problem, we adapt the two-stage approach of Zou (2006) and Wang and Zhu (2011), and apply SIS (Fan and Lv 2008; Fan et al. 2010b) at a first stage to reduce the dimensionality of the problem from p to $d_n = O(n^\kappa)$ where $\kappa \in (0, 1)$. Afterward, adaptive Lasso GLMs with $\text{ERIC}_v(\lambda)$ can be applied on the d_n selected covariates. This two-stage approach, which we call $\text{SIS-ERIC}_v(\lambda)$, is selection consistent. Let $\alpha_{\hat{\lambda}}$ be the model selected from $\text{SIS-ERIC}_v(\lambda)$. Then, we have the following result.

Theorem 2. Suppose the conditions in Theorem 8 of Fan et al. (2010b) hold. Under conditions (A1)–(A5), then $P(\alpha_{\hat{\lambda}} = \alpha_0) \rightarrow 1$.

The above result follows the sure independence screening property of Fan et al. (2010b), that is, the remaining d_n covariates will contain the important covariates with probability tending to one exponentially fast, and Theorem 1 in Section 4. Therefore, the proof is omitted. The result is similar to Theorem 5.1 in Zou and Zhang (2009), with two key differences. First, the result in Zou and Zhang (2009) assumed the tuning parameter can be chosen appropriately for the adaptive elastic net, whereas we focus on how to actually choose the tuning parameter via $\text{ERIC}_v(\lambda)$. Second, the theory of Zou and Zhang (2009) was developed for normal responses, whereas we consider GLMs. Theorem 2 is also similar to the result in Wang and Zhu (2011) where SIS was combined with the adaptive elastic net and the tuning parameters chosen using $\text{BIC}_{\text{wz}}(\lambda)$, although the development there was restricted to linear models.

We conducted a small simulation study comparing the performance of $\text{SIS-ERIC}_v(\lambda)$ to combining SIS with the BIC-type criteria introduced in Section 5. Two models were considered. The first was a logistic regression model with $n = 80, 300, 500$, and $p = 2000$. Covariates $\{x_i; i = 1, \dots, n\}$ were generated

from an AR1 multivariate Gaussian distribution with $\rho = 0.6$. The first 12 coefficients took alternating values of ± 3 , with the remaining coefficients set to zero. We used SIS to reduce the number of covariates from $p = 2000$ to $d_n = \lfloor 3n^{0.5} \rfloor$, and then applied the adaptive Lasso logistic regression with $\gamma = 3$ in the weights. 1000 datasets were generated for each sample size. The second model considered was a linear model with $n = 80, 300, 500$, and $p = 5000$. The covariance, coefficients, the choice of d_n were selected in the same manner as the logistic regression model immediately above. Given the high-dimensionality of the problem in both models, we assessed performance using the mean positive selection rate (PSR) and false discovery rate (FDR), as in done in Chen and Chen (2008, 2012).

In both models, $\text{SIS-ERIC}_v(\lambda)$ had slightly lower PSRs compared to combining SIS with the four BIC-type criteria, that is, it was marginally less successful at detecting true nonzeros (Table 4). On the other hand, $\text{SIS-ERIC}_v(\lambda)$ with $v = 1$ in particular has significantly smaller FDRs, that is, it was substantially better at removing true zeros from the model, meaning it tended choose smaller models compared to the BIC-type criteria. Both these result again underline the aggressive nature of $\text{ERIC}_v(\lambda)$ —it applies greater shrinkage to dramatically reduce the number of false positives, at the risk of missing some truly informative coefficients. The PSRs of all four BIC-type criteria (when used in conjunction with SIS) were similar regardless of sample size. $\text{BIC}_{\text{wll}}(\lambda)$ and $\text{BIC}_{\text{ft}}(\lambda)$, that is, the two BIC-type criteria which penalize more severely for model complexity, had lower FDRs than $\text{BIC}(\lambda)$ and $\text{BIC}_{\text{wz}}(\lambda)$.

7. DISCUSSION

Like all penalized likelihood methods, choosing an appropriate tuning parameter for adaptive Lasso regression is critical to ensuring its good performance. We have proposed a new information criterion, $\text{ERIC}_v(\lambda)$, which extends BIC to account for the effect of applying the adaptive Lasso on the bias-variance tradeoff. We showed that $\text{ERIC}_v(\lambda)$ is selection consistent when the number of covariates increases with sample size, and that this consistency holds in a wider range of cases compared to $\text{BIC}(\lambda)$. Simulations showed $\text{ERIC}_v(\lambda)$, in conjunction with a hybrid estimator, can dramatically outperform several BIC-type criteria for choosing the tuning parameter. For ultra high-dimensional data, we proposed combining $\text{ERIC}_v(\lambda)$ with the SIS procedure, leading to a two-stage approach that is selection consistent and has competitive empirical performance.

While it was proposed as extension of BIC to adaptive Lasso regression only, an obvious question is whether $\text{ERIC}_v(\lambda)$, or some appropriately modified form of it, could be applied to other penalties that are selection consistent under certain regularity conditions, for example, bridge regression. One approach might be to apply the local linear approximation (Zou and Li 2008) to “convert” them to adaptive Lasso penalties and assess the performance of $\text{ERIC}_v(\lambda)$ in such case. More generally, the principle of formulating an information criterion which accounts for the effect of prior information on the bias-variance tradeoff should, in principle, be applicable to any penalty that has a proper prior distribution, such as the Bayesian elastic net (Hans 2011). On the other hand, given penalties such as SCAD and MCP (Zhang 2010) cannot be written as a proper prior distribution since $p'_\lambda(|\beta|) = 0$ for sufficiently large $|\beta|$, then we conjecture $\text{ERIC}_v(\lambda)$ may not be straightforwardly extendable to these penalties.

While the results in Section 6 showed promise, they are relatively brief, and considerably more research needs to be conducted into the use of $\text{ERIC}_v(\lambda)$ in ultra high-dimensional data settings. As one reviewer pointed out, the two-stage approach combining SIS with a penalized regression may not seem appealing, given that penalties such as SCAD could be used directly with selection consistency guaranteed by using $\text{BIC}_{\text{fit}}(\lambda)$ (for instance) to select the tuning parameter (Fan and Tang 2012). With the adaptive Lasso, a more direct approach to handling $p > n$ settings is to construct weights based on fitting univariate GLMs to each covariate (Huang, Ma, and Zhang 2008). Given the strong performance of $\text{ERIC}_v(\lambda)$ in the simulations presented earlier, such a direct approach might also perform very well.

SUPPLEMENTARY MATERIALS

The supplementary materials contain additional proofs and models.

[Received April 2014. Revised June 2014.]

REFERENCES

- Akaike, H. (1974), “A New Look at the Statistical Model Identification,” *IEEE Transactions on Automatic Control*, 19, 716–723. [262]
- Bühlmann, P. L., and van de Geer, S. A. (2011), *Statistics for High-Dimensional Data*, Berlin: Springer. [262]
- Chen, J., and Chen, Z. (2008), “Extended Bayesian Information Criteria for Model Selection With Large Model Spaces,” *Biometrika*, 95, 759–771. [263,264,265,268]
- (2012), “Extended BIC for Small- n -Large- P Sparse GLM,” *Statistica Sinica*, 22, 555–574. [263,268]
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least Angle Regression” (with discussion), *The Annals of Statistics*, 32, 407–499. [262,265]
- Fan, J., and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” *Journal of the American Statistical Association*, 96, 1348–1360. [263]
- Fan, J., and Lv, J. (2008), “Sure Independence Screening for Ultrahigh Dimensional Feature Space,” *Journal of the Royal Statistical Society, Series B*, 70, 849–911. [263,268]
- (2010a), “A Selective Overview of Variable Selection in High Dimensional Feature Space,” *Statistica Sinica*, 20, 101–148. [268]
- Fan, J., and Peng, H. (2004), “Nonconcave Penalized Likelihood With a Diverging Number of Parameters,” *The Annals of Statistics*, 32, 928–961. [264]
- Fan, J., Song, R., et al. (2010b), “Sure Independence Screening in Generalized Linear Models With NP-Dimensionality,” *The Annals of Statistics*, 38, 3567–3604. [263,268]
- Fan, Y., and Tang, C. Y. (2012), “Tuning Parameter Selection in High Dimensional Penalized Likelihood,” *Journal of the Royal Statistical Society, Series B*, 75, 531–552. [262,266,269]
- Frank, I. E., and Friedman, J. H. (1993), “A Statistical View of Some Chemometrics Regression Tools” (with discussion), *Technometrics*, 35, 109–148. [262]
- Friedman, J., Hastie, T., and Tibshirani, R. (2010), “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *Journal of Statistical Software*, 33, 1–22. [266]
- Hans, C. (2011), “Elastic Net Regression Modeling With the Orthant Normal Prior,” *Journal of the American Statistical Association*, 106, 1383–1393. [269]
- Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), New York: Springer-Verlag. [263]
- Hoerl, A. E., and Kennard, R. W. (1970), “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics*, 12, 55–67. [262,265]
- Huang, J., Ma, S., and Zhang, C.-H. (2008), “Adaptive Lasso for Sparse High-Dimensional Regression Models,” *Statistica Sinica*, 18, 1603. [269]
- Kawano, S. (2012), “Selection of Tuning Parameters in Bridge Regression Models via Bayesian Information Criterion,” *arXiv:1203.4326*. [262]
- Kim, Y., Kwon, S., and Choi, H. (2012), “Consistent Model Selection Criteria on High Dimensions,” *The Journal of Machine Learning Research*, 13, 1037–1057. [263]
- Konishi, S., Ando, T., and Imoto, S. (2004), “Bayesian Information Criteria and Smoothing Parameter Selection in Radial Basis Function Networks,” *Biometrika*, 91, 27–43. [262]
- Luo, S., and Chen, Z. (2011), “Selection Consistency of EBIC for GLIM With Non-Canonical Links and Diverging Number of Parameters,” *arXiv:1112.2815*. [263]
- (2013), “Extended BIC for Linear Regression Models With Diverging Number of Relevant Features and High or Ultra-High Feature Spaces,” *Journal of Statistical Planning and Inference*, 143, 494–504. [263]
- McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman & Hall. [262]
- Meier, L., Van De Geer, S., and Bühlmann, P. (2008), “The Group Lasso for Logistic Regression,” *Journal of the Royal Statistical Society, Series B*, 70, 53–71. [265]
- Park, T., and Casella, G. (2008), “The Bayesian Lasso,” *Journal of the American Statistical Association*, 103, 681–686. [263]
- Schellhdorfer, J., Meier, L., Bühlmann, P., Winterthur, A., and Zürich, E. (2014), “GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using ℓ_1 -Penalization,” *Journal of Computational and Graphical Statistics*, 23, 460–477. [265]
- Schwarz, G. (1978), “Estimating the Dimension of a Model,” *The Annals of Statistics*, 6, 461–464. [262,263]
- Sun, W., Wang, J., and Fang, Y. (2013), “Consistent Selection of Tuning Parameters via Variable Selection Stability,” *The Journal of Machine Learning Research*, 14, 3419–3440. [262]
- Wang, H., Li, B., and Leng, C. (2009), “Shrinkage Tuning Parameter Selection With a Diverging Number of Parameters,” *Journal of the Royal Statistical Society, Series B*, 71, 671–683. [262,264,266]
- Wang, H., Li, R., and Tsai, C. (2007), “Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method,” *Biometrika*, 94, 553–568. [263,266]
- Wang, T., and Zhu, L. (2011), “Consistent Tuning Parameter Selection in High Dimensional Sparse Linear Regression,” *Journal of Multivariate Analysis*, 102, 1141–1151. [262,266,268]
- Zhang, C. (2010), “Nearly Unbiased Variable Selection Under Minimax Concave Penalty,” *The Annals of Statistics*, 38, 894–942. [269]
- Zhang, Y., Li, R., and Tsai, C. (2010), “Regularization Parameter Selections via Generalized Information Criterion,” *Journal of the American Statistical Association*, 105, 312–323. [263,265,266]
- Zou, H. (2006), “The Adaptive Lasso and its Oracle Properties,” *Journal of the American Statistical Association*, 101, 1418–1429. [262,264,265,268]
- Zou, H., and Li, R. (2008), “One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models,” *The Annals of Statistics*, 36, 1509–1533. [269]
- Zou, H., and Zhang, H. H. (2009), “On the Adaptive Elastic-Net With a Diverging Number of Parameters,” *The Annals of Statistics*, 37, 1733–1751. [264,268]