Published in final edited form as:

Stat Med. 2012 June 30; 31(14): 1464–1474. doi:10.1002/sim.4484.

L_1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets

K. J. Archer* and **A. A. A. Williams**

Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, U.S.A.

Abstract

Health status and outcomes are frequently measured on an ordinal scale. For high-throughput genomic datasets, the common approach to analyzing ordinal response data has been to break the problem into one or more dichotomous response analyses. This dichotomous response approach does not make use of all available data and therefore leads to loss of power and increases the number of Type I errors. Herein we describe an innovative frequentist approach that combines two statistical techniques, L_1 penalization and continuation ratio models, for modeling an ordinal response using gene expression microarray data. A simulation study was conducted to assess the performance of two computational approaches and two model selection criterion for fitting frequentist L_1 penalized continuation ratio models. Moreover, the approaches were empirically compared using three application datasets, each of which seeks to classify an ordinal class using microarray gene expression data as the predictor variables. We conclude that the L_1 penalized constrained continuation ratio model is a useful approach for modeling an ordinal response for datasets where the number of covariates (p) exceeds the sample size (n), and the decision of whether to use AIC or BIC for selecting the final model should depend upon the similarities between the pathologies underlying the disease states to be classified.

Keywords

Ordinal response; Continuation ratio; L_1 penalization; Gene expression analysis; Microarray

1. Introduction

Health status and outcomes are frequently measured on an ordinal scale. In fact, most histological variables are reported on an ordinal scale data, including scoring methods for liver biopsy specimens from patients with chronic hepatitis, such as the Knodell hepatic activity index, the Ishak score, and the METAVIR score [1]. As another example, histological assessments of breast cancer specimens using the Nottingham Grade results in an ordinal variable indicative of the potential for metastasis, with G1, G2, and G3 representing well, moderately, and poorly differentiated tumors, respectively. For other chronic diseases such as Type II diabetes, subjects are often classified as normal, glucose intolerant, and diabetic, representing three ordinal class categories.

Though there are traditional methods for modeling ordinal response data, such as proportional odds, continuation ratio, and adjacent category models [2, 3], these methods assume independence among the predictor variables and moreover, require that the number of samples

(n) exceed the number of covariates (p) included in the model. For gene expression microarray data, the number of covariates is much larger than the number of samples, so that such models cannot be estimated. Although there are data dimensionality reduction techniques such as principle components, it may still be impossible to satisfactorily reduce the amount of variables to be less than the number of observations without a significant loss of information in the data and encumbrances placed on the interpretability of the mega-gene results. Therefore, in this paper we describe fitting an L_1 penalized continuation ratio model to predict an ordinal class using gene expression data. Penalized models were selected because they have been demonstrated to have excellent performance when applied to high-throughput genomic datasets in fitting linear [4, 5], logistic [6, 7], and Cox proportional hazards models [8, 9, 10, 11]. Continuation ratio models for predicting an ordered phenotypic variable given high throughput genomic data are briefly considered in [12], when describing a general Bayesian approach that incorporates a sparsity prior to enable fitting penalized models; the approach includes the L_1 penalty as a special case. Herein we provide more details about the L_1 penalty when using a likelihood-based approach for fitting penalized continuation ratio models. In Section 2 we review the continuation ratio model; in Section 3 we introduce the L_1 penalized continuation ratio model; in Section 4 we describe the our simulation study and results; in Section 5 we describe the results from applying the methods to three gene expression microarray datasets; and in Section 6 we close with our concluding remarks.

2. The Continuation Ratio Model

The cumulative odds model is perhaps the most widely used, largely because the proportional odds assumption yields a parsimonious model whereby one vector of covariate coefficients is returned. However, the proportional odds assumption may be untenable in many situations. A similar model, the continuation ratio (CR) model, can be fit so as to yield one vector of covariate coefficients (the constrained continuation ratio model) which preserves the parsimony of the cumulative odds model [13]. However, when the proportional odds assumption is not met, a fully unconstrained or partially constrained continuation ratio model may be fit to allow flexibility in the model when the proportional odds assumption is not met [14]. Moreover, the CR model is useful when one considers an outcome such that progression through the ordinal levels cannot be reversed, such as stage of cancer [15].

Suppose for each observation, $i = 1, \dots, n$, the response Y_i belongs to one ordinal class $k = 1, \dots, K$ and $\mathbf{x}_i$ represents a p -length vector of covariates. As with binary logistic regression, rather than modeling the probability we model the logit, and for the CR model we model the logit of the conditional probability or

$$\text{logit}(P(Y=k|Y \leq k, \mathbf{X}=\mathbf{x})) = \log \left(\frac{P(Y=k|Y \leq k, \mathbf{X}=\mathbf{x})}{P(Y < k|Y \leq k, \mathbf{X}=\mathbf{x})} \right) = \alpha_k + \beta_k^T \mathbf{x}. \quad (1)$$

It should be mentioned that there are different ways in which one can set up a continuation ratio model. Here we have used the backward formulation, which is commonly used when progression through disease states from none, mild, moderate, severe is represented by increasing integer values, and interest lies in estimating the odds of more severe disease compared to less severe disease [16]. For observations $i = 1, \dots, n$ the likelihood can be formed by considering the vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iK})^T$ where $y_{ik} = 1$ if the response is in category k and 0 otherwise, so that $n_i = \sum_{k=1}^K y_{ik} = 1$. Using the logit link, the equation representing the conditional probability is

$$\delta_k(\mathbf{x}) = P(Y=k|Y \leq k, \mathbf{X}=\mathbf{x}) = \frac{\exp(\alpha_k + \beta_k^T \mathbf{x})}{1 + \exp(\alpha_k + \beta_k^T \mathbf{x})}. \quad (2)$$

The likelihood for the continuation ratio model is then the product of conditionally independent binomial terms [17], which is given by

$$L(\beta|\mathbf{y}, \mathbf{x}) = \prod_{i=1}^n \left(\prod_{j=2}^K \delta_j^{y_{ij}} (1 - \delta_j)^{1 - \sum_{k=j}^K y_{ik}} \right). \quad (3)$$

We have simplified our notation by not explicitly including the dependence of the conditional probability δ_k on $\mathbf{x}$. Further, simplifying our notation let β represent the vector containing both the thresholds ($\alpha_2, \dots, \alpha_K$) and the log odds ($\beta_1, \dots, \beta_p$) for all $K - 1$ logits, the full parameter vector is

$$\beta = (\alpha_2, \beta_{21}, \beta_{22}, \dots, \beta_{2p}, \dots, \alpha_K, \beta_{K,1}, \beta_{K,2}, \dots, \beta_{K,p})^T \quad (4)$$

which is of length $(K - 1)(p + 1)$. As can be seen from equation 3, the likelihood can be factored into $K - 1$ independent likelihoods, so that maximization of the independent likelihoods will lead to an overall maximum likelihood estimate for all terms in the model [16]. A model consisting of $K - 1$ different β vectors may be overparameterized. To simplify, one commonly fits a constrained continuation model, which includes the $K - 1$ thresholds ($\alpha_2, \dots, \alpha_K$) and one common set of p slope parameters, ($\beta_1, \dots, \beta_p$). To fit a constrained continuation ratio model, the original dataset can be restructured by forming $K - 1$ subsets, where for classes $k = 2, \dots, K$, the subset contains those observations in the original dataset up to class k . Additionally, for the k^{th} subset, the outcome is dichotomized as $y = 1$ if the ordinal class is k and $y = 0$ otherwise. Furthermore, an indicator is constructed for each subset representing subset membership. Thereafter the $K - 1$ subsets are appended to form the restructured dataset, which represents the $K - 1$ conditionally independent datasets in equation 3. Applying a logistic regression model to this restructured dataset yields an L_1 constrained continuation ratio model.

3. L_1 Penalized Constrained Continuation Ratio Model

The least absolute shrinkage and selection operator (LASSO) has been described for ordinary least squares, logistic regression, and Cox proportional hazards models [18, 19]. Generally, the maximization of the likelihood (or minimization of least squares) proceeds as usual subject

to the constraint $\sum_{m=1}^p |\beta_m| \leq t$, where t is a tuning parameter. This penalty corresponds to the L_1 norm, hence is often referred to as L_1 penalized model. The L_1 penalized continuation ratio model can then be expressed as

$$\arg\max_{\beta} \left(L(\beta|\mathbf{y}, \mathbf{x}) - \lambda \sum_{m=1}^p |\beta_m| \right) \quad (5)$$

where $L(\beta|\mathbf{y}, \mathbf{x})$ is given by equation 3. To estimate an L_1 penalized continuation ratio model, we first implemented a function for restructuring the original dataset such that the transformed dataset represents the $K - 1$ conditionally independent datasets needed for the full likelihood as presented in equation 3. Thereafter, L_1 model estimation can proceed using any package capable of fitting a binary logistic regression model, such as `glm` [20], `glmnet` [21], or `lasso2` [22]. The final model can be selected as that attaining the minimum AIC, minimum

BIC, or through cross-validation and the estimated coefficients can be used to obtain the predicted class. That is, for the K class scenario the fitted conditional probabilities from the continuation ratio model can be used to estimate $P(Y = k)$ for all k , where for class K

$$\delta_K = P(Y=K|Y \leq K) = \frac{P(Y=K) \cap P(Y \leq K)}{P(Y \leq K)} = \frac{\pi_K}{\pi_1 + \pi_2 + \dots + \pi_K} = \pi_K. \quad (6)$$

For class k where $1 < k < K$,

$$\delta_k = P(Y=k|Y \leq k) = \frac{P(Y=k) \cap P(Y \leq k)}{P(Y \leq k)} = \frac{\pi_k}{\pi_1 + \dots + \pi_k} \quad (7)$$

so that,

$$\pi_k = \delta_k \prod_{j=k+1}^K (1 - \delta_j). \quad (8)$$

Subsequently,

$$\pi_1 = 1 - \sum_{j=2}^K \pi_j. \quad (9)$$

For classification purposes, the predicted class for observation i can be determined by

$$\hat{\omega}_i = \arg\max_k (\pi_{ik}). \quad (10)$$

4. Simulation

A simulation study was conducted to compare two computational approaches for fitting L_1 penalized constrained continuation ratio models. As previously mentioned, after restructuring the dataset the L_1 continuation ratio model can be estimated using any package capable of fitting a binary logistic regression model, such as `glm` [20], `glmnet` [21], or `lasso2` [22]. However, alternative functions for estimating the AIC, BIC, fitted probabilities, and predicted class must be used for extracting quantities of interest. Therefore, we have developed two new packages, `glm` and `glmnet`, available for download from the Comprehensive R Archive Network for the R programming environment [23]. These two packages were developed because their dependent packages `glm` and `glmnet` fit models along the entire regularization path, whereas the `lasso2` package requires the user to a priori select the penalty parameter. Other packages, such as `ncvreg` and `SIS`, fit a regularization path but do not include options for estimating coefficients that are not included in the penalty term, which is needed for estimation of the α_k terms. One hundred datasets were constructed as follows: first, ninety observations consisting of $p = 1,000$ covariates, where each covariate was independently generated from a standard normal distribution, was taken to be the training data. Ten of the 1,000 covariates were then selected as important predictors and the values for these $j = 1, \dots, 10$ true predictors for all i observations were replaced to represent the $k = 3$ ordinal classes as follows: for observations $i = 1, \dots, 30$, $X_{ij} \sim N(0, 1)$ (class 1); for observations $i = 31, \dots, 60$, $X_{ij} \sim N(\pm 1.5, 1)$ (class 2); and observations $i = 61, \dots, 90$ $X_{ij} \sim N(\pm 3, 1)$ (class 3). Therefore each class consisted of 30 observations. For assessing the impact on the results

due to variation among the covariates, two additional simulation scenarios were performed in the same manner but letting $\sigma = 1.5$ and then $\sigma = 0.5$. In fitting the models, the `glmpath.cr` and `glmnet.cr` functions restructure the training data to represent the $K - 1$ conditionally independent datasets needed for the full likelihood as described in section 2, then fits the L_1 penalized constrained continuation model to this restructured dataset. We note that the default parameter of `alpha=1`, which corresponds to the L_1 penalty for `glmnet`, was used when fitting the `glmnet.cr` models. The default parameter for `glmpath` that controls the proportion of elastic net penalty on the L_2 norm is `lambda2=1e-5`, which we changed to `lambda2=0` to fit an L_1 constrained model for `glmpath.cr`. Models attaining the minimum AIC and minimum BIC were retained. To avoid sensitivities in performance measures due to the random partitioning of data that arise when using five- or ten-fold cross-validation, model performance was assessed using N -fold cross-validation. Results from the 100 datasets were used to compare the four resulting models with respect to: the number of the true predictors having a non-zero coefficient estimate; the number of the noise predictors having a non-zero coefficient estimate; and N -fold cross-validation error. Letting $\hat{\omega}_{(-i)}$ represent the predicted class for observation i when observation i was omitted from the training data, N -fold cross-validation error was

calculated as $\sum_{i=1}^n I(\hat{\omega}_{(-i)} \neq \omega_i)/n$. The association between two ordinal variables X and Y can be estimated by the gamma statistic [3], where given the cross-tabulation matrix T of X and Y having r rows and c columns, the number of concordant pairs for cells $(1, 1)$ to $(r - 1, c - 1)$ is given by

$$C_{kl} = T_{kl} \times \sum_{j=l+1}^c \sum_{i=k+1}^r T_{ij}. \quad (11)$$

Similarly, the number of discordant pairs for cells $(1, 2)$ to $(r - 1, c)$ is given by

$$D_{kl} = T_{kl} \times \sum_{j=1}^{l-1} \sum_{i=k+1}^r T_{ij}. \quad (12)$$

Letting $C = \sum_{j=1}^{c-1} \sum_{i=1}^{r-1} C_{ij}$ and $D = \sum_{j=2}^c \sum_{i=1}^{r-1} D_{ij}$, the gamma statistic of ordinal association is defined as

$$\hat{\gamma} = \frac{C - D}{C + D}. \quad (13)$$

The gamma statistic was estimated as an ordinal measure of association between the true and N -fold predicted class using the cross-tabulation of $\hat{\omega}_{(-i)}$ and ω_i .

The results from our simulation study where $\sigma = 1.0$ or $\sigma = 1.5$ include the median and range calculated over the 100 simulations for: N -fold cross-validation error; the gamma statistic as an ordinal measure of association between the true and N -fold predicted class; the number of the true predictors having a non-zero coefficient estimate (oracle = 10); the number of the noise predictors having a non-zero coefficient estimate (oracle = 0) (Tables 1, 2). When $\sigma = 1.0$, the N -fold cross-validation error for the `glmpath.cr` models was slightly lower compared to the `glmnet.cr` models. When the noise increased ($\sigma = 1.5$), the N -fold cross-validation error increased, and again the N -fold cross-validation error for the `glmpath.cr` models was slightly lower compared to the `glmnet.cr` models. The biggest difference between the four methods was with respect to the number of noise covariates having a non-zero coefficient estimate

(Figure 1 and 2). The number of noise covariates having a non-zero coefficient estimate also increased with increasing σ . The `glmnet` package reported the best specificity for both model selection criteria, with a median of 99.1% (99.0, 99.4) while the specificities for `glm` AIC and `glm` BIC were 98.7% (97.7, 99.0) and 98.9% (98.6, 99.0), respectively for the $\sigma = 1.0$ scenario. For the $\sigma = 1.5$ scenario, `glmnet` had a median specificity of 99.1% (99.0, 99.2) while the specificities for `glm` AIC and `glm` BIC were 98.6% (96.8, 99.0) and 98.9% (98.6, 99.0), respectively. For `glm` the sensitivities were 100% (90%, 100%) for both the $\sigma = 1.0$ and $\sigma = 1.5$ scenario and was 90% (60%, 100%) and 90% (80%, 100%) for `glmnet` for these two scenarios, respectively. As expected, when $\sigma = 0.5$ the N-fold cross-validation error and number of noise predictors having a non-zero coefficient estimate decreased with no real change with respect to the number of true predictors having a non-zero coefficient estimate.

5. Application Datasets

5.1. Classification of normal, impaired fasting glucose, and Type II diabetic samples

Pre-processed gene expression and phenotypic data for GSE21321 were download from Gene Expression Omnibus. Asymptomatic males not previously diagnosed with Type II diabetes were enrolled and subsequently were cross-classified as either normal controls, having impaired fasting glucose, or as Type II diabetics based on a fasting glucose intolerance test. This study included 24,526 probes measured using the IlluminaHumanRef-8 v3.0 Expression BeadChip for 24 unique peripheral blood samples, including 8 controls, 7 with impaired fasting glucose, and 9 type II diabetics. The pre-processed data available had undergone background subtraction which resulted in some negative gene expression values. Therefore, prior to model fitting, genes having expression values less than zero were removed from the analysis, leaving 11,066 probes. Thereafter, the L_1 penalized constrained continuation models were fit using the `glm` and `glmnet` packages. Models attaining the minimum AIC and minimum BIC were retained.

For each model, the number of non-zero parameter estimates, N-fold cross-validation (CV) error, and N-fold gamma statistic as an ordinal measure of association between the true and predicted class appear in Table 3. As can be seen from the table from the N-fold cross-validation procedure, one subject (4.2%) was misclassified in the fitted models, regardless of the fitting algorithm. For the `glmnet` models two probes had a non-zero coefficient estimate whereas for `glm` five probes had a non-zero coefficient estimate (Table 4). The probe ILMN1759232, Insulin receptor substrate 1 (IRS1), had the largest absolute parameter estimate in all models (Table 4). Interestingly, mutations in the IRS1 gene have been associated with Type II diabetes as well as susceptibility to insulin resistance.

5.2. Classification of prostate cancer samples

The pre-processed gene expression and phenotypic data for GSE6099 were download from Gene Expression Omnibus. This study included 20,000 probes measured using the two channel custom spotted cDNA microarrays using 104 prostate tissues samples isolated by laser capture microdissection [24]. Benign prostate tissue was used as the reference and labeled using Cy3 in all hybridizations. Following a previous analysis [12], we restricted attention to 86 samples that included 21 benign, 45 cancer, and 20 metastatic samples and removed probes having a missing value for all samples.

For each model, the number of non-zero parameter estimates, N-fold cross-validation error, and N-fold gamma statistic as an ordinal measure of association between the true and predicted class appear in Table 5. As can be seen from the table, the AIC selected models had the best performance, likely because they include more genes compared to the BIC selected models.

More genes may be needed for correct classification due to the similarity between the two disease states (cancer and metastatic disease). Several genes included in the `glmpathcr` AIC selected model, including CDC25A, ROBO1, WFDC2, AMACR, EP300, BPAG1, DDIT4, ERCC5, PKIB, and PRC1, have been previously studied in association with prostate cancer.

5.3. Classification of normal, ulcerative colitis, and Crohn's disease patients

The pre-processed gene expression and phenotypic data for GSE3365 were download from Gene Expression Omnibus. This study included 22,284 probe sets measured using an Affymetrix GeneChip HG-U133A Array using 127 peripheral blood mononuclear cell samples, including 42 normal, 26 ulcerative colitis, and 59 Crohn's disease samples [25]. Ulcerative colitis and Crohn's disease are both inflammatory bowel diseases, with the distinction that inflammation associated with ulcerative colitis is usually limited to the mucosa whereas inflammation associated with Crohn's disease is deeper within the intestinal wall. Because of their similarity, the two diseases can be difficult to differentiate. In the previous study, the authors compared normal versus inflammatory bowel disease (ulcerative colitis and Crohn's disease patients combined) and ulcerative colitis versus Crohn's disease [25]. Herein we compared the performance of the different modeling techniques to predict the ordinal class, normal < ulcerative colitis < Crohn's disease.

For each model, the number of non-zero parameter estimates, N-fold cross-validation error, and N-fold gamma statistic as an ordinal measure of association between the true and predicted class appear in Table 7. For `glmpathcr`, although the AIC and BIC selected models both misclassified 28 subjects, the subjects misclassified differed hence the gamma ordinal association measure differed. Table 8 displays the probe sets having a non-zero coefficient estimate from the BIC selected `glmpathcr` and the BIC and AIC selected `glmnetcr` models. Generally speaking the same probe sets were included in each of these models. The AIC selected `glmpathcr` model included many more non-zero coefficient estimates. Nevertheless, the AIC selected `glmpathcr` model additionally included HNRNPD and LCN2 which have been previously been associated with Crohn's disease [26, 27] while NFE2L2 and PGDS has been associated with colitis [28, 29].

6. Discussion

For high-throughput genomic datasets, the common approach to analyzing ordinal response data has been to break the problem into one or more dichotomous response analyses. The dichotomous response approach does not make use of all available data and therefore leads to reduced power and an increase in the number of Type I errors [30]. Moreover, treatment of the ordinal outcome as a continuous response with subsequent application of linear modeling techniques is not particularly useful as it results in continuous fitted values, whereas the probability that the observation is from class k is of primary importance [31]. Therefore, in this paper we described a penalized likelihood-based approach, namely, an L_1 penalized constrained continuation ratio model, for modeling an ordinal response. Two computational approaches and two model selection criterion were compared with respect to their classification and variable selection performances using simulated and gene expression data.

The approach presented in section 3 is a likelihood-based method that includes a sparsity penalty. An alternative to this frequentist method is the Bayesian approach, which includes a sparsity prior for the parameter vector β [12]. As previously described, the variance of the parameter for each covariate $m = 1, \dots, p$ is taken to follow some distribution, for example $v_m^2 \sim \text{Gamma}(r, c)$ where r is the shape and c is the scale parameter. Under the independence assumption $p(v^2) = \prod_{m=1}^p p(v_m^2)$. The sparsity prior [12, 32] is the marginal for each β and can be expressed as

$$p(\beta) = \int_{v^2} p(\beta|v^2)p(v^2)dv^2. \quad (14)$$

For example, when $v_m^2 \sim \text{Gamma}(r, c)$ and $p(\beta|v^2) \sim N(0, \Delta(v^2))$, we have the normal gamma sparsity prior [12]. The EM algorithm can be used to obtain the maximum a posteriori estimates for the parameter vector. This methodology has been implemented in the `HGFns` package [33] in the `R` programming environment [23] and is available by request. Because the package fits one model for a user-selected penalty rather than fitting an entire regularization path, a comparison between the frequentist and Bayesian methods was not performed herein.

For the simulated datasets, the `glmnet` models reported the best specificities. For high-throughput genomic datasets where a large number of noise covariates are present, subtle differences in specificities can lead to large differences in the number of false discoveries. For example, assuming 10,000 genes on a microarray are noise covariates, 50 more false discoveries would be expected given the specificities of 98.6% (`glmnet`) and 99.1% (`glmnet`). In our simulation scenarios, although the sensitivities and specificities were both quite high, roughly half of the covariates having non-zero coefficient estimates were false discoveries. Model selection using a smaller penalty may reduce the number of false positives. For the application datasets, there was a high degree of similarity between the `glmnet` and `glmnet` models, particularly when using the BIC selection criteria. The AIC selected `glmnet` model usually resulted in the largest model. For application datasets where the underlying pathology of the disease classes is quite similar, larger models may be necessary. That is sometimes the BIC selected models may be too sparse (Table 5), thus may render poorer performance. For all three applications, we identified genes having a non-zero coefficient estimates that were biologically relevant for the disease process under study.

To explore the differences between the two numeric algorithms, we fit a `glmnet` model with `lasso2=0` to estimate an L1 penalized constrained continuation ratio, then passed the resulting `lambda` vector as the sequence of regularization parameters to the `glmnet` function. We then extracted the estimated coefficients from both fits (`glmnet` and `glmnet`) for the step at which the `glmnet` model had the smallest BIC and calculated the sum of the squared deviations between the estimated coefficients from both models. For the Diabetes dataset, the BIC selected `glmnet` model was at `lambda= 1.891634` at which the sum of the squared deviations between the gene coefficient estimates was 0.002337. Using this same `lambda` trace the best BIC selected `glmnet` model was at `lambda= 0.1078729`; the sum of the squared deviations between the gene coefficients at this λ was 0.00755. However, the sum of the squared deviations from the two BIC selected models when using the algorithms' default calculations for the λ trace was only 0.000936. We note that a larger sum of squared deviations results when including the intercept terms, as they vary more widely between `glmnet` and `glmnet` for the same λ value than the covariate estimates. Similar results were observed for the Prostate cancer and Crohns datasets. Although `glmnet` and `glmnet` are two algorithms designed to yield the same L1 penalized model, the numeric algorithms differ so that the resulting models will have slightly different non-zero coefficient estimates for covariates and very different threshold estimates at the same `lambda` value. Obviously these differences affect the AIC and BIC calculations. When allowing the functions to compute their own `lambda` sequence and then selecting the models using the same criterion (e.g., AIC or BIC), a smaller sum of squared deviations between their estimated coefficients was observed compared to selecting models at the same `lambda` value. Therefore use of a consistent model selection criterion is preferred over the use of a consistent `lambda` value.

We have made the `glmpathcr` and `glmnetcr` R packages available for download from the Comprehensive R Archive Network which are useful because both packages include model fitting functions that include an option to ensure estimation of the α_k parameters. This same option can additionally be used to ensure important confounding variables have non-zero coefficient estimates in the final model. We note that because for `glmnet.cr` the AIC and BIC are calculated each step along the regularization path whereas for `glmpath.cr` the AIC and BIC are estimated in a separate function, `glmpath.cr` has an advantage with respect to computational speed. For example, the elapsed time for the N-fold cross-validation procedure for the diabetes dataset was 5.1 minutes for `glmpath.cr` and 10.3 minutes for `glmnet.cr`.

Covariates having non-zero coefficient estimates in ordinal response models applied to gene expression datasets typically have a monotonic association with the ordinal response. These genes may be involved in the transformation to more severe disease states, and so could potentially be useful prognostic molecular markers or therapeutic targets.

Acknowledgments

This research was supported by the National Institutes of Health/National Institute of Library Medicine grant R03LM009347.

References

1. Theise ND. Liver biopsy assessment in chronic viral hepatitis: a personal, practical approach. *Modern Pathology*. 2007; 20(Suppl 1):S3–S14. [PubMed: 17486049]
2. Hosmer, DW.; Lemeshow, S. *Applied Logistic Regression*, Second Edition. Hoboken NJ: John Wiley & Sons; 2000.
3. Agresti, A. *Categorical Data Analysis*. Hoboken NJ: John Wiley & Sons; 2002.
4. Wu B. Differential gene expression detection using penalized linear regression models: the improved SAM statistics. *Bioinformatics*. 2005; 21:1565–1571. [PubMed: 15598833]
5. Wu B. Differential gene expression detection and sample classification using penalized linear regression models. *Bioinformatics*. 2006; 22:472–476. [PubMed: 16352654]
6. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. *Biostatistics*. 2004; 5:427–443. [PubMed: 15208204]
7. Schimek MG. Penalized binary regression for gene expression profiling. *Methods Inf Med*. 2004; 43:439–444. [PubMed: 15702197]
8. Schumacher M, Binde H, Gerds T. Assessment of survival prediction models based on microarray data. *Bioinformatics*. 2007; 23:1768–1774. [PubMed: 17485430]
9. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*. 2005; 21:3001–3008. [PubMed: 15814556]
10. Li H, Gui J. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*. 2004; 20(Suppl 1):i208–i215. [PubMed: 15262801]
11. Shedden K, Taylor JMG, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell 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 VE, Meyerson M, Kuick R, Dobbin KK, Lively T, Jacobson JW, Beer DG. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*. 2008; 14:822–827.
12. Kiiveri HT. A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations. *BMC Bioinformatics*. 2008; 9:195. [PubMed: 18410693]
13. O'Connell, AA. *Logistic Regression Models for Ordinal Response Variables*. Thousand Oaks CA: Sage Publications; 2006.

14. Cole SR, Ananth CV. Regression models for unconstrained, partially or fully constrained continuation odds ratios. *International Journal of Epidemiology*. 2001; 30:1379–1382. [PubMed: 11821350]
15. Greenland S. Alternative models for ordinal logistic regression. *Statistics in Medicine*. 1994; 13:1665–1677. [PubMed: 7973242]
16. Bender R, Benner A. Calculating ordinal regression models in SAS and S-Plus. *Biometrical Journal*. 2000; 42:677–699.
17. Cox C. Multinomial regression models based on continuation ratios. *Statistics in Medicine*. 1988; 7:435–441. [PubMed: 3358023]
18. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, B*. 1996; 58:267–288.
19. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine*. 1997; 16:385–395. [PubMed: 9044528]
20. Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, B*. 2007; 64:659–677.
21. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010; 33:1–22. [PubMed: 20808728]
22. Lokhorst, J.; Venables, B.; Turlach, B.; Maechler, M. [Date accessed January 10, 2011] lasso2: L1 constrained estimation aka 'lasso'. 2010. <http://CRAN.R-project.org/package=lasso2>
23. R Core Development Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2009.
24. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SD, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM. Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics*. 2007; 39:41–51. [PubMed: 17173048]
25. Burczynski ME, Peterson RL, Twine NC, Zuberek KA, Brodeur BJ, Casciotti L, Maganti V, Reddy PS, Strahs A, Immermann F, Spinelli W, Schwertschlag U, Slager AM, Cotreau MM, Dorner AJ. Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *Journal of Molecular Diagnostics*. 2006; 8:51 s–61 s. [PubMed: 16436634]
26. Noguchi E, Homma Y, Kang X, Netea MG, Ma X. A Crohn's disease-associated NOD2 mutation suppresses transcription of human IL10 by inhibiting activity of the nuclear ribonucleoprotein hnRNP-A1. *Nature Immunology*. 2009; 10:471–479. [PubMed: 19349988]
27. Bolignano D, Della Torea A, Lacquaniti A, Costantino G, Fries W, Buemi M. Neutrophil gelatinase-associated lipocalin levels in patients with Crohn disease undergoing treatment with infliximab. *Journal of Investigational Medicine*. 2010; 58:569–571.
28. Arisawa T, Tahara T, Shibata T, Nagasaka M, Nakamura M, Kamiya Y, Fujita H, Yoshioka D, Okubo M, Sakata M, Wang FY, Hirata I, Nakano H. Nrf2 gene promoter polymorphism is associated with ulcerative colitis in a Japanese population. *Hepatogastroenterology*. 2008; 55:394–397. [PubMed: 18613373]
29. Hokari R, Kurihara C, Nagata N, Aritake K, Okada Y, Watanabe C, Komoto S, Nakamura M, Kawaguchi A, Nagao S, Urade Y, Miura S. Increased expression of lipocalin-type-prostaglandin D synthase in ulcerative colitis and exacerbating role in murine colitis. *American Journal of Physiological: Gastrointestinal and Liver Physiology*. 2010 Dec 16. [Epub ahead of print].
30. Ananth CV, Kleinbaum DG. Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*. 1997; 26:1323–1333. [PubMed: 9447413]
31. Thomson D, Furness RW, Monaghan P. The analysis of ordinal response data in the behavioural sciences. *Animal Behaviour*. 1998; 56:1041–1043. [PubMed: 9790717]
32. Cawley G, Talbot N. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics*. 2006; 22:2348–2355. [PubMed: 16844704]
33. CSIRO Bioinformatics. HGfns: Sparse model fitting. 2010. R package version 3.0.8.

Figure 1. Boxplots from the $\sigma = 1.0$ simulation study of the number of noise predictors having a non-zero coefficient estimate for the `glmpathcr` and `glmnetcr` BIC and AIC selected penalized constrained continuation ratio models.

Figure 2. Boxplots from the $\sigma = 1.5$ simulation study of the number of noise predictors having a non-zero coefficient estimate for the `glmpathcr` and `glmnetcr` BIC and AIC selected penalized constrained continuation ratio models.

Table 1

Results from simulation study for the scenario where $\sigma = 1.0$ include the median and range calculated over the 100 simulations for: N-fold cross-validation error; the gamma statistic as an ordinal measure of association between the true and N-fold predicted class; the number of the true predictors having a non-zero coefficient estimate (oracle = 10); the number of the noise predictors having a non-zero coefficient estimate (oracle = 0).

	N-fold CV error	N-fold CV gamma	Non-zero estimates True Predictors	Non-zero estimates Noise Predictors
glmpathcr BIC	2.2 (0, 8.9)	1 (0.995, 1)	10 (9, 10)	11 (10, 14)
glmpathcr AIC	2.2 (0, 8.9)	1 (0.993, 1)	10 (9, 10)	13 (10, 23)
glmnetcr BIC	3.3 (0, 8.9)	0.999 (0.995, 1)	9 (6, 10)	9 (6, 10)
glmnetcr AIC	3.3 (0, 8.9)	0.999 (0.995, 1)	9 (6, 10)	9 (6, 10)

Table 2

Results from simulation study for the scenario where $\sigma = 1.5$ include the median and range calculated over the 100 simulations for: N-fold cross-validation error; the gamma statistic as an ordinal measure of association between the true and N-fold predicted class; the number of the true predictors having a non-zero coefficient estimate (oracle = 10); the number of the noise predictors having a non-zero coefficient estimate (oracle = 0).

	N-fold CV error	N-fold CV gamma	Non-zero estimates True Predictors	Non-zero estimates Noise Predictors
glmpathcr BIC	5.6 (0, 14.4)	0.998 (0.983, 1)	10 (9, 10)	11 (10, 14)
glmpathcr AIC	5.6 (1.1, 14.4)	0.998 (0.982, 1)	10 (9, 10)	14 (10, 32)
glmnetcr BIC	6.7 (1.1, 15.6)	0.997 (0.978, 1)	9 (8, 10)	9 (8, 10)
glmnetcr AIC	6.7 (1.1, 15.6)	0.997 (0.978, 1)	9 (8, 10)	9 (8, 10)

Table 3

Results from GSE21321, classification of normal, impaired fasting glucose, and Type II diabetic samples. For each model, the number of non-zero parameter estimates, N-fold cross-validation error, and N-fold gamma statistic as an ordinal measure of association between the true and predicted class.

	Number of genes with non-zero coefficient estimates	N-fold CV Error	N-fold CV Gamma
glmpathcr BIC	5	4.2%	1.0
glmpathcr AIC	5	4.2%	1.0
glmnetcr BIC	2	4.2%	1.0
glmnetcr AIC	2	4.2%	1.0

Table 4

Parameter estimates from `glmptchr` and `glmnetcr` for probes having a non-zero coefficient estimate from GSE21321, classification of normal, impaired fasting glucose, and Type II diabetic samples.

Illumina ID	$\hat{\beta}_{\text{glmptchr}}$	$\hat{\beta}_{\text{glmnetcr}}$
ILMN_1705116	0.0334955982	0.00661497
ILMN_1733757	-0.0003751318	0
ILMN_1758311	0.0022607450	0
ILMN_1759232	-0.0347775942	-0.02035163
ILMN_2100437	-0.0001679020	0

Table 5

Results from GSE6099. For each model, the number of non-zero parameter estimates, N-fold cross-validation error, and N-fold gamma statistic as an ordinal measure of association between the true and predicted class.

	Number of genes with non-zero coefficient estimates	N-fold CV Error	N-fold CV Gamma
glmpathcr BIC	8	36.0%	0.927
glmpathcr AIC	40	25.6%	0.942
glmnetcr BIC	8	38.4%	0.909
glmnetcr AIC	21	25.6%	0.969

Table 6

Parameter estimates from BIC selected models using `glmPathcr` and `glmnetcr` for probes having a non-zero coefficient estimate from GSE6099, classification of benign, cancerous, and metastatic prostate tissue.

ID	GENE	$\hat{\beta}_{\text{glmPathcr}}$	$\hat{\beta}_{\text{glmnetcr}}$
Hs6-17-6-12	MEIS2	-0.57052131	-0.5110322
Hs6-18-19-25	CUGBP1	0.15104909	0.0894821
Hs6-19-1-25	LIM	-0.49104049	-0.4751542
Hs6-19-10-4	—	0.07133808	0.0626283
Hs6-19-5-9	ZNF532	-0.50569048	-0.4811424
Hs6-27-18-2	PCYOX1	-0.22084027	-0.1923767
Hs6-32-21-13	CKLFSF3	-0.58128955	-0.5668920
Hs6-5-25-11	P2RY5	-0.25302141	-0.2140612

Table 7

Results from GSE3365, classification of normal, ulcerative colitis, and Crohn's disease. For each model, the number of non-zero parameter estimates, N-fold cross-validation error, and N-fold gamma statistic as an ordinal measure of association between the true and predicted class.

	Number of genes with non-zero coefficient estimates	N-fold CV Error	N-fold CV Gamma
glmpathcr BIC	13	22.0%	0.985
glmpathcr AIC	47	22.0%	0.952
glmnetcr BIC	12	21.3%	0.999
glmnetcr AIC	13	21.3%	0.999

Table 8

Parameter estimates from selected models using `glmpathcr` and `glmnetcr` for probes having a non-zero coefficient estimate from GSE3365, classification of normal, ulcerative colitis, and Crohn's disease samples.

ID	GENE	$\hat{\beta}_{\text{glmpathcr}(BIC)}$	$\hat{\beta}_{\text{glmnetcr}(BIC)}$	$\hat{\beta}_{\text{glmnetcr}(AIC)}$
200080_s_at	H3F3B	0.000085	0.000078	0.000100
201032_at	BLCAP	-0.004141	-0.004114	-0.004297
201121_s_at	PGRMC1	0.000567	0.000564	0.000625
202187_s_at	PPP2R5A	-0.000171	-0.000201	-0.000194
202362_at	RAP1A	0.000638	0.000598	0.000652
202708_s_at	HIST2H2BE	0.001279	0.001164	0.001283
208649_s_at	VCP	-0.002113	-0.002008	-0.002280
210361_s_at	ELF2	-0.001646	-0.001478	-0.001719
212808_at	NFATC2IP	-0.008296	-0.008054	-0.008372
214948_s_at	TMF1	-0.000005	0	0
215071_s_at	HIST1H2AC	0.000446	0.000468	0.000435
217895_at	PTCD3	-0.001498	-0.001384	-0.001499
217917_s_at	DYNLRB1	0	0	0.000003
222147_s_at	ACTR5	-0.007278	-0.007390	-0.007428