The Annals of Applied Statistics
2015, Vol. 9, No. 2, 1053–1075
DOI: 10.1214/15-AOAS825

SAMPLE SIZE DETERMINATION FOR TRAINING CANCER CLASSIFIERS FROM MICROARRAY AND RNA-seq DATA

BY SANDRA SAFO[1,2], XIAO SONG[3] AND KEVIN K. DOBBIN[1,2]

University of Georgia

The objective of many high-dimensional microarray and RNA-seq studies is to develop a classifier of cancer patients based on characteristics of their disease. The germinal center B-cell (GCB) classifier study in lymphoma and the National Cancer Institute's Director's Challenge lung (DC-lung) study are two examples. In recent years, such classifiers are often developed using regularized regression, such as the lasso. A critical question is whether a better classifier can be developed from a larger training set size and, if so, how large the training set should be. This paper examines these two questions using an existing sample size method and a novel sample size method developed here specifically for lasso logistic regression. Both methods are based on pilot data. We reexamine the lymphoma and lung cancer data sets to evaluate the sample sizes, and use resampling to assess the estimation methods. We also study application to an RNA-seq data set. We find that it is feasible to estimate sample size for regularized logistic regression if an adequate pilot data set exists. The GCB and the DC-lung data sets appear adequate, under specific assumptions. Existing human RNA-seq data sets are by and large inadequate, and cannot be used as pilot data. Pilot RNA-seq data can be simulated, and the methods in this paper can be used for sample size estimation. A MATLAB program is made available.

1. Introduction. Regularized regression methods, such as the lasso, are common in the analysis of high-dimensional data [Bi et al. (2014), Moehler et al. (2013), Zhang et al. (2013), Zwiener et al. (2014)]. Regularized logistic regression is often used to classify patients into different groups, such as those who will versus will not respond to a targeted therapy. While development of classifiers is a long process [Dyrskjøt (2003), Hanash, Baik and Kallioniemi (2011), Pfeffer et al. (2009), McShane and Hayes (2012), Simon (2010)], a critical step in that process is determining the sample size necessary to adequately train a classifier from high-dimensional data.

In this paper, we look at three microarray data sets to evaluate whether the sample sizes used were adequate. Also, we examine RNA-seq data and the potential

Received March 2014; revised January 2015.
[1]Supported in part by NIH NCI Grant 1R21CA152460.
[2]Supported in part by Georgia Research Alliance Distinguished Cancer Scientist program.
[3]Supported in part by NSF Grant DMS-11-06816.
Key words and phrases. Sample size, lasso, classification, regularized logistic regression, conditional score, high-dimensional data, measurement error.

to determine sample size for this newer technology. We reexamine the data of Rosenwald et al. (2002). Under the assumption that the germinal center B cell (GCB) lymphoma subtype patients are correctly identified in this study, we examine whether a better classifier can be developed using the lasso logistic regression and, if so, how large a training set would be needed to do so. Similarly, for the lung cancer data set of Shedden et al. (2008), the lasso logistic regression is used to evaluate whether a better outcome predictor could be developed from a larger data set. Third, we examine a more dramatic classification difference comparing prostate tumors to normal prostate tissue [Dettling and Bühlmann (2003)]. We also provide the assumptions on which these sample size estimates are based. These assumptions can be used to evaluate important public health planning questions, such as whether re-running similar studies using RNA-seq technology is likely to yield improved classification. Finally, we examine an RNA-seq data set as a proof of principle to assess the potential of these methods on this new technology. We find that the methods can be effectively applied to RNA-seq data.

What follows is a brief review of the methodology literature. A more extensive review for interested readers appears in the Supplement [Safo, Song and Dobbin (2015)].

The novel approach developed for lasso logistic regression in this paper uses errors-in-variables (EIV) regression. EIV methods for logistic regression include simulation extrapolation [SIMEX, Cook and Stefanski (1994)], conditional score [Stefanski and Carroll (1987)], consistent functional methods [Huang and Wang (2001)], approximate corrected score [Novick and Stefanski (2002)], projected likelihood ratio [Hanfelt and Liang (1995)] and quasi-likelihood [Hanfelt and Liang (1995)]. We discuss each approach briefly. The SIMEX EIV method adds additional measurement error to the data, establishes the trend, and extrapolates back to the no error model using a fitted polynomial regression; as discussed in Cook and Stefanski (1994), evaluating the adequacy of a fitted regression requires judgment and is not automatic. The subjective fitting step can complicate algorithm implementation and Monte Carlo evaluation of performance. So we do not focus on SIMEX, although we do find SIMEX useful in settings where other EIV methods do not perform well. The consistent functional method is most valuable in large-scale studies [Huang and Wang (2000)], which are currently very rare in high dimensions with sequencing or microarrays. The logistic model does not fit the corrected score smoothness assumptions; also, the Monte Carlo corrected score method is not consistent for logistic regression and implementation of the method requires programming software with complex number capabilities [Carroll et al. (2006)]. Quasi-likelihood methods can be challenging to implement for logistic regression. Conditional score methods, on the other hand, are computationally tractable and relatively easy to implement, and have shown good finite-sample performance [Carroll et al. (2006)]. We found the sufficient statistics suggested by Hanfelt and Liang (1995, 1997) to be more stable for this application than the original conditional score, so we used this closely related approach.

A practical question when using a sample size method that will be based on a pilot data set, rather than a parametric model, is whether the pilot data set is large enough. If the pilot data set is too small, then no classifier developed on it may be statistically significantly better than chance, which can be assessed with a permutation test [e.g., Mukherjee et al. (2003)]. But, even if the classifier developed on the pilot data set is better than chance, the pilot data set can still be too small to estimate the asymptotic performance as $n \to \infty$ well. This latter is a more complex question. But because it is practically important, guidelines are developed here for evaluating the pilot data set size.

The sample size method developed in this paper and the one in Mukherjee et al. (2003) are based on resampling from a pilot data set or from a simulated data set if no pilot is available. Resampling is used to estimate the logistic regression slopes for different sample sizes and the prediction error variances. Cross-validation (CV) [e.g., Geisser (1993)] is a well-established method for obtaining nearly unbiased estimates of logistic regression slopes. Because regularized regression already contains a cross-validation step for parameter tuning, estimating the logistic regression slope by cross-validation requires nested (double) cross-validation [e.g., Davison and Hinckley (1997)]. An inner cross-validation loop selects the penalty parameter value, which is then used in the outer loop to obtain the cross-validated classification scores. We also found it necessary to center and rescale individual CV batches, and repeat the CV 20–50 times to denoise the estimates. This process is termed repeated, centered, scaled cross-validation (RCS-CV). To estimate prediction error variances, the leave-one-out bootstrap (LOOBS) [Efron and Tibshirani (1997)] can be used. Modification of standard LOOBS is needed because of the cross-validation step embedded in the regularized regression. To avoid information leak, the prediction error variance is estimated by the leave-one-out nested case-cross-validated (LOO-NCCV-BS) bootstrap [Varma and Simon (2006)]. The same centering and scaling steps added for CV were also added to the LOO-NCCV-BS. We call this CS-LOO-NCCV-BS.

Regularized regression for high-dimensional data is a very active area of current research in statistics. Common methods include the lasso [Tibshirani (1996)], the adaptive lasso [Zou (2006)] and the elastic net [Zou and Hastie (2005)], among many others [Fan and Li (2001), Meier, van de Geer and Bühlmann (2008), Zhu and Hastie (2004)]. In this paper, the focus of the simulation studies is on the lasso logistic regression, with selection of the penalty parameter via the cross-validated error rate. Our sample size methodology can be used with other regularized logistic regression methods, but may require modifications, particularly if additional layers of resampling are involved (e.g., the adaptive lasso).

In the study of lymphoma, we find that there is little room for improvement in the GCB signature. This means that patients who receive treatment based on this signature would be unlikely to have their treatment changed as a result of a much larger study using microarrays being conducted. Similarly, in the lung cancer application, we find that larger studies in lung cancer would not be likely to

yield better survival prediction, despite the relatively poor performance of the classifier. This confirms that, unlike breast cancer, developing clinically useful gene expression-based lung cancer prognostic predictors is probably not feasible. In the prostate cancer data set, we find that the asymptotic accuracy of a classifier that distinguishes tumor from normal tissue is only around 88%. This accuracy may be viewed as lower than expected since tumor and normal tissues tend to be very different in most cancers; whether the low accuracy is due to the high heterogeneity of prostate tissue and prostate cancer tissue or to possible contamination of the normal samples with pre-cancerous or undiagnosed cancers, we could not assess. Finally, we find that these methods can be applied to RNA-seq data, although the novel method appears to give better estimates. But, unfortunately, using our own criteria for the pilot sample size requirements, publicly available human RNA-seq data sets are inadequate. This information supports the notion that there is a critical need to make RNA-seq data more widely accessible to researchers so that they can plan their studies properly. We hope this information will move policy makers to place a priority in finding solutions to the existing privacy concerns with these data sets.

The paper is organized as follows: Section 2 presents the methodology. Section 3 presents the results of simulation studies. Section 4 presents the results of real data analysis and resampling studies. Section 5 presents discussion and conclusions.

2. Methods.

2.1. *The penalized logistic regression model.* Each individual in a population $\mathcal{P}$ belongs to one of two classes, $\mathcal{C}_0$ and $\mathcal{C}_1$. For individual i, let $Y_i = 0$ if $i \in \mathcal{C}_0$ and $Y_i = 1$ if $i \in \mathcal{C}_1$. One wants to predict Y_i based on observed high-dimensional data $g_i \in \Re^p$ and clinical covariates $z_i \in \Re^q$. A widely used model for this setting is the linear logistic regression model,

$$\pi(g_i, z_i) = P(Y_i = 1 | g_i, z_i) = \{1 + \operatorname{Exp}[-\alpha - \delta' z_i - \gamma' g_i]\}^{-1}, \tag{2.1}$$

where $\alpha \in \Re^1$, $\delta \in \Re^q$ and $\gamma \in \Re^p$ are population parameters.

The negative log-likelihood, given observed data (y_i, z_i, g_i) for $i = 1, \ldots, n$, is

$$L(\alpha, \delta, \gamma) = -\sum_{i=1}^{n} \{ y_i \ln[\pi(g_i, z_i)] + (1 - y_i) \ln[1 - \pi(g_i, z_i)] \}.$$

To estimate parameters and reduce the dimension of g_i, a regularized regression is often fit. Coefficients are set to zero using the penalized negative log-likelihood function

$$L_{\text{penalized}}(\alpha, \delta, \gamma) = L(\alpha, \delta, \gamma) + \sum_{k=1}^{p} \lambda_k f(\gamma_k), \tag{2.2}$$

where λ_k are penalty parameters and f is a loss function. If $f(\gamma_k) = |\gamma_k|$ and $\lambda_k \equiv \lambda > 0$, then the result is lasso logistic regression [Tibshirani (1996)]. The first step of the lasso is to estimate the penalty parameter λ, which is typically done by cross-validation. The clinical covariates z_i are not part of the feature selection process in equation (2.2), but they can be added to that process if desired. The regularized regression estimates are the solutions to

$$(\hat{\alpha}, \hat{\delta}, \hat{\gamma}) = \min_{\alpha,\delta,\gamma} L_{\text{penalized}}(\alpha, \delta, \gamma). \tag{2.3}$$

The minimum can be found by the coordinate descent algorithm [Friedman et al. (2008)].

2.2. *Predicted classification scores.* Consider a training set and independent validation set. The training set is

$$T_j = \{(y_1, z_1, g_1), \ldots, (y_n, z_n, g_n)\}$$

and the validation set is

$$V_k = \{(Y_1^v, z_1^v, g_1^v), \ldots, (Y_m^v, z_m^v, g_m^v)\}.$$

The minimization in equation (2.3) based on the data set T_j produces estimates $(\hat{\alpha}_j, \hat{\delta}_j, \hat{\gamma}_j)$. The model is applied to the validation set V_k, resulting in estimated scores

$$\{\hat{\alpha}_j + \hat{\delta}_j' z_i^v + \hat{\gamma}_j' g_i^v\}_{i=1}^m.$$

Let $W_{ij}^u = \hat{\gamma}_j' g_i^v$ be the high-dimensional part of the predicted classification score for individual i in the validation set. That is, the g_i^v is data from a high-dimensional technology, such as RNA-seq expression measurements. Let $X_i^u = \gamma' g_i^v$ and note that we can write $W_{ij}^u = X_i^u + U_{ij}^u$, where $U_{ij}^u = (\hat{\gamma}_j - \gamma)' g_i^v$. (The u superscripts denote unstandardized variables, in contrast to standardized versions presented below.) The model of equation (2.1) can be written in the form

$$P(Y_i^v = 1 | z_i, g_i) = \{1 + \text{Exp}[-\alpha - \delta' z_i^v - X_i^u]\}^{-1}.$$

Note that, unlike the standard logistic regression model, the variable X_i^u does not have a slope parameter multiple. We could develop the model in its present form, but it will simplify presentation if we make it look more like the standard model.

Define $\mu_x = E_{\mathcal{P}}[X_i^u] = \int \gamma' g f(g)\, d\mu$ as the mean of the $\gamma' g_i$ taken across the target population $\mathcal{P}$, where the high-dimensional vectors have density f with respect to a measure μ. Similarly, define $\sigma_x^2 = \text{Var}_{\mathcal{P}}(X_i^u)$. If these exist, then we can standardize the scores

$$X_i = \frac{X_i^u - \mu_x}{\sigma_x}, \qquad U_{ij} = \frac{U_{ij}^u}{\sigma_x}, \qquad W_{ij} = \frac{W_{ij}^u - \mu_x}{\sigma_x} = X_i + U_{ij}, \tag{2.4}$$

resulting in the EIV logistic regression model

$$P(Y_i^v = 1|z_i, X_i) = \{1 + \mathrm{Exp}[-\alpha - \delta' z_i^v - \mu_x - \sigma_x X_i]\}^{-1}$$
$$= \{1 + \mathrm{Exp}[-\alpha_x - \delta' z_i^v - \beta_\infty X_i]\}^{-1},$$

where $\alpha_x = \alpha + \mu_x$, $\beta_\infty = \sigma_x$. Note that $E_{\mathcal{P}}[X_i] = 0$ and $\mathrm{Var}_{\mathcal{P}}(X_i) = 1$. This is the EIV model of Carroll et al. (2006). With these adjustments, we can apply EIV methods in a straightforward way.

Suppose we repeatedly draw training sets T_t at random from the population $\mathcal{P}$, resulting in $T_1, T_2, \ldots$. Each time we apply the developed predictor to the validation set V_k. Each application produces an estimated covariate value vector $\hat{X}_t = W_t$ of length m and corresponding vector of error values U_t, where $U_t = (U_{1t}, \ldots, U_{mt})' = W_t - X_t$. Define $E_n[U_t] = \lim_{t_0\to\infty} \frac{1}{t_0} \sum_{t=1}^{t_0} U_t$ and $\mathrm{Var}_n(U_t) = \lim_{t_0\to\infty} \frac{1}{t_0-1} \sum_{t=1}^{t_0} (U_t - E_n[U_t])(U_t - E_n[U_t])'$, that is, these are the expectation and variance taken across training samples of size n in the population. The derivation of the conditional score method is based on an assumption that the U_t are independent and identically distributed Gaussian with $E_n[U_t] = 0$ and $\mathrm{Var}_n(U_t) = \Sigma_{uu}$, where Σ_{uu} is a positive definite matrix. This assumption can be divided into three component parts:

(1) The $E_n[U_{ij}|g_i] = 0$ for $i = 1, \ldots, m$. Equivalently, $E_n[W_{ij}|g_i] = X_i$, so that the estimated values are unbiased estimates of the population values. Intuitively, if n_{train} is large enough to develop a good classifier, then this assumption should be approximately true. However, if n_{train} is much too small, then the estimated scores may be more or less random and not centered at the true values—so that this assumption would be violated. But the assumption is required for identifiability [Dobbin and Song (2013)]. This shows that some model violation may be expected for our approach as the sample size gets small.

(2) The U_{ij} have finite variance. This would be true if $g_i' \mathrm{Var}_n(\hat{\gamma}_j) g_i < \infty$ for each i. So, if the regularized linear predictor $\hat{\gamma}_j$ has finite second moments for training samples of size n, the condition would be satisfied.

(3) The vector $(U_{1j}, \ldots, U_{mj})$ is multivariate normal. This means that given $G_{\mathrm{mat}} = (g_1, \ldots, g_m)$, $(\hat{\gamma}_j - \gamma)' G_{\mathrm{mat}}$ is multivariate normal. This would be true if $\hat{\gamma}_j$ were multivariate normal, and may be approximately true if conditions under which $\hat{\gamma}_j$ converges to a normal distribution are satisfied [e.g., Bühlmann and van de Geer (2011)].

To further simplify the model, we assume $\mathrm{Var}(U_j) = \sigma_n^2 R_n$ where R_n is a correlation matrix; in other words, we assume the prediction error variance is the same for each individual i.

2.3. *Defining the objective.* Define β_j as the slope (associated with the W_{ij}) from fitting a logistic regression of Y_i on (z_i, W_{ij}) across the entire population $\mathcal{P}$.

In other words, β_j is the true slope from a logistic regression that uses the training-set-derived W_{ij} as predictors; note that there is one well-defined β_j for a particular training set. The tolerance is then

$$\mathrm{Tol}(n) = \left|\beta_\infty - E_n[\beta_j]\right|.$$

Under regularity conditions the tolerance will be finite and $|E_n[\beta_j]| < |\beta_\infty|$, and $\lim_{n\to\infty} \mathrm{Tol}(n) = 0$ [Supplement, Section 5.1, Safo, Song and Dobbin (2015)]. Note that it is possible to have $|E_n[\beta_j]| > \beta_\infty$ in logistic EIV [Stefanski and Carroll (1987)]. Let t_{target} be the targeted tolerance. The targeted sample size n_{target} is the solution to

$$n_{\text{target}} = \min\{n \mid \mathrm{Tol}(n) \le t_{\text{target}}\}.$$

2.4. *Estimation.* Resampling is used to search for n_{target} nonparametrically. This section outlines each step in the estimation process. More detailed descriptions appear in the Supplement [Safo, Song and Dobbin (2015)].

2.4.1. *Estimation for the full pilot data set.* Let n_{pilot} be the size of the pilot data set. The parameter $\beta_{n_{\text{pilot}}} = E_{n_{\text{pilot}}}[\beta_j]$ defined in Section 2.3 can be estimated by cross-validation [e.g., Geisser (1993)]. Regularized logistic regression requires specification of a penalty parameter [λ in equation (2.2)]. Selecting this penalty parameter once using the whole data set results in biased estimates of predicted classification performance [Ambroise and McLachlan (2002), Simon et al. (2003)]. Therefore, a nested (double) cross-validation is required [see, e.g., Davison and Hinkley (1997)]. An inner loop is used to select the penalty parameter λ; then that penalty parameter is used in the outer loops to obtain the cross-validated classification scores. Because the split of the data set into 5 subsets may impact the resulting nested CV slope estimate, we suggest the RCS-CV method; RCS-CV is defined as repeating the cross-validation 20–50 times, centering and scaling each cross-validated batch, and using the mean of these 20–50 cross-validated slopes as the estimate. Centering and scaling of the cross-validated batches is needed to reduce error variance due to instability in the lasso regression parameter estimates (not shown). We recommend 5-fold cross-validation.

The cross-validated scores provide an estimate of the slope for a training sample of size n_{pilot}, which we can denote $\hat{\beta}_{n_{\text{pilot}}}$. We want to apply errors-in-variables regression to estimate the tolerance, $\mathrm{Tol}(n_{\text{pilot}})$, and for that we also need an estimate of the error variance, $\sigma^2_{n_{\text{pilot}}} = \mathrm{Var}_{n_{\text{pilot}}}(U_{ij})$. The leave-one-out bootstrap [e.g., Efron and Tibshirani (1997)] can be used to estimate $\sigma^2_{n_{\text{pilot}}}$. Because tuning parameters must be selected in regularized regression, a nested, case-cross-validated leave-one-out bootstrap (LOO-NCCV-BS) is required [see, e.g., Varma and Simon (2006)]. Letting $W_{ij,bs}$ represent these bootstrap scores for $i = 1, \ldots, n_{\text{pilot}}$ and $j = 1, \ldots, b_0$, where b_0 is the number of bootstraps for each left-out case, then the

estimate of $\sigma^2_{n_{\text{pilot}}}$ is

$$\hat{\sigma}^2_{n_{\text{pilot}}} = \frac{1}{n_{\text{pilot}}(b_0 - 1)} \sum_{i=1}^{n_{\text{pilot}}} \sum_{j=1}^{b_0} (W_{ij,bs} - \overline{W}_{i,\cdot,bs})^2,$$

where $\overline{W}_{i,\cdot,bs} = \frac{1}{b_0} \sum_{j=1}^{b_0} W_{ij,bs}$. As with the CV described in the previous paragraph, one needs to standardize the cross-validated bootstrap batches to have mean zero and variance 1. This is the CS-LOO-NCCV-BS procedure. Note that in practice the leave-one-out bootstrap is performed using a single bootstrap and collating the results appropriately, which reduces the computation cost [Davison and Hinckley (1997)].

Now the $\hat{\sigma}^2_{n_{\text{pilot}}}$ is "plugged in" to a univariate EIV logistic regression which also uses the nested CV predicted classification scores as the W_{ij} in equation (2.4). The conditional score method of Stefanski and Carroll (1987), with the Hanfelt and Liang (1997), equation (3), modification, is used to estimate the asymptotic slope β_∞ associated with the X_i. Briefly, if we write the logistic density of equation (2.3) in the canonical generalized linear model form

$$f(y_i) = \text{Exp}\left\{ \frac{y_i(\alpha + \delta' z_i + \beta_\infty X_i) - b(\alpha + \delta' z_i + \beta_\infty X_i)}{a(\phi)} + c(y_i, \phi) \right\},$$

where the functions are $a(\phi) = 1, b(x) = \ln(x), c(y_i, \phi) = 0$, then letting $\theta = (\alpha, \delta, \beta_\infty)'$ the conditional score function for θ has the form

$$\sum_i \begin{pmatrix} (y_i - E[y_i|A_{\theta i}]) \\ z_i(y_i - E[y_i|A_{\theta i}]) \\ \tilde{x}_i(y_i - E[y_i|A_{\theta i}]) \end{pmatrix},$$

where $A_{\theta_i} = W_{ij} + y_i \Psi \beta_\infty$, $\tilde{x}_i$ is an estimator of X_i based A_{θ_i}, and $\Psi = \text{Var}_{n_{\text{pilot}}}(U_{ij}) a(\phi)$.

The conditional score method produces $\hat{\beta}_\infty$. The tolerance is then estimated with $\widehat{\text{Tol}}(n_{\text{pilot}}) = |\hat{\beta}_\infty - \hat{\beta}_{n_{\text{pilot}}}|$.

2.4.2. *Estimating tolerance for subsets of the pilot data set.* Typically, $\widehat{\text{Tol}}(n_{\text{pilot}})$ will be larger or smaller than t_{target}, the targeted tolerance. In either case, more information about the relationship between $\text{Tol}(n)$ and n is needed to estimate n_{target}. Such information can be obtained by subsampling from the pilot data set. We suggest 7 subsets with a range of sizes be taken from the pilot data set. Each subset should be large enough, as defined in Section 2.5 below. For example, $n_{\text{pilot}} \times k/7$ for $k = 1, \ldots, 7$ can be used. More typically, if the pilot data set is not as large, then one may use $(n_{\text{pilot}}/2) + k/6 * (n_{\text{pilot}}/2)$ for $k = 0, \ldots, 6$. If $n_{\text{pilot}}/2$ is not large enough, then the pilot set is probably inadequate.

For each subset size less than n_{pilot}, call them $n_1^*, \ldots, n_6^*$, take a random sample from the full data set without replacement. Then apply the procedure described

for the full pilot data set to each subset and obtain $\widehat{\text{Tol}}(n_k^*)$, $k = 1, \ldots, 6$. The only modification we suggest to the original RCS-CV procedure is that for the 20–50 repetitions, take a different random sample each time.

2.4.3. *Estimation of* $\hat{n}_{\text{target}}$. Analysis of a pilot or simulated data set produces sample sizes $n_1^* < n_2^* < \cdots < n_7^*$ and corresponding tolerance estimates $\hat{t}_1 = \widehat{\text{Tol}}(n_1^*), \ldots, \hat{t}_7 = \widehat{\text{Tol}}(n_7^*)$. Fit the Box–Cox regression model $(\hat{t}_i^{\kappa} - 1)/\kappa = \delta_0 + \delta_1 n_i^* + \varepsilon_i$ to obtain $\hat{\kappa}$, and define $\hat{h}(x) = (x^{\hat{\kappa}} - 1)/\hat{\kappa}, \hat{\kappa} \neq 0$ and $\hat{h}(x) = \text{Ln}(x), \hat{\kappa} = 0$. Then fit with least squares $n_i^* = \eta + \zeta \hat{h}(\hat{t}_i)$, which produces $\hat{\eta}$ and $\hat{\zeta}$. Finally, if $t_{\text{target}} > 0$ is the desired tolerance, the sample size is

$$\hat{n}_{\text{target}} = \hat{\eta} + \hat{\zeta}\hat{h}(t_{\text{target}}).$$

As discussed above, we recommend estimating each tolerance 20–50 times by repeated random sampling, say, $\hat{t}_{1,1}, \ldots, \hat{t}_{1,20}$ and estimating $\hat{t}_1 = \frac{1}{20}\sum_{i=1}^{20} \hat{t}_{1,i}$. Note that the Box–Cox method requires that only values of $t_{\text{target}} > 0$ be considered. Also, in our modeling context, $t_{\text{target}} \leq 0$ does not make intuitive sense since the tolerance should always be positive. Theoretically, the sample size estimate should be ∞ when $t_{\text{target}} = 0$. But, depending on the estimates $\hat{\eta}, \hat{\zeta}, \hat{\kappa}$, this may not be the case. Our advice is that the estimated sample sizes should increase as t_{target} decreases over the range of sample sizes considered.

2.5. *Are there enough samples in the pilot data set?* The nested resampling methods in our approach require there be adequate numbers in the subsets. If there are n_{pilot} in the pilot data set, then a bootstrap sample will contain on average $0.632 \times n_{\text{pilot}}$ unique samples. A 5-fold case cross-validation of the bootstrap sample will result in $0.2 \times 0.632 \times n_{\text{pilot}} = 0.13 \times n_{\text{pilot}}$ in a validation set. Since the validation set scores will be normalized to have mean zero and variance 1, we recommend at least 80 samples in the training set to ensure at least 10 samples in these cross-validated sets. If the class prevalence is imbalanced, this number should be increased. In particular, we recommend the following:

CONDITION 1. If n_{pilot} is the size of the pilot data set, then $n_{\text{pilot}} \times 0.13 \times \pi_{\text{lowest}} \geq 5$, where π_{lowest} is the proportion from the underrepresented class.

The conditional score methods will not work well as the error variance gets large. Since the conditional score methods are repeated 20–50 times for each subset size in the RCS-CV procedure, the stability of these estimates can be evaluated. Therefore, the following guideline is advised:

PRACTICAL GUIDELINE 1. If the conditional score errors-in-variables regression estimates display instability for any subsample size, use quadratic SIMEX errors-in-variables regression instead. An example of instability would be $|\text{mean}(\hat{\beta}_\infty)/\text{s.d.}(\hat{\beta}_\infty)| < 0.5$, where the mean and standard deviation are taken across the 20–50 replicates.

Resampling-based approaches to sample size estimation require that the relationship between the asymptotic model and the estimated model can be adequately estimated from the pilot data set. Trouble can arise if the learning pattern displayed on the pilot set changes dramatically for sample sizes larger than the pilot data set. For example, there may be no classification signal detectable with 3 samples per class, but one is detectable with 50 samples per class. So a pilot data set of 6 would lead to the erroneous conclusion that the asymptotic error rate is 50%, and any resulting sample size estimates would likely be erroneous. Similarly, the learning process can be uneven, so that the asymptotic error rate estimate increases or decreases as the sample size increases. The latter can happen when some subset of the features have smaller effects than others and are only detected for larger sample sizes. To guard against this in simulations, at least, we found that the following guideline is useful:

CONDITION 2. The predictor needs to find the important features related to the class distinction with power at least 85%.

Our simulation-based software program checks the empirical power for this condition. In the context of resampling from real data, it is not clear how one could verify this assumption empirically. But it may be possible to evaluate the effect size associated with this power by a parametric bootstrap.

2.6. *Translating between logistic slope and misclassification accuracy.* If there are no clinical covariates in the model, then the misclassification error rate for the asymptotic model is [e.g., Efron (1975)]

$$\begin{aligned} &P(Y_i = 1 \text{ and } \alpha + \beta_\infty X_i \le 0) + P(Y_i = 0 \text{ and } \alpha + \beta_\infty X_i > 0) \\ &\quad = \int_{-\infty}^{-\alpha/\beta_\infty} \frac{e^{\alpha+\beta_\infty X_i}}{1+e^{\alpha+\beta_\infty X_i}} f_x(x)\,dx + \int_{-\alpha/\beta_\infty}^{\infty} \frac{1}{1+e^{\alpha+\beta_\infty X_i}} f_x(x)\,dx, \end{aligned}$$

where $f_x(x)$ is the marginal density of the asymptotic scores across the population $\mathcal{P}$. By definition, these scores have mean zero and variance one. If we further assume the scores are Gaussian, then the misclassification rate can be estimated with

$$\sum_{i=1}^{m_0}\left[\frac{e^{\alpha+\beta_\infty x_i}}{1+e^{\alpha+\beta_\infty x_i}}\right]1(x_i \le -\alpha/\beta_\infty) + \sum_{i=1}^{m_0}\left[\frac{1}{1+e^{\alpha+\beta_\infty x_i}}\right]1(x_i > -\alpha/\beta_\infty),$$

where $1(A)$ is the indicator function for event A and m_0 is a number of Monte Carlo simulations, and $x_1, \ldots, x_{m_0}$ are drawn from the distribution $x_i \sim \text{Normal}(0, 1)$. If covariates are added to the model, then the conditional distribution of $x_i|z_i$ needs to be used for the Monte Carlo. If x_i is independent of z_i, then the x_i could be generated from a standard normal, and the Monte Carlo equations modified in the obvious way. The Supplement [Safo, Song and Dobbin (2015)] shows a graph of the no covariate case relationship.

3. Results: Simulation studies. For the simulation studies, high-dimensional data were generated from multivariate normal distributions, both a single multivariate normal and a mixture multivariate normal with homoscedastic variance. Both multivariate normal settings performed similarly [see Supplement, Safo, Song and Dobbin (2015)], so we just present one in the paper. The covariance matrices were identity, compound symmetric (CS) and autoregressive order 1 [AR(1)], as indicated. Class labels were generated from the linear logistic regression model of equation (2.1). Categorical clinical covariate data, when included in simulations, were generated from a distribution with equal probability assigned to each of three categories, where categories are correlated with class labels.

The asymptotic slope parameter β_∞ must be estimated. Table 1 presents a simulation to evaluate the bias and variance of the asymptotic slope parameter estimate $\hat{\beta}_\infty$. Also presented are the corresponding estimates of asymptotic classification accuracy $\widehat{\text{acc}}_\infty$. As can be seen from the table, this approach does well overall at estimating the asymptotic performance for these pilot data set sample sizes (300 and 400), asymptotic slopes $(2, 3, 4, 5)$, multivariate normal high-dimensional data, covariance matrix structures (Identity, CS [Supplement, Safo, Song and Dobbin

TABLE 1

*Estimates of the asymptotic slope β_∞ and corresponding accuracy acc_∞ evaluated by simulations. n_{pilot} is the number of samples in the pilot data set. The covariance structure "Cov" are as follows: AR*1 *is block autoregressive order* 1 *in* 3 *blocks of size* 3 (9 *informative features*) *with parameter* 0.7; *Iden. is identity with* 1 *block of* 1 (1 *informative feature*). *Total of* $p = 500$ *features*; *all noise features independent standard normal. Summary statistics based on* 200 *Monte Carlo. More results appear in the Supplement* [*Safo, Song and Dobbin* (2015)]

		Logist. slope		Class. Acc.			
n_{pilot}	Cov	β_∞	mean $\hat{\beta}_\infty$	acc_∞	$\widehat{\text{acc}}_\infty$ mean	mean $\hat{\sigma}_n^2$	mean $\hat{\beta}_n$
300	AR1	2.0	2.07	0.778	0.783	0.43	1.49
400	AR1	2.0	2.01	0.778	0.779	0.35	1.62
300	AR1	3.0	3.04	0.836	0.838	0.32	2.31
400	AR1	3.0	2.93	0.836	0.834	0.27	2.47
300	AR1	4.0	3.95	0.871	0.869	0.28	3.06
400	AR1	4.0	3.88	0.871	0.868	0.23	3.26
300	AR1	5.0	3.77	0.894	0.865	0.25	3.71
400	AR1	5.0	4.81	0.894	0.891	0.21	3.99
300	Iden.	2.0	2.05	0.778	0.781	0.23	1.87
400	Iden.	2.0	2.01	0.778	0.778	0.19	1.90
300	Iden.	3.0	3.02	0.836	0.836	0.17	2.85
400	Iden.	3.0	2.98	0.836	0.835	0.14	2.87
300	Iden.	4.0	3.97	0.871	0.870	0.14	3.72
400	Iden.	4.0	3.92	0.871	0.869	0.12	3.75
300	Iden.	5.0	4.94	0.894	0.893	0.14	4.50
400	Iden.	5.0	4.86	0.894	0.891	0.12	4.55

(2015)] and AR1) and numbers of informative features (1 and 9). There is some small bias apparent as the slope becomes large ($\beta_\infty = 5$), probably reflecting the fact that large slopes are problematic for EIV logistic regression.

The tolerance associated with the estimated sample size should be within the user-targeted tolerance. To test this, sample sizes were calculated by applying the method to simulated pilot data sets. Then, these sample size estimates were assessed by performing very large pure Monte Carlo studies. Table 2 presents sample size estimates from our method and sample statistics from the Monte Carlo (MC) simulations. The mean tolerances from the MC are all within the targeted tolerance, indicating that the estimated sample sizes do achieve the targeted tolerance. The method tends to produce larger sample size estimates than required with 62%–93% of the true tolerances within the target (rightmost column). Note that our method guarantees that the expected slope is within the tolerance, but not that the actual slope is within the tolerance; this latter would be a stronger requirement.

Implementation of our approach in the presence of clinical covariates was evaluated. Table 3 shows results when a clinical covariate is included into the setting. In this case the clinical covariate is also associated with the class distinction; in particular, in equation (2.1), $\delta = \mathrm{Ln}(2)$ and $z_i \in \{-1, 0, 1\}$, with 1/3rd probability assigned to each value. As can be seen by comparison with Table 2, the addition of the clinical covariate significantly increases the required sample sizes. For exam-

TABLE 2

Evaluation of the sample size estimates from AR(1) and identity covariances. The number in the pilot data set is 400. $\beta_\infty = 4$. *Identity covariance had one informative feature, and AR(1) had nine informative features in a block structure of* 3 *blocks of size* 3 *with correlation parameter* 0.7. *Estimates evaluated using* 400 *Monte Carlo simulations with the estimated sample size. The mean tolerance from the* 400 *simulations and the proportion of the* 400 *within the specified tolerance are given in the rightmost two columns. The dimension is* $p = 500$

Cov.	t_{target}	$\hat{n}$	Mean MC tol.	% of MC within tol.
AR1	0.10	1742	0.09	64%
AR1	0.20	986	0.19	62%
AR1	0.30	715	0.27	67%
AR1	0.40	573	0.34	71%
AR1	0.50	484	0.43	72%
AR1	0.60	424	0.49	75%
AR1	0.70	380	0.57	77%
Identity	0.10	509	0.09	79%
Identity	0.20	322	0.10	87%
Identity	0.30	242	0.15	87%
Identity	0.40	194	0.16	92%
Identity	0.50	162	0.21	90%
Identity	0.60	139	0.25	93%
Identity	0.70	121	0.31	91%

TABLE 3

Clinical covariate simulations. One clinical covariate with 3 *levels which are associated with the class distinction. The identity covariance and an asymptotic true slope of* $\beta_\infty = 4$. *The dimension is* $p = 500$. *See text for more information*

t_{target}	$\hat{n}$	Mean MC tol.	% of MC within tol.
0.1	592	0.07	80%
0.2	416	0.09	88%
0.3	334	0.12	92%
0.4	284	0.14	91%
0.5	249	0.15	95%
0.6	223	0.17	92%

ple, the estimated sample size for a tolerance of 0.20 increases 29%, from 322 to 416. This increase reflects correlation between the clinical covariate and the class labels. The pure Monte Carlo evaluations in Table 2 show that the method does still produce adequate sample size estimates in the presence of the clinical covariate.

Figure 1 is a summarization of results from all the different simulation studies. Negative values on the y-axis mean the sample size was overestimated, and positive values mean the sample size was underestimated. As can be seen in the

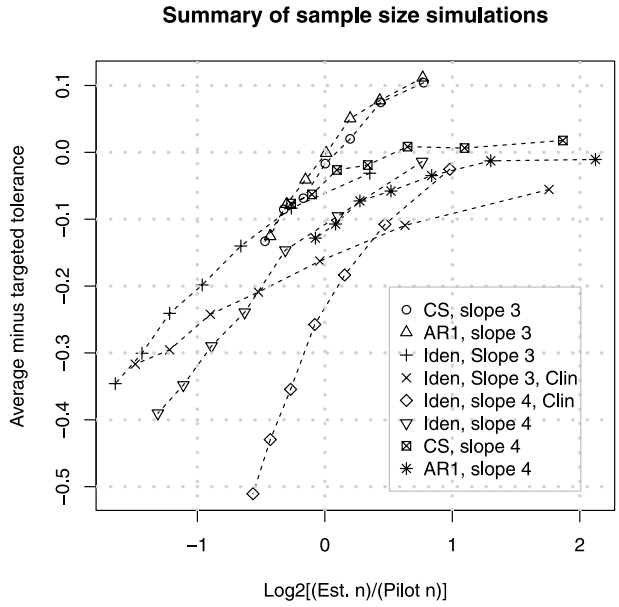

FIG. 1. *Summary of results of simulations. The* x*-axis is the base* 2 *logarithm of the ratio of the estimated training sample size required divided by the pilot training sample size used. The* y*-axis is the average tolerance estimated from pure Monte Carlo simulations minus the targeted tolerance.*

figure, the sample size estimates are mostly adequate or conservative. When the estimated sample size required is smaller than the pilot data set (x-axis values are negative), the resulting tolerance estimates are adequate or conservative; intuitively, identifying a sample size smaller than the pilot data set should be relatively easy. When the estimated sample size required is larger than the pilot data set, the method continues to perform well overall. The exceptions are in the cases of compound symmetric and AR1 covariance with a small slope of 3; in these cases, the y-values are positive, indicating anti-conservative sample size estimates. The problem here seems to be the power to detect the features. For the compound symmetric simulations, the empirical bootstrap power was $7.67/9 = 85.2\%$, and for AR1 simulations, the power was 84.7%. Both are near the cutoff of the 85% power criterion developed in Section 2.5 above. Still, overall, the method seems to perform well.

3.1. *LC and EIV performance in simulations.* We evaluated both our resampling-based method and the resampling-based method of Mukherjee et al. (2003) using pure Monte Carlo estimation of the truth in simulations. We will denote their method by LC (for learning curve) and our method by EIV (for errors-in-variables). Figure 2 shows a comparison of the two methods under a range of simulation settings. In these simulations, our method may have an advantage be-

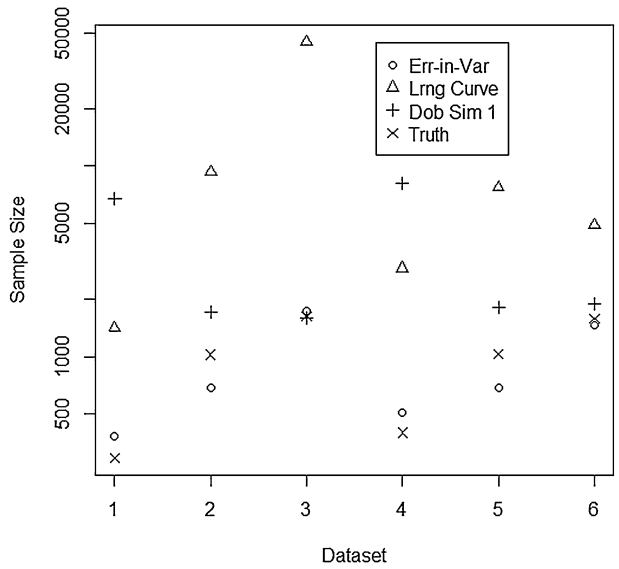

FIG. 2. *Sample size estimates for* 6 *simulated data sets with tolerance set to* 0.1. *Comparison of errors-in-variables*, *learning curve*, *Dobbin and Simon* (2007) (*with false positive rate set to at most* 1) *and the true values* (*from pure Monte Carlo*). *Similar results for tolerance* $= 0.2$ *appear in the Supplement* [*Safo*, *Song and Dobbin* (2015)]. *Y-axis is on the log scale*. *Size of pilot data set is* 300.

cause the logistic regression model was used to generate the response data. Tolerances of 0.1 and 0.2 were considered since these are associated with larger training sample sizes than the pilot data set. Comparing the percentage error of the sample size estimate to an estimate based on pure Monte Carlo, one can see that the learning curve method has an error an order of magnitude or more larger than our method. The LC method tends to consistently overestimate the sample size in these simulations. In sum, the EIV method estimates were closer to the true sample size values than the LC method estimates across all of these simulations.

Finally, we applied the method of Dobbin and Simon (2007) to the same data sets. As background, this method assumes a prespecified significance level cutoff for features, hence, it is not well-suited to the lasso logistic regression. Some ad hoc procedure is needed to approximate the lasso. We examined an approach that (1) picks the optimal significance level cutoff (likely to be anti-conservatively biased because of data re-use), and one that (2) picks a significance level that controls the expected number of false positives to be at most 1 (in other words, p-value cutoff $1/n$). The sample size estimates for approach (1) were 126, 404, 366, 116, 404 and 374, for data sets 1–6, respectively. Perhaps not surprisingly, this approach underestimates the sample size requirement for lasso logistic regression. The sample size estimates for approach (2), using a grid of size 100, were 6700, 1700, 1600, 8100, 1800 and 1900. The results are similar to those for the learning curve.

4. Real data set applications. The methods are applied to four data sets. The purpose of the first three applications is two-fold. First, we determine the adequacy of the sample sizes of these studies using both sample size methods. Second, since microarray data may violate the normality assumption of the simulations, we evaluate the resampling performance of the methods as a check on their performance with this nonnormal data. The purpose of the fourth application is primarily to evaluate the ability of the methods to estimate sample size on RNA-seq data.

A resampling study can be used to compare estimates from a procedure to a resampling-based "truth." Since adequate sample sizes are unknown on these data sets, it was not feasible to compare sample size estimates to any corresponding estimated true values. But we can compare the error rate estimated from a subset of the data set to an independent estimate based on cross-validation on the whole data set.

4.1. *The lymphoma data set*. We applied the EIV and the LC method to the data set of Rosenwald et al. (2002). The classes were germinal center B-cell lymphoma versus all other types of lymphoma. We used both methods to evaluate whether the sample size used in this study was adequate, and we subsetted 5 "pilot data sets" of size 100 at random from this data set. For each of these "pilot data sets," we estimated the performance when $n = 240$ are in the training set. Then, we could compare the estimated performance to a "gold standard" resampling-based performance on the full data set. Results are shown in Table 4. The two methods

TABLE 4

Resampling studies. Data set is the data set used for resampling. Rep is the replication number of 5 *independent random subsamples (without replacement) of size nPilot. nFull is the size of the full data set. Classes for the Shedden data set were Alive versus Dead. Classes for the Rosenwald data set were Germinal-Center B-Cell lymphoma type versus all others. err(nFull) is estimated from* 200 (50 *for Shedden) random cross-validation estimations on the full data set using different partitions each time, and this serves as the gold standard error rate for nFull.* $\widehat{err}(nFull)$ *is the estimated error rate for the full data set based on the LC method or EIV method. Similarly,* $\widehat{err}(\infty)$ *is the asymptotic error rate based on the LC method or EIV method. The first column is the dataset, "R" for Rosenwald and "S" for Shedden. For the Shedden data set, we used conditional score EIV; for the Rosenwald data set, we used quadratic SIMEX EIV because the criterion for the conditional score was violated (Section 2.5)*

					LC method		**EIV method**		**nFull err %**	
Data	**Rep**	**nPilot**	**nFull**	**err(nFull)**	$\widehat{err}(nFull)$	$\widehat{err}(\infty)$	$\widehat{err}(nFull)$	$\widehat{err}(\infty)$	**LC**	**EIV**
R	1	100	240	0.1129	0.0855	0.0729	0.1344	0.1135	−25%	19%
R	2	100	240	0.1129	0.0611	0.0435	0.1078	0.0933	−46%	−5%
R	3	100	240	0.1129	0.0298	0.0089	0.0771	0.0691	−74%	−32%
R	4	100	240	0.1129	0.1443	0.1270	0.1396	0.1379	28%	24%
R	5	100	240	0.1129	0.0682	0.0480	0.0864	0.0783	−40%	−23%
Mean				0.1129	0.0778	0.0601	0.1091	0.0984	−31%	−3%
S	1	200	443	0.4207	0.4638	0.4634	0.4347	0.4347	10%	3%
S	2	200	443	0.4207	0.4496	0.4481	0.4154	0.4151	7%	−1%
S	3	200	443	0.4207	0.4300	0.4258	0.2778	0.2778	2%	−34%
S	4	200	443	0.4207	0.4166	0.4126	0.3550	0.3550	−1%	−16%
S	5	200	443	0.4207	0.4159	0.4117	0.2907	0.2894	−1%	−31%
Mean				0.4207	0.4352	0.4323	0.3548	0.3544	3%	−16%

agree well on this data set, and both indicate that on average the pilot data set size of $n = 240$ provides close to the optimal accuracy possible, within 0.02. Comparing the two approaches in their ability to estimate the error rate for $n = 240$ based on a pilot data set of only $n = 100$, the EIV method has better mean performance in terms of estimating the full data set error than the LC method. Here, the differences are less dramatic than the sample size differences; this may be due to the sensitivity of sample size methods to relatively small changes in asymptotic error rates or to the underlying data distribution. Both methods show some variation in error rate estimates across the five subsets of size 100.

4.2. *The lung cancer data set.* We next applied both methods to the lung cancer data set of Shedden et al. (2008), where the classes were based on survival status at last follow-up. This binary indicator approach was used for predictors developed in the original paper. In this case, the two methods produce similar results. Both methods indicate here that on average the asymptotic performance is extremely close to the full data set performance when $n = 443$, with a difference in

error rate of less than 0.01. Comparing the two methods in terms of their ability to estimate the error rate for $n = 443$ based on a pilot data set of only $n = 200$, the LC method was slightly better on average than the EIV with conditional score based on percentage error (rightmost columns of table); but the conditional score criterion in Section 2.5 was exceeded on 3 of the 5 data sets, and if quadratic SIMEX is used, then the LC and EIV are almost identical [Supplement, Safo, Song and Dobbin (2015)]. This is a very noisy problem and classification accuracy based on a training set of all 443 samples is only estimated to be around 56%–60%.

In order to evaluate the performance of our approach using modifications of the lasso, we next applied our EIV method using the elastic net [Zou and Hastie (2005)] to the lung cancer data set. In 3 of the 5 resampled data sets of size 200, the results were almost identical to the lasso; in one case (data set 5) the ratio of the estimate to its standard deviation was below the 0.5 cutoff for all sample sizes less than 183 and, not surprisingly, our method did not work in that case; in another resampled data set (data set 3), our method produced larger sample size estimates for the elastic net than the lasso. Details are shown in the Supplemental Material [Safo, Song and Dobbin (2015)].

4.3. *Prostate cancer microarray data set.* We next applied our method to the prostate cancer data set described in Dettling and Bühlmann (2003). This data set consisted of a total of 102 human samples, 52 from prostate tumors and 50 from nontumor prostate tissue samples. Intuitively, we may expect that classification of samples into tumor versus nontumor would be a relatively easy problem. For this data set, the EIV sample size estimates for tolerances of 0.10, 0.30, 0.50 were 164, 119 and 103, respectively. Interestingly, these results suggest that the sample size used (102) for the pilot study is inadequate for producing a classifier with a tolerance closer than 0.50 to the optimal value. The cross-validated misclassification rates for the prostate cancer data set reported in the (2003) paper ranged from 4.9% to 13.7%. The average of 20 resamplings from our approach resulted in an estimated $\beta_{102} = 3.55$, corresponding to an error rate of 14%. In this application, the estimated β_∞ averaged over 20 resamplings was 4.16 with standard error 0.12; the 4.16 corresponds to an error rate of 12%. Note that here the relatively large tolerance of 0.5 is associated with a small increase in accuracy because the asymptotic slope estimate is relatively large. In conclusion, our method produces results that are in accord with intuition in that the sample size used produces a classifier with accuracy close to the optimal for this easier classification problem. But we also observed large fluctuations in the β_∞ estimate when the resampled data sets were between 51 and 93, which suggests that the asymptotic estimate of a 12% error rate is quite unstable. We won't speculate as to the cause of this instability, but intuitively one would expect that a smaller error rate would be possible. Importantly, the instability of the asymptotic estimate does not seem to compromise the sample size estimate.

4.4. *The RNA-seq data set*. We performed a proof of principle study to see if these methods could be applied effectively to RNA-seq data. First, note that RNA-seq data after being processed may be in the form of counts (e.g., from the Myrna algorithm), but are more often in the form of continuous values (e.g., normalized Myrna data, or FPKM fragments-per-kilobase of exon per million fragments mapped from Cufflinks or other software). Therefore, linear models with continuous high-dimensional predictors are reasonable to use for RNA-seq data. But it is important to check that the processed data appear reasonably Gaussian and, if not, to transform the data.

We applied the LC and EIV methods to the *Drosophila melanogaster* data of Graveley et al. (2011). Processed data were downloaded from the ReCount database [Frazee, Langmead and Leek (2011)]. Variables with more than 50% missing data were removed. Remaining data were truncated below at 1.5 and log-transformed. Low variance features were filtered out, resulting in $p = 500$ dimensions. Since this was a highly controlled experiment with large biological differences between the fly states, some class distinctions resulted in separable classes. Logistic regression is not appropriate for perfectly separated data. Samples were split into two classes: Class 1 consisted of all the embryos and some adult and white prepupae (WPP); Class 2 consisted of all the larvae and a mix of adults and WPP. The class sizes were 82 and 65. A principal component plot is shown in the Supplement [Safo, Song and Dobbin (2015)]. The data set consisted of a total of $n_{\text{pilot}} = 147$ data points. Technical replicates in the data created a clustering pattern visible in principal components plots. This type of clustering is often observed in cancer patient data sets due to disease subgroupings. We did not attempt to adjust the analysis for the technical replicates. The resulting EIV method equation for the sample size was

$$\hat{n} = 105.73 - 14.25\left(\frac{t^{-0.3434} - 1}{(-0.3434)}\right).$$

For tolerances of 0.1, 0.05 and 0.02, sample size estimates were 156, 180 and 223, respectively. The cross-validated accuracy, averaged over 10 replications, was 91%. Based on the $\hat{\beta}_\infty = 4.55$, the optimal accuracy is 88.5%, and the full data set accuracy is 88%, corresponding to $\hat{\beta}_{147} = 4.42 = 4.55 - \widehat{\text{Tol}}(147)$. The conditional score was used for the EIV method. The LC method curve was $\text{err} = 0.075 + 1.252 \times n^{-0.9122321}$. The asymptotic accuracy estimate is 92.5%, corresponding to $\beta_\infty = 7.2$, and the estimated accuracy when $n = 147$ is 91.2%, corresponding to $\beta_n = 6.1$. The LC sample size estimates for tolerances of 0.10, 0.05 and 0.02 were $1,383$, $8,361$ and $13,342$, respectively. As with the simulation studies, the LC method estimates are much larger than the EIV estimates.

The difference in performance of the two methods on the RNA-seq data is probably attributable to reduced noise on this data set, resulting in a set of genes with large differences between the classes relative to the noise level present. This may

make the fly data set similar to some of the simulated data sets, where large differences were observed between LC and EIV performance. This different structure of the data compared to the microarray data sets may be due to the larger biological variability between the fly states, or a reduction in noise variation due to the RNA-seq platform, or a combination of both.

5. Discussion. In this paper we studied the problem of sample size estimation for regularized logistic regression classification in cancer. Two methods of sample size estimation were studied in simulations and applications. The simulation results suggested that the EIV method works well and that the LC method sometimes works but is sometimes overly conservative. The methods were applied to a lymphoma data set, a lung cancer data set, a prostate cancer data set and an RNA-seq data set. The results in lymphoma and lung cancer suggest that these studies had adequate sample sizes already, and that larger studies are unlikely to yield better classifiers. For the prostate cancer data set, the analysis revealed that the pilot data set size was inadequate, resulting in high variation in the predicted classification scores. The RNA-seq data analysis showed that the EIV method also appears to work well on this type of data, but a critical problem is the lack of publicly available and accessible RNA-seq data sets that could serve as pilot data sets. This observation highlights a critical existing log-jam in medical research, the problem of the lack of availability of RNA-seq data sets (or modified versions thereof) for study planning purposes. In the meantime, existing microarray data sets and/or simulated data sets must be used for RNA-seq study planning.

A new sample size method for training regularized logistic regression-based classifiers was developed. The method exploits a structural similarity between logistic prediction and errors-in-variables regression models. The method was shown to perform well when an adequate pilot data set is available. Methods for assessing the adequacy of a pilot data set were developed. If no adequate pilot data set is available, the method can be used with Monte Carlo samples from a parametric simulation.

An important issue in using either the LC or EIV method is the fitting of the curve that produces the final sample size estimate. In the LC method, as described in Mukherjee et al. (2003), a constrained least squares optimization must be performed on a nonlinear regression model. Constrained optimization methods like the L-BFGS-B algorithm used in the application of the Mukherjee et al. (2003) method in this paper may produce different solutions than standard, unconstrained least squares optimization methods such as Nelder–Mead. In contrast, the Box–Cox algorithm and linear regression fitting used by our approach are more straightforward to implement. Because our method does not need to "extrapolate to infinity" as the typical learning curve method requires, the regression model is chosen that fits the best in the vicinity of the data points. This simplifies the fitting procedure, albeit at the cost of the errors-in-variables regression step. For both methods,

it is advisable to look at the final plot of the fitted line and the data points as a basic regression diagnostic.

The reader may have noted that the variance parameter $\sigma_n^2 = \mathrm{Var}_n(U_{ij})$ is estimated by bootstrapping the pilot data set. But the variance is defined as a variance across independent training sets of size n in the population. Since the bootstrap data sets will have overlap, obviously there is potential bias in the bootstrap estimation procedure. Whether the bootstrap could be modified to reduce this bias is a potential area for future work.

If more than two classes are present in the data, then simple regularized logistic regression is no longer an appropriate analysis strategy. In order to apply our method in that setting, regularized methods for more than two classes would need to be developed, for example, regularized multinomial or ordinal logistic regression methods. Also, corresponding errors-in-variables methods for these multiclass logistic regression methods would be needed. It appears that both these would be prerequisites to such an extension.

If classes are completely separable in the high-dimensional space, then regularized logistic regression is not advisable because the logistic regression slope will be undefined and the logistic fitting algorithms will become unstable. The approach presented in this paper cannot be used in that context.

In this paper we have focused simulations on settings with equal prevalence from each class. If the class prevalences are unequal, then the method can still be applied as presented in the paper—as was done in the applications to the real data sets, for example. However, if the imbalance is large (e.g., 90% versus 10%), then the training set size required by our Condition 1 in Section 2.5 would likely be excessive. But, in this case, it is also less likely that accuracy will be the objective criteria for model selection because a high accuracy may be associated with a classifier that puts most subjects in the majority class. Ideally, positive and negative predictive values and their associated clinical implications would likely be more useful criteria. This is a potential future direction of research.

SUPPLEMENTARY MATERIAL

Supplemental tables, figures, algorithms, details and discussion (DOI: 10.1214/15-AOAS825SUPP; .pdf). Supplemental material for paper by Safo, Song and Dobbin.

REFERENCES

AMBROISE, C. and MCLACHLAN, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **14** 6562–6566.

BI, X., REXER, B., ARTEAGA, C. L., GUO, M. and MAHADEVAN-JANSEN, A. (2014). Evaluating HER2 amplification status and acquired drug resistance in breast cancer cells using Raman spectroscopy. *J. Biomed. Opt.* **19** 25001.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. Chapman & Hall/CRC, Boca Raton, FL. MR2243417

COOK, J. R. and STEFANSKI, L. A. (1994). Simulation-extrapolation estimation in parametric measurement errror models. *J. Amer. Statist. Assoc.* **89** 1314–1328.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. *Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673

DETTLING, M. and BÜHLMANN, P. (2003). Boosting for tumor classification with gene expression. *Bioinformatics* **19** 1061–1069.

DOBBIN, K. K. and SIMON, R. M. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* **8** 101–117.

DOBBIN, K. K. and SONG, X. (2013). Sample size requirements for training high-dimensional risk predictors. *Biostatistics* **14** 639–652.

DYRSKJØT, L. (2003). Classification of bladder cancer by microarray expression profiling: Towards a general clinical use of microarrays in cancer diagnostics. *Expert Rev. Mol. Diagn.* **3** 635–647.

EFRON, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* **70** 892–898. MR0391403

EFRON, B. and TIBSHIRANI, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *J. Amer. Statist. Assoc.* **92** 548–560. MR1467848

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FRAZEE, A. C., LANGMEAD, B. and LEEK, J. T. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics* **12** 449.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GEISSER, S. (1993). *Predictive Inference*: *An Introduction*. Chapman & Hall, New York. MR1252174

GRAVELEY, B. R., BROOKS, A. N., CARLSON, J. W., DUFF, M. O., LANDOLIN, J. M., YANG, L., ARTIERI, C. G., VAN BAREN, M. J., BOLEY, N., BOOTH, B. W., BROWN, J. B., CHERBAS, L., DAVIS, C. A., DOBIN, A., LI, R., LIN, W., MALONE, J. H., MATTIUZZO, N. R., MILLER, D., STURGILL, D., TUCH, B. B., ZALESKI, C., ZHANG, D., BLANCHETTE, M., DUDOIT, S., EADS, B., GREEN, R. E., HAMMONDS, A., JIANG, L., KAPRANOV, P., LANGTON, L., PERRIMON, N., SANDLER, J. E., WAN, K. H., WILLINGHAM, A., ZHANG, Y., ZOU, Y., ANDREWS, J., BICKEL, P. J., BRENNER, S. E., BRENT, M. R., CHERBAS, P., GINGERAS, T. R., HOSKINS, R. A., KAUFMAN, T. C., OLIVER, B. and CELNIKER, S. E. (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471** 473–479.

HANASH, S. M., BAIK, C. L. and KALLIONIEMI, O. (2011). Emerging molecular biomarkers—blood-based strategies to detect and monitor cancer. *Nat. Rev. Clin. Oncol.* **8** 142–150.

HANFELT, J. J. and LIANG, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika* **82** 461–477. MR1366274

HANFELT, J. J. and LIANG, K.-Y. (1997). Approximate likelihoods for generalized linear errors-in-variables models. *J. Roy. Statist. Soc. Ser. B* **59** 627–637. MR1452030

HUANG, Y. and WANG, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *J. Amer. Statist. Assoc.* **95** 1209–1219. MR1804244

HUANG, Y. and WANG, C. Y. (2001). Consistent functional methods for logistic regression with errors in covariates. *J. Amer. Statist. Assoc.* **96** 1469–1482. MR1946591

MCSHANE, L. M. and HAYES, D. F. (2012). Publication of tumor marker research results: The necessity for complete and transparent reporting. *J. Clin. Oncol.* **30** 4223–4232.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

MOEHLER, T. M., SECKINGER, A., HOSE, D., ANDRULIS, M., MOREAUX, J., HIELSCHER, T., WILLLHAUCK-FLECKENSTEIN, M., MERLING, A., BERTSCH, U., JAUCH, A., GOLDSCHMIDT, H., KLEIN, B. and SCHWARTZ-ALBIEZ, R. (2013). The glycome of normal and malignant plasma cells. *PLoS ONE* **8** e83719.

MUKHERJEE, S., TAMAYO, P., ROGERS, S., RIFKIN, R., ENGLE, A., CAMPBELL, C., GOLUB, T. R. and MESIROV, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *J. Comput. Biol.* **10** 119–142.

NOVICK, S. J. and STEFANSKI, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *J. Amer. Statist. Assoc.* **97** 472–481. MR1941464

PFEFFER, U., ROMEO, F., NOONAN, D. M. and ALBINI, A. (2009). Predictin of breast cancer metastasis by genomic profiling: Where do we stand? *Clin. Exp. Metastasis* **26** 547–558.

ROSENWALD, A., WRIGHT, G., CHAN, W. C., CONNORS, J. M., CAMPO, E., FISHER, R. I., GASCOYNE, R. D., MULLER-HERMELINK, H. K., SMELAND, E. B., GILTNANE, J. M., HURT, E. M., ZHAO, H., AVERETT, L., YANG, L., WILSON, W. H., JAFFE, E. S., SIMON, R., KLAUSNER, R. D., POWELL, J., DUFFEY, P. L., LONGO, D. L., GREINER, T. C., WEISENBURGER, D. D., SANGER, W. G., DAVE, B. J., LYNCH, J. C., VOSE, J., ARMITAGE, J. O., MONTSERRAT, E., LÓPEZ-GUILLERMO, A., GROGAN, T. M., MILLER, T. P., LEBLANC, M., OTT, G., KVALOY, S., DELABIE, J., HOLTE, H., KRAJCI, P., STOKKE, T. and STAUDT, L. M. (LYMPHOMA/LEUKEMIA MOLECULAR PROFILING PROJECT) (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346** 1937–1947.

SAFO, S., SONG, X. and DOBBIN, K. K. (2015). Supplement to "Sample size determination for training cancer classifiers from microarray and RNA-seq data." DOI:10.1214/15-AOAS825SUPP.

SHEDDEN, K., TAYLOR, J. M., ENKEMANN, S. A., TSAO, M. S., YEATMAN, T. J., GERALD, W. L., ESCHRICH, S., JURISICA, I., GIORDANO, T. J., MISEK, D. E., CHANG, A. C., ZHU, C. Q., STRUMPF, D., HANASH, S., SHEPHERD, F. A., DING, K., SEYMOUR, L., NAOKI, K., PENELL, N., WEIR, B., VERHAAK, R., LADD-ACOSTA, C., GOLUB, T., GRUIDL, M., SHARMA, A., SZOKE, J., ZAKOWSKI, M., RUSCH, V., KRIS, M., VIALE, A., MOTOI, N., TRAVIS, W., CONLEY, B., SESHAN, V. E., MEYERSON, M., KUICK, R., DOBBIN, K. K., LIVELY, T., JACOBSON, J. W. and BEER, D. G. (2008). Gene expression-based survival prediction in lung adenocarcinoma: A multisite, blinded validation study. *Nat. Med.* **14** 822–827.

SIMON, R. (2010). Clinical trials for predictive medicine: New challenges and paradigms. *Clin. Trials* **7** 516–524.

SIMON, R. M., RADMACHER, M. D., DOBBIN, K. K. and MCSHANE, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.* **95** 14–18.

STEFANSKI, L. A. and CARROLL, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74** 703–716. MR0919838

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

VARMA, S. and SIMON, R. M. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7** 91.

ZHANG, J. X., SONG, W., CHEN, Z. H., WEI, J. H., LIAO, Y. J., LEI, J., HU, M., CHEN, G. Z., LIAO, B., LU, J., ZHAO, H. W., CHEN, W., HE, Y. L., WANG, H. Y., XIE, D. and LUO, J. H. (2013). Prognostic and predictive value of a microRNA signature in stage II colon cancer: A microRNA expression analysis. *Lancet Oncol.* **14** 1295–1306.

ZHU, J. and HASTIE, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5** 427–443.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

ZWIENER, I., FRISCH, B. and BINDER, H. (2014). Transforming RNA-seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **8** e85150.

S. SAFO
DEPARTMENT OF STATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602
USA

X. SONG
DEPARTMENT OF EPIDEMIOLOGY
AND BIOSTATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602
USA

K. K. DOBBIN
DEPARTMENT OF STATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602
USA
AND
DEPARTMENT OF EPIDEMIOLOGY
AND BIOSTATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602
USA
E-MAIL: dobbinke@uga.edu