Robust Logistic Regression with Bounded Data Uncertainties

Patrick L. Harrington Jr., Aimee Zaas, Christopher W. Woods, Geoffrey S. Ginsburg, Lawrence Carin *Fellow, IEEE*, and Alfred O. Hero III. *Fellow, IEEE*

***Abstract*—Building on previous work in robust optimization, we present a formulation of robust logistic regression under bounded data uncertainties. The robust estimates are obtained using block coordinate gradient descent with iterative group thresholding, which zeros out highly uncertain variables. For high dimensional problems with uncertain measurements, we discuss the addition of regularization penalties such that both robustness and block sparsity are imposed in the parameter estimates. An empirical approach to estimate the uncertainty magnitude is presented through the use of quantiles. We compare the results of ℓ_1-Logistic Regression against ℓ_1-Robust Logistic Regression on a real gene expression data set and achieve reductions in the worst-case false alarm rate and probability of error by $10\% - 20\%$, thus illustrating the value added of using robust classifiers in risk sensitive domains when confronted with uncertain measurements.**

***Index Terms*—Robust Optimization, Group Structured Regularization, Logistic Regression, Gene Expression Microarrays.**

I. Introduction

THERE are two common methods for accommodating uncertainty in the observed data in risk minimization problems. The first approach assumes stochastic measurement corruption, centered about the true signal. This method is commonly known as error in variables (EIV) and has a rich history in least-squares and logistic regression (LR) problems [1], [2], [3], [4]. Unfortunately, EIV estimators are optimistic, require solving non-convex optimization problems, and de-regularize Hessian-like matrices making numerical estimation less stable. The latter approach of accommodating measurement uncertainty involves developing estimators that are robust to worst-case perturbations in the data and result in solving well posed convex programs [5], [6], [7], [8], [9], [10], [11].

The work presented throughout this paper builds on the results of [7] and [5]. Specifically, we generalize the robust optimization problem to a variety of different uncertainty sets appropriate for real problems. We present novel group-thresholding conditions which produce block-sparse parameters when confronted with grouped uncertainty. The robust risk functions, resulting from the minimax estimation, are regularized with group structured penalties to accommodate high dimensional data when the underlying signal is both block-sparse and measurements are uncertain. A block coordinate gradient descent with an active shooting speed up algorithm is presented which exploits the iterative grouped thresholding conditions.

An interesting relationship between ridge LR and robust LR (RLR) is presented. The robustness of ridge LR is established by identifying conditions when the uncertainty magnitude of robust can be re-parameterized in terms of the ridge tuning parameter such that both methods yield the same solution. Conditions on the Hessians of each method are established such that the RLR approach converges to this solution faster than ridge LR. We also present an empirical approach to estimating the uncertainty bounds using quantiles. We conclude by presenting a illustrative example using gene expression data and discuss how robust ℓ_1-regularization paths recover "robust genes" that were previously over-looked by standard ℓ_1-regularization. The worst-case probability of errors and false alarm rates of RLR are always less than or equal to those from LR. To the authors knowledge, this is the first application of a robust classifier being applied to gene expression data. The results suggest that "robustification" of logistic classifiers can lead to significant performance gains in gene expression analysis.

The specific contributions of this paper are as follows. 1) We extend the penalized robust logistic regression formulation of [7] to accommodate group structured variable selection (l_1) penalties. 2) We give necessary and sufficient conditions for the solution to this modified robust logistic regression problem and propose an iterative Newton-type optimization algorithm with a group thresholding operator. 3) We obtain a relation between this solution and the solution to ridge logistic regression that uses a l_2 squared penalty. 4) We obtain expressions for the Hessian of the objective functions that are minimized by robust and ridge logistic regression, respectively, which can be used to obtain and compare asymptotic convergence rates of iterative robust and ridge logistic regression. 5) We show that the bounded error approaches advocated here (and in [5], [7]) are natural for bioinformatics applications, or indeed any applications where there are technical replicates that can be used to prescribe estimate error bounds. 6) We illustrate

P. Harrington is with the Bioinformatics Graduate Program and the Department of Statistics, University of Michigan, Ann Arbor, MI, 48109 USA e-mail: (see http://www-personal.umich.edu/∼plhjr/).

Aimee Zaas, Christopher Woods, and Geoffrey S. Ginsburg are with Department of Medicine and the Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, USA.

Lawrence Carin is with the Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina, USA.

A. Hero is with the Department of Electrical Engineering and Computer Science, the Department of Biomedical Engineering, and the Department of Statistics, University of Michigan, Ann Arbor, MI, 48109 USA e-mail: (see http://www.eecs.umich.edu/∼hero/).

concrete benefits of our approach for prediction of symptomatic infection from gene microarray data collected during a human viral challenge study. In particular, our predictor suffers significantly less degradation in probability of error as compared to the standard non-robustified logistic predictor.

The outline of the paper is as follows. In Sec. II we review the robust logistic regression problem and in Sec. III we introduce a group penalized version of this problem. In Sec. IV we describe our active shooting block coordinate descent approach to solving the associated group penalized optimization. In Sec. V we propose a quantile-based method to extract relevant uncertainty bounds when empirical technical replicates of the training data are available. In Sec. VII we evaluate the performance of the proposed method using simulations and a real gene microarray data set collected by our group. Finally, in Sec. VIII we state our principal conclusions.

II. Robust Logistic Regression

The goal of robust logistic regression (RLR) is to extract an estimator by minimizing the worst-case errors in measurements on the LR loss function (binomial deviance) subject to bounded uncertainty. The general case of RLR with spherical uncertainty, previously explored by [7], involves n measured training variables $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1,+1\}$, and adopts a minimax formulation

$$\min_{\beta,\beta_0} \max_{\{\delta_i\}_{i=1}^n} \sum_{i=1}^n \log\left(1+e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right)$$
$$\text{subject to} \quad \|\delta_i\|_{\ell_2} \le \rho \;\forall\; i=1,\ldots,n. \tag{1}$$

Here, the true signals are given by $\{z_i = x_i + \delta_i\}_{i=1}^n$, δ_i is not observed, and the parameter ρ is the magnitude of the worst-case perturbation. The minimax problem in (1) is solved by first solving the inner maximization step analytically, resulting in a convex RLR loss function which is then minimized with respect to β, β_0.

Note that the maximization over each of the n perturbations, δ_i can be moved within the sum loss function in (1)

$$\min_{\beta,\beta_0} \sum_{i=1}^n \max_{\delta_i} \log\left(1+e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right)$$
$$\text{subject to} \quad \|\delta_i\|_{\ell_2} \le \rho \;\forall\; i=1,\ldots,n. \tag{2}$$

We would like to reduce the minimax problem in (2) to a closed form minimization problem over β as performed in [5]. We begin by noting the following

$$-y_i\beta^T\delta_i \le \|\beta\|_{\ell_2}\|\delta_i\|_{\ell_2} \le \|\beta\|_{\ell_2}\rho. \tag{3}$$

Given that the loss function is monotonic in $-y_i\beta^T\delta_i$, we have the following upper-bound on the binomial deviance for the i^{th} sample:

$$\begin{aligned}\Phi &= \log\left(1+e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right) \\ &\le \log\left(1+e^{-y_i\left(\beta^T x_i+\beta_0\right)+\rho\|\beta\|_{\ell_2}}\right).\end{aligned} \tag{4}$$

The upper-bound in both (3) and (4) is achievable for δ_i collinear with β, i.e., $\delta_i = \gamma_i\beta$ for some alignment parameter γ_i, thus yielding the solution to the constrained maximization step for the i^{th} observation

$$\begin{aligned}&\max_{\delta_i, \|\delta_i\|_{\ell_2}\le\rho} \log\left(1+e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right) \\ &\qquad = \log\left(1+e^{-y_i\left(\beta^T x_i+\beta_0\right)+\rho\|\beta\|_{\ell_2}}\right).\end{aligned} \tag{5}$$

We may now proceed with obtaining the robust estimated normal vector β corresponding to the binomial deviance via the following unconstrained minimization problem:

$$\min_{\beta,\beta_0} \sum_{i=1}^n \log\left(1+e^{-y_i\left(\beta^T x_i+\beta_0\right)+\rho\|\beta\|_{\ell_2}}\right). \tag{6}$$

Geometrically, the robust binomial deviance penalizes points based upon their distance to the "uncertainty margins" $\pm\rho\|\beta\|_2$, i.e., a point is penalized more for being the same distance away from the hyperplane under the robust formulation than standard LR. This pessimism is intuitive as observations that are close to the decision boundary could have their true value lying on the misclassified side under worst-case perturbations (see Figure 1).

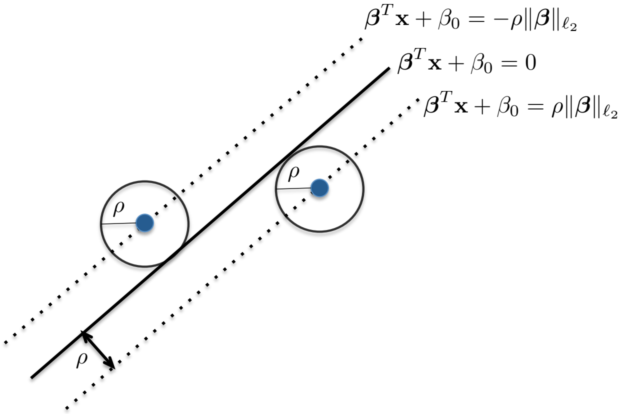

$\boldsymbol{\beta}^T\mathbf{x}+\beta_0 = -\rho\|\boldsymbol{\beta}\|_{\ell_2}$
$\boldsymbol{\beta}^T\mathbf{x}+\beta_0 = 0$
$\boldsymbol{\beta}^T\mathbf{x}+\beta_0 = \rho\|\boldsymbol{\beta}\|_{\ell_2}$
ρ
ρ
ρ

Fig. 1. Bounded uncertainty modification penalizes based on *potentially* mis-classified points translating logistic regression loss to penalize based on margins

III. Robust Logistic Regression with Group Structured Uncertainty Sets

In practice, joint spherical uncertainty may be inappropriate for modeling real data. A more appropriate form of uncertainty occurs when it affects groups of variables. Here, we assume that perturbations have group structure and are applied to G disjoint subsets of the p variables. These assumptions produce the following robust optimization problem

$$\min_{\beta,\beta_0} \sum_{i=1}^n \max_{\delta_i} \log\left(1+e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right)$$
$$\text{subject to} \quad \{\|\delta_{i,g}\|_{\ell_2} \le \rho_g\}_{g=1}^G \;\forall\; i=1,\ldots,n \tag{7}$$

where $\delta_i = \{\delta_{i,g}\}_{g=1}^G$ with $\delta_{i,g} \in \mathbb{R}^{|\mathcal{I}_g|}$ and $\mathcal{I}_g \subset \mathcal{I}$, $\mathcal{I} = \{1,\ldots,p\}$, are the set indices corresponding to the variables in group g. We will assume that $\mathcal{I}_g \cap \mathcal{I}_{g'} = \emptyset$ for $g \neq g'$. Note that that loss function is monotonic in $\sum_{g=1}^G -y_i\beta_g^T\delta_{i,g}$, and therefore, under worst-case perturbations, we have

$$\sum_{g=1}^G -y_i\beta_g^T\delta_{i,g} \leq \sum_{g=1}^G \|\beta_g\|_{\ell_2}\|\delta_{i,g}\|_{\ell_2} \leq \sum_{g=1}^G \rho_g\|\beta_g\|_{\ell_2}. \quad (8)$$

The inner maximization step in (7) is achieved when the upper bounds in (8) is tight, which is when β_g is colinear with $\delta_{i,g}$. Therefore, as in (5), we have analytically computed the inner-maximization step, and thus, our problem reduces to solving the following:

$$\min_{\beta,\beta_0} \sum_{i=1}^n \log\left(1 + e^{-y_i\left(\beta^T x_i + \beta_0\right) + \sum_{g=1}^G \rho_g\|\beta_g\|_{\ell_2}}\right). \quad (9)$$

The term within the argument of the RLR loss function is the "group lasso" penalty (9) which tends to promotes spareness in the groups (or factors) when used to regularize convex risk functions [12], [13].

When the number of groups is $G = p$ (each variable in its own group), the perturbations are interval based, and the group lasso penalty is equivalent to the ℓ_1-penalty. In this case the optimization problem (9) becomes, when $\rho_g = \rho$, for all $g = 1, \ldots, p$,

$$\min_{\beta,\beta_0} \sum_{i=1}^n \log\left(1 + e^{-y_i\left(\beta^T x_i + \beta_0\right) + \rho\|\beta\|_{\ell_1}}\right). \quad (10)$$

The minimization problem in (10) was previously treated in the context of interval perturbations in [7].

A. Regularized Robust Logistic Regression

Penalties such as the ℓ_1-norm are used in high-dimensional data settings, as they tend to zero out many of elements in β, which may better represent the structure of the underlying signal. Many fields of research increasingly involve high-dimensional data measurements that are obtained under noisy measurement conditions, such as gene expression microarrays that measure the activity of thousands of genes by assaying the abundances of mRNA in the sample. Here we develop new logistic classifiers that have the combined advantages of sparsity in variables and robustness to measurement uncertainty.

We will assume that the group structure of the regularization penalties coincide with the structure of uncertainty sets. For an arbitrary set of G disjoint groups, the following regularized robust solution is

$$\min_{\beta_g,\beta_0} \sum_{i=1}^n \log\left(1 + e^{-y_i f_i + \sum_{g=1}^G \rho_g\|\beta_g\|_{\ell_2}}\right) + \sum_{g=1}^G \lambda_g\|\beta_g\|_{\ell_2} \quad (11)$$

with $f_i = \beta^T x_i + \beta_0$. The presence of the additional group-lasso penalty should promote block-sparsity in β while being robust to measurement error affecting the same variables in the group.

IV. Computations for Regularized Robust Logistic Regression

Here, we present a numerical solution to the general regularized RLR problem based on block co-ordinate gradient descent with an active shooting step to speed up the convergence time when confronted with sparse signals or many groups. Since the penalized loss function in (11) is convex, denoted by $L_{\rho,\lambda}$, we can iteratively obtain β_g via block coordinate gradient descent. However, the gradient of (11) does not exist at $\beta_g = 0$, and thus we must resort to sub-gradient methods to identify optimality conditions [14], [13], [15]. The necessary and sufficient conditions for β_g to be a valid solution of (11) require

$$-X_g^T A_\rho y + \left(\rho_g \text{ tr}\left(A_\rho\right) + \lambda_g\right)\frac{\beta_g}{\|\beta_g\|_{\ell_2}} = 0\,, \beta_g \neq 0$$
$$\|X_g^T A_{-\beta_g,\rho} y\|_{\ell_2} \leq \left(\rho_g \text{tr}(A_{-\beta_g,\rho}) + \lambda_g\right), \beta_g = 0$$

with $A_\rho = (I + K_\rho)^{-1}$, $K_\rho = \text{diag}\left(e^{y_i\left(\beta^T x_i + \beta_0\right) - \sum_{g=1}^G \rho_g\|\beta_g\|_{\ell_2}}\right)$. Note that $A_{-\beta_g,\rho}$ means that β_g is set to 0 in K_ρ. The group thresholding that arises from these optimality conditions appears in group lasso regularized problems [12], [13], [15] and to the authors knowledge, has not been previously extracted in the context of RLR [7]. While RLR uses a different loss function than the binomial deviance, the thresholding conditions (above) establish the relationship between regularizing a standard convex risk function with a non-differentiable penalty and the sparse solutions that tend to appear when applying robust linear classifiers to uncertain data [7]. It is intuitive that the thresholding conditions depend on both the uncertainty magnitude ρ_g and the sparseness penalty parameter λ_g.

The proposed block co-ordinate gradient descent consists of updating the g^{th} group parameters by initially computing a Newton-step

$$\delta\beta_g^{(m)} = -\left[\nabla^2_{\beta_g} L_{\rho,\lambda}^{(m)}\right]^{-1} \nabla_{\beta_g} L_{\rho,\lambda}^{(m)} \quad (12)$$

followed by performing a backtracking line search ([16]) for appropriate step size $\nu^{(m)} > 0$, and then updating $\beta_g^{(m+1)}$

$$\beta_g^{(m+1)} \leftarrow \beta_g^{(m)} + \nu_g^{(m)}\delta\beta_g^{(m)}. \quad (13)$$

The numerical solution to solving (11) is outlined below in Algorithm 1. The active shooting [17] step updates the parameters that were non-zero after the initial step until convergence. After this subset has converged, gradient descent is performed over all the variables. This preferential update tends to reach the global minima faster when confronted with many groups.

Algorithm 1:
Active Shooting Block Coordinate Descent with Group Thresholding

```
1) Initialize:
    a) $\beta_0^{(1)} \leftarrow \nu_0^{(0)}\delta\beta_0^{(0)}$ with all parameters set to zero
    b) $\beta_g^{(1)} \leftarrow \nu_g^{(0)}\delta\beta_g^{(0)}$, for $g = 1, \ldots, G$, with $\beta_0$
       evaluated at $\beta_0^{(0)}$ and all other parameters set to
       zero
```

2) Define the active set $\Lambda = \{g : \beta_g^{(0)} \neq 0\}$
3) $\beta_0^{(m+1)} \leftarrow \beta_0^{(m)} + \nu_0^{(m)} \delta\beta_0^{(m)}$ with $\delta\beta_0^{(m)}$ via (12), $\nu_0^{(m)}$ by performing backtracking, and β held at previous value
4) For $g \in \Lambda$
    a) if $\|X_g^T A_{-\beta_g,\rho} y\|_{\ell_2} \leq \rho_g \text{tr}\left(A_{-\beta_g,\rho}\right) + \lambda_g$, $\beta_g^{(m+1)} \leftarrow 0$
    b) else, evaluate $\delta\beta_g^{(m)}$ from (12) while holding all other parameters at previous values, compute step size $\nu_g^{(m)}$ via backtracking, and update $\beta_g^{(m+1)} \leftarrow \beta_g^{(m)} + \nu_g^{(m)} \delta\beta_g^{(m)}$
5) Repeat steps 3 and 4 until some convergence criteria met for active parameters in Λ.
6) If convergence criteria satisfied, define $\Lambda = \{1, \ldots, G\}$ and repeat 3 and 4 until convergence in all parameters.

V. Empirical Estimation of Uncertainty

The magnitude of the potential uncertainty is determined by ρ_g. There are situations when a researcher has prior knowledge on the value of ρ_g but more often this parameter must be estimated empirically. In modern biomedical experiments, in which gene expression microarrays are used to assay the activity of tens of thousands of genes, technical replicates are frequently obtained to assess the effect of of measurement uncertainty.

We will estimate ρ_g using a generalization of the method in [18] based on quantiles. For grouped uncertainty sets, we estimate ρ_g via the following

$$\hat{\rho}_g(\alpha) = \text{inf}_\tau \mathbb{P}(\sqrt{x_g^T x_g} \leq \tau) = \alpha \tag{14}$$

where the distribution in (14) is taken with respect to a data set independent of the training data, such as technical replicates of a biological experiment. Note that (14) reduces to interval based quantile estimates when the number of groups $G = p$. As the cumulative distribution function (CDF) in (14) does not depend on class label y, we obtain (14) by

$$\mathbb{P}(z_g \leq \tau) = \sum_{y \in \{-1,+1\}} \mathbb{P}(z_g \leq \tau | Y = y)\mathbb{P}(Y = y) \tag{15}$$

with $z_g = \sqrt{x_g^T x_g}$ and data centered about their respective class centroids. The class priors are estimated empirically by $\hat{\mathbb{P}}(Y = y) = m_y/m$, where m_y and m are the number of replicate samples with label y and total number of replicate samples, respectively. The estimation in (14) can be assessed with respect to the empirical CDF of the data $\{x_{i,g}\}_{i=1}^n$ or approximated by the inverse-CDF of the χ_{p_g} distribution [18] with $p_g = |\mathcal{I}_g|$ degrees of freedom. The application of the proposed estimation of uncertainty bounds is presented below in the context of high-dimensional gene expression data in which technical replicates are available.

VI. Robust vs. Ridge Regression

A. Robustness of Ridge Logistic Regression

One important question is how the proposed RLR formulation relates to Ridge LR. Ridge LR involves adding a squared ℓ_2-penalty to the binomial deviance loss function:

$$\text{min}_\beta \sum_{i=1}^n \log\left(1 + e^{-y_i \beta^T x_i}\right) + \frac{\lambda}{2}\|\beta\|_{\ell_2}^2. \tag{16}$$

Our goal is to identify values of ρ as a function of λ for which both robust (1) and ridge (16) produce approximately identical estimates of β. We begin by inspecting the optimality conditions for β. The gradients for ridge and robust, respectively, are given as (intercept removed for clarity):

$$\begin{aligned} \nabla_\beta L_\lambda &= -X^T W y + \lambda\beta & (17) \\ \nabla_\beta L_\rho &= -X^T W_\rho y + \rho \text{tr}(W_\rho)\frac{\beta}{\|\beta\|_{\ell_2}}. & (18) \end{aligned}$$

where $W = \text{diag}(\frac{e^{-y_i \beta^T x_i}}{1+e^{-y_i \beta^T x_i}})$ and $W_\rho = \text{diag}(\frac{e^{-y_i \beta^T x_i + \rho\|\beta\|_{\ell_2}}}{1+e^{-y_i \beta^T x_i + \rho\|\beta\|_{\ell_2}}})$. If we linearize the sigmoidal terms in W and W_ρ, the scaled gradients can be approximated by the following:

$$\begin{aligned} n^{-1}\nabla_\beta L_\lambda &\approx \left(\frac{1}{4n}X^T X + (\lambda/n)I\right)\beta - \frac{1}{2}\delta\bar{x}_n \\ &= H_{n,\lambda}\beta - \frac{1}{2}\delta\bar{x}_n & (19) \end{aligned}$$

with $\delta\bar{x}_n = n^{-1}\left(n_+\bar{x}_+ - n_-\bar{x}_-\right)$

$$\begin{aligned} n^{-1}\nabla_\beta L_\rho &\approx \frac{1}{4n}X^T X\beta - \frac{1}{2}\delta\bar{x}_n + \frac{1}{4}\rho^2\beta \\ &+ \frac{\rho}{2\|\beta\|_{\ell_2}}\left((1 - \frac{\beta^T\delta\bar{x}_n}{2})\beta - \frac{1}{2}\|\beta\|_{\ell_2}^2\delta\bar{x}_n\right) \\ &= c_\beta + a_\beta\rho^2 + b_\beta\rho. & (20) \end{aligned}$$

Since the gradient for ridge regression can be approximated by the linear system of equations in (19), we can approximate β_λ via

$$\hat{\beta}_\lambda \approx \frac{1}{2}H_{n,\lambda}^{-1}\delta\bar{x}_n. \tag{21}$$

Inserting (21) into (20), summing (20) over its elements by an inner product with a vector of 1's, and solving for ρ, produces the following quadratic equation of ρ in terms of λ

$$\rho_\lambda \approx \frac{-1^T b_{\hat{\beta}_\lambda} \pm \sqrt{\left(1^T b_{\hat{\beta}_\lambda}\right)^2 - 4 \cdot 1^T a_{\hat{\beta}_\lambda} \cdot 1^T c_{\hat{\beta}_\lambda}}}{2 \cdot 1^T a_{\hat{\beta}_\lambda}}. \tag{22}$$

The positive value of ρ_λ is the uncertainty magnitude. This establishes equivalence between the solutions of ridge and robust logistic regression for a given λ.

B. Convergence Rates of Ridge and Robust

We established conditions of equivalence between ridge and robust LR estimators through a re-parameterization of ρ in terms of λ (22). It is also of interest to investigate the rate at which these methods converge under the proposed conditions of equivalence. For ease of exposition in the analysis, we will assume that the data has been centered and there are balanced sample sizes ($n_+ = n_-$).

We begin by assuming that the gradients of both ridge LR (19) and RLR (20) are in some neighborhood about 0. The structure of the Hessian specify rates of convergence through the eigenvalues. Under the assumptions detailed above, the scaled Hessian for the ridge case is obtained from differentiating (19) is given by

$$n^{-1}\nabla^2_\beta L_\lambda \approx \frac{1}{4n}X^TX + (\lambda/n)I. \tag{23}$$

By differentiating (20) and noting that $\frac{1}{8}\beta^T\delta\bar{x}_n \ll \frac{1}{2}$, we obtain the scaled Hessian for robust logistic regression

$$n^{-1}\nabla^2_\beta L_\rho \approx \left(\frac{1}{4n}X^TX + \frac{1}{4}\rho^2 I\right) + \rho\,(A - B) \tag{24}$$

with positive semi-definite matrix A

$$A = \frac{1}{2\|\beta\|_{\ell_2}}\left(I - \frac{1}{\|\beta\|^2_{\ell_2}}\beta\beta^T\right) \tag{25}$$

and positive semi-definite matrix B

$$B = \frac{1}{8\|\beta\|_{\ell_2}}\delta\bar{x}_n\beta^T. \tag{26}$$

We are interested in extracting conditions on ρ such that the Hessian corresponding to the robust binomial deviance is “larger” than that corresponding to ridge, i.e., $\Delta H = \nabla^2_\beta L_\rho - \nabla^2_\beta L_\lambda \succeq 0$, when both methods yield the same β, or equivalently

$$w^T\Delta H w \geq 0, \forall w. \tag{27}$$

One can directly derive necessary conditions that must be satisfied by ΔH by enforcing that (27) holds for some choice of w. We take $w = \delta\bar{x}$, which is the right eigenvector of B, and substitute $\hat{\beta}_\lambda$ (21) and ρ_λ (22) for β and ρ, respectively in (27). The conditions such that robust logistic regression converges faster to the solution of ridge logistic regression are

$$\rho^2_\lambda + \frac{2(1-\epsilon'_\lambda)}{\|\hat{\beta}_\lambda\|_{\ell_2}}\rho_\lambda \geq 4\lambda/n \tag{28}$$

where ϵ'_λ is given by

$$\epsilon'_\lambda = \left(\hat{\beta}^T_\lambda\delta\bar{x}\right)^2\left(\frac{1}{\|\delta\bar{x}\|^2_{\ell_2}\|\hat{\beta}_\lambda\|^2_{\ell_2}} + \frac{1}{4}\right). \tag{29}$$

VII. Numerical Results

A. Recovery of Regularization Path Under Signal Corruption

Many researchers in machine learning and bioinformatics assess importance of various biomarkers based upon the ordering of ℓ_1-regularization paths. Thus, it is worthwhile to explore how the ordering of the variables within the regularization path change as uncertainty is added. We first show a simple synthetic example there are $n = 100$ samples in each class with $x \in \mathbb{R}^p$, $p = 22$, with $x \sim \mathcal{N}(\mu_y, \Sigma)$. The class centroids are given by $\mu_{+1} = [-\frac{1}{2}, \frac{1}{3}, -\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, 0, \cdots, 0]^T$ and $\mu_{-1} = [\frac{1}{2}, -\frac{1}{3}, 0, \cdots, 0, 0, 0, 0, 0]^T$. The structure of Σ is block diagonal, where diagonal components $\{\sigma_{ii}\}^p_{i=1}$ are equal to one and the block structure affects only the variables $x_1, x_2, x_3, x_4, x_5, x_6$, where the off-diagonal elements in this block $\{\sigma_{ij}\}^6_{j=1,j\neq i}$ are equal to 0.1. The other off-diagonal elements have covariance of zero.

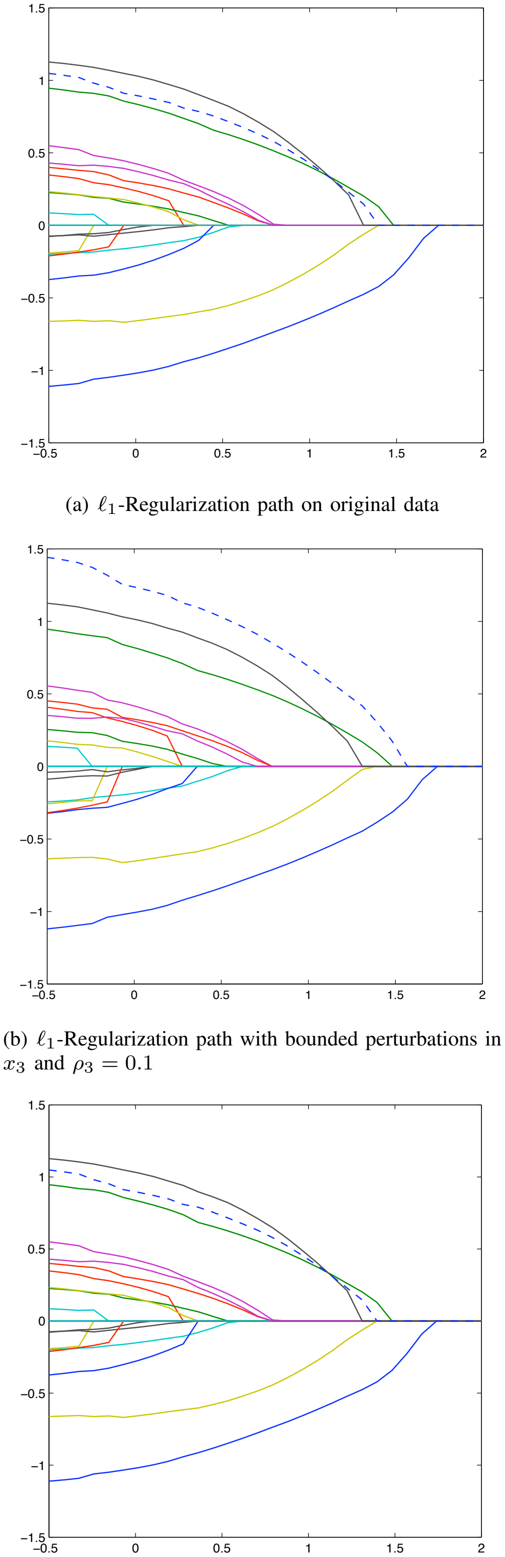
Figure 2(a) represents the ℓ_1-regularization path as a function of $\log_{10}\lambda$ on the original data. In Figure 2(b), x_3 has been perturbed such that $x_3 \leftarrow x_3 + \delta_3$ with $-\rho \leq \delta_3 \leq \rho$, which leads to a shift in the ordering within the regularization path under normal ℓ_1-LR. Using the perturbed data set and knowledge that $\rho = 0.1$ for x_3 and presented with interval uncertainty, the ℓ_1-regularized robust logistic regression recovers the original ordering of the variables (see Figure 2(c)).

B. Human Rhino Virus Gene Expression Data

Here we present numerical results on peripheral blood gene expression data set from a group of $n = 20$ patients inoculated with the Human Rhino Virus (HRV), the typical agent of the common cold [19]. Half of the patients responded with symptoms ($y = +1$) and the other half did not ($y = -1$). The original $12,023$ genes on the Affymetrix oligonucleotide microarray were reduced to $p = 129$ differentially expressed genes controlling for a 20% False Discovery Rate (FDR) using the same methodology as applied in ([20]). We will regularize both the robust binomial deviance and standard binomial deviance with an ℓ_1-penalty to control the sparsity of the resulting model forcing many elements in β to be zero. In this experiment there were approximately 20 microarray chips that were technical replicates. The technical replicates were used to estimate the interval uncertainty bounds using (14).

Since oligonucleotide microarray devices detect the presence of mRNA abundance by including thousands of different probe sets (short sequences) that bind to a particular mRNA molecule, produced from a specific gene, we will adopt interval uncertainty across the p genes. The estimated interval uncertainty bounds, $\{\hat{\rho}_i(\alpha)\}^p_{i=1}$, were obtained from (14) using the empirical CDF of the independent technical replicate data with linear interpolation between sampled data. The CDF in (14) was computed by (15) with each conditional CDF centered about their class dependent sample mean and equal class priors $\hat{\mathbb{P}}(Y = +1) = \hat{\mathbb{P}}(Y = -1) = 1/2$. Technical replicates are generated from the blood sample used to assess gene expression activity in the training data and thus, are commonly used to isolate the effect of measurement error.

We explored the effect of interval uncertainty bounds resulting from quantiles of (14) at levels of $\alpha = \{0.25, 0.50, 0.75, 0.95, 0.99\}$. Figure 3 shows the ℓ_1-regularization paths obtained by solving (11) on the training data for different interval uncertainty (including no uncertainty corresponding to standard ℓ_1-LR) and illustrates how robustness affects the ordering of the genes. We see from Figure 3(a) that the first gene to appear in the regularization path is the anti-viral defense gene RSAD2. RSAD2 persists at the first gene in the regularization path for $\alpha = \{0, 0.25, 0.50, 0.75\}$ (latter three not shown) but disappears from the regularization path completely, along with a few other genes, in Figure 3(a), when $\alpha = \{0.95, 0.99\}$.



(a) ℓ_1-Regularization path on original data

(b) ℓ_1-Regularization path with bounded perturbations in x_3 and $\rho_3 = 0.1$

(c) ℓ_1-Robust regularization path with $\rho_3 = 0.1$

Fig. 2. Regularization paths as a function of $\log_{10} \lambda$: robust recovers original ordering after perturbation

The three "robust genes" that persist across all explored values of α are ADI1, OAS1, and TUBB2A. Of these three, OAS1 (codes for proteins involved in the innate immune response to viral infection) barely appears in Figure 3(a), yet is the first gene to appear in the robust regularization paths that is common to both $\alpha = \{0.95, 0.99\}$ (see Figures 3(b) and 3(c)). These results suggest that when assessing variable importance via regularization paths, one should be aware of the effect of measurement uncertainty on the ordering of the variables.

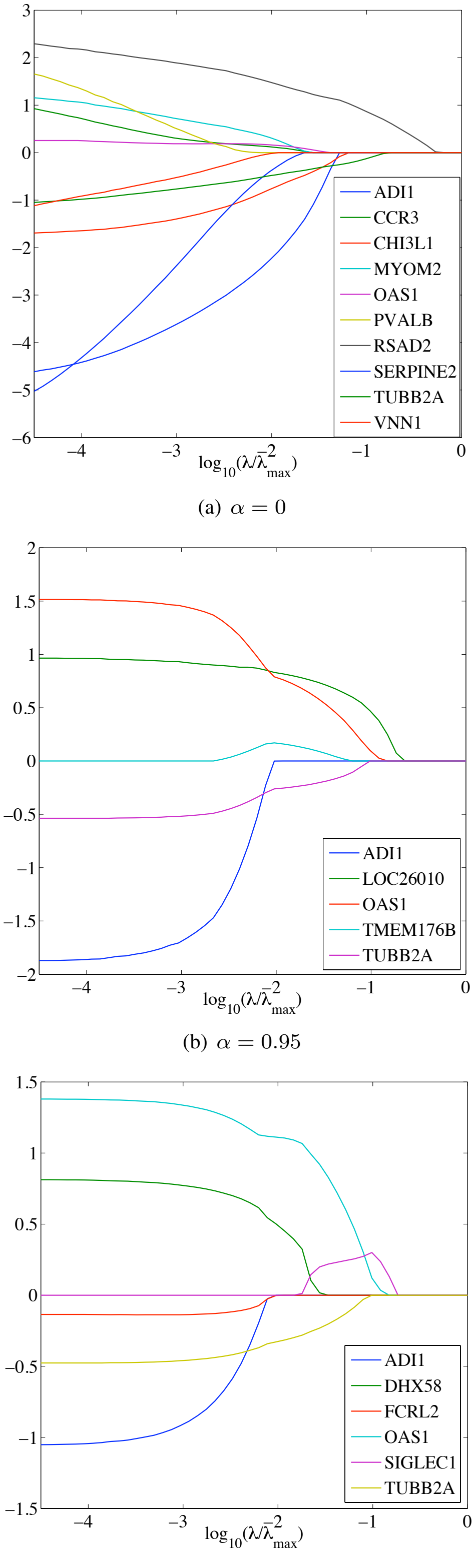
The robust formulation within this paper aims at minimizing the worst-case configuration of perturbations on the logistic regression loss function. As our goal is discriminating between two phenotypes, it is of interest to compare the worst-case probability of error for LR and RLR on perturbed data sets, i.e., $x_i \leftarrow x_i + \delta_i$, $|\delta_i| \leq \hat{\rho}_i(\alpha)$, after training on the original data. The ℓ_1-tuning parameter λ for both standard ℓ_1-LR and ℓ_1-RLR was chosen to minimize the out-of sample probability of error via 5-fold cross validation. Given the cross validated value of λ for both methods, the models were then fit to the entire set of training data. 50,000 perturbed data matrices were then generated subject to interval uncertainty bounds $\{\hat{\rho}_i(\alpha)\}_{i=1}^{p}$ for $\alpha = \{0.25, 0.50, 0.75, 0.95, 0.99\}$. The best-case and worst-case probability of error were recorded in Table I. We see for all values of α explored, the worst-case probability of error corresponding to ℓ_1-RLR is always less than or equal to that of ℓ_1-LR. When $\alpha = \{0.95, 0.99\}$, the worst-case probability of error for ℓ_1-LR is 0.30 but reduces to 0.20 when using ℓ_1-RLR with interval uncertainty estimated from the data using (14). Corresponding to the worst-case probability of errors are the sensitivity and 1-specificity, given in Table II. We see that ℓ_1-RLR achieves the same power as ℓ_1-LR at false alarm rates less than or equal to that of the non-robust method. The proposed method can be used to reduce classifier error sensitivity in large scale classification problems. This can be important in practical applications such as biomarker discovery and predictive health and disease.

TABLE I
BEST AND WORST CASE PROBABILITY OF ERROR, P_e, FOR THE HRV DATA SET

α	ℓ_1-LR		ℓ_1-Robust LR	
	min P_e	max P_e	min P_e	max P_e
0.25	0.10	0.15	0.10	0.15
0.50	0.05	0.20	0.10	0.20
0.75	0.10	0.25	0.10	0.25
0.95	0.00	0.30	0.05	0.20
0.99	0.00	0.30	0.00	0.20

VIII. CONCLUSION

Building on the results of [7] and [5], we have formulated the robust logistic regression problem with group structured uncertainty sets. By adding regularization penalties, one can enforce block-sparsity assumptions of the underlying signal. We have presented a block co-ordinate gradient method with iterative grouped thresholding for solving the penalized RLR problems. The group thresholding is affected by both the group-lasso penalty and the magnitude of the worst-case uncertainty. Thus, RLR tends to promote thresholding of highly



(a) $\alpha = 0$

(b) $\alpha = 0.95$

(c) $\alpha = 0.99$

Fig. 3. Regularization paths on the HRV data set as a function of $\log_{10} \lambda/\lambda_{max}$ for different magnitudes of interval uncertainty, as determined by the α quantile

TABLE II
SENSITIVITY AND 1-SPECIFICITY CORRESPONDING TO WORST CASE PROBABILITY OF ERROR FROM HRV DATA

α	ℓ_1-LR		ℓ_1-Robust LR	
	Sens.	1-Spec.	Sens.	1-Spec.
0.25	0.70	0.00	0.70	0.00
0.50	0.70	0.10	0.70	0.10
0.75	0.70	0.20	0.70	0.20
0.95	0.70	0.20	0.70	0.10
0.99	0.70	0.30	0.70	0.10

uncertain variables, essentially performing an initial step of variable selection. We have proposed an empirical approach to estimating the uncertainty magnitude using quantiles. This approach was applied to a real gene expression data set where quantile estimation was applied to a set of technical replicates. The numerical results on this data set establish that the proposed approach can yield lower worst-case probability of error and lower false alarm rates. Such a gain in worst-case detection performance can improve the performance for predictive health and disease, and in particular for predicting patient phenotype. It will be interesting to explore the situation when the group-structure of the uncertainty differs from that of the regularization penalty. This situation could potentially be solved by modifying the sparse group-lasso solution as in [21]. It is also of interest to explore the kernelization of this method in which one can study the propagation of bounded data uncertainty in a reproducing kernel Hilbert space.

ACKNOWLEDGMENT

The authors sincerely appreciate the thoughtful insight of Ami Wiesel in developing the ideas presented in this paper. The authors also appreciate the valuable suggestions made by Ji Zhu, Kerby Shedden, and Clayton Scott.

The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

”A” (Approved for Public Release, Distribution Unlimited)

REFERENCES

[1] A. Wiesel, Y.C. Eldar, and A. Yeredor, “Linear regression with Gaussian model uncertainty: Algorithms and bounds,” *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2194–2205, 2008.

[2] A. Wiesel, Y.C. Eldar, and A. Beck, “Maximum likelihood estimation in linear models with a Gaussian model matrix,” *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 292, 2006.

[3] R.J. Carroll, C.H. Spiegelman, K.K.G. Lan, K.T. Bailey, and R.D. Abbott, “On errors-in-variables for binary regression models,” *Biometrika*, vol. 71, no. 1, pp. 19–25, 1984.

[4] Sabine Van Huffel, Ivan Markovsky, Richard J. Vaccaro, and Torsten Söderström, “Total least squares and errors-in-variables modeling.,” *Signal Processing*, vol. 2007, 2007.

[5] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, “Parameter estimation in the presence of bounded data uncertainties,” *SIAM J. Matrix Anal. Appl.*, vol. 19, no. 1, pp. 235–252, 1998.

[6] Laurent El Ghaoui and Hervé Lebret, “Robust solutions to least-squares problems with uncertain data,” *SIAM J. Matrix Anal. Appl.*, vol. 18, no. 4, pp. 1035–1064, 1997.

[7] Laurent El Ghaoui, Gert R. G. Lanckriet, and Georges Natsoulis, "Robust classification with interval data," Tech. Rep. UCB/CSD-03-1279, EECS Department, University of California, Berkeley, Oct 2003.

[8] Gert R. G. Lanckriet, Laurent E Ghaoui, Chiranjib Bhattacharyya, and Michael I. Jordan, "A robust minimax approach to classification," *Journal of Machine Learning Research*, vol. 3, 2002.

[9] Seung-Jean Kim, Alessandro Magnani, and Stephen P. Boyd, "Robust fisher discriminant analysis," in *In Advances in Neural Information Processing Systems*. 2006, MIT Press.

[10] Huan Xu, Shie Mannor, and Constantine Caramanis, "Robustness, risk, and regularization in support vector machines," *CoRR*, vol. abs/0803.3490, 2008.

[11] Aharon Ben-Tal, Laurent E. Ghaoui, and Arkadi Nemirovski, *Robust Optimization (Princeton Series in Applied Mathematics)*, Princeton University Press, 2009.

[12] Ming Yuan, Ming Yuan, Yi Lin, and Yi Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, vol. 68, pp. 49–67, 2006.

[13] Lukas Meier, Sara van de Geer, and Peter Bühlmann, "The group lasso for logistic regression," *Journal Of The Royal Statistical Society Series B*, vol. 70, no. 1, pp. 53–71, 2008.

[14] Dimitri P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.

[15] Volker Roth and Bernd Fischer, "The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms," in *ICML '08: Proceedings of the 25th international conference on Machine learning*, 2008, pp. 848–855.

[16] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[17] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu, "Partial correlation estimation by joint sparse regression models," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 735–746, 2009.

[18] A. Wiesel, Y.C. Eldar, and S. Shamai, "Optimization of the mimo compound capacity," *IEEE Transactions on Wireless Communications*, vol. 6, no. 3, pp. 1094, 2007.

[19] A.K. Zaas, M. Chen, J. Varkey, T. Veldman, A.O. Hero, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, et al., "Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans," *Cell Host & Microbe*, vol. 6, no. 3, pp. 207–217, 2009.

[20] A. Swaroop, A.J. Mears, G. Fleury, and A.O. Hero, "Multicriteria Gene Screening for Analysis of Differential Expression with DNA Microarrays," *EURASIP Journal on Advances in Signal Processing*, 2004.

[21] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," arXiv:1001.0736v1, 2010.