Robust Logistic Principal Component Regression for Classification of Data in presence of Outliers

H. C. Wu, S. C. Chan, and K. M. Tsui
Department of Electrical and Electronics Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong.
{andrewhcwu, scchan, kmtsui}@eee.hku.hk

***Abstract*—The Logistic Principal Component Regression (LPCR) has found many applications in classification of high-dimensional data, such as tumor classification using microarray data. However, when the measurements are contaminated and/or the observations are mislabeled, the performance of the LPCR will be significantly degraded. In this paper, we propose a new robust LPCR based on M-estimation, which constitutes a versatile framework to reduce the sensitivity of the estimators to outliers. In particular, robust detection rules are used to first remove the contaminated measurements and then a modified Huber function is used to further remove the contributions of the mislabeled observations. Experimental results show that the proposed method generally outperforms the conventional LPCR under the presence of outliers, while maintaining a performance comparable to that obtained under normal condition.**

I. INTRODUCTION

Remarkable development of data collection and storage capabilities during the past decade has led to an unprecedented increase in data size and complexity for analysis. In many applications, such as microarray studies, the number of variables is much larger than the number of samples [1]. Moreover, variables are very often correlated making the analysis difficult. Classification performance of these datasets may therefore be complicated by groups of possibly correlated variables that are irrelevant for class separation [2]. To handle such correlated variables in large dataset, principal component regression (PCR) is a promising approach to eliminate irrelevant features by reducing effective dimension of the dataset. To apply PCR for classification, one may first invoke the PCR to compute the PC scores, and then employ the Logistic Regression (LR) to regress against the observations and PC scores. An advantage of adopting the LR is that it does not require pairwise or one-against all training for multi-class classification [3]. The resultant algorithm is referred to as the Logistic PCR [4]. In practice, the accuracy of classification can be considerably affected by outliers, which are samples that deviate significantly from other remaining samples of the same group due to errors in data taking, special events, etc. There are mainly two types of outliers in classification: 1.) contaminated measurements, and 2.) mislabeled observations [5]. Due to these outliers, the performance of the LPCR, which is derived from the least squares (LS) criterion, may also degrade considerably.

To overcome this problem, we propose a new robust LPCR method to reduce the effect of these outliers on the classification accuracy. In the proposed LPCR algorithm, robust detection rules derived from the T^2 score, squared prediction error (SPE) are used to remove the contaminated measurements. Then, a modified Huber function is incorporated into the conventional LPCR to remove samples that are detected as mislabeled observations using the logistic error. Though the concept of robust M-estimation based on automatic threshold selection (ATS) and the modified Huber function has been reported in [6] for robust estimation in linear systems, the incorporation of T^2 score, squared prediction error (SPE) and logistic error, and its application to non-linear LR is to our best knowledge new. Experimental results show that the proposed robust LPCR offers much better classification accuracy than the conventional LPCR in the presence of outliers, while the performance is also highly comparable under normal situation.

This paper is organized as follows. In Section II, the the LPCR are revisited. Section III introduces the proposed robust LPCR for suppressing the effect of the outliers in classification/regression. Experimental results are presented in Section IV and conclusions are drawn in Section V.

II. THE LOGISTIC PRINCIPAL COMPONENT REGRESSION

In LPCR, the measurement vectors are first projected into the major subspace spanned by a number of chosen major principal components (PCs). Then, the PC scores are computed, which are the contributions of the measurement vectors to each PC. Finally, the LR is invoked on the PC scores and the observations, which can be binary or multi-class, to perform classification. In this paper, we mainly focus on the commonly encountered binary classification problem.

More specifically, suppose that we have N subjects or samples. The N binary observations of $y = 0,1$ and its corresponding J measurement variables can be respectively grouped into an $(N \times 1)$ vector $\boldsymbol{Y} = [y_1, y_2, ..., y_N]^T$ and an $(N \times J)$ matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]^T$, where each vector $\boldsymbol{x}_i = [x_{i,1}, ,..., x_{i,J}]^T$, $i = 1, \ldots, N$, represents the measurements of the *i-th* sample. Usually, $\boldsymbol{x}_i$ is "centered", i.e. with its

mean removed, before the PCs are computed. In PCA, we wish to express centered data matrix $\overline{\boldsymbol{X}} = [\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2, ..., \overline{\boldsymbol{x}}_N]^T$ in terms of B principal components:

$$\overline{\boldsymbol{X}} = \sum_{m=1}^{B} \boldsymbol{t}_m \boldsymbol{p}_m^T + \boldsymbol{E} = \boldsymbol{\Gamma}\boldsymbol{P}^T + \boldsymbol{E}, \quad (1)$$

where $\boldsymbol{\Gamma} = [\boldsymbol{t}_1, ..., \boldsymbol{t}_B]^T$ is the score matrix, $\boldsymbol{P} = [\boldsymbol{p}_1, ..., \boldsymbol{p}_B]^T$ is the collection of PCs or loading matrix, and B is an appropriately chosen number of PCs to achieve a sufficiently small approximation error $\boldsymbol{E}$. Therefore, irrelevant information can be removed to improve the classification results. A common way to determine the PCs is to compute the eigenvalue decomposition (EVD) of the empirical correlation matrix:

$$\boldsymbol{C}_{\boldsymbol{xx}} = E[\overline{\boldsymbol{x}} \cdot \overline{\boldsymbol{x}}^T] = \tfrac{1}{n-1} \overline{\boldsymbol{X}}^T \overline{\boldsymbol{X}} = \boldsymbol{U}\Lambda\boldsymbol{U}^T, \quad (2)$$

where the columns of $\boldsymbol{U}$ are the eigenvectors and they are also the PCs and $\boldsymbol{\Lambda} = diag\{\lambda_1, ..., \lambda_J\}$ contains the eigenvalues in descending order of magnitude $(\lambda_1 \geq \lambda_2 \ldots \geq \lambda_J)$. If the first B largest eigenvalues and their eigenvectors $\boldsymbol{U}_B$ are retained, then one gets $\boldsymbol{P} = \boldsymbol{U}_{\boldsymbol{B}}$. The PC scores $\boldsymbol{\Gamma}$ can be determined as

$$\boldsymbol{\Gamma} = \overline{\boldsymbol{X}}\boldsymbol{U}_B. \quad (3)$$

After that, the LR is invoked on the observations $\boldsymbol{Y}$ and the PC scores $\boldsymbol{\Gamma}$ to perform classification. Due to the reduced dimension (B versus the original dimension J) after using PCA, the variance of the regression will be reduced. For notation convenience, we use $\boldsymbol{\tau}_i = [t_{i,1}, t_{i,2}, \ldots t_{i,B}]^T$ to represent the PC scores of the chosen B PCs for measurement $\boldsymbol{x}_i$ of each sample, and $\boldsymbol{\Gamma} = [\boldsymbol{\tau}_1^T, \boldsymbol{\tau}_2^T, \ldots, \boldsymbol{\tau}_N^T]$. In the LR, the conditional class probability $P(y = 1 \mid \boldsymbol{\tau})$ for $\boldsymbol{\tau}$ in class 1 is modeled as:

$$P(y = 1 \mid \boldsymbol{\tau}) = e^{\eta(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau})} / (1 + e^{\eta(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau})}) = p(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}), \quad (4)$$

where $\eta(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau}) = \alpha + \boldsymbol{\tau}^T\boldsymbol{\beta}$ is chosen as a linear predictor in terms of $\boldsymbol{\tau}$ with intercept α and regression coefficients $\boldsymbol{\beta}$. The regression coefficients are usually estimated by maximizing the log-likelihood function as follows:

$$\log(L) = \sum_{i=1}^{N} [y_i \log p_i(\boldsymbol{\gamma}) + (1 - y_i) \log(1 - p_i(\boldsymbol{\gamma})], \quad (5)$$

where for notation convenience, we have used $\log(L)$ for $\log L(\boldsymbol{\gamma} \mid \boldsymbol{\tau}_i, y_i)$, $\boldsymbol{\gamma} = [\alpha \quad \boldsymbol{\beta}^T]^T$ and $p_i(\boldsymbol{\gamma}) = p_i(\alpha, \boldsymbol{\beta}, \boldsymbol{\tau})$. The solution of (5) can be obtained by setting the partial derivative of the log likelihood to zero. This yields:

$$\partial \log(L) / \partial \boldsymbol{\gamma} = \boldsymbol{0} \;\Rightarrow\; \tilde{\boldsymbol{\Gamma}}^T (\boldsymbol{Y} - \boldsymbol{p}(\boldsymbol{\gamma})) = \boldsymbol{0}, \quad (6)$$

where $\boldsymbol{p}(\boldsymbol{\gamma}) = [p_1(\boldsymbol{\gamma}), ..., p_{N'}(\boldsymbol{\gamma})]^T$ and $\tilde{\boldsymbol{\Gamma}} = [\boldsymbol{1}_{N'} \quad \boldsymbol{\Gamma}]$. As $\boldsymbol{p}(\boldsymbol{\gamma})$ is non-linear in $\boldsymbol{\gamma}$, a first order Taylor series can be used as an approximation to $\boldsymbol{p}(\boldsymbol{\gamma})$ as follows:

$$p_i(\boldsymbol{\gamma}^{(k)} + \boldsymbol{\delta}^{(k)}) \approx p_i(\boldsymbol{\gamma}^{(k)}) + p'_i(\boldsymbol{\gamma}^{(k)}) \boldsymbol{\delta}^{(k)}, \quad (7)$$

where $\boldsymbol{\gamma}^{(k)}$ is the solution obtained in the $k-th$ iteration, $\boldsymbol{\delta}^{(k)} = \boldsymbol{\gamma}^{(k+1)} - \boldsymbol{\gamma}^{(k)}$, $p'_i(\boldsymbol{\gamma}^{(k)}) = p_i^{(k)}(1 - p_i^{(k)})\tilde{\boldsymbol{\tau}}_i$ is the derivative of $p_i(\boldsymbol{\gamma})$ evaluated at $\boldsymbol{\gamma}^{(k)}$, and $\tilde{\boldsymbol{\tau}}_i = [1 \quad \boldsymbol{\tau}_i^T]^T$. With (7), the normal equation in (6) can be rewritten as

$$\tilde{\boldsymbol{\Gamma}}^T (\boldsymbol{Y} - \boldsymbol{p}^{(k)} - \boldsymbol{W}^{(k)} \tilde{\boldsymbol{\Gamma}} \boldsymbol{\delta}^{(k)}) = \boldsymbol{0}, \quad (8)$$

where $\tilde{\boldsymbol{\Gamma}} = [\tilde{\boldsymbol{\tau}}_1^T, \tilde{\boldsymbol{\tau}}_2^T, \ldots, \tilde{\boldsymbol{\tau}}_N^T]$, $\boldsymbol{W}^{(k)} = diag\{w_1^{(k)}, w_2^{(k)} .... w_N^{(k)}\}$, $w_i^{(k)} = p_i^{(k)}(1 - p_i^{(k)})$, $p_i^{(k)} = p_i(\boldsymbol{\gamma}^{(k)})$ and $\boldsymbol{p}^{(k)} = \boldsymbol{p}(\boldsymbol{\gamma}^{(k)})$. Conventionally, the regression parameters can be solved using iteratively using (8) which gives rise to the Iterated reweighted least squares (IRWLS) algorithm.

III. Robust Logistic Principal Component Regression

The proposed LPCR algorithm is mainly divided into three steps: 1.) The robust L_1 median is computed to center the measurement variables before performing the PCR. This offers improved robustness over the conventional sample mean, which is sensitive to impulsive noise. 2.) PCR is invoked on the centered data and robust detection rules derived from the T^2 score and SPE are used to remove the contaminated measurements in $\boldsymbol{X}$. 3.) A modified Huber function is incorporated into the LR and the resultant robust LR is used to remove mislabeled observations detected using the logistic error.

A. The Robust L1 median

In the proposed robust LPCR, we employ the robust L_1 median reported in [7] for centering:

$$\boldsymbol{\mu}_{L1} = \arg\min_{\boldsymbol{\mu}} \{\textstyle\sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \boldsymbol{\mu} \right\|_2\}, \quad (9)$$

where $\|.\|_2$ denotes the Euclidean distance. The robust L_1 median estimates the robust centre by minimizing the sum of Euclidean distances to all points in the data set. The centered measurement matrix can be computed as follows:

$$\overline{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{1}_N \boldsymbol{\mu}_{L1}^T, \quad (10)$$

where $\boldsymbol{1}_N$ is an $(N \times 1)$ vector with all entries equal to ones.

B. Detection Rules and Automatic Threshold Selection

After robust L_1 median, PCR is invoked on the centered data to compute the PC scores as in (1) to (3). The error measures T^2 score and SPE are computed as

$$T_i^2 = \left\| \overline{\boldsymbol{x}}_i^T \boldsymbol{U}_B \boldsymbol{\Lambda}_B^{-1} \boldsymbol{U}_B^T \overline{\boldsymbol{x}}_i \right\|_2^2 \text{ and } \mathrm{SPE}_i = \left\| \overline{\boldsymbol{x}}_i - \boldsymbol{U}_B \boldsymbol{U}_B^T \overline{\boldsymbol{x}}_i \right\|_2^2, \quad (11)$$

where $\boldsymbol{U}_B$ and $\boldsymbol{\Lambda}_B^{-1}$ are the major PCs and eigenvalues obtained similarly as in (2). They are used to quantify how a subject deviates from the group of samples. To perform outlier detection, the following robust detection rules can be used

$$\left| T_i^2 - \mu_{T^2} \right| \geq \Gamma_{T^2} \text{ and } \left| \mathrm{SPE_i} - \mu_{\mathrm{SPE}} \right| \geq \Gamma_{SPE}, \quad (12)$$

where $\mu_{T^2} = med(T_i^2)$ and $\mu_{\mathrm{SPE}} = med(\mathrm{SPE}_i)$. Here, ξ is a threshold quartile parameter corresponding to the upper $(1 - P\{X > \xi\})$ percentile of the Gaussian distribution. Γ_{T^2} and Γ_{SPE} are robust thresholds for identifying the sample as outlier and they can be selected by the ATS reported in [6]:

$$\Gamma_{T^2} = \xi\sigma_{T^2}, \; \Gamma_{\mathrm{SPE}} = \xi\sigma_{SPE}, \quad (13)$$

where σ_{T^2} and σ_{SPE} are respectively robust scale estimators of T_i^2 and SPE_i in (11) and they can be determined using the following robust scale estimators [6]

$$\sigma_{T^2}^2 = 2.13med((\Delta T_i^2)^2), \sigma_{SPE}^2 = 2.13med(\Delta SPE_i^2), \quad (14)$$

where $\Delta T_i^2 = T_i^2 - \mu_{T^2}$, $\Delta SPE_i = \mathrm{SPE}_i - \mu_{\mathrm{SPE}}$ and 2.13 is a correction factor for Gaussian input [6]. From the robust detection rules in (12), we calculate robust weights $q_{x,i}$ for the measurement $\boldsymbol{x}_i$ of each sample and it is given by

$$q_{x,i} = \begin{cases} 0 & \left|T_i^2 - \mu_{T^2}\right| \geq \Gamma_{T^2} \text{ or } \left|\mathrm{SPE}_i - \mu_{\mathrm{SPE}}\right| \geq \Gamma_{SPE} \\ 1 & otherwise \end{cases}. \quad (15)$$

In other words, if T_i^2 or SPE_i of the $i-th$ sample exceeds the thresholds defined in (12), $\boldsymbol{x}_i$ is identified as an outlier and its robust weight is set to $q_{x,i} = 0$, so that its contribution is removed, otherwise $q_{x,i} = 1$. Afterwards, the proposed robust LR is invoked on the observations $\boldsymbol{Y}$ and the PC scores $\boldsymbol{\Gamma}$ to perform classification. Next, we will discuss how the conventional LR should be modified when mislabelled observations are encountered.

C. Robust Least M-estimate based Logistic Regression

In the conventional LR, the normal equation in (8) can be rewritten as $(\tilde{\boldsymbol{\Gamma}}^T \boldsymbol{W}^{(k)})\tilde{\boldsymbol{\Gamma}}\boldsymbol{\gamma}^{(k+1)} = \tilde{\boldsymbol{\Gamma}}^T \boldsymbol{W}^{(k)} \boldsymbol{Z}$ where $\boldsymbol{Z}^{(k)} = \tilde{\boldsymbol{\Gamma}}\boldsymbol{\gamma}^{(k)} + \tilde{\boldsymbol{W}}^{(k)^{-1}}(\boldsymbol{Y} - \boldsymbol{p}^{(k)})$. This is equivalent to

$$\min . \sum_{i=1}^{N} \left\| w_i^{(k)^{1/2}} (z_i - \tilde{\boldsymbol{\tau}}_i^T \boldsymbol{\gamma}^{(k+1)}) \right\|_2^2, \quad (16)$$

where $w^{(k)}$ and $\tilde{\boldsymbol{\tau}}_i$ are same as those in (8). $z_i^{(k)} = \tilde{\boldsymbol{\tau}}_i^T \boldsymbol{\gamma}^{(k)} + w^{(k)^{-1}}(y_i - p_i^{(k)})$. Here, we refer the term $e_i^{(k)} = w_i^{(k)^{1/2}}(z_i - \tilde{\boldsymbol{\tau}}_i^T \boldsymbol{\gamma}^{(k+1)})$ to as the logistic error and it is the weighted error between the observation $\boldsymbol{y}_i$ and predicted probability $p_i(\boldsymbol{\gamma}^{(k)} + \boldsymbol{\delta}^{(k)}) \approx p_i^{(k)} + w^{(k)} \tilde{\boldsymbol{\tau}}_i^T \boldsymbol{\delta}^{(k)}$ in (7)

$$e_i^{(k)} = w^{(k)^{-1/2}}(y_i - (p_i^{(k)} + w^{(k)} \tilde{\boldsymbol{\tau}}_i^T \boldsymbol{\delta}^{(k)}))). \quad (17)$$

By solving the weighted LS in (16) and using the solution as the new estimate $\boldsymbol{\gamma}^{(k)}$ repeatedly, one obtains the IRWLS algorithm. Usually, the iteration stops when a maximum number of iterations is reached or when the change $\boldsymbol{\delta}^{(k)}$ is sufficiently small. However, a single outlier with large amplitude in $\boldsymbol{x}_i$ or mislabel in $\boldsymbol{y}_i$ can substantially increase the LS error in (17). This affects adversely the estimation of the regression coefficients [5]. To overcome this problem, we employ robust M-estimation [8], [13] where a robust M-estimate function $\rho(.)$ for the objective function in (17) is used:

$$\min . \sum_{i=1}^{N} q_{x,i} \rho(w_i^{(k)^{1/2}} z_i - w_i^{(k)^{1/2}} \tilde{\boldsymbol{\tau}}_i^T \boldsymbol{\gamma}), \quad (18)$$

where $q_{x,i}$ is the robust weight defined in (16). For simplicity, $\rho(e) = \begin{cases} e^2/2 & 0 \leq |e| < \Gamma_e \\ \Gamma_e^2/2 & otherwise \end{cases}$ is chosen here as the modified Huber M-estimate function in [6] and $\Gamma_e = \xi\sigma_e$ is a robust threshold defined similarly as in (13) and $\sigma_e^2 = 2.13med(e_i^2)$. The solution to (18) can be found by differentiating the function in (18) with respect to $\boldsymbol{\gamma}$

$$\begin{aligned} \partial \textstyle\sum_{i=1}^{N} q_{x,i} \rho(e_{\alpha,i}) / \partial \alpha &= \textstyle\sum_{i=1}^{N} q_{x,i} \psi(e_{\alpha,i}) \partial e_{\alpha,i} / \partial \alpha \\ &= \textstyle\sum_{i=1}^{N} -q_{x,i} \psi(e_{\alpha,i}) w_i^{(k)^{1/2}}, \end{aligned} \quad (19a)$$

$$\begin{aligned} \partial \textstyle\sum_{i=1}^{N} q_{x,i} \rho(e_{\beta,i,j}) / \partial \beta_j &= \textstyle\sum_{i=1}^{N} q_{x,i} \psi(e_{\beta,i,j}) \partial e_{\beta,i,j} / \partial \beta_j \\ &= \textstyle\sum_{i=1}^{N} -q_{x,i} \psi(e_{\beta,i,j}) w^{(k)^{1/2}} t_{i,j}, \end{aligned} \quad (19b)$$

where $e_{\alpha,i} = w_i^{(k)^{1/2}} z_i - w_i^{(k)^{1/2}} \alpha$, $e_{\beta,i,j} = w_i^{(k)^{1/2}} z_i - w_i^{(k)^{1/2}} t_{i,j} \beta_j$, $\psi(e) = \begin{cases} e & 0 \leq |e| < \Gamma_e \\ 0 & otherwise \end{cases}$ is the derivative of $\rho(e)$. The normal equations above can be rewritten to the matrix form

$$\nabla_{\gamma} \rho(\boldsymbol{E}^{(k)}) = -\tilde{\boldsymbol{\Gamma}}^T \boldsymbol{W}^{(k)^{1/2}} \boldsymbol{Q}_y^{(k)} \boldsymbol{Q}_x \boldsymbol{E}^{(k)} = \boldsymbol{0}, \quad (20)$$

where $\boldsymbol{E}^{(k)} = [e_1^{(k)}, e_2^{(k)}, \ldots, e_N^{(k)}]^T$, $\boldsymbol{Q}_x = diag\{q_{x,1}, \ldots q_{x,N}\}$, and $\boldsymbol{Q}_y^{(k)} = diag\{q_{y,1}^{(k)}, \ldots q_{y,N}^{(k)}\}$ is the robust weighting matrix for removing the contribution of the mislabeled observations with $q_{y,i}^{(k)} = \begin{cases} \psi(e_i^{(k)})/e_i^{(k)} & e_i^{(k)} \neq 0 \\ 0 & otherwise \end{cases}$. We can see that the modified Huber function simply ignores a sample when it is detected as a mislabeled observation and proceeds with normal updating otherwise. Note that other more complicated M-estimate function such as the Hampel's function [8] can also be used to suppress the contribution of the mislabeled observations to different extents. As $\boldsymbol{W}^{(k)^{1/2}} \boldsymbol{Q}_y^{(k)} \boldsymbol{Q}_x \boldsymbol{W}^{(k)^{1/2}} = \boldsymbol{W}^{(k)} \boldsymbol{Q}_y^{(k)} \boldsymbol{Q}_x$ and $\boldsymbol{E}^{(k)} = \boldsymbol{W}^{(k)^{1/2}}(\boldsymbol{Z}^{(k)} - \tilde{\boldsymbol{\Gamma}}\boldsymbol{\gamma}^{(k+1)})$, the solution of Eqn. (20) can be simplified to

$$\tilde{\boldsymbol{\Gamma}}^T \boldsymbol{W}^{(k)} \boldsymbol{Q}_y^{(k)} \boldsymbol{Q}_x (\boldsymbol{Z} - \tilde{\boldsymbol{\Gamma}}\boldsymbol{\gamma}^{(k+1)}) = \boldsymbol{0}. \quad (21)$$

Hence, the solution to Eqn. (21) is

$$\boldsymbol{\gamma}^{(k+1)} = (\tilde{\boldsymbol{\Gamma}}^T \boldsymbol{W}^{(k)} \boldsymbol{Q}_y^{(k)} \boldsymbol{Q}_x \tilde{\boldsymbol{\Gamma}})^{-1} (\tilde{\boldsymbol{\Gamma}}^T \boldsymbol{W}^{(k)} \boldsymbol{Q}_y^{(k)} \boldsymbol{Q}_x \boldsymbol{Z}). \quad (22)$$

IV. Experimental Results

As an illustration, we consider the Leukemia dataset [9] obtained from the Kent Ridge Bio-medical Data Set Repository (http://datam.i2r.a-star.edu.sg/datasets /krbd/) and compare the classification performance of the proposed robust LPCR and the conventional counterpart. Data-preprocessing such as logarithmic transformation, filtering and standardization of the raw dataset are performed according to the procedures reported in [10]. The pre-processed dataset contains 47 (25) samples of class 0 (1), and 3571 variables. Outliers injected into $\boldsymbol{X}$ are generated using the contaminated Gaussian model : $\boldsymbol{n}_i \sim (1-\eta_{im})N(\boldsymbol{0}, \sigma_g^2 \boldsymbol{I}) + \eta_{im} N(\boldsymbol{0}, \sigma_{\mathrm{im}}^2 \boldsymbol{I})$, where $\boldsymbol{n}_i$ is the impulsive outlier, η_{im} is contamination probability, and $N(\boldsymbol{\mu}, \boldsymbol{R})$ denotes a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and

covariance $\boldsymbol{R}$. σ_g^2 is the variance of the additive Gaussian component and σ_{im}^2 is the variance of the impulsive component and they are chosen as 1 and 10^4 respectively. Labels in $\boldsymbol{Y}$ are randomly perturbed with the same contamination probability η_{im} to simulate mislabeled observations. A 10-fold double loop cross validation (CV) is adopted for parameter tuning and evaluation to avoid the optimistic and selection biases [11], [12]. The classification accuracies of the two algorithms are obtained using 50 Monte Carlo runs[1] . Fig. 1 and Table I shows the classification accuracies of different algorithms under three types of outliers: 1) contaminated measurements, 2) mislabeled observations, and 3) outliers with both contaminated measurements and mislabeled observations of contamination levels $\eta_{im} = 0$, 5/72 and 10/72. We can see that the proposed LPCR generally has higher mean classification accuracy and smaller fluctuation than the conventional LCPR algorithm under the presence of outliers, while the performance of the proposed LPCR is nearly the same as its conventional counterpart under normal condition.

V. Conclusion

A new robust Logistic PCR (R-PLCR) for classification in large dataset with possible outliers is proposed. It aims to reduce the effect of outliers on classification accuracy by detecting and removing the contaminated measurements and employing a modified Huber function to remove the adverse contributions of the mislabeled observations. Experimental results using the Leukemia dataset with injected contaminated measurements and mislabels show that the proposed robust LPCR offers much better classification accuracy than the conventional LPCR in the presence of these outliers, while the performance is also highly comparable to that obtained in normal condition.

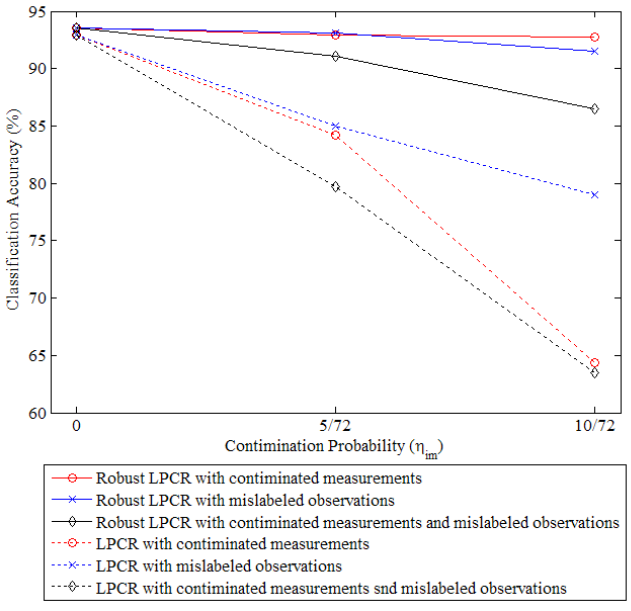

Fig. 1. Mean classification accuracies of the proposed robust LPCR algorithm and the conventional LPCR algorithm under three types of outliers 1.) contaminated measurements, 2.) mislabeled observations, 3.) outliers with mislabeled observation and contaminated measurements

[1] In each Monte Carlo run, a complete run of the 10-fold double loop CV procedure is performed.

References

[1] C. Q. Chang, Z. Ding, Y. S. Hung and P. C. W. Fung, "Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data," *Bioinformatics,* vol. 24, no. 11, pp. 1349, Apr. 2008

[2] J. Fan and Y. Fan, "High-dimensional Classification using Features Annealed Independence Rules," *Annals of Statistics*, vol. 36, no. 6, pp. 2605–2637, 2008.

[3] I. Lubenko and A.D. Ker. , "Steganalysis using logistic regression," in Proc. Society of Photo-Optical Instrumentation Engineers, vol. 7880, pp. 78800K - 78800K-11 , Feb. 2011

[4] B. D. Marx and E. P. Smith, "Principal Component Estimation for Generalized Linear Regression," *Biometrika*, vol. 27, no. 1, pp. 23- 31, Mar. 1990.

[5] P. J. Rousseeuw and B. C. van Zomeren, "Unmasking Multivariate Outliers and Leverage Points," *J. Amer. Stat. Assoc*, vol. 85, no. 411, pp. 633-639, Sep. 1990.

[6] S. C. Chan and Y. Zhou, "On the Performance Analysis of the Least Mean M-Estimate and Normalized Least Mean M-Estimate Algorithms with Gaussian inputs and Additive Gaussian and Contaminated Gaussian Noises," *J. Signal Proces. Syst.*, vol. 60, no. 1, pp. 81-103, July 2010.

[7] H. Fritz, P. Filzmoser and C. Croux, "A comparison of algorithms for the multivariate L1-median," *Computational Statistics*, DOI: 10.1007/s00180-011-0262-4, June 2010.

[8] P. J. Huber, *Robust statistics* , John Wiley and Sons, New York, 1981.

[9] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield and E. Lander, "Molecular classification of cancer: dass discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531 - 537, Oct. 1999.

[10] S. Dudoit, J. Fridlyand and T. Speed, "*Comparison of discrimination methods for the classification of tumors using gene expression data*.," *J. Amer. Stat. Assoc*., vol. 97, no. 457, pp. 77-87, Mar. 2002.

[11] C. Ambroise and G. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *PNAS*, vol. 99, no. 10, pp. 6562 - 6566, Apr. 2002.

[12] H. Ishibuchi, Y. Nakashima and Y. Nojima, "Double cross-validation for performance evaluation of multi-objective genetic fuzzy systems," in *Proc. IEEE GEFS 2011*, pp. 31-38, 11-15 April 2011.

[13] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: the Approach Based on Influence Functions.* Wiley-Interscience, New York, 1987.

TABLE I. Classification Results ($mean \pm std$)

Algorithms	Contaminated Measurements		
	$\eta_{im} = 0$	$\eta_{im} = 5/72$	$\eta_{im} = 10/72$
Robust LPCR	93.55±1.23	92.21±1.51	90.78±2.10
LPCR	92.89±1.02	84.23±2.68	64.38±2.98
Mislabeled Observations			
Robust LPCR	93.55±1.23	93.10±1.74	91.54±2.27
LPCR	92.89±1.02	84.98±2.81	78.99±2.84
Contaminated Measurements and Mislabeled Observations			
Robust LPCR	93.55±1.23	91.06±3.06	86.47±3.74
LPCR	92.89±1.02	79.73±4.17	63.47±4.67