Asymptotic distribution and sparsistency for ℓ_1 penalized parametric M-estimators, with applications to linear SVM and logistic regression

Guilherme Rocha [*], Xing Wang [†] and Bin Yu [‡]

October 24, 2009

Abstract

Since its early use in least squares regression problems, the ℓ_1-penalization framework for variable selection has been employed in conjunction with a wide range of loss functions encompassing regression, classification and survival analysis. While a well developed theory exists for the ℓ_1-penalized least squares estimates, few results concern the behavior of ℓ_1-penalized estimates for general loss functions. In this paper, we derive two results concerning penalized estimates for a wide array of penalty and loss functions. Our first result characterizes the asymptotic distribution of penalized parametric M-estimators under mild conditions on the loss and penalty functions in the classical setting (fixed-p-large-n). Our second result explicits necessary and sufficient generalized irrepresentability (GI) conditions for ℓ_1-penalized parametric M-estimates to consistently select the components of a model (sparsistency) as well as their sign (sign consistency). In general, the GI conditions depend on the Hessian of the risk function at the true value of the unknown parameter. Under Gaussian predictors, we obtain a set of conditions under which the GI conditions can be re-expressed solely in terms of the second moment of the predictors. We apply our theory to contrast ℓ_1-penalized SVM and logistic regression classifiers and find conditions under which they have the same behavior in terms of their model selection consistency (sparsistency and sign consistency). Finally, we provide simulation evidence for the theory based on these classification examples.

[*]Indiana University, gvrocha@indiana.edu, gvrocha.stat@gmail.com
[†]Renmin University of China, wangxingscy@gmail.com
[‡]University of California, Berkeley, binyu@stat.berkeley.edu

1 Introduction

When modeling the a response variable $\mathbf{Y} \in \mathcal{Y}$ as a function of a set of predictors $\mathbf{X} \in \mathbb{R}^p$, statisticians often rely on M-estimators for linear models defined as

$$\left(\hat{\alpha}_n, \hat{\beta}_n\right) \quad := \quad \underset{a\in\mathbb{R}, b\in\mathbb{R}^p}{\arg\min} \left[\tfrac{1}{n} \cdot \textstyle\sum_{i=1}^n L(Y_i, a + X_i^T b, t)\right], \tag{1}$$

where $Z_i = (Y_i, X_i)$, $i = 1, \ldots, n$, are independent observations of $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$ and the loss function $L : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ measures the lack of quality of $a + X_i b$ in representing Y_i. For a given problem, many alternative loss functions can be used. Some recent results are aimed at comparing the properties of estimates obtained from alternative loss functions (Zhang, 2004; Bartlett et al., 2006).

The choice of an appropriate loss function must take the goal of the analysis into account. Often, the estimates in (1) are used as a tool in understanding the effects of $\mathbf{X}$ on $\mathbf{Y}$. In that case, sparse estimates $\hat{\beta}_n$ are desirable as they select which predictors in $\mathbf{X}$ have an effect on the response $\mathbf{Y}$. Sparse estimates are often achieved by a penalized estimate

$$\left(\hat{\alpha}_n(\lambda_n), \hat{\beta}_n(\lambda_n)\right) \quad := \quad \underset{a\in\mathbb{R}, b\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^n L(Y_i, a + X_i b) + \lambda_n \cdot T(b)\right], \tag{2}$$

where $\lambda_n \geq 0$ is a regularization parameter and $T : \mathbb{R}^p \to \mathbb{R}_+$ is a function penalizing non-sparse models. Many alternative sparsity inducing penalties exist and a popular family of such penalties is the set of ℓ_γ norms with $\gamma \in (0, 1]$ used in bridge estimates (Frank and Friedman, 1993). The ℓ_γ norm function given by $\|b\|_\gamma := \left(\sum_{j=1}^p \|b_j\|^\gamma\right)^{\frac{1}{\gamma}}$. Two important particular cases are the ℓ_0-penalty – defined as a penalty on the number of non-zero terms in the estimate used in (Akaike, 1973, 1974; Schwarz, 1978; Rissanen, 1978; Hansen and Yu, 2001), and the ℓ_1-penalty used in the LASSO (Tibshirani, 1996) and basis pursuit (Chen et al., 2001).

Recently, a large number of ℓ_1-penalized estimates based on different loss functions have been proposed in the literature. Some examples are the logistic regression and Cox's proportional hazards model loss (Tibshirani, 1997; Park and Hastie, 2006), the hinge loss function for classification (Zhu et al., 2004), the quantile regression loss (Li and Zhu, 2008), and the log-determinant Bregman divergence of covariance

matrices (Banerjee et al., 2005; Ravikumar et al., 2008). Simultaneously, many families of sparse inducing penalties have been introduced such as the SCAD penalty (Fan and Li, 2001) and the generalized elastic net (Friedman, 2008). In this paper, we present theoretical results allowing the behavior of estimates based on different loss and penalty functions to be compared.

Our first main result is a characterization of the asymptotic distribution of the penalized estimates in (2) for a wide class of penalty and convex loss functions. Our result extends previous results for the squared error loss and ℓ_γ norms by Knight and Fu (2000) and applies to the classical asymptotic setup (large n, fixed p). We state our results in a modular fashion so they encompass several combinations of loss and penalty functions. We provide sufficient conditions on the loss and on the penalty functions for our results to apply. On the loss side, our results depend on convexity of the loss function and on the risk function defined as

$$R(t) := \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[L(\mathbf{Y}, a + b^T\mathbf{X})\right], \text{ for } (a,b) \in \mathbb{R}^{1+p}, \tag{3}$$

to be twice continuously differentiable at the "true" value of the parameters (α, β)

$$(\alpha,\beta) \quad := \quad \arg\min_{t\in\mathbb{R}^p} R(a,b). \tag{4}$$

Our second result obtains necessary and sufficient conditions for the ℓ_1-penalized estimate in (2) to consistently select the zeroes (sparsistency) and signs (sign consistency) in the parameter β. Previous results for ℓ_1-penalized least squares linear regression show that the set of active and inactive predictors must be sufficiently disentangled for sparsistency to hold. This requirement is embodied in "incoherence" or "irrepresentability" conditions (Meinshausen and Bühlmann, 2004; Zhao and Yu, 2006; Zou, 2006; Wainwright, 2006). We call the condition for sparsistency and sign-consistency of general ℓ_1-penalized M-estimators the generalized irrepresentability (GI) condition. Intuitively, the GI condition can be interpreted as a requirement that the the effects of active and inactive predictors on the loss are distinguishable enough (after "controlling" for the intercept term). This second result relies on the quadratic approximation developed on the first result and is thus only applicable on the classical small-p-large-n case.

Our third result shows that, if the predictors are zero-meaned Gaussian and the response variable only depends on $\mathbf{X}$ through an affine function of the predictors, the conditions for ℓ_1-penalized estimates as in (2)

do not depend on the loss function. In that case, the GI condition reduces to the "irrepresentable" condition in Zhao and Yu (2006). This surprising result stems from the properties of the multivariate Gaussian distribution, namely on its linear mean and constant variance when conditioned on one of its linear combinations.

We apply the theory to contrast and compare linear classifiers based on the hinge loss (parametric SVM) and logistic regression. We obtain expressions for the Hessians of the SVM and logistic regression risks and characterize them as weighted averages of the second moment matrices of the predictors conditional on a properly defined "linear predictor" variable $\mathbf{M} = \alpha + \beta^T \mathbf{X}$. Based on this characterization and using our third result, we show that, for a given joint distribution $(\mathbf{Y}, \mathbf{X})$ where the predictors are Gaussian and the response variable only depends on the predictors through an affine transformation, the two classifiers are either both sparsistent or not. For more general joint distributions, one of the classifiers can be sparsistent while the other is not. Over a set of cases where the predictors are mixed Gaussian, we observed logistic regression to be sparsistent more often than SVM classifiers but also observed mixed results in finite samples. The conditionally weighted second moment characterization of the Hessians also evidences that the Hessians of both SVM and logistic regression risk functions emphasize the second moment of the predictors closer to the optimal separating hyperplane. This emphasis on the region close to the margin echoes previous results in the non-parametric works of Audibert and Tsybakov (2007) and Steinwart and Scovel (2007) and help explain the similarities between SVM and logistic regression classifiers.

The remainder of this paper is organized as follows. Section 2 presents our asymptotic results for penalized empirical risk minimizers for general loss and penalty functions. Section 3 presents necessary and sufficient conditions for model selection consistency of ℓ_1-norm penalized empirical risk minimizers for general loss functions. Section 4 applies the results in the previous two sections to the study and comparison of SVM and logistic regression classifiers satisfy the requirements for our results from the previous two sections to apply. Section 5 shows a series of simulations providing empirical support for the model selection consistency theory we developed as well as comparisons between SVM and logistic regression classifiers. Finally, Section 6 concludes with a brief discussion.

2 Asymptotic distribution of penalized parametric M-estimators

In this section, we present the first main result of this paper (Theorem 4) which characterizes the asymptotic distribution of penalized empirical risk minimizers for a broad range of penalty and loss functions for a fixed number of predictors. Theorem 4 extends previous results by Knight and Fu (2000) regarding norm-penalized least squares estimates. In essence, the steps in the proof of 4 closely parallel the ones used by Knight and Fu (2000), but we keep the study of the convergence of loss and penalty functions separate so our results can be applied to any combination of loss and penalty functions satisfying the conditions detailed below.

Before proceeding, we introduce some notation. Our results apply to penalized estimates defined as

$$\hat{\theta}_n(\lambda_n) \quad := \quad \underset{t\in\Theta\subset\mathbb{R}^{p+1}}{\arg\min} \left[\frac{1}{n}\sum_{i=1}^{n} L(Z_i,t) + \lambda_n \cdot T(t)\right]. \tag{5}$$

The definition in (2) is a particular case that encompasses linear models by setting $Z_i = (Y_i, X_i) \in \mathcal{Y} \times \mathbb{R}^p$, $t = (a, b)$ and $L(Z_i, t) = L(Y_i, a + b^T X_i)$. In this extended case, the best model we can select is parameterized by

$$\theta \quad := \quad \underset{t\in\Theta\subset\mathbb{R}^{p+1}}{\arg\min}\ R(t), \tag{6}$$

where the risk function has the usual definition

$$R(t) := \mathbb{E}_{\mathbf{Z}}\left[L(\mathbf{Z}, t)\right], \text{ for } t \in \Theta \subset \mathbb{R}^{p+1}. \tag{7}$$

Let $u \in \mathbb{R}^{p+1}$ and q_n be a sequence of non-negative numbers such that $q_n \to \infty$ as $n \to \infty$. Define

$$\begin{array}{rcll} C_\theta^{(n)}(\mathbf{Z}, u) & := & \sum_{i=1}^{n}\left[L\left(\mathbf{Z}, \theta + \frac{u}{q_n}\right) - L\left(\mathbf{Z}, \theta\right)\right], & \\ G_\theta^{(n)}(u) & := & \left[T\left(\theta + \frac{u}{q_n}\right) - T\left(\theta\right)\right], & \text{and} \\ V_\theta^{(n)}(\mathbf{Z}, \lambda_n, u) & := & C_\theta^{(n)}(\mathbf{Z}, u) + \lambda_n \cdot G_\theta^{(n)}(u). & \end{array}$$

The $V_\theta^{(n)}$ function corresponds to a recentered and rescaled version of the objective function in (5) so

$$q_n \cdot \left(\hat{\theta}_n(\lambda_n) - \theta\right) \quad = \quad \underset{u \in \mathbb{R}^p}{\arg\min}\, V_\theta^{(n)}(Z, \lambda_n, u).$$

The asymptotic behavior of $\hat{\theta}_n(\lambda_n)$ can be characterized in terms of asymptotic results for $V_\theta^{(n)}$ and its minimizer. A close study of the proof used by Knight and Fu (2000) shows that for the most part, the convergences of the loss $\left(C_\theta^{(n)}\right)$ and the penalty $\left(G_\theta^{(n)}\right)$ functions are studied separately. This is reflected in our Theorem 1: a versatile and an important "assembling" tool. Any set of assumptions made on the loss and penalty functions that ensures the conditions required by Theorem 1 can be used to obtain a characterization of the distribution of penalized estimates.

Theorem 1. *Let $\lambda_n \geq 0$ be a sequence of positive (potentially random) real numbers, Z_i, $i = 1, \ldots, n$, be a sequence of i.i.d. realizations from a distribution $\mathbb{P}_{\mathbf{Z}}$, $L : \mathcal{Z} \times \Theta \to \mathbb{R}$ be a loss function and $T : \Theta \to \mathbb{R}$ be a penalty function. Let $\hat{\theta}(\lambda_n)$ be as defined in* (5)*.*

Suppose there exist functions C_θ, G_θ, a constant λ, a random vector $\mathbf{W}$ and a sequence q_n of deterministic positive real numbers with $q_n \to \infty$ as $n \to \infty$ such that, for any compact set $K \subset \mathbb{R}^p$:

i) $\sup_{u \in K} \left| \sum_{i=1}^n \left[L\left(Z_i, \theta + \frac{u}{q_n}\right) - L\left(Z_i, \theta\right) \right] - C_\theta(\mathbf{W}, u) \right| \xrightarrow{p} 0$*;*

ii) $\sup_{u \in K} \left| \lambda_n \left[T\left(\theta + \frac{u}{q_n}\right) - T\left(\theta\right) \right] - \lambda \cdot G_\theta(u) \right| \xrightarrow{p} 0$*;*

iii) $\hat{\theta}_n(\lambda_n)$ *is* $O_p(q_n^{-1})$*.*

Let $V_\theta(\mathbf{W}, u) = C_\theta(\mathbf{W}, u) + \lambda \cdot G_\theta(u)$*.*

If i) and ii) hold, then:

a) $\sup_{u \in K} \left| V_\theta^{(n)}(\mathbf{Z}, \lambda_n, u) - V_\theta(\mathbf{W}, u) \right| \xrightarrow{p} 0$*;*

If i), ii) and iii) hold, then:

b) $q_n \left(\hat{\theta}_n(\lambda_n) - \theta\right) \xrightarrow{d} \underset{u}{\arg\min}\, V_\theta(\mathbf{W}, u)$*.*

Roughly speaking, we can prove Theorem 1 by observing that boundedness in probability of the sequence $\hat{\theta}_n(\lambda_n)$ implies that $\hat{\theta}_n(\lambda_n) \in K$, for some compact set K with probability approaching 1. Given

this condition, it follows that the uniform convergence in probability over compact sets is sufficient to ensure that the minimizer of $V_\theta^{(n)}$ converges in probability to the minimizer of V_θ. A detailed proof is given in Appendix A.

Based on Theorem 1, we now proceed to study the loss and penalty functions separately.

2.1 Loss functions

We now establish sufficient conditions for the loss function to display the convergence required in Theorem 1. Our results use standard approximations for the loss function in terms of the risk function combined with the Convexity Lemma by Pollard (1991), which is used as a tool to upgrade pointwise convergence results to uniform convergence over compact sets.

Loss Assumptions (LA)

L1. The parameter $\theta = \arg\min_{t\in\Theta} \mathbb{E}\left[L(\mathbf{Z}, t)\right]$ *is bounded and unique;*

L2. $\mathbb{E}\left|L(\mathbf{Z}, t)\right| < \infty$ *for each* t*;*

L3. The loss function $L(Z, t)$ *is such that:*

a) $L(Z, t)$ *is differentiable with respect to* t *at* $t = \theta$ *for* $\mathbb{P}_{\mathbf{Z}}$*-almost every* Z *with derivative* $\nabla_t L(Z, \theta)$ *and*

$$J(\theta) \;:=\; \mathbb{E}\left[\nabla_t L(\mathbf{Z}, \theta) \nabla_t L(\mathbf{Z}, \theta)^T\right] < \infty; \tag{8}$$

b) the risk function $R(t) = \mathbb{E}\left[L(\mathbf{Z}, t)\right]$ *is twice differentiable with respect to* t *at* $t = \theta$ *with positive definite Hessian matrix*

$$\left[H(\theta)\right]_{ij} := \left.\frac{\partial^2 R(t)}{\partial t_i \partial t_j}\right|_{t=\theta} = \left.\frac{\partial^2 \left(\mathbb{E}\left[L(\mathbf{Z}, t)\right]\right)}{\partial t_i \partial t_j}\right|_{t=\theta}; \tag{9}$$

L4. The loss function $L(Z, t)$ *is convex with respect to its argument* t *for* $\mathbb{P}_{\mathbf{Z}}$*-almost every* Z*.*

Assumptions L1-L4 – the L being a mnemonic for the loss function – are relatively mild. The first assumption on the loss function (L1) ensures that the parameter θ in (6) is well defined and is thus a minimal

requirement. Assumption L2 yields that a law of large numbers is valid for each value of t, and thus that the risk function equals the pointwise limit of the empirical risk. In our proofs, assumption L3 is used extensively to obtain local quadratic asymptotic approximations to the risk function around the parameter θ that are pointwise valid around θ (i.e., for each $\theta + \frac{u}{q_n}$ for a sequence $0 < q_n \to \infty$ as $n \to \infty$). The requirement that the risk function is twice differentiable does not require differentiability of the loss function itself, as will become evident in our analysis of the hinge loss in Section 4. Finally, assumption L4 is used to upgrade the local approximation for the risk function from pointwise to uniform over compact sets by means of Pollard's convexity lemma (Pollard, 1991). Alternative assumptions can replace L4: any set of conditions yielding uniform convergence over compact sets will do. One could, for instance, replace it by conditions on the local complexity/entropy of the loss function (see, for instance, Dudley, 1999). We stick to convexity here given its computational convenience and widespread use in statistics and machine learning (Bartlett et al., 2006).

Lemma 2. *Under the LA assumptions L1, L2, and L3:*

a) There exists a p-dimensional random vector $\mathbf{W} \sim N(0, J(\theta))$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\left[L\left(Z_i, \theta + \frac{u}{\sqrt{n}}\right) - L\left(Z_i, \theta\right)\right] - \left[u^T \cdot H(\theta) \cdot u + \mathbf{W}^T \cdot u\right] \xrightarrow{p} 0, \text{ for each } u \in \mathbb{R}^p.$$

b) If, in addition, LA assumption L4 holds, then:

b.1) for every compact subset $K \subset \mathbb{R}^p$,

$$\sup_{u \in K}\left\|\frac{1}{n}\sum_{i=1}^{n}\left[L\left(Z_i, \theta + \frac{u}{\sqrt{n}}\right) - L\left(Z_i, \theta\right)\right] - \left[u^T \cdot H(\theta) \cdot u + \mathbf{W}^T \cdot u\right]\right\| \xrightarrow{p} 0, \text{ and}$$

b.2) $\sqrt{n} \cdot \hat{\theta}_n(0) = O_p(1)$.

Our proof of the pointwise convergence (a) and of boundedness of the M-estimator (b.2) is offered in the Appendix A. It can be seen as an extension of the results for the absolute error loss due to Pollard (1991). The upgrade from pointwise convergence to uniform convergence over compact sets is a direct application of the Convexity Lemma in Pollard (1991).

2.2 Penalty functions

Lemma 3 establish conditions for non-adaptive penalties to satisfy the conditions required by Theorem 1.

Penalty Assumptions (PA)

P1. $T : \Theta \to \mathbb{R}$ *is non-random and* $T(t) \geq 0$ *for all* $t \in \Theta$*;*

P2. T *is continuous in* $t \in \Theta$*;*

P3. The function

$$G_\theta(u) := \lim_{h \downarrow 0} \frac{T(\theta + u \cdot h) - T(\theta)}{h} \tag{10}$$

is well defined and continuous for all $u \in \mathbb{R}^p$*;*

P4. The set $\{t \in \Theta : T(t) \leq c\}$ *is compact for all* $c < T(\theta)$*.*

The set of assumptions P1 through P4 on the penalties – P is a mnemonic for penalty function – is broad enough to encompass all ℓ_γ norms with $\gamma > 0$ and the set of generalized elastic net penalties in Friedman (2008). With minor adjustments, the SCAD penalty (Fan and Li, 2001) can also be treated by our theory. We emphasize that convexity is not a requirement. Non-randomness and continuity (assumptions P1 and P2) make it easy to obtain uniform convergence over compact sets. Condition P3 is similar but milder than a differentiability requirement. We prove that the penalty function converges uniformly over compact sets by using conditions P1 through P3. Condition P4 is useful in ensuring that the penalized estimates are bounded in probability. It amounts to a requirement that the penalty function T constrains the penalized estimates to be within a compact set for all $\lambda > 0$.

Lemma 3. *Let* θ *be as defined in* (6)*,* q_n *be a sequence of non-random positive real numbers satisfying* $q_n \to \infty$ *and* λ_n *be a sequence on non-negative (potentially random) real numbers with* $\lambda_n \cdot q_n^{-1} \xrightarrow{p} \lambda$ *as* $n \to \infty$*. Suppose that the* T *is a penalty function satisfying the PA conditions P1 through P3. Then, for all compact subsets* $K \in \mathbb{R}^p$*:*

$$\sup_{u \in K} \left| \lambda_n \cdot \left[T\left(\theta + \frac{u}{q_n}\right) - T(\theta) \right] - \lambda \cdot G_\theta(u) \right| \to 0, \text{ as } n \to \infty.$$

A proof for Lemma 3 is offered in Appendix A.

2.3 Convergence of penalized empirical risk minimizers

We now state our first main result, which characterizes the asymptotic distribution of penalized parametric M-estimators.

Theorem 4. *Assume λ_n be a sequence of non-negative (potentially random) real numbers such that $\lambda_n \cdot n^{-\frac{1}{2}} \xrightarrow{p} \lambda \geq 0$ as $n \to \infty$. Let θ, $\hat{\theta}_n(\lambda_n)$, $J(\theta)$, $H(\theta)$, and $G_\theta(u)$ be as defined in* (6), (5),(8), (9), *and* (10) *respectively. Define:*

$$V_\theta(w,u) \quad = \quad u^T \cdot H(\theta) \cdot u + w^T \cdot u + \lambda \cdot G_\theta(u), \text{ for } w \in \mathbb{R}^p.$$

If the loss function satisfies the LA assumptions and the penalty function satisfies the PA assumptions, then there exists a p-dimensional random vector $\mathbf{W} \sim N(0, J(\theta))$ such that:

$$\sqrt{n}\left(\hat{\theta}_n(\lambda_n) - \theta\right) \xrightarrow{d} \arg\min_u V_\theta(\mathbf{W}, u).$$

Proof. **Theorem 4.**

In the appendix, we prove that $\hat{\theta}_n(0) = O_p(1)$ implies $\hat{\theta}_n(\lambda_n) = O_p(1)$ for all $\lambda_n > 0$ (Lemma 11). Thus, under the assumptions made, Lemma 2 along with Lemma 11 ensures that conditions (i) and (iii) in Theorem 1 are satisfied. Additionally, Lemma 3 ensures that condition (ii) in Theorem 1 is met. The result then follows directly from Theorem 1. □

We emphasize that the approximation afforded by Theorem 4 is valid for the unique minimizer of the risk function θ as defined in (6). As the penalty function is not assumed to be convex, local minima may exist in finite samples. However, the conditions in Theorem 4 ensure that asymptotically the penalty component of the $V_\theta^{(n)}$ function is negligible in comparison to the risk component and asymptotically the minimizer is unique.

In the next section, we use the asymptotic characterization of the distribution of ℓ_1-norm penalized empirical risk minimizers in Theorem 4 to obtain necessary and sufficient conditions for the existence of a

sequence of tuning parameters λ_n for which $\hat{\theta}_n(\lambda_n)$ is model selection consistent.

3 Model selection consistency of ℓ_1-penalized for M-estimators

Our main result concerning ℓ_1-norm penalized estimates gives necessary and sufficient conditions ensuring the existence of a sequence of regularization parameters λ_n such that $\hat{\theta}_n(\lambda_n)$ correctly identify the signs of the entries in the optimal vector of coefficients θ as defined in (6) as the sample size increases. Before we can state this result, we must introduce some notation and terminology. To allow the usual practice of including non-penalized intercepts to linear models, we write the risk minimizer as $\theta = (\alpha, \beta) \in \mathbb{R}^{p+1}$, where only the coefficients in $\beta \in \mathbb{R}^p$ are included in the ℓ_1-penalty. We define a partition of β in terms of its sparsity pattern:

$$\mathcal{A} = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}, \quad \text{and} \quad \mathcal{A}^c = \{j \in \{1, \ldots, p\} : \beta_j = 0\}.$$

We let q denote the number of indices in the set $\mathcal{A}$. We will say that an estimate $\hat{\theta}_n(\lambda)$ is *sign-correct* if $\text{sign}(\hat{\beta}_n(\lambda)) = \text{sign}(\beta)$, where $\text{sign}(t)$ for a vector $t \in \mathbb{R}^p$ is a p-dimensional vector with:

$$[\text{sign}(t)]_j \quad = \quad \begin{cases} 1, \text{ if } t_j > 0, \\ 0, \text{ if } t_j = 0, \text{ and} \\ -1, \text{ if } t_j < 0. \end{cases} \tag{11}$$

We will say that a sequence of estimates of regularization paths $\hat{\theta}(.) : \mathbb{R} \to \Theta$ is *sparsistent and sign consistent* if there exists a sequence λ_n of (potentially random) non-negative values of the regularization parameters such that

$$\lim_{n \to \infty} \mathbb{P}\left(\hat{\beta}_n(\lambda_n) \text{ is sign correct}\right) = 1.$$

We emphasize that the definition requires only the penalized components of θ to be asymptotically sign-correct.

For a risk function satisfying assumption L2 above, rearrange and partition the $(1 + q + (p - q)) \times$

$(1 + q + (p - q))$ Hessian:

$$H(\theta) \quad = \quad \begin{bmatrix} H_{\alpha,\alpha}(\theta) & H_{\alpha,\mathcal{A}}(\theta) & H_{\alpha,\mathcal{A}^c}(\theta) \\ H_{\mathcal{A},\alpha}(\theta) & H_{\mathcal{A},\mathcal{A}}(\theta) & H_{\mathcal{A},\mathcal{A}^c}(\theta) \\ H_{\mathcal{A}^c,\alpha}(\theta) & H_{\mathcal{A}^c,\mathcal{A}}(\theta) & H_{\mathcal{A}^c,\mathcal{A}^c}(\theta) \end{bmatrix}. \tag{12}$$

Theorem 5. *Let $\hat{\theta}_n(\lambda_n) = \left(\hat{\alpha}_n(\lambda_n), \hat{\beta}_n(\lambda_n)\right)$ be as defined in* (5) *above with an ℓ_1-penalty applied only to the terms in $\hat{\beta}_n(\lambda_n)$. Suppose the loss function satisfy the conditions in Assumption Set 1 and define*

$$\eta(\theta) := 1 - \left\| H_{\mathcal{A}^c,\mathcal{A}}(\theta) \left[H_{\mathcal{A},\mathcal{A}}(\theta) - H_{\mathcal{A},\alpha}(\theta) H_{\alpha,\alpha}(\theta)^{-1} H_{\alpha,\mathcal{A}}(\theta) \right]^{-1} \mathit{sign}(\beta_{\mathcal{A}}) \right\|_{\infty} \geq 0. \tag{13}$$

a) Let λ_n is a sequence of non-negative (potentially random) real numbers such that such that $\lambda_n \cdot n^{-1} \xrightarrow{p} 0$, and $\lambda_n \cdot n^{-\frac{1+c}{2}} \xrightarrow{p} \lambda > 0$ for some $0 < c < \frac{1}{2}$ as $n \to \infty$. If $\eta(\theta) > 0$, then:

$$\mathbb{P}\left[\mathit{sign}\left(\hat{\beta}_n(\lambda_n)\right) = \mathit{sign}(\beta) \right] \geq 1 - \exp[-n^c].$$

b) Conversely, if $\eta(\theta) < 0$, then, for any sequence of non-negative numbers λ_n

$$\lim_{n\to\infty} \mathbb{P}\left[\mathit{sign}\left(\hat{\beta}_n(\lambda_n)\right) = \mathit{sign}(\beta) \right] < 1.$$

The result in Theorem 5 extends the model selection consistency results in Zhao and Yu (2006) concerning LASSO estimates (based on L_2-loss) to more general parametric estimates defined as ℓ_1-norm penalized M-estimators based on loss functions satisfying the conditions in Assumption Set 1. We will call the condition in (13) the *generalized irrepresentability condition* (GI condition) which in the case of the L_2-loss with zero-mean predictors recovers Zhao and Yu's irrepresentable condition. Accordingly, we call $\eta(\theta)$ the *GI index*, which can be interpreted as a measure of incoherence between the active and inactive predictors. Positive values of $\eta(\theta)$ imply the effects of active and inactive predictors are distinguishable enough so the ℓ_1-penalized estimate can correctly identify the signs of all coefficients in the optimal model given a sufficiently large sample size.

A condition similar to the generalized irrepresentability condition (13) appears in Ravikumar et al.

(2008). There, the GI condition is used to obtain sufficient conditions for the consistent selection of the terms of an infinite dimensional precision matrix estimate defined as the ℓ_1-norm penalized minimizer of the log-likelihood loss for Gaussian distributions. This suggests it is possible to extend Theorem 5 to the non-parametric setting where the number of regressors p grows with the sample size n (i.e., $p = p_n \to \infty$ as $n \to \infty$). Such extension will be the subject of future research.

Finally, we would like to emphasize that, even if $\eta(\theta) < 0$, it may be possible to correctly recover the signs of β with a relatively high probability. What the converse in Theorem 5 says is that this probability is bounded away from 1 in the limit.

3.1 Simplification of the GI condition under linear models and Gaussian predictors

Our next result gives sufficient conditions for the $\eta(\alpha, \beta)$ to be computable directly from the covariance of the predictors. This result is limited to ℓ_1-penalized linear models as defined in (2). Since the loss function only depends on $\mathbf{X}$ through an affine transformation, the Hessian $H(a, b)$ of the risk function $R(a, b)$ as well as the covariance matrix of scores $J(a, b)$ involves the expected value of an expression involving the second order cross products in the matrix $Q(\mathbf{X})$ defined as

$$Q(\mathbf{X}) \;:=\; \begin{bmatrix} 1 & \mathbf{X}^T \\ \mathbf{X} & \mathbf{X}\mathbf{X}^T \end{bmatrix}. \tag{14}$$

Theorem 6. *Let the coefficients of a linear model (α, β) be as defined in* (4)*. If*

a) $\mathbf{X} \sim N(0, \Sigma)$*, and*

b) the Hessian of the risk function in (3) *can be written in the form*

$$H(\alpha, \beta) \;=\; \mathbb{E}\left[\mathbb{E}\left[Q(\mathbf{X}) \,\middle|\, \alpha + \mathbf{X}^T\beta\right] \cdot w(\alpha + \mathbf{X}^T\beta)\right], \text{ for some function } w : \mathbb{R} \to \mathbb{R}, \tag{15}$$

then $\eta(\alpha, \beta) = 1 - \left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}^c}\mathbf{X}_{\mathcal{A}}^T\right)\right]\left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T\right)\right]^{-1} \text{sign}(\beta_{\mathcal{A}})$.

A proof is given in Appendix A. Theorem 6 tells us that, for zero-mean Gaussian predictors and loss functions whose Hessian can be expressed as a weighted "average" of second moments of $\mathbf{X}$ conditional on

the linear predictor variable $\mathbf{M}_{\alpha,\beta}(\mathbf{X}) := \alpha + \mathbf{X}^T \cdot \beta$, the GI condition can be computed directly from the matrix of second moments $\mathbf{E}\left[\mathbf{X}\mathbf{X}^T\right]$. In Section 4, we will see that Theorem 6 holds for linear SVM and logistic regression classifiers.s Besides the particular cases studied in Section 4, we notice that Theorem 6 can find ample use for ℓ_1 penalized estimates in view of our next result.

Corollary 7. *Suppose that:*

a) $\mathbf{X} \sim N(0, \Sigma)$,

b) $L(\mathbf{Z}, t) = L(\mathbf{Y}, a + b^T\mathbf{X})$,

c) $L(\mathbf{Y}, a + \mathbf{X}^T b)$ *is twice differentiable in its second argument for almost every* $\mathbf{Y}$, *and*

d) $\mathbf{Y} \perp \mathbf{X}|\alpha + \beta^T\mathbf{X}$.

Then, $\eta(\alpha, \beta) = 1 - \left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}^c}\mathbf{X}_{\mathcal{A}}^T\right)\right]\left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^T\right)\right]^{-1}$ *sign*$(\beta_{\mathcal{A}})$.

Proof. Let $\frac{\partial^2 L(\mathbf{Y}, \alpha + \mathbf{X}^T\beta)}{(\partial v)^2}$ denote the second derivative of L with respect to its second argument. Since $\mathbf{Y} \perp \mathbf{X}|\alpha + \beta^T\mathbf{X}$, we get

$$\begin{aligned} H(\alpha, \beta) &= \mathbb{E}\left[\mathbb{E}\left[Q(\mathbf{X}) \cdot \frac{\partial^2 L(\mathbf{Y}, \alpha + \mathbf{X}^T\beta)}{(\partial v)^2}\,\middle|\,\alpha + \mathbf{X}^T\beta\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[Q(\mathbf{X})\,\middle|\,\alpha + \mathbf{X}^T\beta\right] \cdot \mathbb{E}\left[\frac{\partial^2 L(\mathbf{Y}, \alpha + \mathbf{X}^T\beta)}{(\partial v)^2}\,\middle|\,\alpha + \mathbf{X}^T\beta\right]\right]. \end{aligned}$$

Condition (b) in Theorem 6 is thus satisfied with $w(\alpha + \beta^T\mathbf{X}) = \mathbb{E}\left[\frac{\partial^2 L(\mathbf{Y}, \alpha + \mathbf{X}^T\beta)}{(\partial v)^2}\,\middle|\,\alpha + \mathbf{X}^T\beta\right]$. □

Corollary 7 shows that, if the predictors are Gaussian and the response $\mathbf{X}$ only depends on $\mathbf{X}$ through an affine transform, the conditions for model selection consistency of many Generalized Linear Models (GLMs Nelder and Wedderburn, 1972) only depends on the covariance between relevant and irrelevant predictors even if the model is not correctly specified. For canonical GLMs, condition (d) can be relaxed as the weight function can be shown not to depend on the response $\mathbf{Y}$.

As we will see in the case of the hinge loss, twice differentiability of the loss with respect to its second argument is not essential. For condition (b) in Theorem 6 to be satisfied, what seems to be essential is that the loss has the form shown in (2) and that $\mathbf{Y}$ is conditionally independent of $\mathbf{X}$ given $\alpha + \mathbf{X}^T\beta$.

4 Application to SVM and logistic regression classifiers

We now obtain the limiting behavior of some linear classifiers to study the model selection consistency of their ℓ_1-penalized estimates. We will use these results along with Theorem 5 to study the model selection consistency of ℓ_1-penalized SVM and logistic regression classifiers. The response variable $\mathbf{Y} \in \{-1, 1\}$ is modeled in terms of a linear transformation of a set of predictors $\mathbf{X} \in \mathbb{R}^p$. Setting some of the coefficients on the estimates of the β parameter to zero corresponds to eliminating some effects from the model thus leading to more interpretable models.

In what follows, we will characterize the asymptotic behavior of the loss functions associated to logistic regression and support vector machines. Logistic regressions are a particular case of Generalized Linear Models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) and are widely used by statisticians when modeling the outcome of binomial variables. Support vector machines (Cortes and Vapnik, 1995) are amply used for obtaining linear classification rules and is based on the hinge-loss function. For both the logistic regression and support vector machines, the corresponding loss functions are often interpreted as convex surrogates for the $0 - 1$ classification loss (Zhang, 2004; Bartlett et al., 2006). Efficient algorithms exist for obtaining both the ℓ_1-norm penalized SVM (Zhu et al., 2004) and logistic (Park and Hastie, 2006) classifiers. Both SVM classification and logistic regression have been used to select relevant predictors in areas as diverse as genomics (see, for instance Guyon et al., 2002; Meier et al., 2006) and text categorization (see, for instance Joachims, 1998; Genkin et al., 2007).

We now set up terminology and notation we will use in connection with the SVM and logistic classifiers for the remainder of the paper. Given a value for the parameters in the linear classification model $t = (a, b) \in \mathbb{R}^{1+p}$, a *linear classification rule* is defined as

$$\hat{\mathbf{Y}}(\mathbf{X}|t) = \text{sign}\left(a + \mathbf{X}^T b\right). \tag{16}$$

The separating hyperplane $\mathcal{H}(t)$ associated to a linear classification rule as in (16) is defined as

$$\mathcal{H}(t) \quad := \quad \{\mathbf{x} \in \mathbb{R}^p : a + \mathbf{x}b = 0\}, \text{ for } t = (a, b). \tag{17}$$

The set $\mathcal{H}(t)$ defines the boundary in the predictor space between the points where, for the linear classification rule based in t, the response variable is predicted to be 1 (the set $\{\mathbf{x} : \hat{\mathbf{Y}}(\mathbf{x}|t) = 1\} = \{\mathbf{x} : a + \mathbf{x}^t b > 0\}$) from the points where $\mathbf{Y}$ is predicted to be -1 (the set $\{\mathbf{x} : \hat{\mathbf{Y}}(\mathbf{x}|t) = 1\} = \{\mathbf{x} : a + \mathbf{x}^t b < 0\}$). We call *optimal linear classification rule* the classification rule corresponding to setting $t = \theta$ and the *estimated linear classification rules* the classification rule formed by setting $t = \hat{\theta}_n(\lambda_n)$ with $\hat{\theta}_n(\lambda)$ as defined in (5). We define the *linear predictor variable*:

$$\mathbf{M} \quad := \quad \alpha + \mathbf{X}^T \beta, \tag{18}$$

which measures the distance from point $\mathbf{X}$ to the separating hyper-plane defined by the optimal linear classifier. If the distribution of $\mathbf{Y}$ only depends on $\mathbf{X}$ through a linear combination, both the linear SVM and logistic regression are known to recover the optimal Bayes classifier. We also define the true conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ as:

$$p(\mathbf{X}) = \mathbb{P}\left(\mathbf{Y} = 1 | \mathbf{X}\right). \tag{19}$$

4.1 Regularity conditions and model selection consistency for SVM and logistic classifiers

Before we can use the results from Section 3 to study and compare the ℓ_1-penalized SVM and logistic linear classifiers, we must obtain a set of conditions on the joint distribution of $(\mathbf{X}, \mathbf{Y})$ such that the hinge and logistic regression losses satisfy the requirements on loss functions laid out in Assumption Set 1. Conditions C1-C3 – C a mnemonic for classification – gives one such a set of sufficient conditions in terms of the marginal distribution of the predictors $\mathbf{X}$ and the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$.

Classification Assumptions (CA)

C1. $\mathrm{var}\left[\mathbf{X} \,|\mathbf{Y}\,\right] \in \mathbb{R}^{p \times p}$ *is a positive definite matrix for* $\mathbf{Y} \in \{1, -1\}$,

C2. The distribution of $\mathbf{X}$ *has a density* $f_{\mathbf{X}}(\mathbf{x}) > 0$, *for all* $\mathbf{x} \in \mathbb{R}^p$, *and*

C3. $p(\mathbf{X}) \in (0, 1)$ *for almost every* $\mathbf{X}$, *that is, for all values* $\mathbf{X}$ *in the support of the distribution of* $\mathbf{X}$, $\mathbf{Y}$ *can assume any of its two possible values;*

Condition C1 rules out the case of perfectly correlated predictors and is required to ensure uniqueness of the minimizer θ as defined in (6). Assumptions C2 and C3 are used to ensure the SVM and logistic regression loss functions satisfy the assumptions in Lemma 2, but can be relaxed.

The remainder of this section describes linear SVM and logistic classification, shows how Conditions C1-C3 ensure their corresponding loss functions are amenable to the theory laid out in Section 3 and provide expressions for the covariance matrix of scores $J(\theta)$ and the Hessian $H(\theta)$ for the risk functions associated to the linear SVM and logistic regression classifiers.

4.1.1 Logistic Regression

The canonical logistic regression is one instance of Generalized Linear Model (Nelder and Wedderburn, 1972) where the probability of $\mathbf{Y} = 1$ is modeled as:

$$\mathbb{P}\left(\mathbf{Y} = 1 | \mathbf{X}, a, b\right) \quad = \quad \frac{\exp(a + \mathbf{X}^T b)}{1 + \exp(a + \mathbf{X}^T b)}, \tag{20}$$

where $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$ are parameters to be determined. The population parameters α and β are defined as the minimizers of the Kullbach-Leibler divergence between the true conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ and the Bernoulli distribution with parameter given by (20). The corresponding loss function is:

$$L\left(\mathbf{Y}, a + b^T \mathbf{X}\right) \quad = \quad -a \cdot \mathbb{I}\left(\mathbf{Y} = 1\right) - \mathbb{I}\left(\mathbf{Y} = 1\right) \cdot \mathbf{X}^T \cdot b + \log\left[1 - \exp\left(a + \mathbf{X}^T \cdot b\right)\right], \tag{21}$$

where $\mathbb{I}\left(\mathbf{Y} = 1\right)$ is the indicator of $\mathbf{Y} = 1$. An estimate for $\theta = (\alpha, \beta)$ is obtained by minimizing the empirical risk with respect to $t = (a, b)$.

Lemma 8. *Suppose that the conditions in Assumption Set 3 are observed. Then, the logistic regression loss function* (21) *satisfies the conditions in Assumption Set 1 with:*

$$\begin{aligned} J(\theta) \quad &= \quad \mathbb{E}\left[Q(\mathbf{X}) \cdot \left[p\left(\mathbf{X}\right) - 2 \cdot p\left(\mathbf{X}\right) \cdot \frac{\exp\left(\alpha + \mathbf{X}\beta\right)}{1 + \exp\left(\alpha + \mathbf{X}\beta\right)} + \left(\frac{\exp\left(\alpha + \mathbf{X}\beta\right)}{1 + \exp\left(\alpha + \mathbf{X}\beta\right)}\right)^2\right]\right], \text{ and} \\ H(\theta) \quad &= \quad \mathbb{E}\left[Q(\mathbf{X}) \cdot \frac{\exp\left(\alpha + \mathbf{X}\beta\right)}{\left(1 + \exp\left(\alpha + \mathbf{X}\beta\right)\right)^2}\right]. \end{aligned}$$

A proof is given in Appendix A. The expression for the Hessian of the logistic loss can be rewritten as

$$H(\theta) \quad = \quad \mathbb{E}\left[\mathbb{E}\left[Q(\mathbf{X})\,\middle|\,\alpha+\mathbf{X}\beta\right]\cdot\frac{\exp\left(\alpha+\mathbf{X}\beta\right)}{\left(1+\exp\left(\alpha+\mathbf{X}\beta\right)\right)^2}\right], \tag{22}$$

and hence satisfies the conditions of Theorem 6 even if the model is not correctly specified. Indeed, the Hessian for the logistic risk does not depend on the distribution of $\mathbf{Y}$ at all.

In addition, equation (22) tells us that the Hessian for the logistic regression risk function is a weighted average of second moment matrices conditional on the linear predictor variable $\alpha+\mathbf{X}\beta$. Because $\frac{\exp(\alpha+\mathbf{X}\beta)}{(1+\exp(\alpha+\mathbf{X}\beta))^2}$ is an even function of the linear predictor variable, the matrices of conditional second moments at predictor variables that are equally distant from the separating hyperplane are equally weighted. In addition, the higher weight is given to $\mathbb{E}\left[Q(\mathbf{X})\,\middle|\,\alpha+\mathbf{X}\beta=0\right]$ and the weighting is decreasing on the absolute value of the linear predictor variable. As a result, in what concerns asymptotic model selection consistency of ℓ_1-norm penalized logistic coefficient estimates, the correlation structure of the predictors on regions closer to the separating hyperplane have the most importance confirming the margin phenomenon observed earlier in non-parametric works by Audibert and Tsybakov (2007) and Steinwart and Scovel (2007).

4.1.2 The parametric SVM: linear classification with the Hinge loss function

Classification by means of Support Vector Machines with linear kernel was first introduced in the case where it is possible to perfectly separate the space of predictors $\mathbf{X}$ according to the the binomial variable $\mathbf{Y}$. In that setting, the SVM parameters define a hyper-plane (characterized by the parameters α, β) that maximizes the gap between the classes:

$$\begin{aligned}(\hat{\alpha},\hat{\beta}) \quad = \quad & \underset{a,b}{\arg\min} && \|b\|_2 \\ & \text{s.t.} && \mathbf{Y}_i\cdot\left(a-\mathbf{X}_i^T b\right)\geq 1, \quad \text{for all } i=1,\ldots,n.\end{aligned}$$

To adapt this method to the "no perfect-separation" case, non-negative slack variables ξ_i are introduced and the optimization problem becomes

$$\begin{aligned}(\hat{\alpha}, \hat{\beta}) &= \underset{a,b}{\arg\min} && \|\beta\|_2 + C \cdot \textstyle\sum_{i=1} \xi_i \\ &&& \text{s.t.} && \mathbf{Y}_i \cdot \left(a - \mathbf{X}_i^T b\right) \geq 1 - \xi_i, \quad \text{for all } i = 1, \ldots, n, \text{ and} \\ &&&&& \xi_i \geq 0, \qquad\qquad \text{for all } i = 1, \ldots, n,\end{aligned}$$

where C is a constant controlling the trade-off between margin maximization and total amount of slack. The "lack of fit" in SVM is measured by the total distance of the misclassified points to the classification boundary, represented as the sum of the slack variables. The Euclidean norm acts as a penalization term: in the perfect separation case it unsureness uniqueness of the solution. More consistently with the form in (5), the empirical SVM parameter estimates can then be rewritten as:

$$\begin{aligned}(\hat{\alpha}, \hat{\beta}) &= \underset{a,b}{\arg\min} \sum_{i=1} L(\mathbf{Y}_i, a + b^T \mathbf{X}) + \lambda \cdot \|\beta\|_2, \text{ with} \\ L(\mathbf{Y}_i, a + b^T \mathbf{X}) &= \left[1 - \mathbf{Y}_i \left(a - \mathbf{X}_i^T b\right)\right] \cdot \mathbb{I}\left[1 - \mathbf{Y}_i \cdot \left(a - \mathbf{X}_i^T b \geq 0\right)\right], \text{ the } \textit{hinge-loss function.}\end{aligned}$$

Here, we will consider the hinge loss on its own, in the spirit of the "assembling" Lemma 1. The next result establishes that under the conditions of Assumption Set 3, the hinge loss satisfies the assumptions in Theorem 2.

Lemma 9. *Suppose the conditions in Assumption Set 3 hold. If in addition $\beta \neq 0$, then the hinge loss function* (23) *satisfies the conditions in Assumption Set 1 with:*

$$\begin{aligned}J(\theta) &= \mathbb{E}\left[\left[p(\mathbf{X}) \cdot \mathbb{I}(1 - \alpha - \mathbf{X}^T\beta \geq 0) + (1 - p(\mathbf{X})) \cdot \mathbb{I}(1 + \alpha + \mathbf{X}^T\beta \geq 0)\right] \cdot Q(\mathbf{X})\right], \textit{ and} \\ H(\theta) &= \mathbb{E}\left[\left[p(\mathbf{X}) \cdot \delta\left(1 - \alpha - \mathbf{X}^T\beta\right) + (1 - p(\mathbf{X})) \cdot \delta\left(1 + \alpha + \mathbf{X}^T\beta\right)\right] \cdot Q(\mathbf{X})\right],\end{aligned}$$

where δ denotes Dirac delta function.

The expressions for $J(\theta)$ and $H(\theta)$ in Lemma 9 closely parallel results by Koo et al. (2008) concerning the Bahadur representation of the linear support vector machines. In Appendix A, we present an alternative

proof similar in spirit to the construction by Phillips (1991). In Koo et al. (2008) conditions ensuring $\beta \neq 0$ are also obtained.

Borrowing from the terminology for support vector regression, we call the set where $\alpha + \mathbf{X}^T\beta = -1$ the negative "elbow" of the SVM risk. Similarly, the positive "elbow" of the SVM risk is the set where $\alpha + \mathbf{X}^T\beta = 1$. Assuming that $\mathbf{Y}$ is independent of $\mathbf{X}$ given $\alpha + \mathbf{X}^T\beta$, the expression for the Hessian in Lemma 9 can be rewritten in a more revealing form in terms of conditional expectations at these elbows of the SVM risk:

$$\begin{aligned} H(\theta) &= \mathbb{E}\left[\mathbb{E}\left[Q(\mathbf{X})\left|\alpha+\mathbf{X}^T\beta\right.\right]\cdot\mathbb{P}\left(\mathbf{Y}=1\left|\alpha+\mathbf{X}^T\beta\right.\right)\cdot\delta(1-\alpha-\mathbf{X}^T\beta)\right] \\ &\quad +\mathbb{E}\left[\mathbb{E}\left[Q(\mathbf{X})\left|\alpha+\mathbf{X}^T\beta\right.\right]\cdot\left[1-\mathbb{P}\left(\mathbf{Y}=1\left|\alpha+\mathbf{X}^T\beta\right.\right)\right]\cdot\delta(-1-\alpha-\mathbf{X}^T\beta)\right] \\ &= \mathbb{E}\left[Q(\mathbf{X})\left|\alpha+\mathbf{X}^T\beta=1\right.\right]\cdot\mathbb{P}\left(\mathbf{Y}=1\left|\alpha+\mathbf{X}^T\beta=1\right.\right)\cdot\tilde{f}(1) \\ &\quad +\mathbb{E}\left[Q(\mathbf{X})\left|\alpha+\mathbf{X}^T\beta=-1\right.\right]\cdot\mathbb{P}\left(\mathbf{Y}=-1\left|\alpha+\mathbf{X}^T\beta=-1\right.\right)\cdot\tilde{f}(-1), \end{aligned} \tag{23}$$

where $\tilde{f}$ denotes the density of the linear predictor variable $\alpha + \mathbf{X}^T\beta$. This representation for the Hessian of the linear SVM risk (expected value of the hinge loss over $\mathbf{Y}$ and $\mathbf{X}$) shows that if $\mathbf{Y}$ is independent of $\mathbf{X}$ given $\alpha + \mathbf{X}^T\beta$ the hinge loss function is amenable to the results in Theorem 6. It also provides many insights into the behavior of the linear SVM classifier.

Equation (23) tells us that the Hessian of the SVM risk is a weighted sum of the conditional second moments of the predictors given that the linear predictor variable $\alpha + \mathbf{X}^T\beta$ is at the elbows of the SVM risk. According to Theorem 5, the generalized irrepresentability condition is not affected if the Hessian matrix is multiplied by a constant. It follows that, with respect to model selection consistency of ℓ_1-norm penalized linear SVM classifiers, the scalar factors $\mathbb{P}\left(\mathbf{Y}=-1\left|\alpha+\mathbf{X}^T\beta=-1\right.\right)\cdot\tilde{f}(-1)$ and $\mathbb{P}\left(\mathbf{Y}=-1\left|\alpha+\mathbf{X}^T\beta=-1\right.\right)\cdot\tilde{f}(-1)$ only determine the relative importance of the two conditional second moment matrices, $\mathbb{E}\left[Q(\mathbf{X})\left|\alpha+\mathbf{X}^T\beta=1\right.\right]$ and $\mathbb{E}\left[Q(\mathbf{X})\left|\alpha+\mathbf{X}^T\beta=-1\right.\right]$, in the composition of the Hessian. If the two conditional moment matrices happen to be equal, the scalar factors have no bearings in whether the generalized irrepresentable condition is met or not. If the two conditional moment matrices are different, the relative importance of the conditional second moments at the two elbows depends on the density of the linear predictor variable $\alpha + \mathbf{X}^T\beta$ and how well defined a class is at each of the el-

bows. For example, if $\tilde{f}(1) \gg \tilde{f}(-1)$ and $\mathbb{P}\left(\mathbf{Y} = 1 \,\middle|\, \alpha + \mathbf{X}^T\beta = 1\right) \gg \mathbb{P}\left(\mathbf{Y} = -1 \,\middle|\, \alpha + \mathbf{X}^T\beta = -1\right)$, the SVM Hessian will be largely determined by the second moment of the predictor at the positive elbow $\mathbb{E}\left[Q(\mathbf{X}) \,\middle|\, \alpha + \mathbf{X}^T\beta = 1\right]$, which in turn will have the most influence in determining whether ℓ_1-norm penalized SVM classifier is model selection consistent.

In addition to determining the weighting between the conditional covariances, the density of the predictors and the probabilities of $\mathbf{Y}$ belonging to each class on the positive and negative can inflate or deflate the covariance matrix of $\hat{\theta}_n$. Standard results concerning parametric M-estimators (see, for instance Bickel and Docksum, 2001; Casella and Berger, 2001) yield that $\lim_{n\to\infty} \text{var}\left[\sqrt{n} \cdot \left(\hat{\theta}_n - \theta\right)\right] = H^{-1}(\theta)J(\theta)H^{-1}(\theta)$. As a result, the higher the density of the predictors and the easier the separation of the classes at the elbows, the larger the Hessian and the less variable the coefficients in the SVM classifier.

5 Simulations

We now present a series of simulation results which give empirical evidence supporting the theory for model selection consistency for ℓ_1-penalized linear SVM and logistic regression classifiers. In addition, we use the simulations to compare the model selection performance of ℓ_1-penalized linear SVM and logistic regression classifiers asymptotically and in finite samples. To avoid a simulation set-up that is biased in favor of either linear SVMs or logistic regression, we base our conclusions on randomly selected joint distributions for $(\mathbf{Y}, \mathbf{X})$, where $\mathbf{Y}$ is the binomial response variable and $\mathbf{X}$ is the predictor. We start off by detailing how the designs used throughout this section are sampled.

5.1 Randomly constructing joint distributions $(\mathbf{Y}, \mathbf{X})$

Throughout our simulation experiments, we will call a design the joint distribution of $(\mathbf{Y}, \mathbf{X})$ characterized by the parameters of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ and the the distribution of the predictors $\mathbf{X}$.

The conditional distribution of the binomial random variable $\mathbf{Y} \in \{-1, 1\}$ given $\mathbf{X} \in \mathbb{R}^p$ is characterized by a *probability profile function* $g : \mathbb{R} \to (0, 1)$, an intercept ζ and a normal direction to the separating hyperplane $\nu \in \mathbb{R}^p$. Given these elements, we set $\mathbb{P}(\mathbf{Y} = 1|\mathbf{X}) = g(\zeta + \mathbf{X}^T\nu)$, so $\mathbf{Y}$ is independent of $\mathbf{X}$ given any one-to-one transformation of $\zeta + \mathbf{X}^T\nu$, in particular $\alpha + \mathbf{X}^T\beta$. In all designs, we set $\zeta = 0$. Given

a number of non-zero terms q, we partition the normal direction according to $\nu = \begin{bmatrix} \nu_{\mathcal{A}} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^q \times \mathbb{R}^{p-q}$. The non-zero component of the normal direction to the separating hyper-plane $\mathbf{v}_{\mathcal{A}}$ is sampled uniformly on the unit sphere on $\mathbb{R}^q$. One problem with this sampling scheme is that it may result in tiny coefficients which are hard to detect in finite samples, thus complicating the comparison between asymptotic and experimental results. To avoid such tiny coefficients, we discard directions $\nu_{\mathcal{A}}$ having $\max_{1\le j\le q} |\nu_j| / \min_{1\le j\le q} |\nu_j| > 5$. To provide stronger evidence in favor of Theorem 5, we will consider two different probability profile functions g:

$$\begin{array}{rrcll} \text{the logistic function,} & g_1(r) & := & \frac{\exp(r)}{1+\exp(r)}, & \text{and} \\ \text{the ``blip'' function,} & g_2(r) & := & \frac{1}{2}\left(1 + r \cdot \exp\left(\frac{1-r^2}{2}\right)\right). & \end{array}$$

The logistic function (g_1) is the canonical link for Bernoulli GLM models. The "blip" function (g_2) concentrates all the action close to the separation boundary between the classes and is thus expected to favor SVM classifiers.

For the distribution of the predictors, we consider two families of distributions: Gaussian and mixture of Gaussian distributions. For the Gaussian predictors, the mean is fixed at $\mathbf{0} \in \mathbb{R}^p$ and a covariance matrix $\Sigma \in \mathbb{R}^p$ is sampled as follows. First, $\tilde{\Sigma} \in \mathbb{R}^p$ is sampled from a Wishart$(\mathbf{I}_p, p, p)$ distribution, where $\mathbf{I}_p$ is the identity matrix. Then, $\tilde{\Sigma} \in \mathbb{R}^p$ is normalized to have unit diagonal and $\Sigma = \gamma \cdot \tilde{\Sigma}$ with the scalar $\gamma > 0$ chosen so that $\nu^T \Sigma \nu = \sigma^2$, where σ^2 is a parameter controlling the variance of $\mathbf{X}$. The mixed Gaussian predictors are a mixture of two Gaussian distributions with equal proportions, common variance Σ and symmetric means μ and $-\mu$. The parameter μ is randomly selected as $\mu = \frac{4}{5} \cdot \tilde{\mu} \cdot \sigma$, where $\tilde{\mu} = |\chi| \cdot \nu + \mathrm{w}$, with $\chi \sim N(0,1)$ and $\mathrm{w} \sim N(0, \mathbf{I}_p)$. The common variance matrix of the components of the mixture of Gaussian is sampled similarly as the covariance matrix for the Gaussian case, with the difference that γ is chosen so $\nu^T \Sigma \nu = \frac{9}{25} \cdot \sigma^2$. The factors $\frac{4}{5}$ for μ and $\frac{9}{25}$ for Σ are used to ensure that the contribution of the mean and variance for the second moment $\mathbb{E}\mathbf{X}\mathbf{X}^T = \mu\mu^T + \Sigma$ is somewhat balanced.

To obtain the population parameter $\theta = (\alpha, \beta)$ as defined in (6) for each of the sample designs, we first notice that the probability profile functions satisfy $g_j(z) = 1 - g_j(-z)$, for $j = 1, 2$, $z \in \mathbb{R}$ and the distribution of the predictors are symmetric about zero. It thus follows that the optimization problem defining θ is symmetric about $0 \in \mathbb{R}^p$ and we have $\alpha = 0$ for all designs. Then, because $\mathbb{P}(\mathbf{Y} = 1|\mathbf{X})$

only depends on $\mathbf{X}$ through $\mathbf{X}^T\nu$, β has the form $\beta = c^* \cdot \nu$, for some scalar $c^* \in \mathbb{R}$. The value of c^* that minimizes the risk is obtained by numerically minimizing the average of the risk function conditional on $\mathbf{X}$ for a large sample (10^6) from the predictor distribution. For any given design, the value of c^* differs depending on the risk function being used.

5.2 Model selection consistency and the GI condition for linear classifiers

We now provide empirical evidence of the validity of Theorem 5 for ℓ_1-norm penalized linear SVM and logistic regression classifiers. According to Theorem 5, the proportion of paths containing sign-correct estimates should approach 1 as $n \to \infty$ if the GI index $\eta(\theta)$ is positive.

To estimate the probability that a sample regularization path contains a sign-consistent estimate for a given design, we used replicates of the regularization path by sampling from the joint distribution of $(\mathbf{Y}, \mathbf{X})$ and computing the regularization path for ℓ_1-norm penalized linear SVM (Li and Zhu, 2008) and logistic regression (Park and Hastie, 2006). To compute the GI index for a given design, we can use the expressions in equations (22) and (23) in conjunction with the expressions for the conditional second moments of Gaussian and mixed Gaussian random variables shown in Appendix B.

Figures 1 and 2 show plots of the proportion of sample regularization paths containing a sign-correct solution against the GI index $\eta(\theta)$ under various conditions. In all cases considered, the proportion of times the ℓ_1-penalized classifier contains a sign-correct estimate in its regularization path increases as n increases if $\eta(\theta) > 0$.

Figures 1 and 2 also show that that, in most cases, correct recovery of the signs of θ is harder if $\eta(\theta) < 0$. One notable exception occurs for mixed Gaussian predictors under the "blip" conditional probability profile. In that case, it is possible to have a high probability of correct sign recovery even under $\eta(\theta) < 0$. Notice that this result does not contradict Theorem 5. Even though there the probability that the signs will not be recovered correctly is never zero if $\eta(\theta) < 0$, it can be quite small. A more careful analysis of the probability of correct sign recovery must take into account the variance of the estimates $\hat{\beta}_j^{(n)}(\lambda_n)$ with indices in $\mathcal{A}^c = \{j \in 1, \ldots, p : \beta_j = 0\}$.

It also possible to notice that, given the asymptotic nature of the results, the probability of correct sign-recovery can still be small for smaller sample sizes n and for larger number of predictors p especially under

a fainter signal ("blip" conditional probability profile). The extension to the theory in Section 3 to the non-parametric case $p = p_n \to \infty$ can potentially offer more precise answers on the how the total number of predictors affects the chance that the regularization path contains a sign-correct model.

5.3 Comparison of ℓ_1-penalized SVM and logistic regression classifiers

In addition to allowing us to study the model selection consistency of SVM and logistic classifiers, Theorem 5 along with Lemmas 8 and 9 lets us to shed some light onto a question often asked by practitioners: which of SVM and logistic regression classifiers should be used for variable selection? Our theoretical and experimental results suggest that, if variable selection is made through ℓ_1-penalization, the answer depends critically on the sample size available.

5.3.1 Large sample (asymptotic) comparison

If a large enough sample size is available, Theorem 6 suggests that in terms of variable selection by means of ℓ_1-norm penalized estimates logistic and SVM are equally likely to be model selection consistent for the designs sampled as described in 5.1. For non-Gaussian predictors, a comparison of the GI indices $\eta(\theta)$ shows that model selection consistency can be theoretically guaranteed for logistic regression classifiers in more designs than SVM. The results are shown in Figure 3 and Table 1. Interestingly, for the distribution of designs considered, logistic was more likely to be model selection consistent even under the "blip" conditional probability profile function – thought to favor SVM by concentrating must of the class discrimination information on a band around the optimal separating hyperplane.

5.3.2 Finite sample (asymptotic) comparison

Figure 4 shows a comparison of the proportion of times the ℓ_1-penalized logistic and SVM regularization paths contained a model with correctly selected variables. In each plot, each point is obtained by plotting the proportion of paths containing a sign-correct model for logistic (vertical axis) against the same proportion for SVM for a given design. Thus, the further a point sits to the lower right corner, the better was the performance of SVM in comparison to logistic for that specific design. The proportions are obtained from 50 replications of one of the designs sampled as described in Section 5.1.

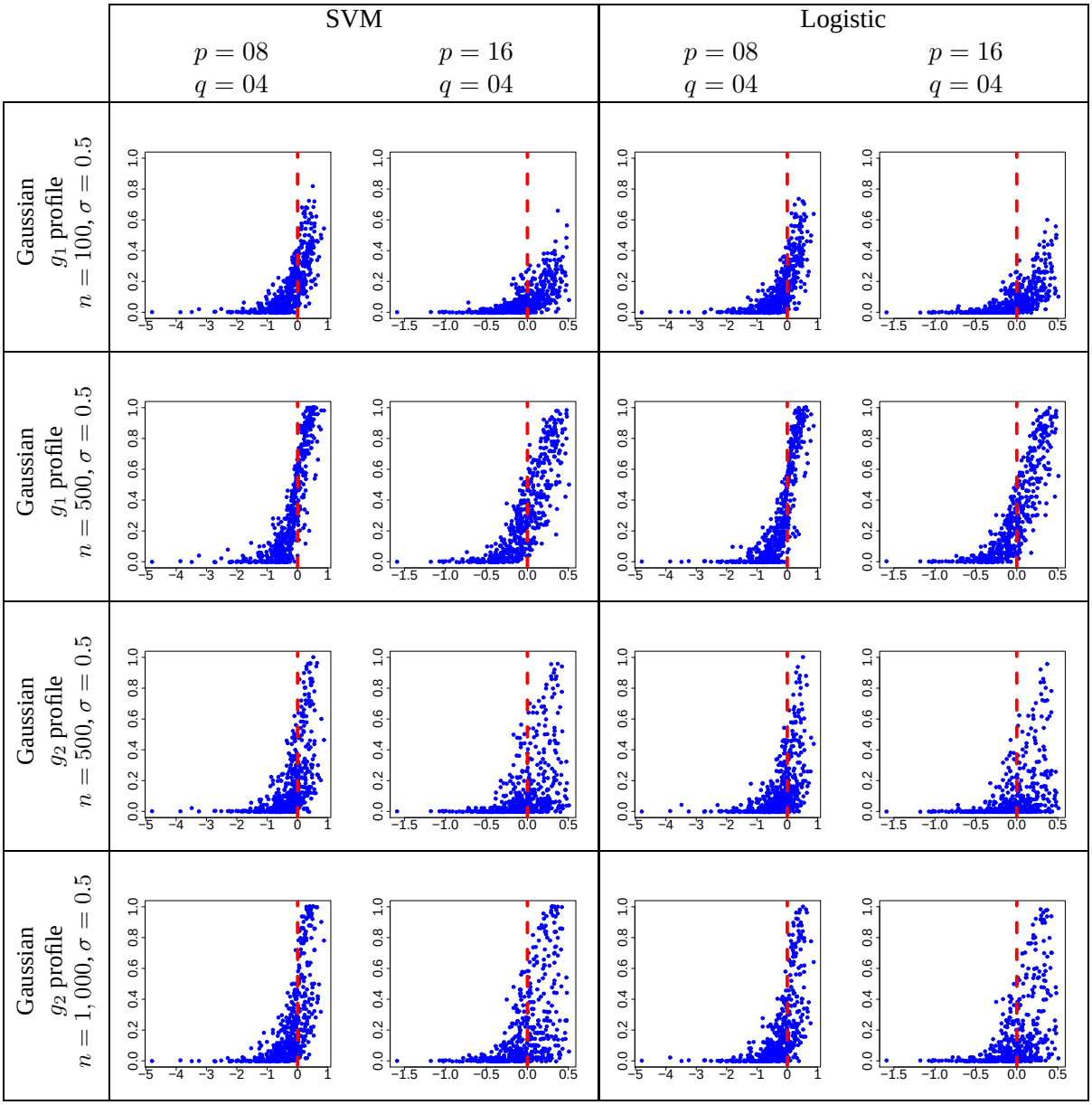

Figure 1: Proportion of sample regularization paths containing a sign-model vs. GI index $\eta(\theta)$ under Gaussian predictors: The proportion at each point is based on 50 replicates of the sample regularization path for the corresponding design. The results displayed in these panels show good agreement with the theory for sign consistency of general ℓ_1-penalized M-estimators developed in Section 3: for increasing sample sizes, the proportion of paths containing sign correct model approaches one as the sample size increases whenever $\eta(\theta) > 0$. For $\eta(\theta) < 0$, the chance of correct sign recovery are low throughout. Not surprisingly, the asymptotic approximation works better for smaller p. Also notice that the fainter signal of the "blip" profile makes the recovery of the correct signs harder.

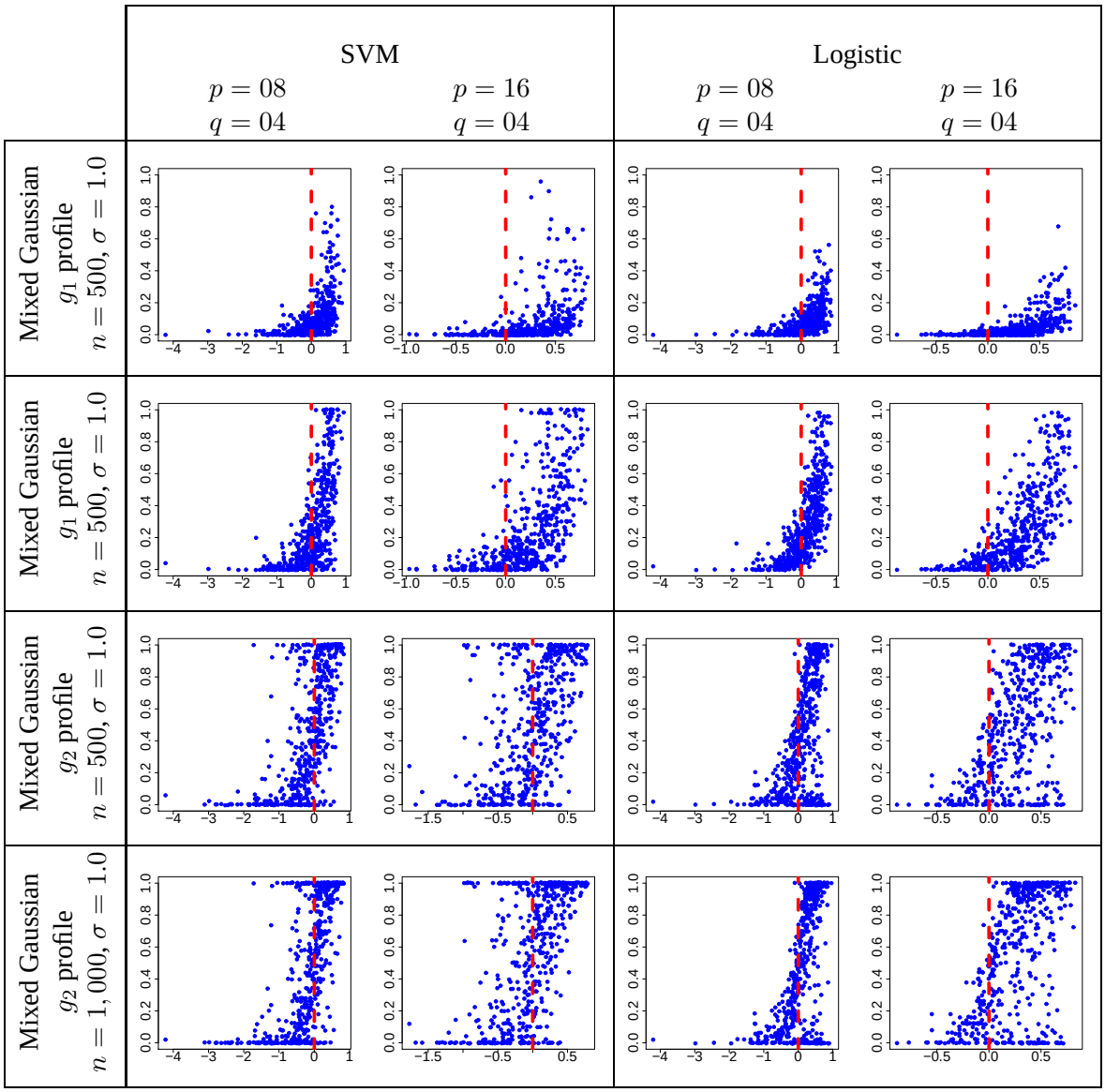

Figure 2: Proportion of sample paths containing a sign-correct model vs. GI index for mixed Gaussian predictors: The proportion at each point is based on 50 replicates of the sample regularization path for the corresponding design. As in Figure 1, the results displayed in these panels show good agreement with the theory for sign consistency of general ℓ_1-penalized M-estimators developed in Section 3: for increasing sample sizes, the proportion of paths containing sign correct model approaches one as the sample size increases whenever $\eta(\theta) > 0$. It is interesting to notice that a high probability of correct sign recovery is possible even if $\eta(\theta) < 0$ (see the SVM estimates under the "blip" profile) but it does not approach one asymptotically. Also notice that the fainter signal of the "blip" profile makes the recovery of the correct signs harder when $\eta(\theta) > 0$, especially for the logistic classifiers.

The results shown in Figure 4 suggest that the comparison of logistic and SVM classification based solely on the GI condition should be taken with a grain of salt. Two factors are involved here: first, the results are based on asymptotic approximations and, second, a negative GI index does not necessarily imply a low probability of correct sign recovery (though such probability is known not to approach one). While in most cases the two methods are comparable in their ability to contain a correct model in their regularization path, SVM does seem to have some advantage over logistic under Gaussian predictors and the "blip" conditional profile even at large sample sizes ($n = 1,000$). For smaller sample sizes ($n = 100$), SVM did perform markedly better than logistic regression under mixed Gaussian predictors and the logistic profile.

6 Discussion and concluding remarks

In this paper, we have extended the asymptotic characterization of the distribution of LASSO estimates (ℓ_1 penalized least squares) given by Knight and Fu (2000) to more general loss and penalty functions in the parametric case. The key to our extension consists of finding conditions under which it is possible to obtain a local quadratic approximation that is uniformly valid on a neighborhood of the risk minimizer. Given the widespread use of convex loss functions use in the literature, the Convexity Lemma by Pollard (1991) was our tool of choice. As we restrict attention to the parametric case, we have been able to keep the study of loss and penalty functions separate. To the possible extent, we have state our results in a modular fashion so they can be applied to various combinations of loss and penalty functions.

We have used the asymptotic characterization of the distribution of ℓ_1 penalized parametric M-estimates to obtain sufficient conditions ensuring the existence of a model selection consistent estimate for some appropriate value of the regularization parameter. Interestingly, the condition involves the Hessian but not the variance of the score function evaluated at the risk minimizer. Ravikumar et al. (2008) have obtained a similar condition in the non-parametric case ($p \gg n$) for the penalized maximum likelihood estimate of Gaussian covariance matrices. That suggests the results we present in this paper can be extended to the non-parametric setting under appropriate conditions, which will be the theme of future research. We also show (Theorem 6) that, under appropriate assumptions, the condition for sign-consistency of ℓ_1 penalized parametric M-estimates can be expressed solely in terms of the matrix of second moments of the predictors.

Our simulations provide ample empirical evidence to the theory we have presented in the context of

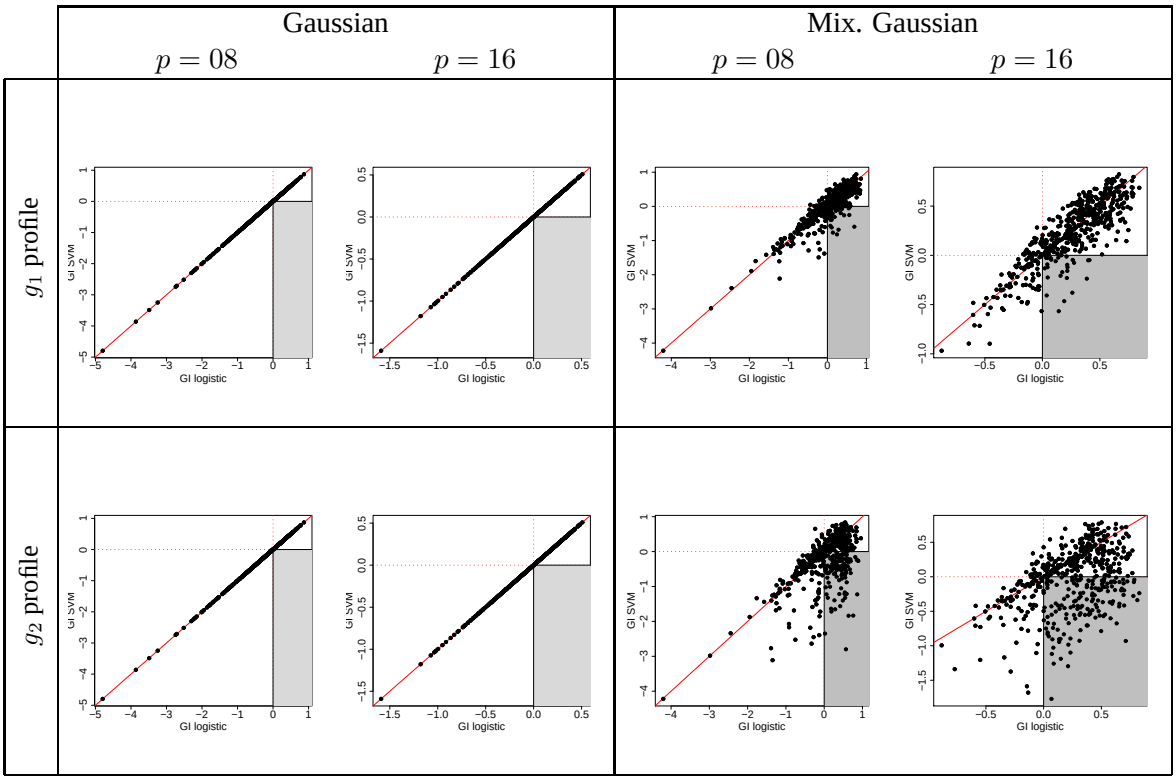

Figure 3: Logistic GI vs. SVM GI for 500 designs: The shaded area shows where logistic regression is model selection consistent and SVM is not. Under Gaussian predictors (the four leftmost panels), the GI indices are exactly the same for the SVM and logistic classifiers, as expected in view of Theorem 6 and Lemmas 8 and 9. For mixed Gaussian predictors, the logistic regression classifier is model selection consistent slightly more often than SVM under the logistic conditional probability profile and, surprisingly, much more often under the "blip" design. Recall, however, that SVM was shown to have high probability of correct sign recovery even in cases with $\eta(\theta) < 0$.

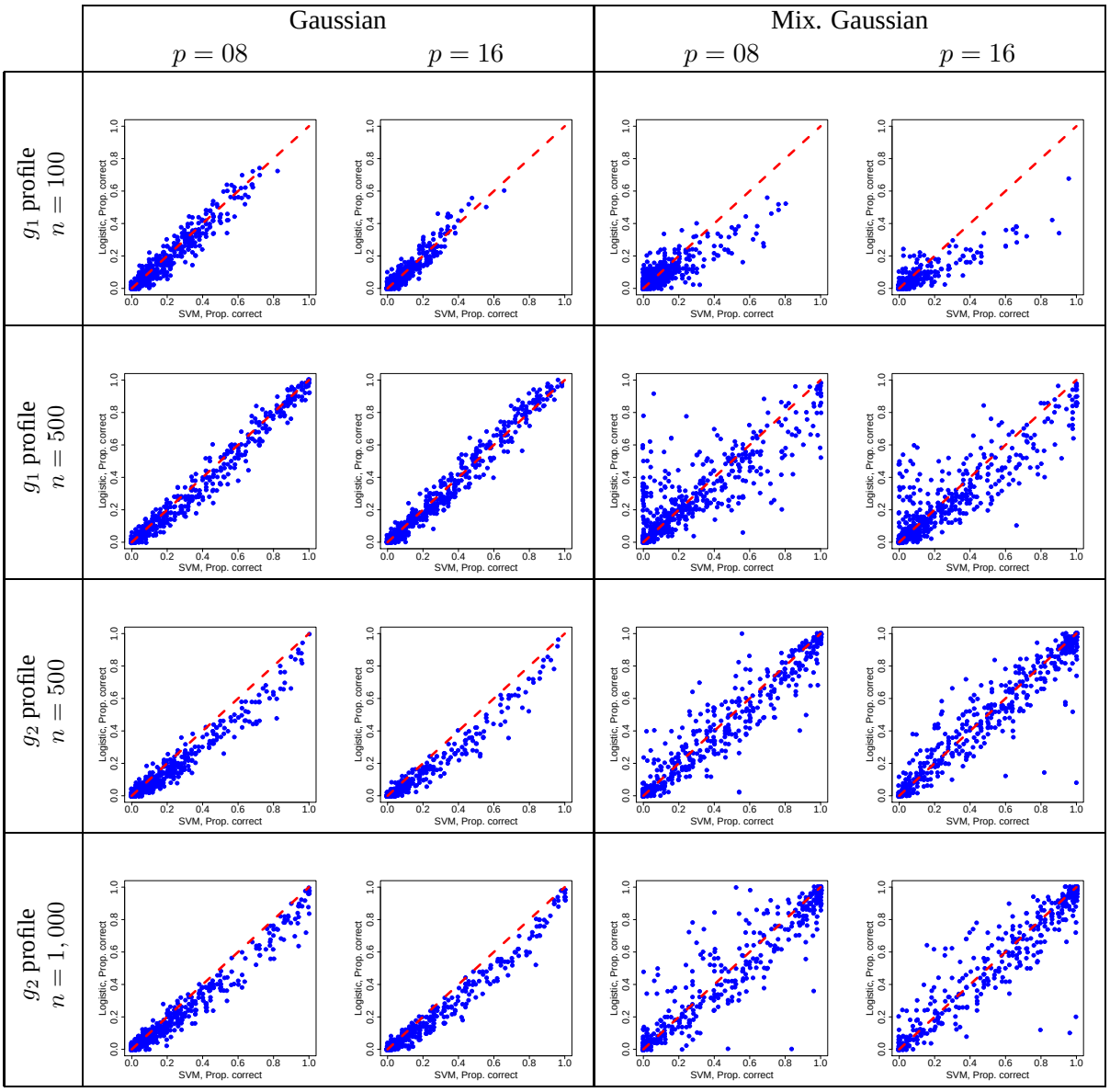

Figure 4: Comparison of the proportion of sample paths containing sign correct estimates in finite samples: The GI condition (Theorem 5) concerns an asymptotic guarantee and does not ensure the probability of correct sign recovery to be low if $\eta(\theta)$. In these plots, we compare SVM and logistic classifiers in terms of probability of correct sign recovery in finite samples. The SVM classifier seems to perform better in terms of the probability of correct sign recovery under Gaussian predictors and the "blip" conditional probability profile. The SVM classifier also performs better in smaller sample sizes under mixed Gaussian predictors and the logistic conditional probability profile.

$p = 08$

g_1 profile

	SVM not MSC	SVM MSC
Logistic not MSC	31.8%	6.0%
Logistic MSC	7.0%	55.2%

g_2 profile

	SVM not MSC	SVM MSC
Logistic not MSC	31.8%	5.2%
Logistic MSC	26.0%	37.0%

$p = 16$

g_1 profile

	SVM not MSC	SVM MSC
Logistic not MSC	15.8%	4.8%
Logistic MSC	5.4%	74.0%

g_2 profile

	SVM not MSC	SVM MSC
Logistic not MSC	16.6%	3.6 %
Logistic MSC	32.2%	47.6 %

Table 1: Frequency at which SVM and logistic are model selection consistent: Each table shows the proportion out of 500 designs with mixed Gaussian predictors in which the ℓ_1-norm penalized SVM and logistic classifiers are model selection consistent (MSC). For most designs, both SVM and logistic would asymptotically contain estimates with all signs correct in their regularization paths. Among the cases where only one of the two classifiers had would asymptotically contain a sign correct estimate in its path, the logistic classifier would be the correct one in most cases.

SVM and logistic regression classification. For Gaussian predictors and a given design, one of the two can happen: both logistic regression and linear SVM classifiers will are sparsistent and sign-consistent or neither of them is. In finite samples, SVM seems to enjoy a slight advantage in picking the correct signs in the cases we simulated. For a set of randomly selected designs with non-Gaussian predictors, logistic regression classifiers were sparsistent and sign-consistent more frequently than SVM classifiers. In finite samples, however, the evidence in favor of either SVM or logistic regression classifiers was mixed.

Acknowledgments

The authors would like to thank Youjuan Li, Wang Li and Ji Zhu for providing code for the ℓ_1-norm penalized estimation of SVMs and to thankfully acknowledge support from grants NSF DMS-0605165 (06-08), NSFC (60628102), a grant from MSRA and a CDI award from NSF. Guilherme Rocha would like to acknowledge helpful discussions with Karen Kafadar, Nicolai Meinshausen and Ram Rajagopal.

A Proofs of theoretical results

A.1 Proof of results in Section 2

We now state and prove the results in Section 2. Before that, we prove technical Lemma 10 which is used in the proof of Theorem 4.

Lemma 10. *Define:*

$$V_\theta^{(n)}(\hat{\mathbb{P}}_n, \lambda_n, u) := \sum_{i=1}^{n} \left[L\left(Z_i, \theta + \frac{u}{q_n}\right) - L\left(Z_i, \theta\right) \right] + \lambda_n \cdot \left[T\left(\theta + \frac{u}{q_n}\right) - T\left(\theta\right) \right].$$

Then:

$$q_n\left(\hat{\theta}_n(\lambda_n) - \theta\right) \;=\; \underset{u \in \mathbb{R}^p}{\arg\min} \left[V_\theta^{(n)}(\hat{\mathbb{P}}_n, \lambda_n, u) \right]. \tag{A-1}$$

Proof of Lemma 10. From the definition of $\hat{\theta}_n(\lambda_n)$, we know that:

$$\begin{aligned}
\hat{\theta}_n(\lambda_n) &= \underset{t \in \Theta}{\arg\min} \left[\sum_{i=1}^{n} \left[L\left(Z_i, \theta + \frac{q_n\left(t-\theta\right)}{q_n}\right)\right] + \lambda_n \cdot T\left(\theta + \frac{q_n\left(t-\theta\right)}{q_n}\right)\right] \\
&= \underset{t \in \Theta}{\arg\min} \left[\left[\sum_{i=1}^{n} \left[L\left(Z_i, \theta + \frac{q_n\left(t-\theta\right)}{q_n}\right) - L\left(Z_i, \theta\right)\right]\right] + \lambda_n \cdot \left[T\left(\theta + \frac{q_n\left(t-\theta\right)}{q_n}\right) - T\left(\theta\right)\right]\right].
\end{aligned}$$

The result follows from making a variable transformation $u(t) = q_n \cdot (t-\theta)$ and letting $\hat{u} = u(\hat{\theta}_n(\lambda_n))$. □

Proof of Theorem 1. a) The conclusion in (a) follows easily from the triangular inequality, since for each compact set K:

$$\begin{aligned}\sup_{u\in K}\left|V_\theta^{(n)}(\hat{\mathbb{P}}_n,\lambda_n,u)-V_\theta(\mathbf{W},u)\right| &\leq \sup_{u\in K}\left|\left[\sum_{i=1}^n\left[L\left(Z_i,\theta+\frac{u}{q_n}\right)-L\left(Z_i,\theta\right)\right]\right]-C_\theta(\mathbf{W},u)\right| \\ &+ \sup_{u\in K}\left|\lambda_n\cdot\left[T\left(\theta+\frac{u}{q_n}\right)-T(\theta)\right]-\lambda\cdot G_\theta(u)\right|.\end{aligned}$$

b) Define:

$$\begin{aligned}\hat{u}_n &= q_n\cdot\left(\hat{\theta}(\lambda_n)-\theta\right), \\ \hat{u} &= \arg\min\left[C_\theta(\mathbf{W},u)-\lambda\cdot G_\theta(u)\right].\end{aligned}$$

For any compact set K_ε and each n, we know:

$$\begin{aligned}\mathbb{P}\left(|\hat{u}_n-\hat{u}|>\delta\right) &= \mathbb{P}\left(|\hat{u}_n-\hat{u}|>\delta\,\middle|\,\hat{u}_n\in K_\varepsilon\right)\cdot\mathbb{P}\left(\hat{u}_n\in K_\varepsilon\right) \\ &\quad +\mathbb{P}\left(|\hat{u}_n-\hat{u}|>\delta\,\middle|\,\hat{u}_n\notin K_\varepsilon\right)\cdot\mathbb{P}\left(\hat{u}_n\notin K_\varepsilon\right). \\ &\leq \mathbb{P}\left(|\hat{u}_n-\hat{u}|>\delta\,\middle|\,\hat{u}_n\in K_\varepsilon\right)+\mathbb{P}\left(\hat{u}_n\notin K_\varepsilon\right).\end{aligned}$$

Since $\hat{u}_n$ is $O_p(1)$, there exists a compact set K_ε such that $\lim_{n\to\infty}\mathbb{P}\left(\hat{u}_n\notin K_\varepsilon\right)=0$ and thus:

$$\lim_{n\to\infty}\mathbb{P}\left(|\hat{u}_n-\hat{u}|>\delta\,\middle|\,\hat{u}_n\notin K_\varepsilon\right)\cdot\mathbb{P}\left(\hat{u}_n\notin K_\varepsilon\right) \leq \lim_{n\to\infty}\mathbb{P}\left(\hat{u}_n\notin K_\varepsilon\right)=0.$$

To show the second term vanish, the uniform convergence over compact sets gives that:

$$\lim_{n\to\infty}\mathbb{P}\left(|\hat{u}_n-\hat{u}|>\delta\,\middle|\,\hat{u}_n\in K_\varepsilon\right)=0,$$

which concludes the proof. □

Proof of Lemma 2. **Proof of a) pointwise convergence:** To establish pointwise convergence, define:

$$\begin{aligned} \delta_{i,n}(u) &= L(Z_i, \theta + \tfrac{u}{\sqrt{n}}) - L(Z_i, \theta) \\ D_i &= \nabla_b L(Z_i, \theta) \\ R_{i,n}(u) &= \delta_{i,n}(u) - D_i \tfrac{u}{\sqrt{n}} \\ W_n &= \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n D_i \\ B_n(u) &= \textstyle\sum_{i=1}^n \delta_{i,n}(u). \end{aligned} \tag{A-2}$$

In terms of these definitions, we have:

$$\begin{aligned} B_n(u) &= n \cdot \left[\frac{1}{n} \sum_{i=1}^n \left[L\left(Z_i, \theta + \frac{u}{\sqrt{n}} \right) - L\left(Z_i, \theta \right) \right] \right] \\ &= W_n^T \cdot u + \sum_{i=1}^n \left[R_{i,n}(u) \right]. \end{aligned} \tag{A-3}$$

Now, because θ is optimal we have $\mathbb{E} D_i = 0$ and, thus:

$$\mathbb{E}\left[B_n(u) \right] = \sum_{i=1}^n \mathbb{E}\left[R_{i,n}(u) \right].$$

Summing $\mathbb{E}\left[B_n(u) \right]$ and subtracting $\sum_{i=1}^n \mathbb{E}\left[R_{i,n}(u) \right]$ from the right hand side of (A-3):

$$B_n(u) = \mathbb{E}\left[B_n(u) \right] + W_n^T u + \sum_{i=1}^n \left[R_{i,n}(u) - \mathbb{E}\left[R_{i,n}(u) \right] \right].$$

Pointwise convergence for each u follows from obtaining a quadratic approximation to $\mathbb{E}\left[B_n(u) \right]$, a weak convergence for $W_n^T u$ and proving that the last term is $o_p(1)$. These facts are established next.

i) Quadratic approximation to $\mathbb{E}\left[B_n(u) \right]$:

First notice that this term is just the difference of the risk function evaluated at θ and $\theta + \frac{u}{\sqrt{n}}$:

$$\begin{aligned} \mathbb{E}\left[B_n(u) \right] &= n \cdot \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \left[L\left(Z_i, \theta + \frac{u}{\sqrt{n}} \right) - L\left(Z_i, \theta \right) \right] \right] \\ &= n \cdot \left[\mathbb{E}\left[L\left(\mathbf{Z}, \theta + \frac{u}{\sqrt{n}} \right) \right] - \mathbb{E}\left[L(\mathbf{Z}, \theta) \right] \right] \\ &= n \cdot \left[R\left(\theta + \frac{u}{\sqrt{n}} \right) - R\left(\theta \right) \right] \end{aligned}$$

Since the risk function R is twice differentiable (L2.c) and θ is optimal, the gradient of the risk with respect to its argument t must be zero at $t = \theta$. In addition, for $H(\theta)$ as defined in assumption L2.c, we can

write the approximation:

$$\begin{aligned}\mathbb{E}\left[B_n(u)\right] &= n \cdot \left[\frac{u^T}{\sqrt{n}} \cdot \left[H(\theta)\right] \cdot \frac{u}{\sqrt{n}} + o\left(\left(\frac{|u|}{\sqrt{n}}\right)^2\right)\right]. \\ &= u^T \cdot H(\theta) \cdot u + o(1).\end{aligned}$$

ii)] Weak convergence of $W_n^T u$:

Optimality of θ and differentiability of the risk function imply that $\mathbb{E}\left[D_i\right] = 0$, thus $\mathbb{E}\left[W_n\right] = 0$.

Since $\nabla_b L(Z, b)$ exists almost everywhere, all terms in the summation defining W_n^T almost surely exist. Since the terms in the summation are i.i.d. and each has finite variance, the Central Limit applies and we can conclude that:

$$W_n \xrightarrow{d} N\left(0, J(\theta)\right), \text{ with } J(\theta) = \mathbb{E}\left[\nabla_t L(\mathbf{Z}, \theta) \nabla_t L(\mathbf{Z}, \theta)^T\right].$$

iii) $\sum_{i=1}^n \left[R_{i,n}(u) - \mathbb{E}\left[R_{i,n}(u)\right]\right]$ **is** $o_p(1)$**:**

Let $\xi_i = R_{i,n}(u) - \mathbb{E}\left[R_{i,n}(u)\right]$. Since convergence in quadratic mean implies convergence in probability, it is enough to prove that $\mathbb{P}\left|\sum_{i=1}^n \xi_i\right|^2 = o(1)$.

Clearly $\mathbb{E}\xi_i = 0$ for all i. That, along with independence across the observed samples, yields:

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{i=1}^n \xi_i\right|^2\right] &= \operatorname{var}\left[\sum_{i=1}^n \xi_i\right] \\ &= \sum_{i=1}^n \operatorname{var}\left[\xi_i\right] \\ &= \sum_{i=1}^n \left(\mathbb{E}\left[\left|R_{i,n}(u)\right|^2\right] - \left[\mathbb{E}\left(R_{i,n}(u)\right)\right]^2\right) \\ &\leq \sum_{i=1}^n \left(\mathbb{E}\left[\left|R_{i,n}(u)\right|^2\right]\right).\end{aligned}$$

Because $L(Z_i, t)$ is differentiable at $t = \theta$ for almost every Z_i, we have that:

$$\left|R_{i,n}(u)\right|^2 = \left|L\left(Z_i, \theta + \frac{u}{\sqrt{n}}\right) - L(\mathbf{Z}_i, \theta) - \nabla_b L(Z_i, \theta) \cdot \frac{u}{\sqrt{n}}\right|^2 = o\left(\frac{1}{n}\right), \text{ for almost all } Z_i.$$

We conclude that $\mathbb{E}\left[\sum_{i=1}^n \left|R_{i,n}(u)\right|^2\right] = o(1)$.

Proof of b.1) uniform convergence over compact sets:

Uniform convergence of $B_n(u)$ over compact sets follows from the pointwise convergence just proven and the Convexity Lemma due to Pollard (1991).

Proof of b.2) boundedness of $\sqrt{n}\cdot\hat{\theta}(0)$: Our proof of $\sqrt{n}$-boundedness of the un-penalized estimate is an adaptation of an argument due to Pollard (1991). As a first step, we "complete the squares" in the quadratic approximation by letting C be a decomposition of the (non-singular) Hessian matrix, i.e. $C^TC = H(\theta)$. We then write $B_n(u)$ as:

$$B_n(u) \quad = \quad \left\| Cu + \frac{1}{2}(C^{-1})^T\mathbf{W}_n \right\|^2 - \left\| \frac{1}{2}(C^{-1})^T\mathbf{W}_n \right\|^2 + r_n(u),$$

Let A_n denote the ball with center $-\frac{1}{2}(C^{-1})^T\mathbf{W}_n$ and radius $\delta > 0$. Since $\mathbf{W}_n$ converges in distribution, it is stochastically bounded and, hence, a compact set K^* with probability arbitrarily close to one can be chosen to contain A_n. Thus:

$$\Delta_n \quad := \quad \sup_{u\in A_n} |r_n(u)| \xrightarrow{p} 0.$$

We now study the behavior of B_n outside of A_n to conclude that $\hat{\theta}_n(0)$ is consistent. To do that, let z be a point outside the ball and define:

$$\begin{aligned} m &= \|z - \frac{1}{2}(C^{-1})^T\mathbf{W}_n\|_2 \\ v &= \frac{z - \frac{1}{2}(C^{-1})^T\mathbf{W}_n}{m} \end{aligned}$$

Because of convexity, we have that for $u^* = -\frac{1}{2}C^{-1}(C^{-1})^T\mathbf{W}_n + \delta\cdot v$ on the boundary of the A_n ball:

$$\begin{aligned} \left(\frac{\delta}{m}\right) B_n(z) + \left(1 - \frac{\delta}{m}\right) B_n\left(\frac{1}{2}(C^{-1})^T\mathbf{W}_n\right) &\geq B_n(u^*) \\ &\geq \inf_{|v|\leq 1}\left(v^TH(\theta)v\right) - \frac{1}{4}\mathbf{W}_n^T\left[H(\theta)\right]^{-1}\mathbf{W}_n - \Delta_n \\ &\geq \delta^2 \inf_{|v|\leq 1}\left(v^TH(\theta)v\right) - \frac{1}{4}\mathbf{W}_n^T\left[H^{-1}(\theta)\right]\mathbf{W}_n - \Delta_n \\ &\geq \delta^2\Lambda_p(H(\theta)) - B_n\left(-\frac{1}{2}C^{-1}(C^{-1})^T\mathbf{W}_n\right) - 2\Delta_n, \end{aligned}$$

where $\Lambda_p(H(\theta))$ is the smallest eigenvalue of $H(\theta)$. We then conclude that:

$$\inf_{\left|u+\frac{1}{2}(C^{-1})^T\mathbf{W}_n\right|>\delta} B_n(u) \geq B_n(-\frac{1}{2}(C^{-1})^T\mathbf{W}_n) + \frac{m}{\delta}\left[\delta^2\Lambda_p(H(\theta)) - 2\Delta_n\right].$$

Since $\Delta_n \xrightarrow{p} 0$, we have that with probability approaching one that $\left|\hat{u} + \frac{1}{2}(C^{-1})^T\mathbf{W}_n\right| < \delta$ and the result follows from recalling that $\hat{u} = \sqrt{n}(\hat{\theta}_n(0) - \theta)$. □

Proof of Lemma 3. We first brake the problem into two easier to handle pieces:

$$\begin{aligned}\sup_{u\in K\subset\mathbb{R}^p}\left\|\lambda_n\cdot\left[T(\theta+\frac{u}{q_n})-T(\theta)\right]-\lambda\cdot G_\theta(u)\right\| \;\leq\; & \sup_{u\in K\subset\mathbb{R}^p}\left\|\left(\frac{\lambda_n}{q_n}-\lambda\right)\cdot\left[\frac{T(\theta+u\cdot q_n^{-1})-T(\theta)}{q_n^{-1}}\right]\right\| \\ & +\sup_{u\in K\subset\mathbb{R}^p}\left\|\lambda\cdot\left(\frac{T(\theta+u\cdot q_n^{-1})-T(\theta)}{q_n^{-1}}-G_\theta(u)\right)\right\|\end{aligned}$$

Since T is continuous, for the compact set $K\subset\mathbb{R}^p$ there exists $0<M_K<\infty$ such that:

$$\sup_{u\in K\subset\mathbb{R}^p}\left\|\left(\frac{\lambda_n}{q_n}-\lambda\right)\cdot\left[\frac{T(\theta+u\cdot q_n^{-1})-T(\theta)}{q_n^{-1}}\right]\right\| \;\leq\; \left\|\left(\frac{\lambda_n}{q_n}-\lambda\right)\right\|\cdot M_K\xrightarrow{p}0,\text{ as } n\to\infty.$$

For the second term, we know from condition P4 in Assumption Set 2:

$$\lim_{n\to\infty}\frac{T(\theta+u\cdot q_n^{-1})-T(\theta)}{q_n^{-1}} \;=\; \lim_{h\downarrow 0}\frac{T(\theta+h\cdot u)-T(\theta)}{h}=G_\theta(u).$$

Because G_θ is assumed continuous, the pointwise convergence can be strengthened to uniform convergence over compact sets. □

Lemma 11. *Let $\hat{\theta}_n(\lambda_n)$ be as defined in* (5)*, q_n be a sequence such that $q_n\to\infty$ as $n\to\infty$ and λ_n be a sequence of (potentially random) non-negative real numbers. Assume T is a penalty function satisfying condition P4 in PA. If $q_n\cdot\hat{\theta}_n(0)=O_p(1)$, then $q_n\cdot\hat{\theta}_n(\lambda_n)=O_p(1)$.*

Proof. First, we use a contradiction to prove that $T(\hat{\theta}_n(\lambda_n))\leq T(\hat{\theta}_n(0))$. From the definition of $\hat{\theta}_n(0)$, we have:

$$\frac{1}{n}\cdot\sum_{i=1}^n L\left(Z_i,\hat{\theta}_n(0)\right) \;\leq\; \frac{1}{n}\cdot\sum_{i=1}^n L\left(Z_i,\hat{\theta}_n(\lambda_n)\right).$$

Supposing that $T(\hat{\theta}_n(\lambda_n))>T(\hat{\theta}_n(0))$, we get

$$\frac{1}{n}\cdot\sum_{i=1}^n L\left(Z_i,\hat{\theta}_n(0)\right)+\lambda_n\cdot T(\hat{\theta}_n(0)) \;<\; \frac{1}{n}\cdot\sum_{i=1}^n L\left(Z_i,\hat{\theta}_n(\lambda_n)\right)+\lambda_n\cdot T(\hat{\theta}_n(\lambda_n)),$$

a contradiction with the definition of $\hat{\theta}_n(\lambda_n)$ as the minimizer of $f(t)=\left[\frac{1}{n}\sum_{i=1}^n L\left(Z_i,t\right)\right]+\lambda_n\cdot T(t)$.

Now, from $q_n\cdot\hat{\theta}_n(0)=O_p(1)$, we have that, for any $\delta>0$, there exists compact $K_0\subset\Theta$ such that $\mathbb{P}\left(q_n\cdot\hat{\theta}_n(0)\in K_0\right)>1-\delta$. Let $U=\max_{t\in K_0}q_n\cdot T(t)$ and define $\tilde{K}_0=\{t\in\Theta: q_n\cdot T(t)\leq U\}$. Since $T(\hat{\theta}_n(\lambda_n))<T(\hat{\theta}_n(0))$, it follows that $\mathbb{P}\left(\hat{\theta}(\lambda_n)\in\tilde{K}_0\right)\geq\mathbb{P}\left(\hat{\theta}(0)\in\tilde{K}_0\right)>1-\delta$. □

A.2 Proof of results in Section 3

Proof of Theorem 5. For the ℓ_1-penalty, the difference between $\lambda_n \left(\|\beta + \frac{u}{\sqrt{n}}\|_1 - \|\beta\|_1 \right) - \frac{\lambda_n}{\sqrt{n}} \left(u_{\mathcal{A}} - \|u_{\mathcal{A}^c}\|_1 \right) \to 0$, uniformly as $n \to \infty$. Using Theorem 1 and Lemma 2,

$$\sqrt{n} \left(\hat{\theta}_n(\lambda_n) - \theta \right) \xrightarrow{d} \hat{u} := \arg\min_u \left[u^T H(\theta) u + \mathbf{W}^T u + \frac{\lambda_n}{\sqrt{n}} \left(u_{\mathcal{A}} + \|u_{\mathcal{A}^c}\|_1 \right) \right],$$

with $\mathbf{W} \sim N(0, J(\theta))$. We assume, without loss of generality that $\beta_{\mathcal{A}} > 0$ with the inequality holding element-wise. In that case:

$$\begin{array}{rlll} \text{sign}(\hat{\beta}_j(\lambda_n)) = \text{sign}(\beta_j) & \Leftrightarrow & \frac{\hat{u}_j}{\sqrt{n}} \geq -\beta_j, & \text{for } j \in \mathcal{A}, \\ \text{sign}(\hat{\beta}_j(\lambda_n)) = \text{sign}(\beta_j) & \Leftrightarrow & \hat{u}_j = 0, & \text{for } j \in \mathcal{A}^c. \end{array}$$

For the remainder of this proof, we drop denote $H(\theta)$ by H. In terms of the α, $\mathcal{A}$, $\mathcal{A}^c$ partition, the Karush-Kuhn-Tucker (KKT) conditions for optimization defining $\hat{u}$ above are

$$\begin{array}{rcll} H_{\alpha,\mathcal{A}} \cdot \hat{u}_{\mathcal{A}} + H_{\alpha,\mathcal{A}^c} \cdot \hat{u}_{\mathcal{A}^c} + H_{\alpha,\alpha} \cdot \hat{u}_\alpha + \mathbf{W}_\alpha & = & 0, & \\ H_{\mathcal{A},\mathcal{A}} \cdot \hat{u}_{\mathcal{A}} + H_{\mathcal{A},\mathcal{A}^c} \cdot \hat{u}_{\mathcal{A}^c} + H_{\mathcal{A},\alpha} \cdot \hat{u}_\alpha + \mathbf{W}_{\mathcal{A}} - \frac{\lambda_n}{\sqrt{n}} & = & 0, & \\ H_{j,\mathcal{A}} \cdot \hat{u}_{\mathcal{A}} + H_{j,\mathcal{A}^c} \cdot \hat{u}_{\mathcal{A}^c} + H_{j,\alpha} \cdot \hat{u}_\alpha + \mathbf{W}_j - \frac{\lambda_n}{\sqrt{n}} \cdot \text{sign}(\hat{u}_j) & = & 0, & \text{for } j \in \mathcal{A}^c \text{ s.t. } \hat{u}_j \neq 0 \\ |H_{j,\mathcal{A}} \cdot \hat{u}_{\mathcal{A}} + H_{j,\mathcal{A}^c} \cdot \hat{u}_{\mathcal{A}^c} + H_{j,\alpha} \cdot \hat{u}_\alpha + \mathbf{W}_j| & \leq & \frac{\lambda_n}{\sqrt{n}}, & \text{for } j \in \mathcal{A}^c \text{ s.t. } \hat{u}_j = 0. \end{array}$$

To select the zero terms in β correctly, we must have $\hat{u}_{\mathcal{A}^c = 0}$. In that case,

$$\begin{bmatrix} \hat{u}_\alpha \\ \hat{u}_{\mathcal{A}} \end{bmatrix} = \begin{bmatrix} H_{\alpha,\alpha} & H_{\mathcal{A},\alpha} \\ H_{\mathcal{A},\alpha} & H_{\mathcal{A},\mathcal{A}} \end{bmatrix}^{-1} \cdot \begin{bmatrix} -\mathbf{W}_\alpha \\ \frac{\lambda_n}{\sqrt{n}} \cdot \mathbf{1}_q - \mathbf{W}_{\mathcal{A}} \end{bmatrix}.$$

Using Schur's inversion formula for partitioned matrices, we get:

$$\hat{u}_{\mathcal{A}} = \left[H_{\mathcal{A},\mathcal{A}} - H_{\mathcal{A},\alpha} H_{\alpha,\alpha}^{-1} H_{\alpha,\mathcal{A}} \right]^{-1} \cdot \left[\frac{\lambda_n}{\sqrt{n}} \cdot \mathbf{1}_q - \mathbf{W}_{\mathcal{A}} - \left[H_{\mathcal{A},\alpha} H_{\alpha,\alpha}^{-1} \right] \cdot \mathbf{W}_\alpha \right].$$

Define a zero mean Gaussian random vector $\tilde{\mathbf{W}} = \begin{bmatrix} \tilde{\mathbf{W}}_{\mathcal{A}}, & \tilde{\mathbf{W}}_{\mathcal{A}^c} \end{bmatrix}$:

$$\begin{array}{rcl} \tilde{\mathbf{W}}_{\mathcal{A}} & := & -\mathcal{H} \cdot \left[\mathbf{W}_{\mathcal{A}} + H_{\mathcal{A},\alpha} \cdot H_{\alpha,\alpha}^{-1} \cdot \mathbf{W}_\alpha \right], \text{ and} \\ \tilde{\mathbf{W}}_{\mathcal{A}^c} & := & \mathbf{W}_{\mathcal{A}^c} - \begin{bmatrix} H_{\mathcal{A}^c,\alpha} \\ H_{\mathcal{A}^c,\mathcal{A}} \end{bmatrix}^T \begin{bmatrix} H_{\alpha,\alpha} & H_{\alpha,\mathcal{A}} \\ H_{\mathcal{A},\alpha} & H_{\mathcal{A},\mathcal{A}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{W}_\alpha \\ \mathbf{W}_{\mathcal{A}} \end{bmatrix}, \text{ with} \\ \mathcal{H} & := & \left[H_{\mathcal{A},\mathcal{A}} - H_{\mathcal{A},\alpha} H_{\alpha,\alpha}^{-1} H_{\alpha,\mathcal{A}} \right]^{-1} \end{array}$$

The M-estimated parameter fails to have the correct signs if:

$$\begin{array}{rcll}
\tilde{\mathbf{W}}_j & \geq & \sqrt{n}\cdot\beta_j - \frac{\lambda_n}{\sqrt{n}}\cdot\sum_{k\in\mathcal{A}}\mathcal{H}_{jk}, & \text{for some } j\in\mathcal{A}, \quad \text{OR,} \\
\tilde{\mathbf{W}}_j & > & \frac{\lambda_n}{\sqrt{n}}\left(1 - H_{\mathcal{A}^c,\mathcal{A}}\left[H_{\mathcal{A},\mathcal{A}} - H_{\mathcal{A},\alpha}H_{\alpha,\alpha}^{-1}H_{\alpha,\mathcal{A}}\right]^{-1}\mathbf{1}_q\right), & \text{for some } j\in\mathcal{A}^c, \quad \text{OR,} \\
\tilde{\mathbf{W}}_j & < & \frac{\lambda_n}{\sqrt{n}}\left(-1 - H_{\mathcal{A}^c,\mathcal{A}}\left[H_{\mathcal{A},\mathcal{A}} - H_{\mathcal{A},\alpha}H_{\alpha,\alpha}^{-1}H_{\alpha,\mathcal{A}}\right]^{-1}\mathbf{1}_q\right), & \text{for some } j\in\mathcal{A}^c.
\end{array}$$

For the remainder of the proof, let Φ denote the standard normal cumulative distribution function, and define $\varsigma_j^2 := \mathrm{var}(\tilde{\mathbf{W}}_j)$.

Proof of a):

We prove that if $\|H_{\mathcal{A}^c,\mathcal{A}}\left[H_{\mathcal{A},\mathcal{A}} - H_{\mathcal{A},\alpha}H_{\alpha,\alpha}^{-1}H_{\alpha,\mathcal{A}}\right]^{-1}\mathbf{1}_q\|_\infty < 1$ and there exists $c > 0$ and $\lambda\in\mathbb{R}$ sucg that $\frac{\lambda_n}{n^{\frac{1+c}{2}}} \xrightarrow{p} \lambda$ and $\frac{\lambda_n}{n} \xrightarrow{p} 0$, then the probability of each of these three events decreases to zero exponentially fast.

For the first event, use the union bound and the inequality $1-\Phi(r) \leq \exp\left[-\frac{r^2}{2}\right]$ for large enough r, to get

$$\begin{array}{rcl}
\mathbb{P}\left(\exists j\in\mathcal{A} : \tilde{\mathbf{W}}_j > \sqrt{n}\cdot\left(\beta_j - \frac{\lambda_n}{n}\right)\right) & \leq & \sum_{j\in\mathcal{A}}\mathbb{P}\left(\tilde{\mathbf{W}}_j > \sqrt{n}\left(\beta_j - \frac{\lambda_n}{n}\right)\right) \\
& \leq & \sum_{j\in\mathcal{A}}\mathbb{P}\left(\frac{\tilde{\mathbf{W}}_j}{\varsigma_j} > \sqrt{n}\left(\frac{\beta_j}{\varsigma_j} - \frac{\lambda_n}{n\cdot\varsigma_j}\right)\right) \\
& \leq & q\cdot\left[1-\Phi\left(\sqrt{n}\cdot\left(\min_{j\in\mathcal{A}}\left(\frac{\beta_j}{\varsigma_j}\right) - \frac{\lambda_n}{n\cdot\max_{j\in\mathcal{A}}\varsigma_j}\right)\right)\right] \\
& \leq & q\cdot\dfrac{\exp\left[-\frac{n}{2}\cdot\left(\min_{j\in\mathcal{A}}\left(\frac{\beta_j}{\varsigma_j}\right)^2 - \frac{\lambda_n^2}{n^2\cdot\max_{j\in\mathcal{A}}\varsigma_j^2}\right)\right]}{\sqrt{n}\cdot\left(\min_{j\in\mathcal{A}}\left(\frac{\beta_j}{\varsigma_j}\right) - \frac{\lambda_n}{n\cdot\max_{j\in\mathcal{A}}\varsigma_j}\right)} \\
& \sim & q\cdot\dfrac{\exp\left[-\frac{n}{2}\min_{j\in\mathcal{A}}\left(\frac{\beta_j}{\varsigma_j}\right)^2\right]}{\sqrt{n}\cdot\min_{j\in\mathcal{A}}\left(\frac{\beta_j}{\varsigma_j}\right)}.
\end{array}$$

To tackle the second and third events, define $\eta = 1 - \|H_{\mathcal{A}^c,\mathcal{A}}\left[H_{\mathcal{A},\mathcal{A}} - H_{\mathcal{A},\alpha}H_{\alpha,\alpha}^{-1}H_{\alpha,\mathcal{A}}\right]^{-1}\mathbf{1}_q\|_\infty$ and notice that

$$\begin{array}{rcl}
\mathbb{P}\left(\tilde{W}_j > \frac{\lambda_n}{\sqrt{n}}\left(1 - H_{\mathcal{A}^c,\mathcal{A}}\left[H_{\mathcal{A},\mathcal{A}} - H_{\mathcal{A},\alpha}H_{\alpha,\alpha}^{-1}H_{\alpha,\mathcal{A}}\right]^{-1}\mathbf{1}_q\right)\right) & \leq & \mathbb{P}\left(\tilde{W}_j > \eta\cdot\frac{\lambda_n}{\sqrt{n}}\right), \text{ for all } j\in\mathcal{A}^c, \text{ AND} \\
\mathbb{P}\left(\tilde{W}_j < -\frac{\lambda_n}{\sqrt{n}}\left(1 - H_{\mathcal{A}^c,\mathcal{A}}\left[H_{\mathcal{A},\mathcal{A}} - H_{\mathcal{A},\alpha}H_{\alpha,\alpha}^{-1}H_{\alpha,\mathcal{A}}\right]^{-1}\mathbf{1}_q\right)\right) & \leq & \mathbb{P}\left(-\tilde{W}_j > -\eta\cdot\frac{\lambda_n}{\sqrt{n}}\right).
\end{array}$$

As a result, using the union bound gives that the probability of the second or third event happening is bounded above by $\sum_{j\in\mathcal{A}^c}\mathbb{P}\left(\left|\tilde{W}_j\right| > \eta\cdot\frac{\lambda_n}{\sqrt{n}}\right)$. To prove this probability vanishes exponentially fast, we

use the same inequality as above:

$$\begin{aligned}
\sum_{j\in\mathcal{A}^c}\mathbb{P}\left(\left|\tilde{W}_j\right|>\eta\cdot\frac{\lambda_n}{\sqrt{n}}\right) &\leq \sum_{j\in\mathcal{A}^c}\mathbb{P}\left(\left|\frac{\tilde{W}_j}{\varsigma_j}\right|>\frac{\eta}{\varsigma_j}\cdot\frac{\lambda_n}{\sqrt{n}}\right)\\
&\leq 2\cdot\sum_{j\in\mathcal{A}^c}\left[1-\Phi\left(\frac{\eta}{\varsigma_j}\cdot\frac{\lambda_n}{\sqrt{n}}\right)\right]\\
&\leq 2(p-q)\cdot\left[1-\Phi\left(\frac{\eta}{\max\limits_{j\in\mathcal{A}^c}\varsigma_j}\cdot\frac{\lambda_n}{\sqrt{n}}\right)\right]\\
&\leq 2(p-q)\cdot\exp\left[-\frac{1}{2}\left(\frac{\eta}{\max\limits_{j\in\mathcal{A}^c}\varsigma_j}\cdot\frac{\lambda_n}{\sqrt{n}}\right)^2\right]\\
&\sim 2(p-q)\cdot\exp\left[-\frac{1}{2}\cdot n^c\cdot\left(\min_{j\in\mathcal{A}^c}\frac{\eta}{\varsigma_j}\right)^2\right].
\end{aligned}$$

Proof of b):

To prove the converse in part (b), first notice that $\mathcal{H}_{\mathcal{A},\mathcal{A}}$ is a positive definite matrix. It follows that $\sum_{j\in\mathcal{A}}\sum_{k\in\mathcal{A}}\mathcal{H}_{jk}>0$, and thus that there must exists $j\in\mathcal{A}$ with $\sum_{k\in\mathcal{A}}\mathcal{H}_{jk}>0$. Thus, if $\frac{\lambda_n}{n}\to\infty$, the first event takes place with probability approaching one (exponentially fast) as long as $\mathcal{A}$ is non-empty. On the other hand, if $\frac{\lambda_n}{\sqrt{n}}\to 0$, then the union of the second and third event occurs with probability approaching one (exponentially fast). Thus, we only need to consider the case $\frac{\lambda_n}{n^{\frac{1+c}{2}}}\to\lambda$, for some finite $\lambda\in\mathbb{R}$ and $c\in[0,1)$.

As before, let $\eta=1-\|H_{\mathcal{A}^c,\mathcal{A}}\left[H_{\mathcal{A},\mathcal{A}}-H_{\mathcal{A},\alpha}H_{\alpha,\alpha}^{-1}H_{\alpha,\mathcal{A}}\right]^{-1}\mathbf{1}_q\|_\infty$. If $c>0$ and $\eta<0$, the probability of the second or third events converges to one (exponentially fast). If $\eta=0$, the second or third events have a positive probability of taking place regardless of λ_n. Likewise, if $c=0$, $\eta<0$, the second or third events happen with strictly positive probability. □

Proof of Theorem 6. Throughout this proof we denote $\nu:=\frac{\beta}{\|\beta\|}\in\mathbb{R}^p$, the unit vector in the direction of β. Using the properties of Gaussian distributions and the condition $\nu^T\mu=0$, we get

$$\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\left|\alpha+\mathbf{X}^T\beta\right.\right] = \mu\mu^T+\Sigma+\tfrac{\Sigma\nu\nu^T\Sigma}{(\nu^T\Sigma\nu)^2}\cdot\left[\left(\tfrac{\left(\alpha+\mathbf{X}^T\beta\right)-\left(\alpha+\beta^T\mu\right)}{\|\beta\|}\right)^2-1\right].$$

For details, we refer the reader to Appendix B.1. Letting $f_{\mathbf{M}}$ denote the density of the random variable $\mathbf{M}=\alpha+\mathbf{X}^T\beta$ and defining

$$\kappa := \int\left[\left(\frac{m-\left(\alpha+\beta^T\mu\right)}{\|\beta\|}\right)^2-1\right]\cdot w(m)\cdot\tilde{f}_{\mathbf{M}}(m)\cdot\mathrm{d}m,$$

the Hessian of the risk function becomes:

$$H(\theta) \;=\; \mu\mu^T + \Sigma + \kappa \cdot \frac{\Sigma\nu\nu^T\Sigma}{\left(\nu^T\Sigma\nu\right)^2}.$$

Partition the vectors ν, μ, and the matrix Σ according to the sparsity pattern in ν:

$$\begin{aligned}
\nu &= \left[\begin{array}{cc} \nu_{\mathcal{A}}^T & \mathbf{0}^T \end{array}\right]^T, \\
\mu &= \left[\begin{array}{cc} \mu_{\mathcal{A}}^T & \mu_{\mathcal{A}^c}^T \end{array}\right]^T, \text{ and} \\
\Sigma &= \left[\begin{array}{cc} \Sigma_{\mathcal{A},\mathcal{A}} & \Sigma_{\mathcal{A},\mathcal{A}^c} \\ \Sigma_{\mathcal{A}^c,\mathcal{A}} & \Sigma_{\mathcal{A}^c,\mathcal{A}^c} \end{array}\right].
\end{aligned}$$

The partitioned Hessian becomes

$$H(\theta) \;=\; \left[\begin{array}{cc} \left(\mu_{\mathcal{A}}\mu_{\mathcal{A}}^T + \Sigma_{\mathcal{A},\mathcal{A}}\right)\cdot\left(\mathbf{I}_q + \kappa\cdot\nu_{\mathcal{A}}\nu_{\mathcal{A}}^T\cdot\Sigma_{\mathcal{A},\mathcal{A}}\right) & \left(\mathbf{I}_q + \kappa\cdot\Sigma_{\mathcal{A},\mathcal{A}}\cdot\nu_{\mathcal{A}}\nu_{\mathcal{A}}^T\right)\cdot\left(\mu_{\mathcal{A}}\mu_{\mathcal{A}^c}^T + \Sigma_{\mathcal{A},\mathcal{A}^c}\right) \\ \left(\mu_{\mathcal{A}^c}\mu_{\mathcal{A}}^T + \Sigma_{\mathcal{A}^c,\mathcal{A}}\right)\cdot\left(\mathbf{I}_q + \kappa\cdot\nu_{\mathcal{A}}\nu_{\mathcal{A}}^T\cdot\Sigma_{\mathcal{A},\mathcal{A}}\right) & \mu_{\mathcal{A}^c}\mu_{\mathcal{A}^c}^T + \Sigma_{\mathcal{A}^c,\mathcal{A}^c} + \kappa\cdot\Sigma_{\mathcal{A}^c,\mathcal{A}}\cdot\nu_{\mathcal{A}}\nu_{\mathcal{A}}^T\cdot\Sigma_{\mathcal{A},\mathcal{A}^c} \end{array}\right].$$

Defining $\mathbf{A} := \left(\mathbf{I}_q + \kappa\cdot\nu_{\mathcal{A}}\nu_{\mathcal{A}}^T\cdot\Sigma_{\mathcal{A},\mathcal{A}}\right)$, we get

$$\begin{aligned}
H_{\mathcal{A}^c,\mathcal{A}}(\theta)\left[H_{\mathcal{A},\mathcal{A}}(\theta)\right]^{-1} &= \left(\mu_{\mathcal{A}^c}\mu_{\mathcal{A}}^T + \Sigma_{\mathcal{A}^c,\mathcal{A}}\right)\cdot\mathbf{A}\times\left[\left(\mu_{\mathcal{A}}\mu_{\mathcal{A}}^T + \Sigma_{\mathcal{A},\mathcal{A}}\right)\cdot\mathbf{A}\right]^{-1} \\
&= \left(\mu_{\mathcal{A}^c}\mu_{\mathcal{A}}^T + \Sigma_{\mathcal{A}^c,\mathcal{A}}\right)\mathbf{A}\mathbf{A}^{-1}\left(\mu_{\mathcal{A}}\mu_{\mathcal{A}}^T + \Sigma_{\mathcal{A},\mathcal{A}}\right)^{-1} \\
&= \left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}^c,\mathcal{A}}\right)\right]\left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A},\mathcal{A}}\right)\right]^{-1}.
\end{aligned}$$

The result follows from post-multiplying both sides of this last equation by $\text{sign}(\beta_{\mathcal{A}})$. □

Proof of Lemma 8. Throughout the proof of Lemma 8, we define $\tilde{\mathbf{X}} = \left[\begin{array}{cc} 1 & \mathbf{X}^T \end{array}\right]^T$.

L1) We first prove the existence of a minimizer. Given that the risk function is continuous, it is enough to prove that the closed set $\mathcal{S}(\mathbf{M}) = \left\{t \in \mathbb{R}^p : \mathbb{E}\left[-\mathbb{I}(\mathbf{Y}=1)\cdot\tilde{\mathbf{X}}^T t + \log\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)\right] \leq \mathbf{M}\right\}$ for large enough $\mathbf{M}$ is bounded. We establish boundedness of $\mathcal{S}(\mathbf{M})$ by proving that $\mathcal{S}(\mathbf{M})$ is contained on a finite sphere around the origin which can be established by proving that, for any $\mathbf{M}$ there exists γ such that:

$$|\langle t,t\rangle| > \gamma \quad\Rightarrow\quad \mathbb{E}\left[-\mathbb{I}(\mathbf{Y}=1)\cdot\tilde{\mathbf{X}}^T t + \log\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)\right] > \mathbf{M}. \tag{A-4}$$

To prove the assertion in (A-4), let t be a non-zero vector with $u = \sqrt{\langle t,t\rangle} \neq 0$, so t can be written

as $t = uv$, for some $v \neq 0$. For any $u_0 \in \mathbb{R}$, we can write:

$$\begin{aligned}
\mathbb{E}\left[-u \cdot \mathbb{I}(\mathbf{Y}=1) \cdot \tilde{\mathbf{X}}^T v + \log\left(1+\exp\left(u \cdot \tilde{\mathbf{X}}^T v\right)\right)\right] &= \\
-u \cdot \mathbb{E}\left[p(\mathbf{X}) \cdot \tilde{\mathbf{X}}^T v\right] + \mathbb{E}\left[\log\left(1+\exp\left(u \cdot \tilde{\mathbf{X}}^T v\right)\right)\right] &\geq \\
-u \cdot \mathbb{E}\left[p(\mathbf{X}) \cdot \tilde{\mathbf{X}}^T v\right] + \mathbb{E}\left[\log\left(1+\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)\right)\right] + \mathbb{E}\left[\frac{\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)}{1+\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)} \cdot \tilde{\mathbf{X}}^T v\right] \cdot (u - u_0) &= \\
c(u_0, v) + \mathbb{E}\left[\left(\frac{\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)}{1+\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)} - p(\mathbf{X})\right) \cdot \tilde{\mathbf{X}}^T v\right] \cdot u,&
\end{aligned}$$

where the inequality follows from convexity of the mapping $s \mapsto \log\left(1+\exp\left(s \cdot \tilde{\mathbf{X}}^T v\right)\right)$ and $c(u_0, v) = \mathbb{E}\left[\log\left(1+\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)\right)\right] - \mathbb{E}\left[\frac{\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)}{1+\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)} \cdot \tilde{\mathbf{X}}^T v\right] \cdot u_0$, which does not involve u.

Under the assumptions made, we can use the dominated convergence theorem to get:

$$\begin{aligned}
\lim_{u_0 \to \infty} \mathbb{E}\left[\left(\frac{\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)}{1+\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)} - p(\mathbf{X})\right) \cdot \tilde{\mathbf{X}}^T v\right] &= \mathbb{E}\left[(1-p(\mathbf{X})) \cdot \tilde{\mathbf{X}}^T v\right], \text{ and} \\
\lim_{u_0 \to -\infty} \mathbb{E}\left[\left(\frac{\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)}{1+\exp\left(u_0 \cdot \tilde{\mathbf{X}}^T v\right)} - p(\mathbf{X})\right) \cdot \tilde{\mathbf{X}}^T v\right] &= -\mathbb{E}\left[p(\mathbf{X}) \cdot \tilde{\mathbf{X}}^T v\right].
\end{aligned}$$

Since the density is everywhere positive, the hyperplane $\{\mathbf{s} \in \mathbb{R}^{p+1} : \mathbf{s}^T v = 0\}$ has probability zero for any $v \neq 0$ and thus we have either $\mathbb{E}\left(\tilde{\mathbf{X}} v\right) > 0$ or $\mathbb{E}\left(\tilde{\mathbf{X}} v\right) < 0$.

If $\mathbb{E}\left(\tilde{\mathbf{X}} v\right) > 0$, pick u_0 large enough so $\mathbb{E}\left[(1-p(\mathbf{X})) \cdot \tilde{\mathbf{X}}^T v\right] > 0$ to conclude that

$$\lim_{u \to \infty} \mathbb{E}\left[-\mathbb{I}(\mathbf{Y}=1) \cdot \tilde{\mathbf{X}}^T t + \log\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)\right] = \infty.$$

If, on the other hand, $\mathbb{E}\left(\tilde{\mathbf{X}} v\right) < 0$, pick u_0 small enough so $\mathbb{E}\left[p(\mathbf{X}) \cdot \tilde{\mathbf{X}}^T v\right] < 0$ to conclude:

$$\lim_{u \to \infty} \mathbb{E}\left[-\mathbb{I}(\mathbf{Y}=1) \cdot \tilde{\mathbf{X}}^T t + \log\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)\right] = \infty.$$

This establishes that for any $\mathbf{M}$ there exists γ such that the risk function exceeds $\mathbf{M}$ and completes the proof of existence.

The proof of uniqueness follows from strict convexity of the risk function. We prove that below by showing that the Hessian matrix of the risk function is everywhere strictly positive definite under the assumptions made.

L2) For the canonical logistic regression loss function, we have:

$$\begin{aligned}\mathbb{E}\left[|L(\mathbf{Y},\mathbf{X},t)|\right] &= \mathbb{E}\left[\left|\mathbb{I}(\mathbf{Y}=1)\cdot\tilde{\mathbf{X}}^T t-\log\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)\right|\right]\\ &\leq \mathbb{E}\left[\left|\tilde{\mathbf{X}}\right|\right]^T t+\mathbb{E}\left[\left|\log\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)\right|\right]\\ &= \mathbb{E}\left[\left|\tilde{\mathbf{X}}\right|\right]^T t+\mathbb{E}\left[\log\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)\right],\end{aligned}$$

where the equality follows from $\exp(\tilde{\mathbf{X}}^T t)\geq 0$ for all t. Because $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right]<\infty$, there exists C such that $\mathbb{E}\left[|\mathbf{X}_j|\right]<C$ for all $j=1,\ldots,p$ and the first term of the sum is bounded above. To bound the second term, write:

$$\begin{aligned}\mathbb{E}\left[\log\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)\right] &\leq \mathbb{E}\left[\log\left(1+\exp\left(\left|\tilde{\mathbf{X}}^T t\right|\right)\right)\right]\\ &\leq \log(2)+2\cdot\mathbb{E}\left[|\tilde{\mathbf{X}}^T t|\right],\end{aligned}$$

where the first inequality follows from $h(u):=\log\left(1+\exp\left(u\right)\right)$ being non-decreasing and the second stems from h having derivatives bounded above by 1. The result now follows from $\mathbb{E}\left|\mathbf{X}\right|$ being bounded.

L3) The canonical logistic regression loss function is twice differentiable everywhere, with:

$$\begin{aligned}\nabla_t L(\mathbf{Y},\mathbf{X},t) &= \left[\mathbb{I}\left(\mathbf{Y}=1\right)-\frac{\exp\left(\tilde{\mathbf{X}}^T t\right)}{1+\exp\left(\tilde{\mathbf{X}}^T t\right)}\right]\cdot\tilde{\mathbf{X}},\text{ and}\\ \nabla_t^2 L(\mathbf{Y},\mathbf{X},t) &= \frac{\exp\left(\tilde{\mathbf{X}}^T t\right)}{\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)^2}\cdot\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T.\end{aligned}$$

For all $\mathbf{Y}$, $\mathbf{X}$ and $t\in\mathbb{R}^{p+1}$, we have:

$$\begin{aligned}\left[(2\mathbf{Y}-1)-\frac{\exp\left(\tilde{\mathbf{X}}^T t\right)}{1+\exp\left(\tilde{\mathbf{X}}^T t\right)}\right] &\leq 2,\text{ and}\\ \frac{\exp\left(\tilde{\mathbf{X}}^T t\right)}{\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)^2} &\leq 1,\end{aligned}$$

so, using the assumptions on the moments of $\mathbf{X}$, we know that:

$$\begin{aligned}\mathbb{E}\left[|\nabla_t L(\mathbf{Y},\mathbf{X},t)|\right] &\leq 2\max\left[1,\max_{1\leq j\leq p}\mathbb{E}\left|\mathbf{X}_j\right|\right]<\infty,\text{ and}\\ \mathbb{E}\left[\left|\nabla_t^2 L(\mathbf{Y},\mathbf{X},t)\right|\right] &\leq \mathbb{E}\left[\mathbf{Q}(\mathbf{X})\right]<\infty.\end{aligned}$$

Using the Dominated Convergence Theorem we get that:

$$\begin{aligned}
\nabla_t R(t) &= \mathbb{E}\left[\nabla_t L(\mathbf{Y},\mathbf{X},t)\right] = \mathbb{E}\left[\left(\mathbb{E}\left(\mathbb{I}(\mathbf{Y}=1)|\mathbf{X}\right) - \frac{\exp(\tilde{\mathbf{X}}^T t)}{1+\exp(\tilde{\mathbf{X}}^T t)}\right)\cdot\tilde{\mathbf{X}}\right] \\
&= \mathbb{E}\left[\left(p(\mathbf{X}) - \frac{\exp(\tilde{\mathbf{X}}^T t)}{1+\exp(\tilde{\mathbf{X}}^T t)}\right)\cdot\tilde{\mathbf{X}}\right], \text{ and} \\
\nabla_t^2 R(t) &= \mathbb{E}\left[\nabla_t^2 L(\mathbf{Y},\mathbf{X},t)\right] = \mathbb{E}\left[\frac{\exp\left(\tilde{\mathbf{X}}^T t\right)}{\left(1+\exp\left(\tilde{\mathbf{X}}^T t\right)\right)^2}\cdot\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\right].
\end{aligned}$$

We now prove that the population risk minimizer θ for the logistic regression is unique under the conditions of Assumption Set 3, by proving that the Hessian $\nabla_t^2 R(\theta)$ is a strictly positive definite matrix.

From the assumption that $\mathbb{E}\left[Q(\mathbf{X})\right]$ is strictly positive definite and bounded, we get that:

$$\mathbb{E}\left[Q(\mathbf{X})\right] = \lim_{s\to\infty}\mathbb{E}\left[Q(\mathbf{X})\cdot\mathbb{I}(\|\mathbf{X}\|\leq s)\right],$$

and thus, there must exist large enough S such that $\mathbb{E}\left[Q(\mathbf{X})\cdot\mathbb{I}(\|\mathbf{X}\|\leq S)\right]$ is strictly positive definite. Let $\xi = \inf_{\|\mathbf{X}\|\leq S}\frac{\exp(\alpha+\mathbf{X}^T\beta)}{\left[1+\exp(\alpha+\mathbf{X}^T\beta)\right]^2}$. Because $\|\mathbf{X}\|\leq S$ is a compact set and $\frac{\exp(\alpha+\mathbf{X}^T\beta)}{\left[1+\exp(\alpha+\mathbf{X}^T\beta)\right]^2} > 0$ for all $\mathbf{X}$, we get that $\xi > 0$.

In what follows, the binary relationship between matrices A and B indicated by $A \succeq B$ means $A-B$ is positive semi-definite and its strict version $A \succ B$ means $A-B$ is strictly positive definite. Now,

$$\begin{aligned}
\nabla_t^2 R(\theta) &= \mathbb{E}\left[\frac{\exp(\tilde{\mathbf{X}}^T t)}{\left(1+\exp(\tilde{\mathbf{X}}^T t)\right)^2}\cdot Q(\mathbf{X})\right] \\
&= \mathbb{E}\left[\frac{\exp(\tilde{\mathbf{X}}^T t)}{\left(1+\exp(\tilde{\mathbf{X}}^T t)\right)^2}\cdot Q(\mathbf{X})\cdot\mathbb{I}\left(\|\mathbf{X}\|\leq S\right)\right] + \mathbb{E}\left[\frac{\exp(\tilde{\mathbf{X}}^T t)}{\left(1+\exp(\tilde{\mathbf{X}}^T t)\right)^2}\cdot Q(\mathbf{X})\cdot\mathbb{I}\left(\|\mathbf{X}\|> S\right)\right] \\
&\succeq \xi\cdot\mathbb{E}\left[Q(\mathbf{X})\cdot\mathbb{I}\left(\|\mathbf{X}\|\leq S\right)\right] + \mathbb{E}\left[\frac{\exp(\tilde{\mathbf{X}}^T t)}{\left(1+\exp(\tilde{\mathbf{X}}^T t)\right)^2}\cdot Q(\mathbf{X})\cdot\mathbb{I}\left(\|\mathbf{X}\|> S\right)\right] \succ 0,
\end{aligned}$$

where the last generalized inequality follows from $\mathbb{E}\left[Q(\mathbf{X})\cdot\frac{\exp(\tilde{\mathbf{X}}^T t)}{\left(1+\exp(\tilde{\mathbf{X}}^T t)\right)^2}\cdot\mathbb{I}\left(\|\mathbf{X}\|> S\right)\right] \succeq 0$, $\xi > 0$, and $\mathbb{E}\left[Q(\mathbf{X})\cdot\mathbb{I}\left(\|\mathbf{X}\|\leq S\right)\right] \succ 0$.

L4) The loss function corresponds to the neg-loglikelihood function of a canonical exponential family and is thus convex. As the risk is an expected value of convex functions, it is also convex.

□

Proof of Lemma 9. Throughout the proof of Lemma 9, we define $\tilde{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{X}^T \end{bmatrix}^T$.

L1) We first prove that a minimizer exist. Given that the risk function is continuous, it is enough to prove that for large enough M the closed set $\mathcal{S}(M) = \left\{ t \in \mathbb{R}^p : \mathbb{E}\left[\left| 1 - \mathbf{Y}\tilde{\mathbf{X}}^T t \right|_+ \right] \leq M \right\}$ is bounded. We establish boundedness of $\mathcal{S}(M)$ by proving that $\mathcal{S}(M)$ is contained on a finite box around the origin. Letting e_j be a unit vector with a 1 in its j-th entry and zeroes in all other components, it is sufficient to prove that, for any M and each $j = 1, \ldots, p+1$, there exist $\gamma_{j,M}$ such that:

$$|\langle t, e_j \rangle| > \gamma_{j,M} \quad \Rightarrow \quad \mathbb{E}\left[\left| 1 - \mathbf{Y}\tilde{\mathbf{X}}^T t \right|_+ \right] > M. \tag{A-5}$$

To prove the assertion in (A-5), let t be a non-zero vector with $u = \langle t, e_j \rangle \neq 0$, so t can be written as $t = ue_j + v$, for some v with $\langle v, e_j \rangle = 0$. The risk function at t becomes:

$$\begin{aligned}
\mathbb{E}\left[\left| 1 - \mathbf{Y}\tilde{\mathbf{X}}^T t \right|_+ \right] &= \mathbb{E}\left[\left| 1 - u \cdot \tilde{\mathbf{X}}^T e_j - \tilde{\mathbf{X}}^T v \right| \mathbb{I}(\mathbf{Y} = 1) \right] + \mathbb{E}\left[\left| 1 + u \cdot \tilde{\mathbf{X}}^T e_j - \tilde{\mathbf{X}}^T v \right| \mathbb{I}(\mathbf{Y} = -1) \right] \\
&\geq \mathbb{E}\left[\left| u \cdot \tilde{\mathbf{X}}^T e_j + \tilde{\mathbf{X}}^T v \right| \right] - 1 \\
&\geq \inf_{v \in \mathcal{O}(e_j)} \mathbb{E}\left[\left| u \cdot \tilde{\mathbf{X}}^T e_j + \tilde{\mathbf{X}}^T v \right| \right] - 1 \\
&= |u| \cdot \inf_{v \in \mathcal{O}(e_j)} \mathbb{E}\left[\left| \tilde{\mathbf{X}}^T e_j + \tilde{\mathbf{X}}^T v \right| \right] - 1,
\end{aligned}$$

where $\mathcal{O}(e_j)$ is the set of all vectors orthogonal to e_j. Because e_j has unit norm, $\|e_j + v\| \geq 1$ for all $v \in \mathcal{O}(e_j)$ and it follows that $\{e_j + v : v \in \mathcal{O}(e_j)\} \subset \{v : \|v\| \geq 1\}$, yielding:

$$\mathbb{E}\left[\left| 1 - \mathbf{Y}\tilde{\mathbf{X}}^T t \right|_+ \right] \geq |u| \cdot \inf_{v : \|v\| \geq 1} \mathbb{E}\left[\left| \tilde{\mathbf{X}}^T v \right| \right] - 1 = |u| \cdot \inf_{v : \|v\| = 1} \mathbb{E}\left[\left| \tilde{\mathbf{X}}^T v \right| \right] - 1,$$

where the equality follows from noticing that $\left| \tilde{\mathbf{X}}^T v \right|$ is increasing in $|v|$. If we can find $c > 0$, such that $\inf_{v:\|v\|=1} \mathbb{E}\left[\left| \tilde{\mathbf{X}}^T v \right| \right] > c$, it is possible to find the $\gamma_{j,M}$ we want. To find such a positive lower bound, define the compact set $\mathbf{K} = \{\tilde{x} \in \mathbb{R}^{p+1} : \|\tilde{x}\|_2 \leq \mathbf{C}\}$ for some constant $\mathbf{C}$. Since it is assumed that $f_{\mathbf{X}}(\mathbf{x}) > 0$ is continuous for all $\mathbf{x} \in \mathbb{R}^p$, we get that $f^* = \min_{\mathbf{x} \in \mathbf{K}} > 0$. Now, letting $\mu(\mathbf{A})$ denote the Lebesgue measure of a set $\mathbf{A} \subset \mathbb{R}^{p+1}$, we have:

$$\begin{aligned}
\inf_{v:\|v\|=1} \mathbb{E}\left[\left| \tilde{\mathbf{X}}^T v \right| \right] &\geq \eta \cdot \inf_{v:\|v\| \geq 1} \mathbb{P}\left[\left| \tilde{\mathbf{X}}^T v \right| > \eta \right] \\
&\geq \inf_{v:\|v\| \geq 1} \mathbb{P}\left[\tilde{\mathbf{X}}^T v > \eta, \text{ and } \tilde{\mathbf{X}} \in \mathbf{K} \right] \\
&\geq f^* \cdot \inf_{v:\|v\| \geq 1} \mu\left(\{\mathbf{x} : \tilde{\mathbf{x}}^T v > \eta\} \right). \\
&= f^* \cdot \mu\left(\{\mathbf{x} : \tilde{\mathbf{x}}^T e_1 > \eta\} \right) =: c_\eta > 0,
\end{aligned}$$

where the last equality follows from noticing that, because of symmetry:

$$\mu\left(\{\mathbf{x} : \tilde{\mathbf{x}}^T v > \eta\}\right) = \mu\left(\{\mathbf{x} : \tilde{\mathbf{x}}^T e_1 > \eta\}\right), \text{ for all } v \in \mathbb{R}^{p+1} : \|v\|_2 = 1.$$

Using the strictly positive lower bound afforded by c_η, we get:

$$|u| = |\langle t, e_j\rangle| > \gamma_{j,M} := \frac{M+1}{f^* c_\eta}, \text{ for some } j = 1, \ldots, p \quad \Rightarrow \quad \mathbb{E}\left[\left|1 - \mathbf{Y}\tilde{\mathbf{X}}^T t\right|_+\right] > M.$$

Uniqueness of the minimizer follows from strict convexity of the risk function under the assumptions made. Strict convexity of the risk function in its turn is proved below, by showing the Hessian matrix for the risk is everywhere strictly positive definite.

L2) For all $t \in \mathbb{R}^{p+1}$:

$$\mathbb{E}\left[|L(\mathbf{Z}, t)|\right] \leq \mathbb{E}\left[\max\left\{\left|1 - \tilde{\mathbf{X}}t\right|, \left|1 + \tilde{\mathbf{X}}t\right|\right\}\right] \leq 1 + \mathbb{E}\left[\left|\tilde{\mathbf{X}}t\right|\right] \leq 1 + \sqrt{t^t \cdot \mathbb{E}\left[\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right] \cdot t},$$

which is bounded given the assumptions on the distribution of $\mathbf{X}$.

L3) The hinge loss is not differentiable on the set $\{\mathbf{X} : \tilde{\mathbf{X}}t = 1 \text{ or } \tilde{\mathbf{X}}t = -1\}$, which under the assumed conditions has zero probability. At all other points, hinge loss function has derivative with respect to t

$$\nabla_t L(\mathbf{Z}, t) \quad = \quad \tilde{\mathbf{X}} \cdot \left[\mathbb{I}(\mathbf{Y} = 1) \cdot \mathbb{I}(\tilde{\mathbf{X}}t - 1 < 0) - \mathbb{I}(\mathbf{Y} = -1) \cdot \mathbb{I}(\tilde{\mathbf{X}}t + 1 > 0)\right]$$

To obtain the Hessian, write the SVM risk as $R(t) = R_1(t) + R_2(t)$, with

$$\begin{aligned} R_1(t) &:= \mathbb{E}\left[p(\mathbf{X}) \cdot \left(1 - \tilde{\mathbf{X}}t\right) \cdot \mathbb{I}\left(1 - \tilde{\mathbf{X}}t > 0\right)\right], \text{ and} \\ R_2(t) &:= \mathbb{E}\left[(1 - p(\mathbf{X})) \cdot \left(\tilde{\mathbf{X}}t - 1\right) \cdot \mathbb{I}\left(\tilde{\mathbf{X}}t - 1 > 0\right)\right]. \end{aligned}$$

Let $\nabla_t R_1(t) := \mathbb{E}\left[p(\tilde{\mathbf{X}}) \cdot \mathbb{I}\left(1 - \tilde{\mathbf{X}}t > 0\right) \cdot \tilde{\mathbf{X}}\right]$ and $\nabla_t^2 R_1(t) := \mathbb{E}\left[p(\tilde{\mathbf{X}}) \cdot \delta(1 - \tilde{\mathbf{X}}t) \cdot \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right]$. We first show that $r_1(t) := R_1(t + \Delta t) - R_1(t) = \nabla_t R_1(t) \cdot \Delta t = o(\|\Delta t\|)$. To do that, let $\mathbf{d}_k = \frac{\Delta t_k}{|\Delta t_k|}$ and write

$$\begin{aligned} \frac{R_1(t + \Delta t_k) - R_1(t)}{\Delta t_k} \quad = \quad & \mathbb{E}\left[\frac{p(\tilde{\mathbf{X}})\left(1 - \tilde{\mathbf{X}}t\right)\left[\mathbb{I}\left(1 - \tilde{\mathbf{X}}t > \tilde{\mathbf{X}}\Delta t\right) - \mathbb{I}\left(1 - \tilde{\mathbf{X}}t > 0\right)\right]}{\tilde{\mathbf{X}}\Delta t} \cdot \tilde{\mathbf{X}}\right] \mathbf{d}_k \\ & -\mathbb{E}\left[p(\tilde{\mathbf{X}}) \cdot \mathbb{I}\left(1 - \tilde{\mathbf{X}}t > \tilde{\mathbf{X}}\Delta t\right) \cdot \tilde{\mathbf{X}}\right] \mathbf{d}_k. \end{aligned}$$

Using the Dominated Convergence Theorem to take the limit as $\Delta t \downarrow 0$ and collecting the limit of

the multiplier of $\mathbf{d}_k$ yields

$$\begin{aligned}
\nabla_t R_1(t) &= \lim_{|\Delta t|\to 0} \mathbb{E}\left[\frac{p(\tilde{\mathbf{X}})\left(1-\tilde{\mathbf{X}}t\right)\left[\mathbb{I}\left(1-\tilde{\mathbf{X}}t>\tilde{\mathbf{X}}\Delta t\right)-\mathbb{I}\left(1-\tilde{\mathbf{X}}t>0\right)\right]}{\tilde{\mathbf{X}}\Delta t}\cdot\tilde{\mathbf{X}}\right] \\
&\quad - \lim_{|\Delta t|\to 0}\mathbb{E}\left[p(\tilde{\mathbf{X}})\cdot\mathbb{I}\left(1-\tilde{\mathbf{X}}t>\tilde{\mathbf{X}}\Delta t\right)\cdot\tilde{\mathbf{X}}\right] \\
&= \mathbb{E}\left[p(\tilde{\mathbf{X}})\left(1-\tilde{\mathbf{X}}t\right)\delta(1-\tilde{\mathbf{X}}^T t)\cdot\tilde{\mathbf{X}}\right]-\mathbb{E}\left[p(\tilde{\mathbf{X}})\cdot\mathbb{I}\left(1-\tilde{\mathbf{X}}t>0\right)\cdot\tilde{\mathbf{X}}\right] \\
&= -\mathbb{E}\left[p(\tilde{\mathbf{X}})\cdot\mathbb{I}\left(1-\tilde{\mathbf{X}}t>0\right)\cdot\tilde{\mathbf{X}}\right].
\end{aligned}$$

To obtain the second differential for $R_1(t)$, write the residuals from the approximation from the first differential:

$$\begin{aligned}
\frac{r_1(\Delta t)}{|\Delta t_k|^2} &= \mathbf{d}_k\mathbb{E}\left[\frac{p(\tilde{\mathbf{X}})\cdot\tilde{\mathbf{X}}^T\cdot\left[\mathbb{I}\left(1-\tilde{\mathbf{X}}t>\tilde{\mathbf{X}}\Delta t\right)-\mathbb{I}\left(1-\tilde{\mathbf{X}}t>0\right)\right]\cdot\left(1-\tilde{\mathbf{X}}t\right)\cdot\tilde{\mathbf{X}}}{\Delta t^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\Delta t}\right]\mathbf{d}_k \\
&\quad -\mathbf{d}_k\cdot\mathbb{E}\left[\frac{p(\tilde{\mathbf{X}})\cdot\tilde{\mathbf{X}}^T\cdot\left[\mathbb{I}\left(1-\tilde{\mathbf{X}}t>\tilde{\mathbf{X}}\Delta t\right)-\mathbb{I}\left(1-\tilde{\mathbf{X}}t>0\right)\right]\cdot\tilde{\mathbf{X}}}{\tilde{\mathbf{X}}\Delta t}\right]\cdot\mathbf{d}_k
\end{aligned}$$

The second derivative is obtained using the Dominated Convergence Theorem to compute the limits of the terms in the sum. For the second term, the limit follows directly from pointwise convergence to a Dirac delta function:

$$\lim_{|\Delta t|\to 0}\mathbb{E}\left[\frac{p(\tilde{\mathbf{X}})\cdot\tilde{\mathbf{X}}\cdot\left[\mathbb{I}\left(1-\tilde{\mathbf{X}}^T t>\tilde{\mathbf{X}}^T\Delta t\right)-\mathbb{I}\left(1-\tilde{\mathbf{X}}^T t>0\right)\right]\cdot\tilde{\mathbf{X}}^T}{\tilde{\mathbf{X}}^T\Delta t}\right]=-\mathbb{E}\left[p(\tilde{\mathbf{X}})\cdot\delta(1-\tilde{\mathbf{X}}^T t)\cdot\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\right].$$

To obtain the limit for the other term, let $W=R\tilde{\mathbf{x}}=\begin{bmatrix}\mathbf{w}_1 & \mathbf{w}_2\end{bmatrix}\in\mathbb{R}\times\mathbb{R}^{p-1}$ be a linear rotation of $\tilde{x}$ such that $\mathbf{w}_1=\tilde{\mathbf{x}}^T t$, and let $F_{\tilde{\mathbf{x}}}$, $F_{\mathbf{w}_2}$, and $f_{\mathbf{w}_1|\mathbf{w}_2}$ denote the distributions of $\tilde{X}$, W_2 and the conditional distribution of W_1 given W_2 respectively. Then write

$$\begin{aligned}
&\mathbb{E}\left[\frac{p(\tilde{\mathbf{X}})\cdot\tilde{\mathbf{X}}^T\cdot\left[\mathbb{I}\left(1-\tilde{\mathbf{X}}t>\tilde{\mathbf{X}}\Delta t_k\right)-\mathbb{I}\left(1-\tilde{\mathbf{X}}t>0\right)\right]\cdot\left(1-\tilde{\mathbf{X}}t\right)\cdot\tilde{\mathbf{X}}}{\Delta t_k^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\Delta t_k}\right] &&= \\
&\int\left[\frac{(1-\tilde{\mathbf{x}}t)\cdot[\mathbb{I}(1-\tilde{\mathbf{x}}t>\tilde{\mathbf{x}}\Delta t_k)-\mathbb{I}(1-\tilde{\mathbf{x}}t>0)]}{\Delta t_k^T\tilde{\mathbf{x}}^T\tilde{\mathbf{x}}^T\Delta t_k}\cdot p(\tilde{\mathbf{x}})\cdot\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\right]\mathrm{d}F_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}) &&= \\
&\int\int\frac{(1-\mathbf{w}_1)\cdot[\mathbb{I}(1-\mathbf{w}_1>\tilde{\mathbf{x}}(\mathbf{w})\Delta t_k)-\mathbb{I}(1-\mathbf{w}_1>0)]}{\Delta t_k^T\tilde{\mathbf{x}}(\mathbf{w})\tilde{\mathbf{x}}^T(\mathbf{w})\Delta t_k}s(\mathbf{w}_1,\mathbf{w}_2)dF_{\mathbf{w}_1|\mathbf{w}_2}(\mathbf{w}_1|\mathbf{w}_2)dF_{\mathbf{w}_2}(\mathbf{w}_2), &&
\end{aligned}$$

where we used the notation $s(\mathbf{w}_1,\mathbf{w}_2):=p(\tilde{\mathbf{x}}(\mathbf{w}_1,\mathbf{w}_2))\mathbf{x}(\mathbf{w}_1,\mathbf{w}_2)\mathbf{x}(\mathbf{w}_1,\mathbf{w}_2)^T$.

To obtain the limit, write the inner integral as:

$$\begin{aligned}
&\int\left[\frac{(1-\mathbf{w}_1)\cdot[\mathbb{I}(1-\mathbf{w}_1>\tilde{\mathbf{x}}(\mathbf{w}_1,\mathbf{w}_2)\Delta t_k)-\mathbb{I}(1-\mathbf{w}_1>0)]}{\Delta t_k^T\tilde{\mathbf{x}}(\mathbf{w}_1,\mathbf{w}_2)\tilde{\mathbf{x}}^T(\mathbf{w})\Delta t_k}\cdot s(\mathbf{w}_1,\mathbf{w}_2)\right]dF_{\mathbf{w}_1|\mathbf{w}_2}(\mathbf{w}_1|\mathbf{w}_2) &&= \\
&\int\left[\frac{v\cdot[\mathbb{I}(v>\tilde{\mathbf{x}}(\mathbf{w}_1,\mathbf{w}_2)\Delta t_k)-\mathbb{I}(v>0)]}{\Delta t_k^T\tilde{\mathbf{x}}(1-v,\mathbf{w}_2)\tilde{\mathbf{x}}^T(\mathbf{w})\Delta t_k}\cdot s(1-v,\mathbf{w}_2)\right]dF_{\mathbf{w}_1|\mathbf{w}_2}(1-v|\mathbf{w}_2) &&= \\
&-\int\left[\tfrac{1}{2}\cdot s(\mathbf{w}_1,\mathbf{w}_2)\cdot\delta(1-\mathbf{w}_1)\right]dF_{\mathbf{w}_1|\mathbf{w}_2}(\mathbf{w}_1|\mathbf{w}_2)+o(\|\Delta t\|). &&
\end{aligned}$$

Plugging that back into the expression for the expected value, we get:

$$\begin{array}{ll} \lim_{|\Delta t_k|\to 0} \mathbb{E}\left[\frac{p(\tilde{\mathbf{X}})\cdot\tilde{\mathbf{X}}^T\cdot\left[\mathbb{I}\left(1-\tilde{\mathbf{X}}t>\tilde{\mathbf{X}}\Delta t_k\right)-\mathbb{I}\left(1-\tilde{\mathbf{X}}t>0\right)\right]\cdot\left(1-\tilde{\mathbf{X}}t\right)\cdot\tilde{\mathbf{X}}}{\Delta t_k^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\Delta t_k}\right] & = \\ -\int\left[\int\left[\frac{1}{2}\cdot s(\mathbf{w}_1,\mathbf{w}_2)\cdot\delta(1-\mathbf{w}_1)\right]dF_{\mathbf{w}_1|\mathbf{w}_2}(\mathbf{w}_1\,|\mathbf{w}_2)\right]dF\mathbf{w}_2(\mathbf{w}_2) & = \\ -\frac{1}{2}\cdot\mathbb{E}\left[p(\tilde{\mathbf{X}})\cdot\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\cdot\delta(1-\tilde{\mathbf{X}}^Tt)\right]. & \end{array}$$

Summing the two terms (and taking into account the factor $\frac{1}{2}$ in the Taylor expansion) yield

$$\nabla_t^2 R_1(t) = \mathbb{E}\left[p(\tilde{\mathbf{X}})\cdot\delta(1-\tilde{\mathbf{X}}^Tt)\cdot\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\right].$$

For R_2, analogous steps yield

$$\begin{array}{rcl} \nabla_t R_2(t) & = & \mathbb{E}\left[\left(1-p(\tilde{\mathbf{X}})\right)\cdot\mathbb{I}\left(1+\tilde{\mathbf{X}}t>0\right)\cdot\tilde{\mathbf{X}}\right], \text{ and} \\ \nabla_t^2 R_2(t) & = & \mathbb{E}\left[\left(1-p(\tilde{\mathbf{X}})\right)\cdot\delta(1+\tilde{\mathbf{X}}^Tt)\cdot\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\right]. \end{array}$$

The result follows from summing the differentials for $R_1(t)$ and $R_2(t)$.

$$\begin{array}{rcl} \nabla_t R(t) & = & \mathbb{E}\left[\left(\left(1-p(\tilde{\mathbf{X}})\right)\cdot\mathbb{I}\left(1+\tilde{\mathbf{X}}t>0\right)-p(\tilde{\mathbf{X}})\cdot\mathbb{I}\left(1+\tilde{\mathbf{X}}t>0\right)\right)\cdot\tilde{\mathbf{X}}\right], \text{ and} \\ \nabla_t^2 R(t) & = & \mathbb{E}\left[\left(\left(1-p(\tilde{\mathbf{X}})\right)\cdot\delta(1+\tilde{\mathbf{X}}^Tt)+p(\tilde{\mathbf{X}})\cdot\delta(1-\tilde{\mathbf{X}}^Tt)\right)\cdot\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\right]. \end{array}$$

Finally, we prove that the minimizer of the SVM risk is unique by establishing that $\nabla_t^2 R(\theta)$ is strictly positive definite. To do that, first write:

$$\begin{array}{rcl} \nabla_t^2 R(t) & = & \mathbb{E}\left[Q(\mathbf{X})\cdot\mathbb{I}\left(\mathbf{Y}=1\right)\left|\alpha+\mathbf{X}^T\beta=1\right.\right]\cdot\tilde{f}(1) \\ & & +\mathbb{E}\left[Q(\mathbf{X})\cdot\mathbb{I}\left(\mathbf{Y}=-1\right)\left|\alpha+\mathbf{X}^T\beta=-1\right.\right]\cdot\tilde{f}(-1) \\ & = & \mathbb{E}\left[Q(\mathbf{X})\left|\mathbf{Y}=1,\alpha+\mathbf{X}^T\beta=1\right.\right]\cdot\mathbb{P}\left(\mathbf{Y}=1\left|\alpha+\mathbf{X}^T\beta=1\right.\right)\cdot\tilde{f}(1) \\ & & +\mathbb{E}\left[Q(\mathbf{X})\left|\mathbf{Y}=-1,\alpha+\mathbf{X}^T\beta=-1\right.\right]\cdot\mathbb{P}\left(\mathbf{Y}=-1\left|\alpha+\mathbf{X}^T\beta=-1\right.\right)\cdot\tilde{f}(-1), \end{array}$$

where $\tilde{f}$ denote the density of the random variable $\alpha+\mathbf{X}^T\beta$. Given assumption C2, $\tilde{f}(-1)>0$ and $\tilde{f}(1)>0$. In addition, assumption C3 gives that $\mathbb{P}\left(\mathbf{Y}=-1\left|\alpha+\mathbf{X}^T\beta=1\right.\right)>0$ and $\mathbb{P}\left(\mathbf{Y}=1\left|\alpha+\mathbf{X}^T\beta=1\right.\right)>0$. It is thus, enough to prove that either $\mathbb{E}\left[Q(\mathbf{X})\left|\mathbf{Y}=1,\alpha+\mathbf{X}^T\beta=-1\right.\right]$ or $\mathbb{E}\left[Q(\mathbf{X})\left|\mathbf{Y}=-1,\alpha+\mathbf{X}^T\beta=-1\right.\right]$ is strictly positive definite (or both).

Define $\mathbf{v}_\beta=\frac{\beta}{\|\beta\|}$, the unit vector in the direction of β. The condition $\alpha+\mathbf{X}^T\beta=\kappa$ is equivalent to

$\mathbf{X}^T\mathbf{v}_\beta = \frac{\kappa-\alpha}{\|\beta\|}$, so for any scalar κ:

$$\mathbb{E}\left[Q(\mathbf{X})\,\middle|\mathbf{Y}, \alpha + \mathbf{X}^T\beta = \kappa\right] \;=\; \mathbb{E}\left[Q(\mathbf{X})\,\middle|\mathbf{Y}, \mathbf{X}^T\mathbf{v}_\beta = \frac{\kappa-\alpha}{\|\beta\|}\right].$$

Then notice that:

$$\mathbb{E}\left[Q(\mathbf{X})\,\middle|\mathbf{Y}, \mathbf{X}^T\mathbf{v}_\beta = \frac{\kappa-\alpha}{\|\beta\|}\right] \succeq \left(\frac{\kappa-\alpha}{\|\beta\|}\right)^2 \left(\mathbf{v}_\beta\mathbf{v}_\beta^T\right) + \mathrm{var}\left(\mathbf{X}\,\middle|\mathbf{Y}, \mathbf{X}^T\mathbf{v}_\beta = \frac{\kappa-\alpha}{\|\beta\|}\right),$$

where $A \succeq B$ denotes that $A - B$ is positive semi definite. Because $\mathrm{var}\left[\mathbf{X}\,|\mathbf{Y}\right]$ is assumed to be non-singular, $\mathrm{var}\left(\mathbf{X}\,\middle|\mathbf{Y}, \mathbf{X}^T\mathbf{v}_\beta = \frac{\kappa-\alpha}{\|\beta\|}\right)$ has rank $p-1$ and $\mathrm{var}\left(\mathbf{X}\,\middle|\mathbf{Y}, \mathbf{X}^T\mathbf{v}_\beta = \frac{\kappa-\alpha}{\|\beta\|}\right)\mathbf{v}_\beta = 0$. Thus, as long as $\kappa \neq \alpha$, $\mathbb{E}\left[Q(\mathbf{X})\,\middle|\mathbf{Y}, \mathbf{X}^T\mathbf{v}_\beta = \frac{\kappa-\alpha}{\|\beta\|}\right]$ is strictly positive definite.

If $\alpha \notin \{-1, 1\}$, both terms in the sum defining $\nabla_t^2 R(\theta)$ are strictly positive definite. If $\alpha \in \{-1, 1\}$, one of the terms in the sum defining $\nabla_t^2 R(\theta)$ is singular, but the other is necessarily strictly positive definite. Thus, it follows that $\nabla_t^2 R(\theta)$ is strictly positive definite as stated.

L4) We can write $\|1 - \tilde{\mathbf{X}}t\|_+$ as the maximum between the constant function 0 and the function $\mathbf{Y} - \tilde{\mathbf{X}}t$ which is linear – thus, convex – on t. Since it is the maximum between two convex functions on t, $\|\mathbf{Y} - \tilde{\mathbf{X}}t\|_+$ is convex on t. A similar argument yields that $\|\tilde{\mathbf{X}}t - 1\|_-$ is convex on t.

The loss function $L(\mathbf{Z}, t)$ is written as the sum (with positive weights) of convex functions, which proves that AL.IV holds for the SVM loss function.

□

B Calculations for SVM and logistic risk Hessians in selected cases

In this section, we first obtain expressions of the second moment of the predictors given the value of the margin variable $\alpha + \mathbf{X}^T\beta$ in the case of predictors $\mathbf{X}$ having a Gaussian and a mixture of Gaussian distributions. Given the characterization of the SVM and logistic Hessians as a "weighted average" of such conditional second moments in Equations (22) and (23), the expressions for such conditional moments are useful in analytically comparing ℓ_1-penalized SVM and logistic classifiers with respect to their model selection properties. We then give explicit analytical expressions for the Hessian and logistic regression risk functions in the case of Gaussian and mixed Gaussian predictors.

For the duration of this section, $\mathbf{Z}$ denotes a rotated version of $\mathbf{X}$ whose first component is the projection of $\mathbf{X}$ along the direction normal to the optimal separating hyperplane $\mathcal{H}(\theta)$.

B.1 Conditional moments of Gaussian predictors given the value of one of its projections

To obtain the conditional second moments used in the expressions for Hessians of the SVM and logistic regression risk functions, we first construct an orthogonal matrix S according to

$$S \;:=\; \left[\begin{array}{cc} \nu & U \end{array}\right],$$

with U a $p \times (p-1)$ matrix constructed using a Gram-Schmidt orthogonalization (as long as $\beta \neq \mathbf{0}$). By construction, $U^T\nu = \mathbf{0}$ and $U^TU = \mathbf{I}_{p-1}$. The random vector $\mathbf{Z} = S^T\mathbf{X} \in \mathbb{R}^p$ is partitioned into a random scalar $\mathbf{Z}_1 = \nu^T\mathbf{X}$ in the direction of ν and a $p-1$ dimensional random vector $\mathbf{Z}_2 = U^T\mathbf{X}$ orthogonal to ν,

$$\mathbf{Z}^T \;:=\; \left[\begin{array}{cc} \mathbf{Z}_1 & \mathbf{Z}_2^T \end{array}\right] = \left[\begin{array}{cc} \mathbf{X}^T\nu & \mathbf{X}^TU \end{array}\right]^T.$$

Conditioning on the *margin variable* $\mathbf{M} = \alpha + \mathbf{X}^T\beta$ – defined in (18) – is equivalent to conditioning on the $\mathbf{Z}_1 = \nu^T\mathbf{X}$ since:

$$\alpha + \mathbf{X}^T\beta = \mathbf{M} \Leftrightarrow \nu^T\mathbf{X} = \frac{\mathbf{M}-\alpha}{\|\beta\|} \Leftrightarrow \mathbf{Z}_1 = \frac{\mathbf{M}-\alpha}{\|\beta\|}.$$

Since S is orthogonal, $\mathbf{X} = S\mathbf{Z}$ and

$$\mathbb{E}\left[\mathbf{X}\mathbf{X}^T|\nu^T\mathbf{X}\right] \;=\; S\mathbb{E}\left[\mathbf{Z}\mathbf{Z}^T|\mathbf{Z}_1\right]S^T \;=\; \left[S\,\mathbb{E}\left[\mathbf{Z}|\mathbf{Z}_1\right]\right]\cdot\left[S\,\mathbb{E}\left[\mathbf{Z}|\mathbf{Z}_1\right]\right]^T + S\cdot\mathrm{var}\left[\mathbf{Z}\,|\mathbf{Z}_1\right]\cdot S^T.$$

For $\mathbf{X} \sim \mathcal{N}(\mu,\Sigma)$, $\mathbf{Z}$ is also Gaussian with expected value $S^T\mu$ and variance $S^T\Sigma S$. Partitioning the expressions for the expected value and variance of $\mathbf{Z}$ we get

$$\mathbb{E}\mathbf{Z} \;=\; \left[\begin{array}{c} \nu^T\mu \\ U^T\mu \end{array}\right], \quad \text{and} \quad \mathrm{var}\left[\mathbf{Z}\right] \;=\; \left[\begin{array}{cc} \nu^T\Sigma\nu & \nu^T\Sigma U \\ U^T\Sigma\nu & U^T\Sigma U \end{array}\right].$$

Based on these expressions and standard results on multivariate Gaussian distributions, we get:

$$\begin{array}{rclrcll} \mathbb{E}\left[\mathbf{Z}_1\,|\mathbf{Z}_1\right] & = & \mathbf{Z}_1, & \mathbb{E}\left[\mathbf{Z}_2\,|\mathbf{Z}_1\right] & = & U^T\left[\mu + \frac{\Sigma\nu}{\nu^T\Sigma\nu}\left(\mathbf{Z}_1 - \nu^T\mu\right)\right], & \\ \mathrm{var}\left[\mathbf{Z}_1\,|\mathbf{Z}_1\right] & = & 0, & \mathrm{var}\left[\mathbf{Z}_2\,|\mathbf{Z}_1\right] & = & U^T\left[\Sigma - \frac{\Sigma\nu\nu^T\Sigma}{\nu^T\Sigma\nu}\right]U, & \text{and} \quad \mathrm{cov}\left[\mathbf{Z}_1,\mathbf{Z}_2\,|\mathbf{Z}_1\right] \;=\; 0. \end{array}$$

It thus follows that:

$$\begin{array}{rcl} \mathbb{E}\left[\mathbf{X}\,\middle|\nu^T\mathbf{X}\right] & = & S\cdot\mathbb{E}\left[\mathbf{Z}\,\middle|\mathbf{Z}_1 = \nu^T\mathbf{X}\right] = \nu\cdot\mathbb{E}\left[\mathbf{Z}_1\,|\mathbf{Z}_1\right] + U\cdot\mathbb{E}\left[\mathbf{Z}_2\,|\mathbf{Z}_1\right] \\ & = & \left(\nu^T\mathbf{X}\right)\cdot\left[\mathbf{I}_p + UU^T\frac{\Sigma}{\nu^T\Sigma\nu}\right]\cdot\nu + UU^T\left[\mathbf{I}_p - \frac{\Sigma}{\nu^T\Sigma\nu}\nu\nu^T\right]\cdot\mu, \text{and} \\ \mathrm{var}\left[\mathbf{X}\,\middle|\nu^T\mathbf{X}\right] & = & S\cdot\mathrm{var}\left[\mathbf{Z}\,\middle|\mathbf{Z}_1 = \nu^T\mathbf{X}\right]\cdot S^T \\ & = & \left(UU^T\right)\cdot\left(\Sigma - \frac{\Sigma\nu\nu^T\Sigma}{\nu^T\Sigma\nu}\right)\cdot\left(UU^T\right). \end{array}$$

By noticing that $UU^T = U(U^TU)^{-1}U^T$ is a projection matrix on the orthogonal complement of the space spanned by ν, UU^T can be rewritten as $UU^T = \mathbf{I}_p - \nu\left(\nu^T\nu\right)^{-1}\nu^T = \mathbf{I}_p - \nu\nu^T$. Using this expression for UU^T and some algebra,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}\left|\nu^T\mathbf{X}\right.\right] &= \mu + \left[\mathbf{I}_p + \left(\mathbf{I}_p - \nu\nu^T\right)\frac{\Sigma\nu}{\nu^T\Sigma\nu}\right]\nu\nu^T\left(\mathbf{X}-\mu\right) \\
&= \mu + \frac{\Sigma\nu}{\nu^T\Sigma\nu}\left(\nu^T\mathbf{X} - \nu^T\mu\right), \text{and} \\
\text{var}\left[\mathbf{X}\left|\nu^T\mathbf{X}\right.\right] &= \Sigma - \frac{\Sigma\nu\nu^T\Sigma}{\nu^T\Sigma\nu}.
\end{aligned} \tag{A-6}$$

From (A-6), the second moment of $\mathbf{X}$ given $\nu^T\mathbf{X}$ becomes

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\left|\nu^T\mathbf{X}\right.\right] &= \mathbb{E}\left[\mathbf{X}\left|\nu^T\mathbf{X}\right.\right]\mathbb{E}\left[\mathbf{X}\left|\nu^T\mathbf{X}\right.\right]^T + \text{var}\left[\mathbf{X}\left|\nu^T\mathbf{X}\right.\right] \\
&= \left(\mu\mu^T + \Sigma\right) + \frac{\Sigma\nu\nu^T\Sigma}{(\nu^T\Sigma\nu)^2}\cdot\left[\left(\nu^T\mathbf{X} - \nu^T\mu\right)^2 - 1\right] + \left[\frac{\Sigma\nu\mu^T}{\nu^T\Sigma\nu} + \frac{\mu\nu^T\Sigma}{\nu^T\Sigma\nu}\right]\cdot\left[\nu^T\mathbf{X} - \nu^T\mu\right].
\end{aligned} \tag{A-7}$$

Given the linear predictor variable as defined in (18), the conditional first and second moments of $\mathbf{X}$ are

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}\left|\mathbf{M}\right.\right] &= \mu + \left(\frac{\mathbf{M}-\alpha-\mu^T\beta}{\|\beta\|}\right)\cdot\frac{\Sigma}{\nu^T\Sigma\nu}\cdot\nu, \text{and} \\
\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\left|\mathbf{M}\right.\right] &= \mathbb{E}\left(\mathbf{X}\mathbf{X}^T\right) + \frac{\Sigma\nu\nu^T\Sigma}{(\nu^T\Sigma\nu)^2}\cdot\left[\left(\frac{\mathbf{M}-\alpha-\beta^T\mu}{\|\beta\|}\right)^2 - 1\right] + \left[\frac{\Sigma\nu\mu^T}{\nu^T\Sigma\nu} + \frac{\mu\nu^T\Sigma}{\nu^T\Sigma\nu}\right]\cdot\left[\frac{\mathbf{M}-\alpha-\beta^T\mu}{\|\beta\|}\right].
\end{aligned}$$

B.2 Hessians for SVM and Logistic regression risk functions

With the expression for the conditional second moments of a multivariate Gaussian variable given the value of one of its projections along the direction $\nu = \frac{\beta}{\|\beta\|}$, equations (23) and (22) give expressions for the Hessian of SVM and logistic regression risk functions.

B.2.1 Hessians for Gaussian predictors

To simplify the expressions, we partition the Hessian according to the intercept and the predictors $\mathbf{X}$ as

$$H_\theta(t) = \begin{bmatrix} \left[H_\theta(t)\right]_{\alpha,\alpha} & \left[H_\theta(t)\right]_{\alpha,\beta} \\ \left[H_\theta(t)\right]_{\beta,\alpha} & \left[H_\theta(t)\right]_{\beta,\beta} \end{bmatrix}, \text{ with}$$

$\left[H_\theta(t)\right]_{\alpha,\beta} = \left[H_\theta(t)\right]_{\beta,\alpha}^T$. Throughout this section $\tilde{f}$ denotes the density of the linear predictor variable $\mathbf{M}$.

Hessian for the Logistic regression classifier: Using the expressions derived in Section 4.1.1, we get

$$\begin{aligned}
\left[H_\theta(\theta)\right]_{\alpha,\alpha} &= \kappa_0 \\
\left[H_\theta(\theta)\right]_{\beta,\alpha} &= \kappa_0\cdot\mu + \kappa_1\cdot\frac{\Sigma}{\nu^T\Sigma\nu}\cdot\nu \\
\left[H_\theta(\theta)\right]_{\beta,\beta} &= \left(\mu\mu^T + \Sigma\right)\cdot\kappa_0 + \left[\frac{\Sigma\nu\mu^T + \mu\nu^T\Sigma}{\nu^T\Sigma\nu}\right]\cdot\kappa_1 + \left[\frac{\Sigma\nu\nu^T\Sigma}{\nu^T\Sigma\nu}\right]\cdot\kappa_2
\end{aligned} \tag{A-8}$$

where κ_0, κ_1 and κ_2 are scalars given by

$$\begin{aligned}
\kappa_0 &= \int \left[\frac{\exp(m)}{\left(1+\exp(m)\right)^2}\right] \cdot \tilde{f}(m) \cdot \mathrm{d}m \,, \\
\kappa_1 &= \int \left(\frac{m-\alpha-\mu^T\beta}{\|\beta\|}\right) \cdot \left[\frac{\exp(m)}{\left(1+\exp(m)\right)^2}\right] \cdot \tilde{f}(m) \cdot \mathrm{d}m \,, \text{ and} \\
\kappa_2 &= \int \left[\left(\frac{m-\alpha-\mu^T\beta}{\|\beta\|}\right)^2 - 1\right] \cdot \left[\frac{\exp(m)}{\left(1+\exp(m)\right)^2}\right] \cdot \tilde{f}(m) \cdot \mathrm{d}m \,.
\end{aligned} \tag{A-9}$$

Hessian for the SVM classifier: Using the expressions derived in Section 4.1.2, we get

$$\begin{aligned}
\left[H_\theta(\theta)\right]_{\alpha,\alpha} &= \kappa_0 \\
\left[H_\theta(\theta)\right]_{\beta,\alpha} &= \kappa_0 \cdot \mu + \kappa_1 \cdot \tfrac{\Sigma}{\nu^T\Sigma\nu} \cdot \nu \\
\left[H_\theta(\theta)\right]_{\beta,\beta} &= \left(\mu\mu^T + \Sigma\right) \cdot \kappa_0 + \left[\tfrac{\Sigma\nu\mu^T + \mu\nu^T\Sigma}{\nu^T\Sigma\nu}\right] \cdot \kappa_1 + \left[\tfrac{\Sigma\nu\nu^T\Sigma}{\nu^T\Sigma\nu}\right] \cdot \kappa_2
\end{aligned} \tag{A-10}$$

where κ_0, κ_1 and κ_2 are scalars given by

$$\begin{aligned}
\kappa_0 &= \tilde{f}(1) \cdot \mathbb{P}\left(\mathbf{Y} = 1 \,|\mathbf{M} = 1\right) + \tilde{f}(-1) \cdot \mathbb{P}\left(\mathbf{Y} = -1 \,|\mathbf{M} = -1\right), \\
\kappa_1 &= \left(\tfrac{1-\alpha-\mu^T\beta}{\|\beta\|}\right) \cdot \tilde{f}(1) \cdot \mathbb{P}\left(\mathbf{Y} = 1 \,|\mathbf{M} = 1\right) \\
&\quad + \left(\tfrac{-1-\alpha-\mu^T\beta}{\|\beta\|}\right) \cdot \tilde{f}(-1) \cdot \mathbb{P}\left(\mathbf{Y} = -1 \,|\mathbf{M} = -1\right), \text{ and} \\
\kappa_2 &= \left[\left(\tfrac{1-\alpha-\mu^T\beta}{\|\beta\|}\right)^2 - 1\right] \cdot \tilde{f}(1) \cdot \mathbb{P}\left(\mathbf{Y} = 1 \,|\mathbf{M} = 1\right) \\
&\quad + \left[\left(\tfrac{-1-\alpha-\mu^T\beta}{\|\beta\|}\right)^2 - 1\right] \cdot \tilde{f}(-1) \cdot \mathbb{P}\left(\mathbf{Y} = -1 \,|\mathbf{M} = -1\right).
\end{aligned} \tag{A-11}$$

B.2.2 Hessians for mixed Gaussian predictors

When $\mathbf{X}$ is distributed according to a mixture of K multivariate Gaussians, the conditional moments of $\mathbf{X}$ involved in the expression for the risk Hessian can be written as a weighted sum of the corresponding conditional moments for each of the individual Gaussian components as detailed next. Letting π_k denote the proportion of the mixture sampled from a multivariate Gaussian with mean μ_k and covariance matrix Σ_k, for $k = 1, \ldots, K$, the density function of $\mathbf{X}$ is:

$$f\left(\mathbf{x}\right) = \sum_{k=1}^{K} \pi_k \cdot \left(\frac{1}{2\pi|\Sigma_k|}\right)^{\frac{p}{2}} \exp\left[-\frac{1}{2} \cdot (\mathbf{x} - \mu_k)\Sigma_k^{-1}(\mathbf{x} - \mu_k)^T\right].$$

The conditional second moment $\mathbb{E}\left[\mathbf{X}\mathbf{X}^T \,|\mathbf{M}, \mu_k, \Sigma_k\right]$ given the margin variable $\mathbf{M}$ and that $\mathbf{X}$ was sampled from the component with mean μ_k and covariance Σ_k follows from (A-7) above. The first and second moment conditional solely on the margin variable $\mathbf{M}$ can then be computed as:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X} \,|\mathbf{M}\right] &= \textstyle\sum_{k=1}^{K} \mathbb{E}\left[\mathbf{X} \,|\mathbf{M}, \mu_k, \Sigma_k\right] \cdot \mathbb{P}(\mu_k, \Sigma_k \,|\mathbf{M}), \qquad \text{and} \\
\mathbb{E}\left[\mathbf{X}\mathbf{X}^T \,|\mathbf{M}\right] &= \textstyle\sum_{k=1}^{K} \mathbb{E}\left[\mathbf{X}\mathbf{X}^T \,|\mathbf{M}, \mu_k, \Sigma_k\right] \cdot \mathbb{P}(\mu_k, \Sigma_k \,|\mathbf{M}),
\end{aligned}$$

where $\mathbb{P}(\mu_k, \Sigma_k \,|\mathbf{M})$ denotes the probability of a point having been sampled from the Gaussian component with center μ_k and variance Σ_k given the margin variable $\mathbf{M}$. The distribution of $\mathbf{M} = \alpha + \mathbf{X}^T\beta$ is itself a mixture of Gaussians whose density $\tilde{f}$ is

$$\tilde{f}\,(m) \;=\; \sum_{k=1}^{K} \pi_k \cdot \left(\frac{1}{2\pi|\beta^T\Sigma_k\beta|}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2}\cdot\frac{(m-\alpha-\beta^T\mu_k)^2}{\beta^T\Sigma_k\beta}\right].$$

An expression for $\mathbb{P}(\mu_k, \Sigma_k \,|\mathbf{M})$ then follows from using Bayes's theorem:

$$\mathbb{P}\,(\mu_j, \Sigma_j \,|\mathbf{M}) \;=\; \frac{\pi_k \cdot \left(\beta^T\Sigma_k\beta\right)^{-\frac{1}{2}} \cdot \exp\left[-\frac{1}{2}\cdot\frac{\left(\mathbf{M}-\alpha-\beta^T\mu_k\right)^2}{\beta^T\Sigma_k\beta}\right]}{\sum\limits_{\tilde{k}=1}^{K} \pi_{\tilde{k}} \cdot \left(\beta^T\Sigma_{\tilde{k}}\beta\right)^{-\frac{1}{2}} \cdot \exp\left[-\frac{1}{2}\cdot\frac{\left(\mathbf{M}-\alpha-\beta^T\mu_{\tilde{k}}\right)^2}{\beta^T\Sigma_{\tilde{k}}\beta}\right]}.$$

Using (A-8) and (A-10), we have that the Hessian for SVM and logistic regression risks are given by:

$$\begin{aligned}
[H_\theta(\theta)]_{\alpha,\alpha} &= \sum_{k=1}^{K} \kappa_{k,0} \\
[H_\theta(\theta)]_{\beta,\alpha} &= \sum_{k=1}^{K} \left(\kappa_{k,0}\cdot\mu_k + \kappa_{k,1}\cdot\tfrac{\Sigma_k\nu}{\nu^T\Sigma_k\nu}\right), \qquad \text{and} \\
[H_\theta(\theta)]_{\beta,\beta} &= \sum_{k=1}^{K} \left[\kappa_{k,0}\cdot\left(\mu_k\mu_k^T+\Sigma_k\right) + \kappa_{k,1}\cdot\left(\tfrac{\Sigma_k\nu\mu_k^T+\mu_k\nu^T\Sigma_k}{\nu^T\Sigma_k\nu}\right) + \kappa_{k,2}\cdot\left(\tfrac{\Sigma_k\nu\nu^T\Sigma_k}{\nu^T\Sigma_k\nu}\right)\right],
\end{aligned} \tag{A-12}$$

where the scalars $\kappa_{k,0}$, $\kappa_{k,1}$ and $\kappa_{k,2}$ for $k = 1, \ldots, K$ are computed according to the risk function. For each Gaussian component, the $\kappa_{k,0}$, $\kappa_{k,1}$ and $\kappa_{k,2}$ correspond to the κ_0, κ_1 and κ_2 scalars in (A-9) and (A-11) multiplied by the conditional probability of that component given the margin variable as indicated next.

For the logistic risk and mixed Gaussian predictors, the κ scalars are:

$$\begin{aligned}
\kappa_{k,0} &= \int \mathbb{P}\,(\mu_k, \Sigma_k \,|\mathbf{M}=m\,)\cdot\left[\tfrac{\exp(m)}{\left(1+\exp(m)\right)^2}\right]\cdot\tilde{f}_k(m)\cdot\mathrm{d}m\;, \\
\kappa_{k,1} &= \int \mathbb{P}\,(\mu_k, \Sigma_k \,|\mathbf{M}=m\,)\cdot\left(\tfrac{m-\alpha-\mu_k^T\beta}{\|\beta\|}\right)\cdot\left[\tfrac{\exp(m)}{\left(1+\exp(m)\right)^2}\right]\cdot\tilde{f}_k(m)\cdot\mathrm{d}m\;, \text{ and} \\
\kappa_{k,2} &= \int \mathbb{P}\,(\mu_k, \Sigma_k \,|\mathbf{M}=m\,)\cdot\left[\left(\tfrac{m-\alpha-\mu_k^T\beta}{\|\beta\|}\right)^2-1\right]\cdot\left[\tfrac{\exp(m)}{\left(1+\exp(m)\right)^2}\right]\cdot\tilde{f}_k(m)\cdot\mathrm{d}m\;.
\end{aligned} \tag{A-13}$$

For the SVM risk and mixed Gaussian predictors, the κ scalars are:

$$\begin{aligned}
\kappa_{k,0} &= \mathbb{P}\left(\mu_k, \Sigma_k \,|\mathbf{M}=-1\right) \cdot \tilde{f}_k(1) \cdot \mathbb{P}\left(\mathbf{Y}=1 \,|\mathbf{M}=1\right) \\
&\quad + \mathbb{P}\left(\mu_k, \Sigma_k \,|\mathbf{M}=1\right) \cdot \tilde{f}_k(-1) \cdot \mathbb{P}\left(\mathbf{Y}=-1 \,|\mathbf{M}=-1\right), \\
\kappa_{k,1} &= \mathbb{P}\left(\mu_k, \Sigma_k \,|\mathbf{M}=1\right) \cdot \left(\tfrac{1-\alpha-\mu_k^T\beta}{\|\beta\|}\right) \cdot \tilde{f}_k(1) \cdot \mathbb{P}\left(\mathbf{Y}=1 \,|\mathbf{M}=1\right) \\
&\quad + \mathbb{P}\left(\mu_k, \Sigma_k \,|\mathbf{M}=-1\right) \cdot \left(\tfrac{-1-\alpha-\mu_k^T\beta}{\|\beta\|}\right) \cdot \tilde{f}_k(-1) \cdot \mathbb{P}\left(\mathbf{Y}=-1 \,|\mathbf{M}=-1\right), \text{ and} \qquad \text{(A-14)} \\
\kappa_{k,2} &= \mathbb{P}\left(\mu_k, \Sigma_k \,|\mathbf{M}=1\right) \cdot \left[\left(\tfrac{1-\alpha-\mu_k^T\beta}{\|\beta\|}\right)^2 - 1\right] \cdot \tilde{f}_k(1) \cdot \mathbb{P}\left(\mathbf{Y}=1 \,|\mathbf{M}=1\right) \\
&\quad + \mathbb{P}\left(\mu_k, \Sigma_k \,|\mathbf{M}=-1\right) \cdot \left[\left(\tfrac{-1-\alpha-\mu_k^T\beta}{\|\beta\|}\right)^2 - 1\right] \cdot \tilde{f}_k(-1) \cdot \mathbb{P}\left(\mathbf{Y}=-1 \,|\mathbf{M}=-1\right).
\end{aligned}$$

References

AKAIKE, H. 1973. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, B. N. Petrov and F. Csáki, Eds. Akadémia Kiadó, Budapest, 267–281.

AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 6, 716.

AUDIBERT, J.-Y. AND TSYBAKOV, A. B. 2007. Fast learning rates for plug-in classifiers. *Annals of Statistics 35*, 2, 608–633.

BANERJEE, O., D'ASPREMONT, A., AND ELGHAOUI, L. 2005. Sparse covariance selection via robust maximum likelihood estimation. Tech. rep., arXiv http://arxiv.org/abs/cs.CE/0506023.

BARTLETT, P., JORDAN, M., AND MCAULIFFE, J. 2006. Convexity, classification and risk bounds. *Journal of the American Statistical Association 101*, 138–156.

BICKEL, P. AND DOCKSUM, K. 2001. *Mathematical Statistics: Basic Ideas and Selected Topics*, 2nd ed. Vol. 1. Prentice Hall.

CASELLA, G. AND BERGER, J. 2001. *Statistical Inference*. Duxbury Press.

CHEN, S. S., DONOHO, D. L., AND SAUNDERS, M. A. 2001. Atomic decomposition by basis pursuit. *SIAM Review 43*, 1, 129–159.

CORTES, C. AND VAPNIK, V. 1995. Support-vector networks. *Machine Learning 30*, 273–297.

DUDLEY, R. M. 1999. *Uniform central limit theorems*. Cambridge University Press.

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.

FRANK, I. E. AND FRIEDMAN, J. 1993. A statistical view of some chemometrics regression tools. *Technometrics 35*, 109–148.

FRIEDMAN, J. H. 2008. Fast sparse regression and classification. Tech. rep., Stanford University Department of Statistics. July.

GENKIN, A., LEWIS, D. D., AND MADIGAN, D. 2007. Large-scale bayesian logistic regression for text categorization. *Technometrics 49*, 3 (August), 291–304.

GUYON, I., WESTON, J., BARNHILL, S., AND VAPNIK, V. 2002. Gene selection for cancer classification using

support vector machines. *Machine Learning 46*, 389–422.

HANSEN, M. AND YU, B. 2001. Model selection and the principle of minimum description length. *Journal of the American Statistical Association 96*, 454, 746–774.

JOACHIMS, T. 1998. *Text categorization with Support Vector Machines: Learning with many relevant features*. Lecture Notes in Computer Science, vol. 1398. Springer Berlin / Heidelberg, Berlin/Heidelberg, 137–142.

KNIGHT, K. AND FU, W. J. 2000. Asymptotics for lasso type estimators. *The Annals of Statistics 28*, 5, 1356–1378.

KOO, J.-Y., LEE, Y., KIM, Y., AND PARK, C. 2008. A bahadur representation of the linear support vector machine. *Journal of Machine Learning Research 9*, 1343–1368.

LI, R. AND LIANG, H. 2008. Variable selection in semiparametric regression modeling. *Annals of Statistics 36*, 261–286.

LI, Y. AND ZHU, J. 2008. The l1-norm quantile regression. *Journal of Computational and Graphical Statistics 17*, 163–185.

MCCULLAGH, P. AND NELDER, J. A. 1989. *Generalized Linear Models*. Chapman & Hall, London ; New York.

MEIER, L., VAN DER GEER, S., AND BÜHLMANN, P. 2006. The group lasso for logistic regression. Tech. rep., ETHZ.

MEINSHAUSEN, N. AND BÜHLMANN, P. 2004. Consistent neighborhood selection for sparse high-dimensional graphs with the lasso. Tech. rep., ETHZ.

MEINSHAUSEN, N. AND YU, B. 2006. Lasso-type recovery of sparse representations for high-dimensional data. Tech. rep., Department of Statistics, UC Berkeley.

NELDER, J. A. AND WEDDERBURN, R. W. M. 1972. Generalized linear models. *Journal of the Royal Statistical Society, Series A 135*, 3, 370–384.

PARK, M.-Y. AND HASTIE, T. 2006. l_1 regularization path algorithms for generalized linear models. Tech. rep., Stanford University Department of Statistics.

PHILLIPS, P. C. B. 1991. A shortcut to lad estimator asymptotics. *Econometric Theory 7*, 4, 450–463.

POLLARD, D. 1991. Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*, 2, 186–199.

RAVIKUMAR, P., RASKUTTI, G., WAINWRIGHT, M., AND YU, B. 2008. High-dimensional covariance estimation minimizing ℓ_1-penalized log-determinant divergence. Tech. Rep. 767, UC Berkeley Department of Statistics.

RISSANEN, J. 1978. Modeling by shortest data description. *Automatica 14*, 465–471.

SCHWARZ, G. 1978. Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

STEINWART, I. AND SCOVEL, C. 2007. Fast rates for support vector machines using gaussian kernels. *Annals of Statistics 35*, 2, 575–607.

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B 58*, 1, 267–288.

TIBSHIRANI, R. 1997. The lasso method for variable selection in the cox model. *Statistics in Medicine 16*, 4, 385–395.

WAINWRIGHT, M. 2006. Sharp thresholds for high-dimensional and noisy recovery of sparsity. Tech. rep., Department of Statistics, UC Berkeley.

ZHANG, T. 2004. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics 32*, 1, 56–134.

ZHANG, T. 2009. Some sharp performance bounds for least squares regression with l1 regularization. *Annals of Statistics Forthcoming*.

ZHAO, P. AND YU, B. 2006. On model selection consistency of LASSO. *Journal of Machine Learning Research 7*, 2541–2563.

ZHU, J., ROSSET, S., HASTIE, T., AND TIBSHIRANI, R. 2004. 1-norm svms. In Advances in Neural Information Processing Systems. *NIPS*.

ZOU, H. 2006. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.