Biometrika (2010), **97**, 3, *pp*. 551–566
doi: 10.1093/biomet/asq033
Advance Access publication 30 June 2010

Penalized Bregman divergence for large-dimensional regression and classification

BY CHUNMING ZHANG, YUAN JIANG AND YI CHAI

Department of Statistics, University of Wisconsin–Madison, Wisconsin 53706, U.S.A.
cmzhang@stat.wisc.edu jiangy@stat.wisc.edu chaiyi@stat.wisc.edu

SUMMARY

Regularization methods are characterized by loss functions measuring data fits and penalty terms constraining model parameters. The commonly used quadratic loss is not suitable for classification with binary responses, whereas the loglikelihood function is not readily applicable to models where the exact distribution of observations is unknown or not fully specified. We introduce the penalized Bregman divergence by replacing the negative loglikelihood in the conventional penalized likelihood with Bregman divergence, which encompasses many commonly used loss functions in the regression analysis, classification procedures and machine learning literature. We investigate new statistical properties of the resulting class of estimators with the number p_n of parameters either diverging with the sample size n or even nearly comparable with n, and develop statistical inference tools. It is shown that the resulting penalized estimator, combined with appropriate penalties, achieves the same oracle property as the penalized likelihood estimator, but asymptotically does not rely on the complete specification of the underlying distribution. Furthermore, the choice of loss function in the penalized classifiers has an asymptotically relatively negligible impact on classification performance. We illustrate the proposed method for quasilikelihood regression and binary classification with simulation evaluation and real-data application.

Some key words: Consistency; Divergence minimization; Exponential family; Loss function; Optimal Bayes rule; Oracle property; Quasilikelihood.

1. INTRODUCTION

Regularization is used to obtain well-behaved solutions to overparameterized estimation problems, and is particularly appealing in high dimensions. The topic is reviewed by Bickel & Li (2006). Regularization estimates a vector parameter of interest $\beta \in \mathbb{R}^{p_n}$ by minimizing the criterion function,

$$\ell_n(\beta) = L_n(\beta) + P_{\lambda_n}(\beta) \quad (\lambda_n > 0),$$

consisting of a data fit functional L_n, which measures how well β fits the observed set of data; a penalty functional P_{λ_n}, which assesses the physical plausibility of β; and a regularization parameter λ_n, which regulates the penalty. Depending on the nature of the output variable, the term L_n quantifies the error of an estimator by different error measures. For example, the quadratic loss function has nice analytical properties and is usually used in regression analysis. However, it is not always adequate in classification problems, where the misclassification loss, deviance loss, hinge loss for the support vector machine (Vapnik, 1996) and exponential loss for boosting (Hastie et al., 2001) are more realistic and commonly used in classification procedures.

Currently, most research on regularization methods is devoted to variants of penalty methods in conjunction with linear models and likelihood-based models in regression analysis. For linear model estimation with a fixed number p of parameters, Tibshirani (1996) introduced the L_1-penalty for the proposed lasso method, where the quadratic loss is in use. Theoretical properties related to the lasso have been intensively studied; see Knight & Fu (2000), Meinshausen & Buhlmann (2006) and Zhao & Yu (2006). Zou (2006) mentioned that the lasso is in general not variable selection consistent, but the adaptive lasso via combining appropriately weighted L_1-penalties is consistent. Huang et al. (2008) extended the results in Zou (2006) to high-dimensional linear models. Using the smoothly clipped absolute deviation penalty, Fan & Li (2001) showed that the penalized likelihood estimator achieved the oracle property: the resulting estimator is asymptotically as efficient as the oracle estimator. In their treatment, the number p_n of model parameters is fixed at p, and the loss function equals the negative loglikelihood. Fan & Peng (2004) extended the result to p_n diverging with n at a certain rate.

On the loss side, the literature on penalization methods includes much less discussion of either the role of the loss function in regularization for models other than linear or likelihood-based models, or the impact of different loss functions on classification performance. The least angle regression algorithm (Efron et al., 2004) for L_1-penalization was developed for linear models using the quadratic loss. Rosset & Zhu (2007) studied the piecewise linear regularized solution paths for differentiable and piecewise quadratic loss functions with L_1 penalty. It remains desirable to explore whether penalization methods using other types of loss functions can potentially benefit from the efficient least-angle regression algorithm. Moreover, theoretical results on the penalized likelihood are not readily translated into results for approaches, such as quasilikelihood (Wedderburn, 1974; McCullagh, 1983; Strimmer, 2003), where the distribution of the observations is unknown or not fully specified. Accordingly, a discussion of statistical inference for penalized estimation using a wider range of loss functions is needed.

In this study, we broaden the scope of penalization by incorporating loss functions belonging to the Bregman divergence class which unifies many commonly used loss functions. In particular, the quasilikelihood function and all loss functions mentioned previously in classification fall into this class. We introduce the penalized Bregman divergence by replacing the quadratic loss or the negative loglikelihood in penalized least-squares or penalized likelihood with Bregman divergence, and call the resulting estimator a penalized Bregman divergence estimator. Nonetheless, the Bregman divergence in general does not fulfill assumptions specifically imposed on the likelihood function associated with penalized likelihood.

We investigate new statistical properties of large-dimensional penalized Bregman divergence estimators, with dimensions dealt with separately in two cases:

$$\text{Case I}: \quad p_n \text{ is diverging with } n; \tag{1}$$

$$\text{Case II}: \quad p_n \text{ is nearly comparable with } n. \tag{2}$$

Zhang & Zhang (2010) give an application of the penalization method developed in this paper to estimating the hemodynamic response function for brain fMRI data where p_n is as large as n. The current paper shows that the penalized Bregman divergence estimator, combined with appropriate penalties, achieves the same oracle property as the penalized likelihood estimator, but the asymptotic distribution does not rely on the complete specification of the underlying distribution. From the classification viewpoint, our study elucidates the applicability and consistency of various classifiers induced by penalized Bregman divergence estimators. Technical details of this paper are in the online Supplementary Material.

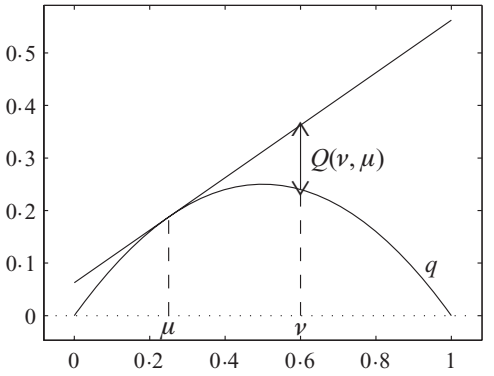

Fig. 1. Illustration of $Q(\nu, \mu)$ as defined in (3). The concave curve is q; the two dashed lines indicate locations of ν and μ; the solid straight line is $q(\mu) + (\nu - \mu)q'(\mu)$; the length of the vertical line with arrows at each end is $Q(\nu, \mu)$.

2. The penalized Bregman divergence estimator

2·1. *Bregman divergence*

We give a brief overview of Bregman divergence. For a given concave function q with derivative q', Bregman (1967) introduced a device for constructing a bivariate function,

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu). \tag{3}$$

Figure 1 displays Q and the corresponding q. It is readily seen that the concavity of q ensures the nonnegativity of Q. Moreover, for a strictly concave q, $Q(\nu, \mu) = 0$ is equivalent to $\nu = \mu$. However, since $Q(\nu, \mu)$ is not generally symmetric in ν and μ, Q is not a metric or distance in the strict sense. Hence, we call Q the Bregman divergence and call q the generating function of Q. See Efron (1986), Lafferty et al. (1997), Lafferty (1999), Kivinen & Warmuth (1999), Grünwald & Dawid (2004), Altun & Smola (2006) and references therein.

The Bregman divergence is suitable for a broad array of error measures Q. For example, $q(\mu) = a\mu - \mu^2$ with some constant a yields the quadratic loss $Q(Y, \mu) = (Y - \mu)^2$. For a binary response variable Y, $q(\mu) = \min\{\mu, (1 - \mu)\}$ gives the misclassification loss $Q(Y, \mu) = I\{Y \neq I(\mu > 1/2)\}$, where $I(\cdot)$ denotes the indicator function; $q(\mu) = -\{\mu \log(\mu) + (1 - \mu)\log(1 - \mu)\}$ gives the Bernoulli deviance loss $Q(Y, \mu) = -\{Y \log(\mu) + (1 - Y)\log(1 - \mu)\}$; $q(\mu) = 2\min\{\mu, (1 - \mu)\}$ results in the hinge loss; and $q(\mu) = 2\{\mu(1 - \mu)\}^{1/2}$ yields the exponential loss $Q(Y, \mu) = \exp[-(Y - 0{\cdot}5)\log\{\mu/(1 - \mu)\}]$.

Conversely, for a given Q, Zhang et al. (2009) provided necessary and sufficient conditions for Q being a Bregman divergence, and in that case derived an explicit formula for q. Applying this inverse approach from Q to q, they illustrated that the quasilikelihood function, the Kullback–Leibler divergence or the deviance loss for the exponential family of probability functions, and many margin-based loss functions (Shen et al., 2003) are Bregman divergences. To our knowledge, there is little theoretical work in the literature on thoroughly examining the penalized Bregman divergence, via methods of regularization, for large-dimensional model building, variable selection and classification problems.

2·2. *The model and penalized Bregman divergence estimator*

Let (X, Y) denote a random realization from some underlying population, where $X = (X_1, \ldots, X_{p_n})^{\mathrm{T}}$ is the input vector and Y is the output variable. The dimension p_n follows

the assumption in (1) or (2). We assume the parametric model,

$$m(x) = E(Y \mid X = x) = F^{-1}(b_{0;0} + x^{\mathrm{T}}\beta_0), \tag{4}$$

where F is a known link function, $b_{0;0} \in \mathbb{R}^1$ and $\beta_0 = (\beta_{1;0}, \ldots, \beta_{p_n;0})^{\mathrm{T}} \in \mathbb{R}^{p_n}$ are the unknown true parameters. Throughout the paper, it is assumed that some entries in β_0 are exactly zero. Write $\beta_0 = \{\beta_0^{(\mathrm{I})\mathrm{T}}, \beta_0^{(\mathrm{II})\mathrm{T}}\}^{\mathrm{T}}$, where $\beta_0^{(\mathrm{I})}$ collects all nonzero coefficients, and $\beta_0^{(\mathrm{II})} = 0$.

Our goal is to estimate the true parameters via penalization. Let $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a sample of independent random pairs from (X, Y), where $X_i = (X_{i1}, \ldots, X_{ip_n})^{\mathrm{T}}$. The penalized Bregman divergence estimator $(\hat{b}_0, \hat{\beta})$ is defined as the minimizer of the criterion function,

$$\ell_n(b_0, \beta) = \frac{1}{n}\sum_{i=1}^{n} Q\{Y_i, F^{-1}(b_0 + X_i^{\mathrm{T}}\beta)\} + \sum_{j=1}^{p_n} P_{\lambda_n}(|\beta_j|), \tag{5}$$

where $\beta = (\beta_1, \ldots, \beta_{p_n})^{\mathrm{T}}$, the loss function $Q(\cdot, \cdot)$ is a Bregman divergence, and $P_{\lambda_n}(\cdot)$ represents a nonnegative penalty function indexed by a tuning constant $\lambda_n > 0$. Set $\widetilde{\beta} = (b_0, \beta^{\mathrm{T}})^{\mathrm{T}}$, and correspondingly $\widetilde{X}_i = (1, X_i^{\mathrm{T}})^{\mathrm{T}}$. Then (5) can be written as

$$\ell_n(\widetilde{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Q\{Y_i, F^{-1}(\widetilde{X}_i^{\mathrm{T}}\widetilde{\beta})\} + \sum_{j=1}^{p_n} P_{\lambda_n}(|\beta_j|). \tag{6}$$

The penalized Bregman divergence estimator is $\widetilde{\beta}_{\mathrm{E}} = (\hat{b}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p_n})^{\mathrm{T}} = \arg\min_{\widetilde{\beta}} \ell_n(\widetilde{\beta})$.

Regarding the uniqueness of $\widetilde{\beta}_{\mathrm{E}}$, assume that the quantities

$$q_j(y; \theta) = (\partial^j/\partial\theta^j) Q\{y, F^{-1}(\theta)\} \quad (j = 0, 1, \ldots), \tag{7}$$

exist finitely up to any order required. Provided that for all $\theta \in \mathbb{R}$ and all y in the range of Y,

$$q_2(y; \theta) > 0, \tag{8}$$

it follows that $L_n(\widetilde{\beta}) = n^{-1}\sum_{i=1}^{n} Q\{Y_i, F^{-1}(\widetilde{X}_i^{\mathrm{T}}\widetilde{\beta})\}$ in (6) is convex in $\widetilde{\beta}$. In that case, if convex penalties are used in (6), then $\ell_n(\widetilde{\beta})$ is necessarily convex in $\widetilde{\beta}$, and hence the local minimizer $\widetilde{\beta}_{\mathrm{E}}$ is the unique global penalized Bregman divergence estimator. For nonconvex penalties, however, the local minimizer may not be globally unique.

3. PENALIZED BREGMAN DIVERGENCE WITH NONCONVEX PENALTIES: $p_n \ll n$

3·1. *Consistency*

We start by introducing some notation. Let s_n denote the number of nonzero coordinates of β_0, and set $\widetilde{\beta}_0 = (b_{0;0}, \beta_0^{\mathrm{T}})^{\mathrm{T}}$. Define

$$a_n = \max_{j=1,\ldots,s_n} |P'_{\lambda_n}(|\beta_{j;0}|)|, \quad b_n = \max_{j=1,\ldots,s_n} |P''_{\lambda_n}(|\beta_{j;0}|)|,$$

where $P_\lambda^{(j)}(|\beta|)$ is shorthand for $(d^j/dx^j)P_\lambda(x)|_{x=|\beta|}$, $j = 1, 2$. Unless otherwise stated, $\|\cdot\|$ denotes the L_2-norm. Theorem 1 guarantees the existence of a consistent local minimizer for (6), and states that the local penalized Bregman divergence estimator $\widetilde{\beta}_{\mathrm{E}}$ is $(n/p_n)^{1/2}$-consistent.

THEOREM 1 (Existence and consistency). *Assume Condition* A *in the Appendix,* $a_n = O(1/n^{1/2})$ *and* $b_n = o(1)$. *If* $p_n^4/n \to 0$, $(p_n/n)^{1/2}/\lambda_n \to 0$ *and* $\min_{j=1,\ldots,s_n} |\beta_{j;0}|/\lambda_n \to \infty$ *as* $n \to \infty$, *then there exists a local minimizer* $\widetilde{\beta}_{\mathrm{E}}$ *of* (6) *such that* $\|\widetilde{\beta}_{\mathrm{E}} - \widetilde{\beta}_0\| = O_P\{(p_n/n)^{1/2}\}$.

3·2. *Oracle property*

Following Theorem 1, the oracle property of the local minimizer is given in Theorem 2 below. Before stating it, we need some notation. Write $X = (X^{(\mathrm{I})\mathrm{T}}, X^{(\mathrm{II})\mathrm{T}})^{\mathrm{T}}$, $\widetilde{X}^{(\mathrm{I})} = (1, X^{(\mathrm{I})\mathrm{T}})^{\mathrm{T}}$, and $\widetilde{\beta}^{(\mathrm{I})} = (b_0, \beta^{(\mathrm{I})\mathrm{T}})^{\mathrm{T}}$. For the penalty term, let

$$d_n = \{0, P'_{\lambda_n}(|\beta_{1;0}|)\mathrm{sign}(\beta_{1;0}), \ldots, P'_{\lambda_n}(|\beta_{s_n;0}|)\mathrm{sign}(\beta_{s_n;0})\}^{\mathrm{T}},$$

$$\Sigma_n = \mathrm{diag}\{0, P''_{\lambda_n}(|\beta_{1;0}|), \ldots, P''_{\lambda_n}(|\beta_{s_n;0}|)\}.$$

For the q function, define $F_n = q^{(2)}\{m(X)\}/[F^{(1)}\{m(X)\}]^2 \widetilde{X}^{(\mathrm{I})}\widetilde{X}^{(\mathrm{I})\mathrm{T}}$ and

$$\Omega_n = E\big[\mathrm{var}(Y \mid X) q^{(2)}\{m(X)\}F_n\big], \quad H_n = -E(F_n).$$

THEOREM 2 (Oracle property). *Assume Condition* B *in the Appendix.*

(i) *If* $p_n^2/n = O(1)$, $(p_n/n^{1/2})/\lambda_n \to 0$ *and* $\liminf_{n\to\infty} \liminf_{\theta\to 0+} P'_{\lambda_n}(\theta)/\lambda_n > 0$ *as* $n \to \infty$, *then any* $(n/p_n)^{1/2}$*-consistent local minimizer* $\widetilde{\beta}_{\mathrm{E}} = (\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})\mathrm{T}}, \hat{\beta}^{(\mathrm{II})\mathrm{T}})^{\mathrm{T}}$ *satisfies* $\mathrm{pr}(\hat{\beta}^{(\mathrm{II})} = 0) \to 1$.

(ii) *Moreover, if* $a_n = O(1/n^{1/2})$, $p_n^5/n \to 0$ *and* $\min_{j=1,\ldots,s_n} |\beta_{j;0}|/\lambda_n \to \infty$, *then for any fixed integer* k *and any* $k \times (s_n + 1)$ *matrix* A_n *such that* $A_n A_n^{\mathrm{T}} \to G$ *with* G *being a* $k \times k$ *nonnegative-definite symmetric matrix,* $n^{1/2} A_n \Omega_n^{-1/2}\{(H_n + \Sigma_n)(\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})} - \widetilde{\beta}_0^{(\mathrm{I})}) + d_n\} \to N(0, G)$ *in distribution.*

Theorem 2 has some useful consequences: First, the p_n-dimensional penalized Bregman divergence estimator, combined with appropriate penalties, achieves the same oracle property as the penalized likelihood estimator of Fan & Peng (2004): the estimators of the zero parameters take exactly zero values with probability tending to 1, and the estimators of the nonzero parameters are asymptotically normal with the same means and variances as if the zero coefficients were known in advance. Second, the asymptotic distribution of the penalized Bregman divergence estimator relies on the underlying distribution of $Y \mid X$ through $E(Y \mid X)$ and $\mathrm{var}(Y \mid X)$, but does not require a complete specification of the underlying distribution. Third, the asymptotic distribution depends on the choice of the Q-loss only through the second derivative of its generating q function. This enables us to evaluate the impact of loss functions on the penalized Bregman divergence estimators and to derive an optimal loss function in certain situations.

According to Theorem 2, the asymptotic covariance matrix of $\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})}$ is $V_n = (H_n + \Sigma_n)^{-1}\Omega_n(H_n + \Sigma_n)^{-1}$. In practice, V_n is unknown and needs to be estimated. Typically, the sandwich formula can be exploited to form an estimator of V_n by

$$\hat{V}_n = (\hat{H}_n + \hat{\Sigma}_n)^{-1}\hat{\Omega}_n(\hat{H}_n + \hat{\Sigma}_n)^{-1}, \tag{9}$$

where $\hat{\Omega}_n = n^{-1}\sum_{i=1}^n q_1^2(Y_i; \widetilde{X}_i^{(\mathrm{I})\mathrm{T}}\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})})\widetilde{X}_i^{(\mathrm{I})}\widetilde{X}_i^{(\mathrm{I})\mathrm{T}}$, $\hat{H}_n = n^{-1}\sum_{i=1}^n q_2(Y_i; \widetilde{X}_i^{(\mathrm{I})\mathrm{T}}\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})})\widetilde{X}_i^{(\mathrm{I})}\widetilde{X}_i^{(\mathrm{I})\mathrm{T}}$ and $\hat{\Sigma}_n = \mathrm{diag}\{0, P''_{\lambda_n}(|\hat{\beta}_1|), \ldots, P''_{\lambda_n}(|\hat{\beta}_{s_n}|)\}$.

Proposition 1 below demonstrates that for any $(n/p_n)^{1/2}$-consistent estimator $\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})}$ of $\widetilde{\beta}_0^{(\mathrm{I})}$, $\hat{V}_n$ is a consistent estimator for the covariance matrix V_n, in the sense that $A_n(\hat{V}_n - V_n)A_n^{\mathrm{T}} \to 0$ in probability for any $k \times (s_n + 1)$ matrix A_n satisfying $A_n A_n^{\mathrm{T}} \to G$, where k is any fixed integer.

PROPOSITION 1 (Covariance matrix estimation). *Assume Condition* B *in the Appendix, and* $b_n = o(1)$. *If* $p_n^4/n \to 0$, $(p_n/n)^{1/2}/\lambda_n \to 0$ *and* $\min_{j=1,\ldots,s_n} |\beta_{j;0}|/\lambda_n \to \infty$ *as* $n \to \infty$, *then for any* $\|\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})} - \widetilde{\beta}_0^{(\mathrm{I})}\| = O_P\{(p_n/n)^{1/2}\}$, *we have that* $A_n(\hat{V}_n - V_n)A_n^{\mathrm{T}} \to 0$ *in probability for any* $k \times (s_n + 1)$ *matrix* A_n *satisfying* $A_n A_n^{\mathrm{T}} \to G$, *where* G *is a* $k \times k$ *matrix.*

Is there an optimal choice of q such that the corresponding V_n matrix achieves its lower bound? We have that $V_n = H_n^{-1}\Omega_n H_n^{-1}$ in two special cases. One is $\Sigma_n = 0$ for large n and large $\min_{j=1,\ldots,s_n}|\beta_{j;0}|$, which results from the smoothly clipped absolute deviation and hard thresholding penalties; another one is $\Sigma_n = 0$ for all n, which results from the weighted L_1-penalties in Theorem 6 below. In these cases, it can be shown via matrix algebra that the optimal q satisfies the generalized Bartlett identity in (11) below. On the other hand, for an arbitrary $\Sigma_n \neq 0$, the complication rises; the optimal q is generally not available in closed-form.

3·3. *Hypothesis testing*

We consider hypothesis testing about $\widetilde{\beta}_0^{(\mathrm{I})}$ formulated as

$$H_0 : A_n\widetilde{\beta}_0^{(\mathrm{I})} = 0 \quad \text{versus} \quad H_1 : A_n\widetilde{\beta}_0^{(\mathrm{I})} \neq 0, \tag{10}$$

where A_n is a given $k \times (s_n + 1)$ matrix such that $A_n A_n^{\mathrm{T}} = G$ with G being a $k \times k$ positive-definite matrix. This form of linear hypothesis allows one to test simultaneously whether a subset of variables used are statistically significant by taking some specific form of the matrix A_n; for example, $A_n = [I_k, 0_{k,s_n+1-k}]$ yields $A_n A_n^{\mathrm{T}} = I_k$.

We propose a generalized Wald-type test statistic of the form

$$W_n = n\big(A_n\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})}\big)^{\mathrm{T}}\big(A_n\hat{H}_n^{-1}\hat{\Omega}_n\hat{H}_n^{-1}A_n^{\mathrm{T}}\big)^{-1}\big(A_n\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})}\big),$$

where $\hat{\Omega}_n$ and $\hat{H}_n$ are as defined in (9). This test is asymptotically distribution-free, as Theorem 3 justifies that, under the null, W_n would for large n be distributed as χ_k^2.

THEOREM 3 (Wald-type test under H_0). *Assume Condition* C *in the Appendix, and let* $a_n = o\{1/(ns_n)^{1/2}\}$ *and* $b_n = o(1/p_n^{1/2})$. *If* $p_n^5/n \to 0$, $(p_n/n)^{1/2}/\lambda_n \to 0$ *and* $\min_{j=1,\ldots,s_n}|\beta_{j;0}|/\lambda_n \to \infty$ *as* $n \to \infty$, *then under* H_0 *in* (10), $W_n \to \chi_k^2$ *in distribution.*

Remark 1. To appreciate the discriminating power of W_n in assessing the significance, the asymptotic power can be analyzed. It can be shown that under H_1 in (10) where $\|A_n\widetilde{\beta}_0\|$ is independent of n, $W_n \to +\infty$ in probability at the rate n. Hence W_n has power function tending to 1 against fixed alternatives. Besides, W_n has a nontrivial local power detecting contiguous alternatives approaching the null at the rate $n^{-1/2}$. We omit the lengthy details.

In the context of penalized likelihood estimator $\widetilde{\beta}_{\mathrm{E}}$, Fan & Peng (2004) showed that the likelihood-ratio-type test statistic

$$\Lambda_n = 2n\left\{\min_{\widetilde{\beta}\in\mathbb{R}^{p_n+1}:A_n\widetilde{\beta}^{(\mathrm{I})}=0}\ell_n(\widetilde{\beta}) - \ell_n(\widetilde{\beta}_{\mathrm{E}})\right\}$$

follows an asymptotic χ^2 distribution under the null hypothesis. Theorem 4 below explores the extent to which this result can feasibly be extended to Λ_n constructed from the broad class of penalized Bregman divergence estimators.

THEOREM 4 (Likelihood-ratio-type test under H_0). *Assume* (8) *and Condition* D *in the Appendix,* $a_n = o\{1/(ns_n)^{1/2}\}$ *and* $b_n = o(1/p_n^{1/2})$. *If* $p_n^5/n \to 0$, $(p_n/n)^{1/2}/\lambda_n \to 0$ *and* $\min_{j=1,\ldots,s_n}|\beta_{j;0}|/\lambda_n \to \infty$ *as* $n \to \infty$, *then under* H_0 *in* (10), *provided that* q *satisfies the generalized Bartlett identity,*

$$q^{(2)}\{m(\cdot)\} = -\frac{c}{\mathrm{var}(Y \mid X = \cdot)}, \tag{11}$$

for a constant $c > 0$, *we have that* $\Lambda_n/c \to \chi_k^2$ *in distribution.*

Curiously, the result in Theorem 4 indicates that in general, condition (11) on q restricts the application domain of the test statistic Λ_n. For instance, in the case of binary responses, the Bernoulli deviance loss satisfies (11), but the quadratic loss and exponential loss violate (11). This limitation reflects that the likelihood-ratio-type test statistic Λ_n may not be straightforwardly valid for the penalized Bregman divergence estimators.

Remark 2. For a Bregman divergence Q, condition (11) with $c = 1$ is equivalent to the equality $E[\partial^2 Q\{Y, m(\cdot)\}/\partial m(\cdot)^2 \mid X = \cdot] = E([\partial Q\{Y, m(\cdot)\}/\partial m(\cdot)]^2 \mid X = \cdot)$, which includes the Bartlett identity (Bartlett, 1953) as a special case, when Q is the negative loglikelihood. Thus, we call (11) the generalized Bartlett identity. It is also seen that the quadratic loss satisfies (11) for homoscedastic regression models even without knowing the error distribution.

4. Penalized Bregman divergence with convex penalties: $p_n \approx n$

4·1. *Consistency, oracle property and hypothesis testing*

For the nonconvex penalties discussed in § 3, the condition $p_n^4/n \to 0$ or $p_n^5/n \to 0$ can be relaxed to $p_n^3/n \to 0$ in the particular situation where the Bregman divergence is a quadratic loss and the link is an identity link. It remains unclear whether p_n can be relaxed in other cases.

This section aims to improve the rate of consistency of the penalized Bregman divergence estimators and to relax conditions on p_n using certain convex penalties, the weighted L_1-penalties, under which the penalized Bregman divergence estimator $\widetilde{\beta}_{\mathrm{E}} = (\hat{b}_0, \hat{\beta}^{\mathrm{T}})^{\mathrm{T}}$ is defined to minimize the criterion function,

$$\ell_n(\widetilde{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Q\{Y_i, F^{-1}(\widetilde{X}_i^{\mathrm{T}}\widetilde{\beta})\} + \lambda_n \sum_{j=1}^{p_n} w_j|\beta_j|, \tag{12}$$

with $w_1, \ldots, w_{p_n}$ representing nonnegative weights. Define

$$w_{\max}^{(\mathrm{I})} = \max_{j=1,\ldots,s_n} w_j, \quad w_{\min}^{(\mathrm{II})} = \min_{s_n+1 \leqslant j \leqslant p_n} w_j.$$

Lemma 1 obtains the existence of a $(n/p_n)^{1/2}$-consistent local minimizer of (12). This rate is identical to that in Theorem 1 but, unlike Theorem 1, Lemma 1 includes the L_1-penalty. Other results parallel to those in § 3 can similarly be obtained.

Lemma 1 (Existence and consistency). *Assume Conditions* A1–A7 *in the Appendix and* $w_{\max}^{(\mathrm{I})} = O_P\{1/(\lambda_n n^{1/2})\}$. *If* $p_n^4/n \to 0$ *as* $n \to \infty$, *then there exists a local minimizer* $\widetilde{\beta}_{\mathrm{E}}$ *of* (12) *such that* $\|\widetilde{\beta}_{\mathrm{E}} - \widetilde{\beta}_0\| = O_P\{(p_n/n)^{1/2}\}$.

Lemma 1 imposes a condition on the weights of nonzero coefficients alone, but ignores the weights on zero coefficients. Theorem 5 below reflects that incorporating appropriate weights to the zero coefficients can improve the rate of consistency from $(p_n/n)^{1/2}$ to $(s_n/n)^{1/2}$.

Theorem 5 (Existence and consistency). *Assume Conditions* A1–A7 *in the Appendix,* $w_{\max}^{(\mathrm{I})} = O_P\{1/(\lambda_n n^{1/2})\}$ *and there exists a constant* $M \in (0, \infty)$ *such that* $\lim_{n\to\infty} \mathrm{pr}(w_{\min}^{(\mathrm{II})}\lambda_n > M) = 1$. *If* $s_n^4/n \to 0$ *and* $s_n(p_n - s_n) = o(n)$, *then there exists a local minimizer* $\widetilde{\beta}_{\mathrm{E}}$ *of* (12) *such that* $\|\widetilde{\beta}_{\mathrm{E}} - \widetilde{\beta}_0\| = O_P\{(s_n/n)^{1/2}\}$.

More importantly, conditions on the dimension p_n are much relaxed. For example, Theorem 5 allows $p_n = o(n^{(3+\delta)/(4+\delta)})$ for any $\delta > 0$, provided $s_n = O(n^{1/(4+\delta)})$, whereas Theorem 1 requires $p_n = o(n^{1/4})$ for any $s_n \leqslant p_n$. This implies that p_n can indeed be relaxed to the case (2) of being

nearly comparable with n. On the other hand, the proof of Theorem 5 relies on the flexibility of the weights $\{w_j\}$, as seen in an $I_{2,1}^{(\mathrm{II})}$ term. Thus, directly carrying the proof of Theorem 5 through to either the nonconvex penalties in Theorem 1 or the L_1-penalty is not feasible.

Theorem 6 gives an oracle property for the $(n/s_n)^{1/2}$-consistent local minimizer.

THEOREM 6 (Oracle property). *Assume Conditions* A1, A2, B3, A4, B5, A6–A7 *in the Appendix.*

(i) *If* $s_n^2/n = O(1)$ *and* $w_{\min}^{(\mathrm{II})}\lambda_n n^{1/2}/(s_n p_n)^{1/2} \to \infty$ *in probability as* $n \to \infty$, *then any* $(n/s_n)^{1/2}$*-consistent local minimizer* $\widetilde{\beta}_{\mathrm{E}} = (\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})\mathrm{T}}, \hat{\beta}^{(\mathrm{II})\mathrm{T}})^{\mathrm{T}}$ *satisfies* $\mathrm{pr}(\hat{\beta}^{(\mathrm{II})} = 0) \to 1$.

(ii) *Moreover, if* $w_{\max}^{(\mathrm{I})} = O_P\{1/(\lambda_n n^{1/2})\}$, $s_n^5/n \to 0$ *and* $\min_{j=1,\ldots,s_n} |\beta_{j;0}|/(s_n/n)^{1/2} \to \infty$, *then for any fixed integer* k *and any* $k \times (s_n + 1)$ *matrix* A_n *such that* $A_n A_n^{\mathrm{T}} \to G$ *with* G *being a* $k \times k$ *nonnegative-definite symmetric matrix,* $n^{1/2} A_n \Omega_n^{-1/2}\{H_n(\widetilde{\beta}_{\mathrm{E}}^{(\mathrm{I})} - \widetilde{\beta}_0^{(\mathrm{I})}) + \lambda_n \mathrm{W}_n \mathrm{sign}(\widetilde{\beta}_0^{(\mathrm{I})})\} \to N(0, G)$ *in distribution, where* $\mathrm{W}_n = \mathrm{diag}(0, w_1, \ldots, w_{s_n})$ *and* $\mathrm{sign}\{\widetilde{\beta}_0^{(\mathrm{I})}\} = \{\mathrm{sign}(b_{0;0}), \ldots, \mathrm{sign}(\beta_{s_n;0})\}^{\mathrm{T}}$.

For testing hypotheses of the form (10), the generalized Wald-type test statistic W_n proposed in § 3·3 continues to be applicable. Theorem 7 derives the asymptotic distribution of W_n.

THEOREM 7 (Wald-type test under H_0). *Assume Conditions* A1, A2, B3, C4, B5, A6–A7 *in the Appendix, and that* $w_{\max}^{(\mathrm{I})} = o_P[1/\{\lambda_n (n s_n)^{1/2}\}]$. *If* $s_n^5/n \to 0$ *and* $\min_{j=1,\ldots,s_n} |\beta_{j;0}|/(s_n/n)^{1/2} \to \infty$ *as* $n \to \infty$, *then under* H_0 *in* (10), $W_n \to \chi_k^2$ *in distribution.*

4·2. *Weight selection*

We propose a penalized componentwise regression method for selecting weights by

$$\hat{w}_j = |\hat{\beta}_j^{\mathrm{PCR}}|^{-1} \quad (j = 1, \ldots, p_n), \tag{13}$$

based on some initial estimator, $\hat{\beta}^{\mathrm{PCR}} = (\hat{\beta}_1^{\mathrm{PCR}}, \ldots, \hat{\beta}_{p_n}^{\mathrm{PCR}})^{\mathrm{T}}$, minimizing

$$\ell_n^{\mathrm{PCR}}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p_n} Q\{Y_i, F^{-1}(X_{ij}\beta_j)\} + \kappa_n \sum_{j=1}^{p_n} |\beta_j|, \tag{14}$$

with some sequence $\kappa_n > 0$. Theorem 8 indicates that under assumptions on the correlation between the predictor variables and the response variable, the weights selected by the penalized componentwise regression satisfy the conditions in Theorem 5.

THEOREM 8 (Penalized componentwise regression for weights: $p_n \approx n$). *Assume Conditions* A1, A2, B3, A4, A6, A7 *and* E. *Assume that in Condition* E, $\mathcal{A}_n = \lambda_n n^{1/2}$, $\mathcal{A}_n/\kappa_n \to \infty$ *and* $\mathcal{B}_n/\kappa_n = O(1)$ *for* κ_n *in* (14). *Suppose* $\lambda_n n^{1/2} = O(1)$, $\lambda_n = o(\kappa_n)$ *and* $\log(p_n) = o(n\kappa_n^2)$. *Assume that* $E(X) = 0$ *in model* (4). *Then there exist local minimizers* $\hat{\beta}_j^{\mathrm{PCR}}$ $(j = 1, \ldots, p_n)$, *of* (14) *such that the weights* $\hat{w}_j$ $(j = 1, \ldots, p_n)$, *defined in* (13) *satisfy that* $\hat{w}_{\max}^{(\mathrm{I})} = O_P\{1/(\lambda_n \sqrt{n})\}$ *and* $\hat{w}_{\min}^{(\mathrm{II})}\lambda_n \to \infty$ *in probability in Theorem 5, where* $\hat{w}_{\max}^{(\mathrm{I})} = \max_{j=1,\ldots,s_n} \hat{w}_j$ *and* $\hat{w}_{\min}^{(\mathrm{II})} = \min_{s_n+1 \leqslant j \leqslant p_n} \hat{w}_j$.

5. CONSISTENCY OF THE PENALIZED BREGMAN DIVERGENCE CLASSIFIER

This section deals with the binary response variable Y, which takes values 0 and 1. In this case, the mean regression function $m(x)$ in (4) becomes the class label probability, $\mathrm{pr}(Y = 1 \mid X = x)$.

From the penalized Bregman divergence estimator $(\hat{b}_0, \hat{\beta}^{\mathrm{T}})^{\mathrm{T}}$ proposed in either § 3 or § 4, we can construct the penalized Bregman divergence classifier, $\hat{\phi}(x) = I\{\hat{m}(x) > 1/2\}$, for a future input variable x, where $\hat{m}(x) = F^{-1}(\hat{b}_0 + x^{\mathrm{T}}\hat{\beta})$.

In the classification literature, the misclassification loss of a classification rule ϕ at a sample point (x, y) is $l\{y, \phi(x)\} = I\{y \neq \phi(x)\}$. The risk of ϕ is the expected misclassification loss, $R(\phi) = E[l\{Y, \phi(X)\}] = \mathrm{pr}\{\phi(X) \neq Y\}$. The optimal Bayes rule, which minimizes the risk with respect to ϕ, is $\phi_{\mathrm{B}}(x) = I\{m(x) > 1/2\}$. For a test sample (X^o, Y^o), which is an independent and identically distributed copy of samples in the training set $\mathcal{T}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$, the optimal Bayes risk is then $R(\phi_{\mathrm{B}}) = \mathrm{pr}\{\phi_{\mathrm{B}}(X^o) \neq Y^o\}$. Meanwhile, the conditional risk of the penalized Bregman divergence classification rule $\hat{\phi}$ is $R(\hat{\phi}) = \mathrm{pr}\{\hat{\phi}(X^o) \neq Y^o \mid \mathcal{T}_n\}$. For $\hat{\phi}$ induced by the penalized Bregman divergence regression estimation using a range of loss functions combined with either the smoothly clipped absolute deviation, L_1 or weighted L_1-penalties, Theorem 9 verifies the classification consistency attained by $\hat{\phi}$.

THEOREM 9 (Consistency of the penalized Bregman divergence classifier). *Assume Conditions* A1 *and* A4 *in the Appendix. Suppose that* $\|\widetilde{\beta}_{\mathrm{E}} - \widetilde{\beta}_0\| = O_P(r_n)$. *If* $r_n p_n^{1/2} = o(1)$, *then the classification rule* $\hat{\phi}$ *constructed from* $\widetilde{\beta}_{\mathrm{E}}$ *is consistent in the sense that* $E\{R(\hat{\phi})\} - R(\phi_{\mathrm{B}}) \to 0$ *as* $n \to \infty$.

6. SIMULATION STUDY

6·1. *Set-up*

For illustrative purposes, four procedures for penalized estimators are compared: (I) the smoothly clipped absolute deviation penalty, with an accompanying parameter $a = 3{\cdot}7$, combined with the local linear approximation; (II) the L_1 penalty; (III) the weighted L_1-penalties with weights selected by (13); and (IV) the oracle estimator using the set of significant variables. Throughout the numerical work in the paper, methods (I)–(III) utilize the least angle regression algorithm, F is the log link for count data and the logit link for binary response variables.

6·2. *Penalized quasilikelihood for overdispersed count data*

A quasilikelihood function Q relaxes the distributional assumption on a random variable Y via the specification $\partial Q(Y, \mu)/\partial\mu = (Y - \mu)/V(\mu)$, where $\mathrm{var}(Y \mid X = x) = V\{m(x)\}$ for a known continuous function $V(\cdot) > 0$. Zhang et al. (2009) verified that the quasilikelihood function belongs to the Bregman divergence and derived the generating q function,

$$q(\mu) = \int_{-\infty}^{\mu} \frac{s - \mu}{V(s)} ds. \tag{15}$$

We generate overdispersed Poisson counts Y_i satisfying $\mathrm{var}(Y_i \mid X_i = x_i) = 2m(x_i)$. In the predictor $X_i = (X_{i1}, \ldots, X_{ip_n})^{\mathrm{T}}$, $p_n = n/8$, $n/2$ and $n - 10$, and $X_{i1} = i/n - 0{\cdot}5$. For $j = 2, \ldots, p_n$, $X_{ij} = \Phi(Z_{ij}) - 0{\cdot}5$, where Φ is the standard normal distribution function, and $(Z_{i2}, \ldots, Z_{ip_n})^{\mathrm{T}} \sim N\{0, \rho 1_{p_n-1} 1_{p_n-1}^{\mathrm{T}} + (1 - \rho)\mathrm{I}_{p_n-1}\}$, with 1_d a $d \times 1$ vector of ones and I_d a $d \times d$ identity matrix. Thus $(X_{i2}, \ldots, X_{ip_n})$ are correlated $\mathrm{Un}(0, 1)$ if $\rho \neq 0$. The link function is $\log\{m(x)\} = b_{0;0} + x^{\mathrm{T}}\beta_0$, where $b_{0;0} = 5$ and $\beta_0 = (2, 2, 0, 0, \ldots, 0)^{\mathrm{T}}$.

First, to examine the effect of penalized regression estimates on model fitting, we generate 200 training sets of size n. For each training set, the model error is calculated by $\sum_{l=1}^{L}\{\hat{m}(x_l) - m(x_l)\}^2/L$, at a randomly generated sequence $\{x_l\}_{l=1}^{L=5000}$, and the relative model error is the ratio of model error using penalized estimators and that using nonpenalized estimators. The tuning constants λ_n for the training set in each simulation for methods (I)–(II) are

Table 1. *Simulation results from penalized quasilikelihood estimates, with dependent predictors.* $n = 200$, $\rho = 0{\cdot}2$

Loss	p_n	Method	Regression MRME	Variable selection CZ (SD)	IZ (SD)
Quasilikelihood	$n/8$	SCAD	0·2428	17·74 (5·46)	0 (0)
		L_1	0·3503	14·21 (4·91)	0 (0)
		Weighted L_1	0·1077	21·32 (2·48)	0 (0)
		Oracle	0·0861	23	–
Quasilikelihood	$n/2$	SCAD	0·0409	91·73 (12·56)	0 (0)
		L_1	0·0712	88·00 (14·94)	0 (0)
		Weighted L_1	0·0161	94·84 (5·89)	0 (0)
		Oracle	0·0105	98	–
Quasilikelihood	$n - 10$	SCAD	0·0010	184·37 (8·52)	0 (0)
		L_1	0·0019	181·13 (13·87)	0 (0)
		Weighted L_1	0·0004	185·25 (4·97)	0 (0)
		Oracle	0·0002	188	–

SCAD, smoothly clipped absolute deviation; MRME, mean of relative model errors obtained from the training sets; CZ, average number of coefficients that are correctly estimated to be zero when the true coefficients are zero; IZ, average number of coefficients that are incorrectly estimated to be zero when the true coefficients are nonzero; SD, standard deviation.

selected separately by minimizing the quasilikelihood on a test set of size equal to that of the training set; λ_n and κ_n for method (III) are searched on a surface of grid points. The mean relative model error can be obtained from those 200 training sets. Table 1 summarizes the penalized quasilikelihood estimates of parameters by means of (15). It is clearly seen that if the true model coefficients are sparse, the penalized estimators reduce the function estimation error compared with the nonpenalized estimators.

Second, to study the utility of penalized estimators in revealing the effects in variable selection under quasilikelihood, Table 1 gives the average number of coefficients that are correctly estimated to be zero when the true coefficients are zero, and the average number of coefficients that are incorrectly estimated to be zero when the true coefficients are nonzero. The standard deviations of the corresponding estimations across 200 training sets are given in brackets. Overall, the penalized estimators help yield a sparse solution and build a sparse model. These results lend support to the theoretical results in § 3 and § 4.

In summary, the smoothly clipped absolute deviation and weighted L_1 penalties outperform the L_1 penalty in terms of regression estimation and variable selection. As expected, the oracle estimator, which is practically infeasible, performs better than the three penalized estimators.

6·3. *Penalized Bregman divergence for binary classification*

We generate data with two-classes from the model,

$$X = (X_1, \ldots, X_{p_n})^{\mathrm{T}} \sim N(0, \Sigma), \quad Y \mid X = x \sim \mathrm{Ber}\{m(x)\},$$

where $p_n = n/8$, $n/2$, $n - 10$, $\Sigma = \rho 1_{p_n} 1_{p_n}^{\mathrm{T}} + (1 - \rho)\mathrm{I}_{p_n}$ and $\mathrm{logit}\{m(x)\} = b_{0;0} + x^{\mathrm{T}}\beta_0$ with $b_{0;0} = 3$ and $\beta_0 = (1{\cdot}5, 2, -2, -2{\cdot}5, 0, 0, \ldots, 0)^{\mathrm{T}}$. Table 2 summarizes the penalized estimates of parameters. The results reinforce the conclusion drawn in § 6·2.

Moreover, to investigate the performance of penalized classifiers, we evaluate the average misclassification rate for 10 independent test sets of size 10 000. Table 2 reports the mean of the average misclassification rates calculated from 100 training sets. Evidently, all penalized classifiers perform as well as the optimal Bayes classifier. This agrees with results of Theorem 9 on the

Table 2. *Simulation results from penalized Bregman divergence estimates for binary classification, with dependent predictors.* $n = 200$, $\rho = 0{\cdot}2$

Loss	p_n	Method	Regression MRME	Variable selection CZ (SD)	Variable selection IZ (SD)	Classification MAMR
Deviance	$n/8$	SCAD	0·2504	18·86 (4·37)	0·01 (0·10)	0·1153
		L_1	0·3774	11·31 (5·48)	0·00 (0·00)	0·1218
		Weighted L_1	0·2409	18·11 (2·26)	0·01 (0·10)	0·1160
		Oracle	0·1164	21	0	0·1042
Exponential	$n/8$	SCAD	0·2566	18·92 (4·13)	0·00 (0·00)	0·1162
		L_1	0·3356	12·28 (5·54)	0·00 (0·00)	0·1232
		Weighted L_1	0·2176	19·07 (1·66)	0·01 (0·10)	0·1175
		Oracle	0·1276	21	0	0·1042
Deviance	$n/2$	SCAD	0·0612	94·74 (2·32)	0·03 (0·17)	0·1166
		L_1	0·1148	76·39 (12·97)	0·00 (0·00)	0·1313
		Weighted L_1	0·0782	89·00 (6·38)	0·04 (0·19)	0·1235
		Oracle	0·0240	96	0	0·1043
Exponential	$n/2$	SCAD	0·0915	94·37 (2·91)	0·05 (0·21)	0·1209
		L_1	0·1141	76·05 (11·99)	0·00 (0·00)	0·1315
		Weighted L_1	0·0723	90·60 (4·70)	0·04 (0·19)	0·1222
		Oracle	0·0310	96	0	0·1043
Deviance	$n - 10$	SCAD	0·0230	185·09 (1·53)	0·02 (0·14)	0·1136
		L_1	0·0847	158·19 (17·26)	0·00 (0·00)	0·1401
		Weighted L_1	0·0539	176·51 (8·17)	0·03 (0·17)	0·1273
		Oracle	0·0121	186	0	0·1044
Exponential	$n - 10$	SCAD	0·0360	184·62 (2·20)	0·01 (0·10)	0·1170
		L_1	0·0746	161·15 (14·73)	0·00 (0·00)	0·1386
		Weighted L_1	0·0489	178·70 (5·91)	0·04 (0·19)	0·1271
		Oracle	0·0150	186	0	0·1044

MAMR, mean of the average misclassification rates calculated from training sets.

asymptotic classification consistency. Furthermore, the choice of loss functions in the penalized classifiers has an asymptotically relatively negligible impact on classification performance.

7. Real data

The Arrhythmia dataset (Güvenir et al., 1997) consists of 452 patient records in the diagnosis of cardiac arrhythmia. Each record contains 279 clinical measurements, from electrocardiography signals and other information such as sex, age and weight, along with the decision of an expert cardiologist. In the data, class 01 refers to normal electrocardiography, class 02–class 15 each refers to a particular type of arrhythmia, and class 16 refers to the unclassified remainder.

We intend to predict whether a patient can be categorized as having normal electrocardiography or not. After deleting missing values and class 16, the remaining 430 patients with 257 attributes are used in the classification. To evaluate the performance of the penalized estimates of model parameters in $\text{logit}\{\text{pr}(Y = 1 \mid X_1, \ldots, X_{257})\} = b_0 + \sum_{j=1}^{257} \beta_j X_j$, we randomly split the data into a training set and a test set in the ratio 2:1. For each training set, the tuning constant is selected by minimizing a 3-fold crossvalidated estimate of the misclassification rate; λ_n and κ_n for the penalized componentwise regression are found on a grid of points. We calculate the mean of the misclassification rates and the average number of selected variables over 100 random splittings. It is seen from Table 3 that the penalized classifier using the deviance loss and that using

Table 3. *Arrhythmia data: mean misclassification rate and the average number of selected variables*

Loss	Method	MMR	# Selected variables
Deviance	Nonpenalized	0·4265	257·00
	SCAD	0·2550	16·13
	L_1	0·2358	45·46
	Weighted L_1	0·2340	26·44
Exponential	Nonpenalized	0·4323	257·00
	SCAD	0·2666	15·83
	L_1	0·2397	43·79
	Weighted L_1	0·2366	18·77

MMR, mean of the misclassification rates.

the exponential loss have similar values of misclassification rates. In contrast, the nonpenalized classifiers select all attributes, yielding much higher misclassification rates.

ACKNOWLEDGEMENT

The authors thank the editor, associate editor and two referees for insightful comments and suggestions. The research was supported by grants from the National Science Foundation and National Institutes of Health, U.S.A.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Biometrika* online.

APPENDIX

For a matrix M, its eigenvalues, minimum eigenvalue, maximum eigenvalue and trace are labelled by $\lambda_j(M)$, $\lambda_{\min}(M)$, $\lambda_{\max}(M)$ and $\mathrm{tr}(M)$, respectively. Let $\|M\| = \sup_{\|x\|=1} \|Mx\| = \{\lambda_{\max}(M^{\mathrm{T}}M)\}^{1/2}$ be the matrix L_2 norm; let $\|M\|_F = \{\mathrm{tr}(M^{\mathrm{T}}M)\}^{1/2}$ be the Frobenius norm. See Golub & Van Loan (1996) for details. Throughout the proof, C is used as a generic finite constant.

We first impose some regularity conditions, which are not the weakest possible.

Condition A consists of the following.

A1. Assume $\sup_{n\geqslant 1} \|\widetilde{\beta}_0^{(\mathrm{I})}\|_1 < \infty$ and $\|X\|_\infty$ is bounded;
A2. the matrix $E(\widetilde{X}\widetilde{X}^{\mathrm{T}})$ exists and is nonsingular;
A3. assume $E(Y^2) < \infty$;
A4. there is a large enough open subset of $\mathbb{R}^{p_n+1}$, which contains the true parameter point $\widetilde{\beta}_0$, such that $F^{-1}(\widetilde{X}^{\mathrm{T}}\widetilde{\beta})$ is bounded for all $\widetilde{\beta}$ in the subset;
A5. the eigenvalues of the matrix $-E(q^{(2)}\{m(X)\}/[F^{(1)}\{m(X)\}]^2\widetilde{X}\widetilde{X}^{\mathrm{T}})$ are uniformly bounded away from 0;
A6. the function $q^{(4)}(\cdot)$ is continuous, and $q^{(2)}(\cdot) < 0$;
A7. the function $F(\cdot)$ is a bijection, $F^{(3)}(\cdot)$ is continuous and $F^{(1)}(\cdot) \neq 0$; and finally
A8. assume $P_{\lambda_n}(0) = 0$. There are constants C and D such that when $\theta_1 > C\lambda_n$ and $\theta_2 > C\lambda_n$, $|P'_{\lambda_n}(\theta_1) - P'_{\lambda_n}(\theta_2)| \leqslant D|\theta_1 - \theta_2|$.

Condition B: These are identical to Condition A except that A3 and A5 are replaced by B3 and B5:

B3. there exists a constant $C \in (0, \infty)$ such that $E\{|Y - m(X)|^j\} \leqslant j!C^j$ for all $j \geqslant 3$. Also, $\inf_{n \geqslant 1,\, 1 \leqslant j \leqslant p_n} E\{\mathrm{var}(Y \mid X)X_j^2\} > 0$; and

B5. assume $\lambda_j(\Omega_n)$ and $\lambda_j(H_n)$ are uniformly bounded away from 0; $\|H_n^{-1}\Omega_n\|$ is bounded away from ∞.

Condition C: These are identical to Condition B except that B4 is replaced by:

C4. there is an open subset of $\mathbb{R}^{p_n+1}$ which contains the true parameter point $\widetilde{\beta}_0$, such that $F^{-1}(\widetilde{X}^{\mathrm{T}}\widetilde{\beta})$ is bounded for all $\widetilde{\beta}$ in the subset. Moreover, the subset contains the origin.

Condition D: This is identical to Condition C except that C5 is replaced by:

D5. assume $\lambda_j(H_n)$ are uniformly bounded away from 0; $\|H_n^{-1/2}\Omega_n^{1/2}\|$ is bounded away from ∞.

Condition E is as follows.

E1. Assume $\min_{j=1,\ldots,s_n} |E(X_jY)| \succeq \mathcal{A}_n$ and $\max_{s_n+1 \leqslant j \leqslant p_n} |E(X_jY)| = o(\mathcal{B}_n)$ for some positive sequences $\mathcal{A}_n$ and $\mathcal{B}_n$, where $s_n \succeq t_n$, for two nonnegative sequences s_n and t_n, denotes that there exists a constant $c > 0$ such that $s_n \geqslant c\, t_n$ for all $n \geqslant 1$.

Proof of Theorem 1. Let $r_n = (p_n/n)^{1/2}$ and $\widetilde{u} = (u_0, u_1, \ldots, u_{p_n})^{\mathrm{T}} \in \mathbb{R}^{p_n+1}$. Similar to Fan & Peng (2004), it suffices to show that for any given $\epsilon > 0$, there is a large constant C_ϵ such that, for large n,

$$\mathrm{pr}\left\{\inf_{\|\widetilde{u}\|=C_\epsilon} \ell_n(\widetilde{\beta}_0 + r_n\widetilde{u}) > \ell_n(\widetilde{\beta}_0)\right\} \geqslant 1 - \epsilon. \tag{A1}$$

Define $\widetilde{\beta}_L = \widetilde{\beta}_0 + r_n\widetilde{u}$. To show (A1), consider

$$\begin{aligned} D_n(\widetilde{u}) = {} & \frac{1}{n}\sum_{i=1}^{n}\left[Q\left\{Y_i, F^{-1}\left(\widetilde{X}_i^{\mathrm{T}}\widetilde{\beta}_L\right)\right\} - Q\left\{Y_i, F^{-1}\left(\widetilde{X}_i^{\mathrm{T}}\widetilde{\beta}_0\right)\right\}\right] \\ & + \sum_{j=1}^{p_n}\left\{P_{\lambda_n}(|\beta_{j;0} + r_nu_j|) - P_{\lambda_n}(|\beta_{j;0}|)\right\} \equiv I_1 + I_2. \end{aligned} \tag{A2}$$

First, we consider I_1. For $\mu = F^{-1}(\theta)$, obtain $q_j(y;\theta)\,(j = 1, 2, 3)$, from (7). By Taylor's expansion,

$$I_1 = I_{1,1} + I_{1,2} + I_{1,3}, \tag{A3}$$

where $I_{1,1} = r_n/n\sum_{i=1}^n q_1(Y_i; \widetilde{X}_i^{\mathrm{T}}\widetilde{\beta}_0)\widetilde{X}_i^{\mathrm{T}}\widetilde{u}$, $I_{1,2} = r_n^2/(2n)\sum_{i=1}^n q_2(Y_i; \widetilde{X}_i^{\mathrm{T}}\widetilde{\beta}_0)(\widetilde{X}_i^{\mathrm{T}}\widetilde{u})^2$ and $I_{1,3} = r_n^3/(6n)$ $\sum_{i=1}^n q_3(Y_i; \widetilde{X}_i^{\mathrm{T}}\widetilde{\beta}^*)(\widetilde{X}_i^{\mathrm{T}}\widetilde{u})^3$ for $\widetilde{\beta}^*$ located between $\widetilde{\beta}_0$ and $\widetilde{\beta}_0 + r_n\widetilde{u}$. Hence $|I_{1,1}| \leqslant O_P\{r_n(p_n/n)^{1/2}\}\|\widetilde{u}\|$ and $I_{1,2} = -(r_n^2/2)\widetilde{u}^{\mathrm{T}}E(q^{(2)}\{m(X)\}/[F^{(1)}\{m(X)\}]^2\widetilde{X}\widetilde{X}^{\mathrm{T}})\widetilde{u} + O_P(r_n^2p_n/n^{1/2})\|\widetilde{u}\|^2$. Conditions A1 and A4 give $|I_{1,3}| \leqslant O_P(r_n^3p_n^{3/2})\|\widetilde{u}\|^3$.

Next, we consider I_2. By Taylor's expansion, $I_2 \geqslant r_n\sum_{j=1}^{s_n} P'_{\lambda_n}(|\beta_{j;0}|)\mathrm{sign}(\beta_{j;0})u_j + (r_n^2/2)\sum_{j=1}^{s_n}$ $P^{(2)}_{\lambda_n}(|\beta_j^*|)u_j^2 \equiv I_{2,1} + I_{2,2}$, for β_j^* between $\beta_{j;0}$ and $\beta_{j;0} + r_nu_j$. Thus, $|I_{2,1}| \leqslant r_na_n\|u^{(\mathrm{I})}\|_1$ and $|I_{2,2}| \leqslant$ $r_n^2b_n\|u^{(\mathrm{I})}\|^2 + Dr_n^3\|u^{(\mathrm{I})}\|^3$, where $u^{(\mathrm{I})} = (u_1, \ldots, u_{s_n})^{\mathrm{T}}$. Since $p_n^4/n \to 0$, we can choose some large C_ϵ such that $I_{1,1}$, $I_{1,3}$, $I_{2,1}$ and $I_{2,2}$ are all dominated by $I_{1,2}$, which is positive by Condition A5. This implies (A1). □

Proof of Lemma 1. Analogous to the proof of Theorem 1, it suffices to show (A1). Note that (A2) continues to hold with $I_2 = \lambda_n\sum_{j=1}^{p_n} w_j(|\beta_{j;0} + r_nu_j| - |\beta_{j;0}|)$ and I_1 is unchanged. Clearly, $I_2 \geqslant -\lambda_nr_n\sum_{j=1}^{s_n} w_j|u_j| \equiv I_{2,1}$, in which $|I_{2,1}| \leqslant \lambda_nr_nw^{(\mathrm{I})}_{\max}\|u^{(\mathrm{I})}\|_1$. The rest of the proof resembles that of Theorem 1 and is omitted. □

Proof of Theorem 5. Write $\widetilde{u} = \{\widetilde{u}^{(\mathrm{I})\mathrm{T}}, u^{(\mathrm{II})\mathrm{T}}\}^{\mathrm{T}}$, where $\widetilde{u}^{(\mathrm{I})} = (u_0, u_1, \ldots, u_{s_n})^{\mathrm{T}}$ and $u^{(\mathrm{II})} = (u_{s_n+1}, \ldots,$ $u_{p_n})^{\mathrm{T}}$. Following the proof of Lemma 1, it suffices to show (A1) for $r_n = (s_n/n)^{1/2}$.

For $I_{1,1}$ in (A3), $I_{1,1} = I_{1,1}^{(\mathrm{I})} + I_{1,1}^{(\mathrm{II})}$ according to $\widetilde{u}^{(\mathrm{I})}$ and $u^{(\mathrm{II})}$. It follows that $|I_{1,1}^{(\mathrm{I})}| \leqslant r_n O_P\{(s_n/n)^{1/2}\} \|\widetilde{u}^{(\mathrm{I})}\|_2$ and $|I_{1,1}^{(\mathrm{II})}| \leqslant r_n O_P(1/n^{1/2})\|u^{(\mathrm{II})}\|_1$.

For $I_{1,2}$ in (A3), similar to the proof of Theorem 1, $I_{1,2} = I_{1,2,1} + I_{1,2,2}$. Define $d_i = q^{(2)}\{m(X_i)\}/[F^{(1)}\{m(X_i)\}]^2$. This yields

$$I_{1,2,1} \geqslant -\frac{r_n^2}{2n}\sum_{i=1}^{n} d_i\big(X_i^{(\mathrm{I})\mathrm{T}}\widetilde{u}^{(\mathrm{I})}\big)^2 - \frac{r_n^2}{n}\sum_{i=1}^{n} d_i\big(X_i^{(\mathrm{I})\mathrm{T}}\widetilde{u}^{(\mathrm{I})}\big)\big(X_i^{(\mathrm{II})\mathrm{T}}u^{(\mathrm{II})}\big) = I_{1,2,1}^{(\mathrm{I})} - I_{1,2,1}^{(\mathrm{cross})}.$$

Then there exists a constant $C > 0$ such that $I_{1,2,1}^{(\mathrm{I})} \geqslant C r_n^2\{1 + o_P(1)\}\|\widetilde{u}^{(\mathrm{I})}\|_2^2$ and $|I_{1,2,1}^{(\mathrm{cross})}| \leqslant O_P(r_n^2 s_n^{1/2})\|\widetilde{u}^{(\mathrm{I})}\|_2 \cdot \|u^{(\mathrm{II})}\|_1$. For $I_{1,2,2}$, partitioning $\widetilde{u}$ into $\widetilde{u}^{(\mathrm{I})}$ and $u^{(\mathrm{II})}$ gives

$$I_{1,2,2} \equiv I_{1,2,2}^{(\mathrm{I})} + I_{1,2,2}^{(\mathrm{cross})} + I_{1,2,2}^{(\mathrm{II})},$$

where $|I_{1,2,2}^{(\mathrm{I})}| \leqslant r_n^2 O_P(s_n/n^{1/2})\|\widetilde{u}^{(\mathrm{I})}\|_2^2$, $|I_{1,2,2}^{(\mathrm{cross})}| \leqslant r_n^2 O_P\{(s_n/n)^{1/2}\}\|\widetilde{u}^{(\mathrm{I})}\|_2\|u^{(\mathrm{II})}\|_1$ and $|I_{1,2,2}^{(\mathrm{II})}| \leqslant r_n^2 O_P(n^{-1/2})\|u^{(\mathrm{II})}\|_1^2$.

For $I_{1,3}$ in (A3), since $s_n p_n = o(n)$, $\|\widetilde{\beta}^*\|_1$ is bounded and thus $|I_{1,3}| \leqslant O_P(r_n^3)\|\widetilde{u}^{(\mathrm{I})}\|_1^3 + O_P(r_n^3)\|u^{(\mathrm{II})}\|_1^3 \equiv I_{1,3}^{(\mathrm{I})} + I_{1,3}^{(\mathrm{II})}$, where $|I_{1,3}^{(\mathrm{I})}| \leqslant O_P(r_n^3 s_n{}^{3/2})\|\widetilde{u}^{(\mathrm{I})}\|_2^3$ and $|I_{1,3}^{(\mathrm{II})}| \leqslant O_P(r_n^3)\|u^{(\mathrm{II})}\|_1^3$.

For I_2 in (A2), $I_2 \geqslant I_{2,1}^{(\mathrm{I})} + I_{2,1}^{(\mathrm{II})}$, where $I_{2,1}^{(\mathrm{I})} = -\lambda_n r_n \sum_{j=1}^{s_n} w_j|u_j|$ and $I_{2,1}^{(\mathrm{II})} = \lambda_n r_n \sum_{j=s_n+1}^{p_n} w_j|u_j|$. Hence $|I_{2,1}^{(\mathrm{I})}| \leqslant \lambda_n r_n w_{\max}^{(\mathrm{I})} s_n^{1/2}\|u^{(\mathrm{I})}\|_2$ and $I_{2,1}^{(\mathrm{II})} \geqslant \lambda_n r_n w_{\min}^{(\mathrm{II})}\|u^{(\mathrm{II})}\|_1$.

It can be shown that either $I_{1,2,1}^{(\mathrm{I})}$ or $I_{2,1}^{(\mathrm{II})}$ dominates all other terms in groups, $\mathcal{G}_1 = (I_{1,2,2}^{(\mathrm{I})}, I_{1,3}^{(\mathrm{I})})$, $\mathcal{G}_2 = (I_{1,1}^{(\mathrm{II})}, I_{1,2,2}^{(\mathrm{II})}, I_{1,3}^{(\mathrm{II})}, I_{1,2,1}^{(\mathrm{cross})}, I_{1,2,2}^{(\mathrm{cross})})$ and $\mathcal{G}_3 = (I_{1,1}^{(\mathrm{I})}, I_{2,1}^{(\mathrm{I})})$. Namely, $I_{1,2,1}^{(\mathrm{I})}$ dominates $\mathcal{G}_1$, and $I_{2,1}^{(\mathrm{II})}$ dominates $\mathcal{G}_2$. For $\mathcal{G}_3$, if $\|u^{(\mathrm{II})}\|_1 \leqslant C_\epsilon/2$, then $\mathcal{G}_3$ is dominated by $I_{1,2,1}^{(\mathrm{I})}$, which is positive; if $\|u^{(\mathrm{II})}\|_1 > C_\epsilon/2$, then $\mathcal{G}_3$ is dominated by $I_{2,1}^{(\mathrm{II})}$, which is positive. □

Proof of Theorem 8. Minimizing (14) is equivalent to minimizing $\ell_{n,j}^{\mathrm{PCR}}(\alpha) = n^{-1}\sum_{i=1}^{n} Q\{Y_i, F^{-1}(X_{ij}\alpha)\} + \kappa_n|\alpha|$, for $j = 1, \ldots, p_n$. The proof may be separated into two parts.

Part 1. To show $\hat{w}_{\max}^{(\mathrm{I})} = O_P\{1/(\lambda_n n^{1/2})\}$, it suffices to show that for $\mathcal{A}_n = \lambda_n n^{1/2}$, there exist local minimizers $\hat{\beta}_j^{\mathrm{PCR}}$ of $\ell_{n,j}^{\mathrm{PCR}}(\alpha)$ such that $\lim_{\delta\to 0}\inf_{n\geqslant 1}\mathrm{pr}(\min_{1\leqslant j\leqslant s_n}|\hat{\beta}_j^{\mathrm{PCR}}| > \mathcal{A}_n\delta) = 1$. It suffices to prove that for $j = 1, \ldots, s_n$ there exist some b_j with $|b_j| = 2\delta$ such that

$$\lim_{\delta\to 0}\inf_{n\geqslant 1}\mathrm{pr}\left[\min_{1\leqslant j\leqslant s_n}\left\{\inf_{|\alpha|\leqslant\delta}\ell_{n,j}^{\mathrm{PCR}}(\mathcal{A}_n\alpha) - \ell_{n,j}^{\mathrm{PCR}}(\mathcal{A}_n b_j)\right\} > 0\right] = 1, \tag{A4}$$

and there exists some large enough $C_n > 0$ such that

$$\lim_{\delta\to 0}\inf_{n\geqslant 1}\mathrm{pr}\left[\min_{1\leqslant j\leqslant s_n}\left\{\inf_{|\alpha|\geqslant C_n}\ell_{n,j}^{\mathrm{PCR}}(\mathcal{A}_n\alpha) - \ell_{n,j}^{\mathrm{PCR}}(\mathcal{A}_n b_j)\right\} > 0\right] = 1. \tag{A5}$$

Note that (A5) holds, since for every $n \geqslant 1$, when $|\alpha| \to \infty$, $\min_{1\leqslant j\leqslant s_n}\{\ell_{n,j}^{\mathrm{PCR}}(\mathcal{A}_n\alpha) - \ell_{n,j}^{\mathrm{PCR}}(\mathcal{A}_n b_j)\} \geqslant \kappa_n\mathcal{A}_n|\alpha| - \max_{j=1,\ldots,s_n}\ell_{n,j}^{\mathrm{PCR}}(\mathcal{A}_n b_j) \to \infty$ in probability. To prove (A4), note that $|\mathcal{A}_n\alpha| \leqslant \mathcal{A}_n\delta = O(1)\delta \to 0$ as $\delta \downarrow 0$. By Taylor's expansion,

$$\begin{aligned}
&\min_{j=1,\ldots,s_n}\left\{\inf_{|\alpha|\leqslant\delta}\ell_{n,j}^{\mathrm{PCR}}(\mathcal{A}_n\alpha) - \ell_{n,j}^{\mathrm{PCR}}(\mathcal{A}_n b_j)\right\} \geqslant \mathcal{A}_n\min_{j=1,\ldots,s_n}\inf_{|\alpha|\leqslant\delta}\left\{(\alpha - b_j)\frac{1}{n}\sum_{i=1}^{n} q_1(Y_i;0)X_{ij}\right\} \\
&\quad + \frac{\mathcal{A}_n^2}{2}\min_{j=1,\ldots,s_n}\inf_{|\alpha|\leqslant\delta}\left\{\alpha^2\frac{1}{n}\sum_{i=1}^{n} q_2(Y_i; X_{ij}\mathcal{A}_n\alpha_j^*)X_{ij}^2 - b_j^2\frac{1}{n}\sum_{i=1}^{n} q_2(Y_i; X_{ij}\mathcal{A}_n b_j^*)X_{ij}^2\right\} \\
&\quad + \mathcal{A}_n\min_{1\leqslant j\leqslant s_n}\inf_{|\alpha|\leqslant\delta}\{\kappa_n(|\alpha| - |b_j|)\} \\
&\equiv I_1 + I_2 + I_3,
\end{aligned}$$

with α_j^* between 0 and α and b_j^* between 0 and b_j. Let $\mu_0 = F^{-1}(0)$ and $C_0 = q''(\mu_0)/F'(\mu_0) \neq 0$. Then

$$\begin{aligned} I_1 &= \mathcal{A}_n \min_{j=1,\ldots,s_n} \inf_{|\alpha| \leqslant \delta} \{C_0(\alpha - b_j)E(YX_j)\} + \mathcal{A}_n \min_{j=1,\ldots,s_n} \inf_{|\alpha| \leqslant \delta} \left[C_0(\alpha - b_j)\frac{1}{n}\sum_{i=1}^n \{Y_i X_{ij} - E(YX_j)\} \right] \\ &\quad - \mathcal{A}_n \max_{1 \leqslant j \leqslant s_n} \sup_{|\alpha| \leqslant \delta} \left\{ C_0\mu_0(\alpha - b_j)\frac{1}{n}\sum_{i=1}^n X_{ij} \right\} \\ &\equiv I_{1,1} + I_{1,2} + I_{1,3}. \end{aligned}$$

We see that $|I_{1,3}| \leqslant O_P[\mathcal{A}_n\{\log(s_n)/n\}^{1/2}]\delta$, by Bernstein's inequality (Lemma 2.2.11 in van der Vaart & Wellner 1996). Again $|I_{1,2}| = O_P[\mathcal{A}_n\{\log(s_n)/n\}^{1/2}]\delta$ by an argument similar to that of Theorem 2. Choosing $b_j = -2\delta \operatorname{sign}\{C_0 E(YX_j)\}$, which satisfies $|b_j| = 2\delta$, gives $I_{1,1} \geqslant |C_0| c \mathcal{A}_n^2 \delta$. For I_2 and I_3, we observe that $|I_2| \leqslant O_P(\mathcal{A}_n^2)\delta^2$ and $|I_3| = O(\mathcal{A}_n\kappa_n)\delta$. By the assumptions, we can choose a small enough $\delta > 0$ such that with probability tending to 1, $I_{1,2}$, $I_{1,3}$, I_2 and I_3 are dominated by $I_{1,1}$, which is positive. Thus (A4) is proved.

Part 2. To verify that $\hat{w}_{\min}^{(\mathrm{II})}\lambda_n \to \infty$ in probability, it suffices to prove that for any $\epsilon > 0$, there exist local minimizers $\hat{\beta}_j^{\mathrm{PCR}}$ of $\ell_{n,j}^{\mathrm{PCR}}(\alpha)$ such that $\lim_{n\to\infty} \mathrm{pr}(\max_{s_n+1 \leqslant j \leqslant p_n} |\hat{\beta}_j^{\mathrm{PCR}}| \leqslant \lambda_n \epsilon) = 1$. Similar to the proof of Theorem 1, we will prove that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{pr}\left[\min_{j=s_n+1,\ldots,p_n} \left\{ \inf_{|\alpha|=\epsilon} \ell_{n,j}^{\mathrm{PCR}}(\lambda_n\alpha) - \ell_{n,j}^{\mathrm{PCR}}(0) \right\} > 0 \right] = 1. \tag{A6}$$

For $j = 1, \ldots, s_n$, by Taylor's expansion,

$$\begin{aligned} &\min_{j=s_n+1,\ldots,p_n} \left\{ \inf_{|\alpha|=\epsilon} \ell_{n,j}^{\mathrm{PCR}}(\lambda_n\alpha) - \ell_{n,j}^{\mathrm{PCR}}(0) \right\} \geqslant \lambda_n \min_{j=s_n+1,\ldots,p_n} \inf_{|\alpha|=\epsilon} \left\{ \alpha \frac{1}{n}\sum_{i=1}^n q_1(Y_i;0)X_{ij} \right\} \\ &\quad + \frac{\lambda_n^2}{2} \min_{j=s_n+1,\ldots,p_n} \inf_{|\alpha|=\epsilon} \left\{ \alpha^2 \frac{1}{n}\sum_{i=1}^n q_2(Y_i; X_{ij}\lambda_n\alpha_j^*)X_{ij}^2 \right\} + \lambda_n \inf_{|\alpha|=\epsilon} (\kappa_n|\alpha|) \\ &\equiv I_1 + I_2 + I_3, \end{aligned}$$

where α_j^* is between 0 and α. Similar to the proof in Part 1, $|I_1| \leqslant O_P[\lambda_n\{\log(p_n - s_n + 1)/n\}^{1/2}]\epsilon + o(\lambda_n\mathcal{B}_n)\epsilon$. Note that $|I_2| \leqslant O_P(\lambda_n^2)\epsilon^2$ and $I_3 = \lambda_n\kappa_n\epsilon$. By assumptions, with probability tending to 1, I_1 and I_2 are dominated by $I_3 > 0$. So (A6) is proved. □

Proof of Theorem 9. We first need to show Lemma A1.

LEMMA A1. *Suppose that (X^o, Y^o) follows the distribution of (X, Y) and is independent of the training set $\mathcal{T}_n$. If Q satisfies (3), then $E[Q\{Y^o, \hat{m}(X^o)\}] = E[Q\{Y^o, m(X^o)\}] + E[Q\{m(X^o), \hat{m}(X^o)\}]$.*

Proof. Let q be the generating function of Q. We deduce from Corollary 3, p. 223 of Chow & Teicher (1988) that $E\{q(Y^o) \mid \mathcal{T}_n, X^o\} = E\{q(Y^o) \mid X^o\}$ and $E[Y^o q'\{\hat{m}(X^o)\} \mid \mathcal{T}_n, X^o] = E(Y^o \mid X^o)q'\{\hat{m}(X^o)\} = m(X^o)q'\{\hat{m}(X^o)\}$. □

We now show Theorem 9. Setting Q in Lemma A1 to be the misclassification loss gives

$$\begin{aligned} 1/2[E\{R(\hat{\phi})\} - R(\phi_{\mathrm{B}})] &\leqslant E[|m(X^o) - 0{\cdot}5| I\{m(X^o) \leqslant 0{\cdot}5,\ \hat{m}(X^o) > 0{\cdot}5\}] \\ &\quad + E[|m(X^o) - 0{\cdot}5| I\{m(X^o) > 0{\cdot}5,\ \hat{m}(X^o) \leqslant 0{\cdot}5\}] \\ &\equiv I_1 + I_2. \end{aligned}$$

For any $\epsilon > 0$, $I_1 \leqslant \mathrm{pr}\{|\hat{m}(X^o) - m(X^o)| > \epsilon\} + \epsilon$ and $I_2 \leqslant \epsilon + \mathrm{pr}\{|\hat{m}(X^o) - m(X^o)| \geqslant \epsilon\}$. The proof completes by showing $I_1 \to 0$ and $I_2 \to 0$. □

REFERENCES

ALTUN, Y. & SMOLA, A. (2006). Unifying divergence minimization and statistical inference via convex duality. In *Learning Theory: 19th Ann. Conf. Learn. Theory*, Ed. G. Lugosi and H. U. Simon, pp. 139–53. Berlin: Springer.

BARTLETT, M. S. (1953). Approximate confidence intervals. *Biometrika* **40**, 12–19.

BICKEL, P. & LI, B. (2006). Regularization in statistics (with discussion). *Test* **15**, 271–344.

BREGMAN, L. M. (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *USSR Comp. Math. Math. Phys.* **7**, 620–31.

CHOW, Y. S. & TEICHER, H. (1988). *Probability Theory*, 2nd ed. New York: Springer.

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Assoc.* **81**, 461–70.

EFRON, B., HASTIE, T., JOHNSTONE, I., & TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–99.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

FAN, J. & PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–61.

GOLUB, G. H. & VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins University Press.

GRÜNWALD, P. D. & DAWID, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Statist.* **32**, 1367–433.

GÜVENIR, H. A., ACAR, B., DEMIRÖZ, G. & ÇEKIN, A. (1997). A supervised machine learning algorithm for arrhythmia analysis. *Comp. Cardiol.* **24**, 433–6.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

HUANG, J., MA, S. G. & ZHANG, C. H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603–18.

KIVINEN, J. & WARMUTH, M. K. (1999). Boosting as entropy projection. In *Proc. 12th Ann. Conf. Comp. Learn. Theory*, pp. 134–44. New York: ACM Press.

KNIGHT, K. & FU, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–78.

LAFFERTY, J. D., DELLA PIESTRA, S., & DELLA PIESTRA, V. (1997). Statistical learning algorithms based on Bregman distances. In *Proc. 5th Can. Workshop Info. Theory*.

LAFFERTY, J. (1999). Additive models, boosting, and inference for generalized divergences. In *Proc. 12th Ann. Conf. Comp. Learn. Theory*, pp. 125–33. New York: ACM Press.

MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.

MEINSHAUSEN, N. & BUHLMANN, P. (2006). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.

ROSSET, S. & ZHU, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35**, 1012–30.

SHEN, X., TSENG, G. C., ZHANG, X. & WONG, W. H. (2003). On ψ-learning. *J. Am. Statist. Assoc.* **98**, 724–34.

STRIMMER, K. (2003). Modeling gene expression measurement error: a quasi-likelihood approach. *BMC Bioinformatics* **4**, 10.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

VAPNIK, V. (1996). *The Nature of Statistical Learning Theory*. New York: Springer.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* New York: Springer.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–47.

ZHANG, C. M., JIANG, Y. & SHANG, Z. (2009). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Can. J. Statist.* **37**, 119–39.

ZHANG, C. M. & ZHANG, Z. J. (2010). Regularized estimation of hemodynamic response function for fMRI data. *Statist. Interface* **3**, 15–32.

ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–67.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

[*Received February* 2009. *Revised February* 2010]