 doi:10.20944/preprints202203.0182.v1

Article

ordinalbayes: Fitting Ordinal Bayesian Regression Models to High-Dimensional Data using R

Kellie J. Archer [1,†*], Anna Eames Seffernick [2], Shuai Sun [3],and Yiran Zhang [4]

[1] Division of Biostatistics, College of Public Health, The Ohio State University; archer.43@osu.edu
[2] Division of Biostatistics, College of Public Health, The Ohio State University; seffernick.13@buckeyemail.osu.edu
[3] Division of Biostatistics, College of Public Health, The Ohio State University; sun.2694@buckeyemail.osu.edu
[4] Amgen, Thousand Oaks, CA; rennyzhang77@gmail.com
* Correspondence: archer.43@osu.edu; Tel.: +1-614-247-6167

Abstract: Stage of cancer is a discrete ordinal response that indicates aggressiveness of disease and is often used by physicians to determine the type and intensity of treatment to be administered. For example, the FIGO stage in cervical cancer is based on the size and depth of the tumor as well as the level of spread. It may be of clinical relevance to identify molecular features from high-throughput genomic assays that are associated with stage of cervical cancer, to elucidate pathways related to tumor aggressiveness, identify improved molecular features that may be useful for staging, and identify therapeutic targets. High-throughput RNA-Seq data and corresponding clinical data (including stage) for cervical cancer patients has been made available through The Cancer Genome Atlas Project (TCGA). We recently described penalized Bayesian ordinal response models that can be used for variable selection for over-parameterized datasets such as the TCGA-CESC dataset. Herein, we describe our `ordinalbayes` R package, available from the Comprehensive R Archive Network (CRAN), which is capable of fitting cumulative logit models when the outcome is ordinal and the number of predictors exceeds the sample size, $P > N$, such as for TCGA data. We demonstrate use of this package through application to TCGA cervical cancer dataset. Our `ordinalbayes` package can be used to fit models to high-dimensional dataset and effectively performs variable selection.

Keywords: cumulative logit; penalized models; LASSO; variable inclusion indicators; spike-and-slab

1. Introduction

Despite the advent of HPV vaccinations and effective screening programs, globally, cervical cancer is the fourth most common cancer among women [1]. The estimated number of new cases in 2020 is 604,127 with 341,831 deaths [2]. Stage of cervical cancer, as outlined in the International Federation of Gynecology and Obstetrics (FIGO) guidelines, is based on physical examination, endoscopic procedures, and imaging. Specifically, FIGO stage is based on the size and depth of the tumor as well as the level of spread [3]. It is important that stage, a discrete ordinal response, be correct as it is used to guide treatment planning, counsel patients with respect to prognosis, and to determine whether the patient meets eligibility criteria for available clinical trials or other research studies [4,5]. Unfortunately, there is still debate as to whether surgical or non-invasive radiological modalities for identifying parametrial and lymph node involvement is preferred when staging a patient [4]. Thus it is clinically relevant to identify molecular features from high-throughput genomic assays that are associated with stage of cervical cancer, to elucidate pathways related to tumor aggressiveness, identify improved molecular features that may be useful for staging, and identify therapeutic targets.

Penalized frequentist models have been widely applied when analyzing high-dimensional data. Such models were initially described for linear [6] and logistic [7] regression, and subsequently for ordinal response models [8–10]. However, when applying penalized frequentist models the penalty parameter, or vector of parameters in the case of elastic net, must be selected by the analyst. As a result, the coefficient estimates from the resulting model are conditional on that penalty parameter. For that reason, penalized Bayesian

Citation: Archer, K.J.; Seffernick, A.E.; Sun, S.; Zhang, Y. ordinalbayes. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2022 doi:10.20944/preprints202203.0182.v1

models were developed for the linear [11–14] and logistic [15–17] regression settings. We also recently described penalized Bayesian models for the ordinal response setting [18] and demonstrated that our penalized Bayesian cumulative logit model has improved variable selection performance when compared to penalized frequentist cumulative logit models [19].

Herein we describe our `ordinalbayes` R package that can be used for fitting penalized Bayesian cumulative logit models. The `ordinalbayes` function can be used to fit LASSO, normal spike-and-slab, double exponential spike-and-slab, and regression-based variable inclusion indicator Bayesian models. Variable selection can be performed using Bayes factor or using the posterior distributions of the variable inclusion indicators directly. In the following sections, we describe our implementation and describe the syntax required for each of our Bayesian models. We then illustrate the functions in the `ordinalbayes` R package using a dataset where we were interested in identifying transcripts important to predicting FIGO stage in cervical cancer patients using high-throughput gene expression data.

2. Materials and Methods

2.1. ordinalbayes Models and Syntax

The primary function for model fitting in the `ordinalbayes` package is `ordinalbayes`. The function arguments are

```
function (formula, data, x = NULL, subset, center = TRUE, scale = TRUE,
    a = 0.1, b = 0.1, model = "regressvi", gamma.ind = "fixed",
    pi.fixed = 0.05, c.gamma = NULL, d.gamma = NULL, alpha.var = 10,
    sigma2.0 = NULL, sigma2.1 = NULL, coerce.var=10, lambda0 = NULL,
    adaptSteps = 5000, burnInSteps = 5000, nChains = 3, numSavedSteps = 9999,
    thinSteps = 3, parallel = TRUE, seed = NULL, quiet = FALSE)
```

This function accepts a model `formula` that specifies the ordinal outcome on the left-hand side of the equation and any unpenalized predictor variable(s) from the phenotypic dataset on the right-hand side of the $\sim$ equation; if no unpenalized predictor variables are included, the model formula includes 1 (the intercept) on the right-hand side of the equation. Unpenalized predictors are those that we want to coerce into the model (e.g., age) so that no penalty is applied. When unpenalized predictors are included (or coerced) into the model, the user can specify the variance associated with those model parameters (default `coerce.var=10`). When analyzing data processed using the `DESeq2` Bioconductor package, the genomic feature object is of class `DESeqTransform` which is a `SummarizedExperiment`, and therefore the phenotypic data are accessed using the `colData` extractor function. When analyzing data processed using packages that structure the genomic feature object as a `Biobase ExpressionSet`, the phenotypic data are accessed using `pData` extractor function. Therefore, in the `ordinalbayes` call, `data` should ideally be `colData` or `pData` calls to the genomic feature object, though a `data.frame` name can be passed. Note that when passing a `data.frame` to `data` that is not connected to the penalized variables (`x`), the user needs to carefully verify that the observations in the `data.frame` are appropriate aligned to the genomic feature data in `x`. For `SummarizedExperiment` objects, the user should pass to `x` the genomic feature data (e.g., expression of genes from high-throughput assays) to be penalized in the fitted model, which can be accessed using the `assay` extractor function. For `ExpressionSet` objects, the genomic features to be penalized can be accessed using the `exprs` extractor function. The user can also pass a matrix to `x`, however, the user needs to carefully verify that the observations in the `x` matrix are appropriately aligned to the phenotypic data. Note that the number of rows in both `data` and `x` should be the same, such that the transpose of `assay` or `exprs` should be supplied to `x`. The user can `subset` the data set prior to model fitting, for example, `subset=(race=="white")`. By default the genomic features are centered (`center=TRUE`) and scaled (`scale=TRUE`).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2022 doi:10.20944/preprints202203.0182.v1

Selected parameters are initialized prior to updating through MCMC. For one chain, the $k-1$ ordinal thresholds, α_k, are initialized to the logit of the cumulative class probabilities, which is equivalent to the estimated $k-1$ thresholds in an intercept-only model

$$\alpha_k = \log\left(\frac{\sum_{i=1}^{n}\sum_{m=1}^{k} y_{ik}/n}{1-\sum_{i=1}^{n}\sum_{m=1}^{k} y_{ik}/n}\right).$$

For chains beyond this first one, initial values for the α_k terms are sampled from a Normal(0, 0.5) distribution and sorted to impose the $\alpha_1 < \cdots < \alpha_{k-1}$ order restriction. Within the MCMC, the α_k terms are sampled from a Normal$(0, \sigma^2_{\alpha_k})$ and users can adjust the variance by specifying `alpha.var` (default 10 such that the precision is 0.10). All penalized coefficients (β_j for $j = 1, \ldots, P$) are initialized to zero.

Other relevant parameters common to all model types include: `nChains`, the number of parallel chains for the model (default 3); `adaptSteps`, the number of iterations for adaptation (default 5000); `burnInSteps`, the number of iterations of the Markov chain to run (default 5000); `numSavedSteps`, the number of saved steps per chain (default 9999); and `thinSteps`, the thinning interval for monitors (default 3). Provided the user will be running the model on a machine with multiple processors, computational speed can be improved by running the chains in parallel, by specifying `parallel = TRUE`. When `parallel = TRUE`, `runjags` executes the MCMC sampling using `nChains` parallel processors. To ensure the user can obtain reproducible results, `seed` accepts an integer and is used to set the random seed. Output from JAGS can be suppressed by specifying `quiet = TRUE`. The user can fit one of four available Bayesian models. A list of the parameters the user can set for all four models is provided in Table A1. Next, each of the four models is described along with the relevant arguments that must be specified by the user. A list of the parameters the user needs to set for each specific model is provided in Table A2.

2.1.1. Regression-Based Variable Inclusion Indicator Ordinal Model

By default the model that is fit is the regression-based variable inclusion indicator Bayesian model, specified by `model="regressvi"`. This model takes the form

$$\begin{aligned}
&\log\left[\frac{Pr(Y_i \le k|\boldsymbol{x}_i)}{Pr(Y_i > k|\boldsymbol{x}_i)}\right] = \alpha_k - \sum_{j=1}^{p}\gamma_j\beta_j x_{ij}, \quad \text{for } k = 1, 2, ..., K-1\\
&\beta_j|\lambda \sim \text{DE}(0, 1/\lambda), \quad \text{for } j = 1, ..., p\\
&\lambda \sim \text{Gamma}(a, b)\\
&\alpha_k \sim \text{Normal}(0, \sigma^2_{\alpha_k}),\ \alpha_1 < \alpha_2 < ... < \alpha_{K-1}, \quad \text{for} \quad k = 1, 2, ..., K-1\\
&\gamma_j \sim \text{Bernoulli}(\pi_j), \quad \text{for } j = 1, ..., p\\
&\pi_j = t \text{ or } \pi_j \sim \text{Beta}(c, d), \quad \text{for } j = 1, ..., p
\end{aligned}$$

and assumes the penalized coefficients are from a Laplace (or double exponential) distribution with parameter λ and that λ is from a Gamma distribution with parameters `a` and `b`. Based on our extensive simulations [19], model performance is not affected by choices of `a` and `b` so we provide defaults of 0.1 for both. The variable inclusion indicator γ_j is assumed to follow a Bernoulli distribution with parameter π_j. The user can select to use a fixed constant prior for $\pi_j = 1, \ldots, P$ by specifying both `gamma.ind="fixed"` and specifying some constant in the (0, 1) interval for `pi.fixed` (default is 0.05). Alternatively, a random prior for π_j is acheived by specifying both `gamma.ind="random"` and parameter values for the Beta distribution `c.gamma` and `d.gamma`. If unpenalized coefficients are included in the model, their coefficients are $\zeta \sim \text{Normal}(0, \sigma^2_{coerce})$.

2.1.2. LASSO Ordinal Model

The LASSO Bayesian ordinal model can be fit by specifying `model="lasso"`. This model assumes the penalized coefficients β_j for $j = 1, \ldots, P$ are from independent Laplace

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2022 doi:10.20944/preprints202203.0182.v1

(or double exponential) distributions with parameter λ and that λ is from a Gamma distribution with parameters `a` and `b`.

$$\log\left[\frac{Pr(Y_i \leq k|\boldsymbol{x}_i)}{Pr(Y_i > k|\boldsymbol{x}_i)}\right] = \alpha_k - \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{for } k = 1, 2, ..., K-1$$
$$\beta_j|\lambda \sim \text{DE}(0, 1/\lambda), \quad \text{for } j = 1, ..., p$$
$$\lambda \sim \text{Gamma}(a, b)$$
$$\alpha_k \sim \text{Normal}(0, \sigma^2_{\alpha_k}),\ \alpha_1 < \alpha_2 < ... < \alpha_{K-1}, \quad \text{for} \quad k = 1, 2, ..., K-1$$

As previously mentioned, model performance is not affected by choices of `a` and `b` so we provide defaults of 0.1 for both. If unpenalized coefficients are included in the model, their coefficients are $\zeta \sim \text{Normal}(0, \sigma^2_{coerce})$.

2.1.3. Normal Spike-and-Slab Ordinal Model

The normal spike-and-slab Bayesian ordinal model can be fit by specifying `model="normalss"`. This model is given by

$$\log\left[\frac{Pr(Y_i \leq k|\boldsymbol{x}_i)}{Pr(Y_i > k|\boldsymbol{x}_i)}\right] = \alpha_k - \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{for } k = 1, 2, ..., K-1$$
$$\beta_j|\gamma_j \sim (1-\gamma_j) \times \text{Normal}(0, \sigma_0^2) + \gamma_j \times \text{Normal}(0, \sigma_1^2), \quad \text{for } j = 1, ..., p$$
$$\alpha_k \sim \text{Normal}(0, \sigma^2_{\alpha_k}),\ \alpha_1 < \alpha_2 < ... < \alpha_{K-1}, \quad \text{for} \quad k = 1, 2, ..., K-1$$
$$\gamma_j \sim \text{Bernoulli}(\pi_j), \quad \text{for } j = 1, ..., p$$
$$\pi_j = t \text{ or } \pi_j \sim \text{Beta}(c, d), \quad \text{for } j = 1, ..., p.$$

When fitting this model the user is required to specify the variance for the spike (σ_0^2) by setting `sigma2.0` to a small positive value (e.g., 0.01) and variance for the slab (σ_1^2) by setting `sigma2.1` to a large positive value (e.g., 10). As with the regression-based variable inclusion indicator Bayesian model, the variable inclusion indicator γ_j is assumed to follow a Bernoulli distribution with parameter π_j. The user can select to use a fixed constant prior for $j = 1, \ldots, P$ by specifying both `gamma.ind="fixed"` and specifying some constant in the (0, 1) interval for `pi.fixed` (default is 0.05). Alternatively, a random prior for π_j is acheived by specifying both `gamma.ind="random"` and parameter values for the Beta distribution `c.gamma` and `d.gamma`. If unpenalized coefficients are included in the model, their coefficients are $\zeta \sim \text{Normal}(0, \sigma^2_{coerce})$.

2.1.4. Double Exponential Spike-and-Slab Ordinal Model

The double exponential spike-and-slab ordinal model can be fit by specifying `model="dess"` and is given by

$$\log\left[\frac{Pr(Y_i \leq k|\boldsymbol{x}_i)}{Pr(Y_i > k|\boldsymbol{x}_i)}\right] = \alpha_k - \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{for } k = 1, 2, ..., K-1$$
$$\beta_j|\lambda, \gamma_j \sim (1-\gamma_j) \times \text{DE}(0, 1/\lambda_0) + \gamma_j \times \text{DE}(0, 1/\lambda), \quad \text{for } j = 1, ..., p$$
$$\lambda \sim \text{Gamma}(a, b)$$
$$\alpha_k \sim \text{Normal}(0, \sigma^2_{\alpha_k}),\ \alpha_1 < \alpha_2 < ... < \alpha_{K-1}, \quad \text{for} \quad k = 1, 2, ..., K-1$$
$$\gamma_j \sim \text{Bernoulli}(\pi_j), \quad \text{for } j = 1, ..., p$$
$$\pi_j = t \text{ or } \pi_j \sim \text{Beta}(c, d), \quad \text{for } j = 1, ..., p$$

When fitting this model the user is required to specify the parameter for the spike (λ_0) using `lambda0` while the slab is taken to be a double exponential distribution with parameter λ where that λ is from a Gamma distribution with parameters `a` and `b`. As

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2022 doi:10.20944/preprints202203.0182.v1

with the regression-based variable inclusion indicator and Normal spike-and-slab models, the variable inclusion indicator γ_j is assumed to follow a Bernoulli distribution with parameter π_j. The user can select to use a fixed constant prior for $j = 1, \ldots, P$ by specifying both `gamma.ind="fixed"` and specifying some constant in the (0, 1) interval for `pi.fixed` (default is 0.05). Alternatively, a random prior for π_j is acheived by specifying both `gamma.ind="random"` and parameter values for the Beta distribution `c.gamma` and `d.gamma`. If unpenalized coefficients are included in the model, their coefficients are $\zeta \sim \text{Normal}(0, \sigma^2_{coerce})$.

2.1.5. Other Package Functions

The `ordinalbayes` function yields an object of class `ordinalbayes`. Generic functions have been specifically tailored to extract meaningful results from the resulting MCMC chain. The `print` function returns several summaries from the MCMC output for each parameter monitored including: the 95th lower confidence limit for the highest posterior density (HPD) credible interval (Lower95), the median value (Median), the 95th upper confidence limit for the HPD credible interval (Upper95), the mean value (Mean), the sample standard deviation (SD), the mode of the variable (Mode), the Monte Carlo standard error (MCerr,) percent of SD due to MCMC (MC%ofSD), effective sample size (SSeff), autocorrelation at a lag of 30 (AC.30), and the potential scale reduction factor (psrf). The `plot` function provides a trace of the sampled output and optionally the density estimate for each variable in the chain. This function additionally adds the appropriate `beta` and `gamma` labels for each penalized variable name.

When identifying important genomic features, the regression-based variable inclusion indicator, normal spike-and-slab, and double exponential spike-and-slab Bayesian ordinal models all incorporate a variable inclusion indicator, γ_j, in the model. Variable selection can be based on whether the posterior mean of γ_j exceeds a pre-specified threshold. Alternatively, we can use Bayes factor to test the hypotheses $H_{0j} : \gamma_j = 0$ versus $H_{aj} : \gamma_j = 1$, where the null hypothesis is rejected for feature j if Bayes factor exceeds a pre-specified threshold. For the LASSO, normal spike-and-slab, and double exponential spike-and-slab Bayesian ordinal models, Bayes factor can be used to test an interval null hypothesis $H_{0j} : |\beta_j| \leq \epsilon$ versus $H_{aj} : |\beta_j| > \epsilon$, where ϵ is a small positive value that is close to 0. For the regression-based variable inclusion indicator Bayesian ordinal model, Bayes factor can be used to test $H_{0j} : |\gamma_j \beta_j| \leq \epsilon$ versus $H_{aj} : |\gamma_j \beta_j| > \epsilon$. Note that for the Bayesian LASSO, no variable inclusion indicators are incorporated so variable selection can only be performed using Bayes Factor for β. The `summary` function requires an `ordinalbayes` object and the user can specify `epsilon` (default 0.1) for testing the null hypothesis that $H_{0j} : |\beta_j| \leq \epsilon$. The output from `summary` is a list containing the following components: `alphamatrix`, the MCMC output for the threshold parameters; `betamatrix`, the MCMC output for the penalized parameters; `zetamatrix`, The MCMC output for the unpenalized parameters (if included); `gammamatrix`, the MCMC output for the variable inclusion parameters (not available when `model = "lasso"`); `gammamean`, the posterior mean of the variable inclusion indicators (not available when `model = "lasso"`); `gamma.BayesFactor`, Bayes factor for the variable inclusion indicators (not available when `model = "lasso"`); `Beta.BayesFactor`, Bayes factor for the penalized parameters; and `lambdamatrix`, the MCMC output for the penalty parameter (not available when `model="normalss"`). The `coef` function also accepts an `ordinalbayes` object and returns a function (default is `method=mean`) of the posterior distribution of the penalized parameter estimates and variable inclusion indicators.

The `predict` function acccepts an `ordinalbayes` object and optionally allows to user to specify new data for unpenalized predictors and the penalized predictors, by invoking `neww =` and `newx =`, respectively. If `neww` and `newx` are not supplied, the original data are used for prediction. The `model.select` parameter allows the user to obtain model predictions through one of three different methods. When `model.select = "average"` (default), the mean coefficient values over the MCMC chain are used to estimate fitted

 doi:10.20944/preprints202203.0182.v1

probabilities; the predicted class is that attaining the maximum fitted probability. When `model.select = "median"`, the median coefficient values over the MCMC chain are used to estimate fitted probabilities; the predicted class is that attaining the maximum fitted probability. When `model.select = "max.predicted.class"`, each step in the chain is used to calculate fitted probabilities and the class, then the final predicted class is taken as that class that is most frequently predicted. The function `fitted` is synonymous with `predict`.

2.2. Analysis of Cervical Cancer Dataset

We downloaded the transcript-level HTSeq count data for the 309 subjects from the The Cancer Genome Atlas Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA-CESC) project [22] having transcriptome profiling performed using the TCGAbiolinks Bioconductor package [21]. We then restricted attention to the 253 cervical cancer subjects with a primary diagnosis of squamous cell carcinoma. Subsequently, we removed one subject whose sample was FFPE preserved, one subject with metastatic disease, two subjects who contributed only solid normal tissue, and seven subjects lacking FIGO stage. This left 242 subjects is Stage I ($N = 124$), II ($N = 61$), and III-IV ($N = 57$). Using the `DESeq2` Bioconductor package [20], we performed differential expression analysis using stage as the independent predictor in the negative binomial model. We then applied the regularized log transformation to robustly transform the count data to a $\log_2$ scale to stabilize the variance, and then filtered the resulting dataset to retain transcripts having a mean expression > 0.5 and FDR< 0.10 from the stage I versus stages III/IV contrast.

We fit a regression-based variable inclusion indicator Bayesian ordinal model using a Beta$(0.01, 0.19)$ hyperprior for the π_j using the `runjags` package to run three parallel chains with 5,000 burn-in, 5,000 tuning steps, and thinned to keep every third step in the sampling process to reduce auto-correlation in our posterior samples, and kept 9,999 saved steps per chain. Convergence was assessed using Gelman and Rubin's potential scale reduction factor (PSRF).

3. Results

There were 1,137 transcripts that were differentially expressed at a Benjamini-Hochberg FDR< 0.05 and 2,009 transcripts that were differentially expressed at a Benjamini-Hochberg FDR< 0.10 when examining the contrast between stage I and stages III/IV. These 2,009 transcripts were retained for Bayesian modeling. Forty transcripts had a Bayes factor > 4 when testing $H_{0j} : |\gamma_j \beta_j| \leq 0.1$ versus $H_{aj} : |\gamma_j \beta_j| > 0.1$. Forty-one transcripts had a Bayes factor > 5 when testing $H_{0j} : \gamma_j = 0$ versus $H_{aj} : \gamma_j = 1$ (Table 1). Notably, the features were the same with exception that Bayes factor testing $\gamma_j = 0$ additionally identified ENSG00000115548 (Gene symbol *KDM3A*).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2022 doi:10.20944/preprints202203.0182.v1

Table 1. Transcripts significant from the regression-based variable inclusion indicator Bayesian ordinal model when testing $H_{0j} : \gamma_j = 0$ versus $H_{aj} : \gamma_j = 1$ using Bayes Factor and a threshold of 4. Annotation information obtained from https://www.ncbi.nlm.nih.gov/gene, https://www.genecards.org, and https://lncipedia.org.

Ensemble ID	Gene symbol	Chr	$\bar{\gamma}$
ENSG00000076344	*RGS11*	16	0.179
ENSG00000077274	*CAPN6*	X	0.264
ENSG00000101888	*NXT2*	X	0.194
ENSG00000115548	*KDM3A*	2	0.174
ENSG00000122884	*P4HA1*	10	0.186
ENSG00000125430	*HS3ST3B1*	17	0.286
ENSG00000131370	*SH3BP5*	3	0.175
ENSG00000135443	*KRT85*	12	0.334
ENSG00000136457	*CHAD*	17	0.179
ENSG00000138398	*PPIG*	2	0.240
ENSG00000150636	*CCDC102B*	18	0.281
ENSG00000161277	*THAP8*	19	0.283
ENSG00000163510	*CWC22*	2	0.301
ENSG00000164485	*IL22RA2*	6	0.196
ENSG00000164651	*SP8*	7	0.231
ENSG00000166091	*CMTM5*	14	0.215
ENSG00000166342	*NETO1*	18	0.197
ENSG00000171121	*KCNMB3*	3	0.186
ENSG00000177173	Pseudogene, parent *NAP1L4P1*	1	0.258
ENSG00000180229	*HERC2P3*	15	0.196
ENSG00000188817	*SNTN*	3	0.236
ENSG00000197360	*ZNF98*	19	0.214
ENSG00000203601	LINC00970	1	0.316
ENSG00000225449	*RAB6C-AS1*	2	0.235
ENSG00000230201	Pseudogene, parent *ATP6V0CP1*	17	0.286
ENSG00000233996	Pseudogene, parent *KDM3AP1*	2	0.248
ENSG00000236138	*DUX4L26*	3	0.247
ENSG00000236819	LINC01563	17	0.311
ENSG00000250602	lnc-ALDH7A1-1	5	0.246
ENSG00000253923	Pseudogene, parent *HSPE1*	8	0.302
ENSG00000256980	*KHDC1L*	6	0.207
ENSG00000259083	lnc-TRAPPC6B-1	14	0.263
ENSG00000259134	LINC00924	15	0.352
ENSG00000260484	lnc-OPRK1-2	8	0.263
ENSG00000263612	lnc-ZNF517-4	8	0.228
ENSG00000264049	MIR4737	17	0.266
ENSG00000264954	*PRR29-AS1*	17	0.221
ENSG00000265579	lnc-CBLN2-1	18	0.227
ENSG00000271711	Pseudogene, parent *SAP30*	3	0.264
ENSG00000272071	lnc-PAPD7-2	5	0.279
ENSG00000276517	Lnc-TTC27-9	2	0.221

Many genes listed in Table 1 are relevant to cervical cancer, related cancers of the female reproductive system, or cancer in general. For example, in a tissue-based study, *CAPN6* was not detected in normal cervical squamous epithelium but its expression was observed in low-grade and increased further in high-grade squamous cervical intraepithelial lesions [23]. *KDM3A* is an epigenetic regulator that has been found to be highly expressed in cervical cancer tissues and involved in cervical cancer progression [24]. *P4HA1* was included in a five-gene signature to predict cervical cancer prognosis [25]. A previous study suggested that *CMTM5* is a tumor suppressor that is frequently methylated and thus loses function in cancer [26], including cervical cancer [27]. *RAB6C* has been shown to be aberrantly methylated in cervical cancer compared to normal tissues [28]. *ALDH7A1* was among 30 genes that demonstrated a dose-response pattern with NNK, a tobacco

 doi:10.20944/preprints202203.0182.v1

carcinogen, in cervical cancer samples [29], implicating tobacco may be a causative factor in cervical cancer development in addition to HPV infection.

Other genes, while not yet described in cervical cancer, have been found to be prognostic in ovarian cancer (*RGS11* [30], *CHAD* and *CBLN2* [31], *NETO1* [32], *HSPE1* [33], and *BIRC6* which Lnc-TTC27-9 is intronic to [34]). Expression of *SH3BP5* is reduced in ovarian cancer samples compared to normal tissue and that silencing of Sab protein expression may lead to chemo-resistance [35]. Expression of *SNTN* has high discriminatory power to differentiate between normal tissue, serous borderline ovarian tumors, and serous ovarian carcinoma [36]. *IL22RA2* is highly expressed in various tissues including those in the female reproductive system [37]. With respect to genes associated with other cancers, *NXT2* was among 12 genes used to define prognostic risk groups in melanoma [38]. A review article described that aberrant expression of *HS3ST3B1* is observed in many cancers and the authors posited that *HS3ST3B1* may act as a tumor-promoting enzyme [39]. Expression of *KRT85* was found to be associated with overall survival in subjects with colon cancer [40].

When using the fitted model using the 2,009 transcripts only 16.9% of subjects were misclassified, with all misclassifications in Stage II. However, when fitting a parsimonious model including only the 41 transcripts in Table 1, the misclassification rate decreased to 11.6%.

4. Discussion

The `ordinalbayes` package is capable of fitting penalized ordinal Bayesian cumulative logit models to high-dimensional datasets. The package includes methods for monitoring the mixing of chains (`plot`) and convergence (`print`). It also includes a `summary` function that permits the user to estimate Bayes factor for testing an interval null hypothesis for β_j and for testing the null that $\gamma_j = 0$, to assist the user with variable selection. The `coef` function uses the posterior distribution to return summary estimates of the penalized β_j and the γ_j indicators. The `predict` (or equivalently, `fitted`) can be used to obtain the estimated class probabilities as well as the predicted class for each observation.

When applied to The Cancer Genome Atlas cervical cancer dataset, predictive performance was excellent. When restricting attention to only the 41 transcripts having Bayes Factor > 4, predictive performance yielded an overall misclassification error 11.6%, though misclassification error increased from 0% for Stage I and III/VI in the full model to 3.2% and 14.0%, respectively, in the reduced model. Interestingly, transcripts that were identified have known associations with cervical cancer, cancers of the female reproductive system, and other cancer in general. The syntax we used to analyze this dataset appears in the Appendix.

Author Contributions: Conceptualization, K.J.A.; methodology, Y.Z. and A.E.S.; software, K.J.A., Y.Z., and A.E.S.; validation, A.E.S., Y.Z., and S.S.; formal analysis, K.J.A.; writing—original draft preparation, K.J.A.; writing—review and editing, A.E.S., Y.Z., and S.S.; funding acquisition, K.J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Cancer Institute of the National Institutes of Health under Award Number R03CA245771. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The TCGA-CESC high-throughput gene expression data can be downloaded using the TCGAbiolinks Bioconductor package.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2022 doi:10.20944/preprints202203.0182.v1

MDPI	Multidisciplinary Digital Publishing Institute
FIGO	International Federation of Gynecology and Obstetrics
TCGA	The Cancer Genome Atlas Project
CRAN	Comprehensive R Archive Network
HPV	Human Papilloma Virus
LASSO	Least Absolute Shrinkage and Selection Operator
MCMC	Markov Chain Monte Carlo
DE	Double exponential
PSRF	potential scale reduction factor
HTSeq	Hign-throughput sequencing
FFPE	Formalin-Fixed Paraffin-Embedded
FDR	False discovery rate

Appendix A

Appendix A.1

The data used in this example are stored in the `finalSet` object. Because this object was derived using the `DESeq2` BioConductor package, we load it first. Please note that due to the use of the default parameters for the number of saved steps per chain (9,999) and the large size of this dataset, the model took 3.2 days to run on a 13 inch MacBook Pro with four cores and 16GB RAM. For those interested in running examples using this package, a smaller version of these data, `reducedSet` which includes the 41 transcripts, may be used instead. Alternatively, parameters relating to the number of steps can be reduced.

The regression-based variable inclusion model with random prior to π was fit after loading the `ordinalbayes R` using the syntax:

```
library("DESeq2")
library("ordinalbayes")
data(finalSet)
fitted.regressvi.random<-ordinalbayes(Stage~1, data=colData(finalSet),
     x=t(assay(finalSet)), model="regressvi",
     gamma.ind="random", c.gamma=0.01, d.gamma=0.19, seed=26)
```

You can evaluate various aspects of the MCMC results of the `ordinalbayes` object by issuing the `print` command.

```
print(fitted.regressvi.random)
```

including the `psrf` to assess model convergence. Please note that to foster reproducibility of our output, we set the random seed. Subsequent runs using different seeds will produce different results due to the random nature of the MCMC sampling.

To summarize the fitted model object,

```
summary.model.fit<-summary(fitted.regressvi.random)
```

To identify which transcripts had a Bayes factor > 4 when testing $H_{0j} : |\gamma_j \beta_j| \leq 0.1$ versus $H_{aj} : |\gamma_j \beta_j| > 0.1$,

```
names(which(summary.model.fit$Beta.BayesFactor>4))
```

Similarly, to identify which transcripts had a Bayes factor > 4 when testing $H_{0j} : \gamma_j = 0$ versus $H_{aj} : \gamma_j = 1$,

```
names(which(summary.model.fit$gamma.BayesFactor>4))
```

To obtain the $\bar{\gamma}$ estimates we used the following code:

```
coefficients<-coef(fitted.regressvi.random)
coefficients$gamma[which(summary.model.fit$gamma.BayesFactor>4)]
```

To obtain model predictions,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2022 doi:10.20944/preprints202203.0182.v1


```
phat<-predict(fitted.regressvi.random)
table(phat$class, colData(finalSet)$Stage)
```

```
     1   2   3
  1 124  28   0
  2   0  20   0
  3   0  13  57
```

To determine the adequacy of a more parsimonious model, we then restricted attention to 41 transcripts having `gamma.BayesFactor>4`. The `reducedSet` object is provided in the `ordinalbayes` package though due to the random nature of the MCMC sampling, the number of transcripts having Bayes Factor for γ could differ so we demonstrate how we derived our object.

```
reducedSet<-finalSet[which(summary.model.fit$gamma.BayesFactor>4),]
fitted.regressvi.reduced<-ordinalbayes(Stage~1, data=colData(reducedSet),
     x=t(assay(reducedSet)), model="regressvi",
     gamma.ind="random", c.gamma=100, d.gamma=1, seed=26)
```

Because we were using `gamma.ind="random"`, we changed the parameter values for the variable inclusion indicator hyperprior to `c.gamma=100, d.gamma=1` ensure virtually all transcripts would be included in each model. If fitting a model using `gamma.ind="fixed"`, the hyperprior `pi.fixed=0.99` would accomplish the same thing. This smaller model only took 9.1 minutes to complete.

```
phat.reduced<-predict(fitted.regressvi.reduced)
table(phat.reduced$class, colData(reducedSet)$Stage)
```

```
     1   2   3
  1 120   9   1
  2   4  45   7
  3   0   7  49
```

This more parsimonious model that included 41 transcripts had a misclassification rate of 11.6%. The class-specific misclassification rates [Stage I (3.2%), Stage II (26.2%), Stage III/IV (14.0%)] may indicate that smaller classes are more difficult to predict .

Table A1. `ordinalbayes` parameters available for all models.

Parameter	Description and default values
`alpha.var`	Variance for α_k in the MCMC chain (default 10)
`coerce.var`	Variance associated with any unpenalized predictors in the MCMC chain (default 10)
`adaptSteps`	Number of iterations for adaptation (default 5,000)
`burnInSteps`	Number of iterations of the Markov chain to run (default 5,000)
`nChains`	Number of parallel chains to run (default 3)
`numSavedSteps`	Number of saved steps for each chain (default 9,999)
`thinSteps`	The thinning interval for monitors (default 3)
`parallel`	Run the MCMC on multiple processors (default `TRUE`)
`model`	Specify which penalized ordinal model to fit (default `regressvi`)
`center`	If `TRUE` (default), center the variables to be penalized in the model
`scale`	If `TRUE` (default), scale the variables to be penalized in the model
`seed`	An integer value for the random seed to ensure reproducibility
`quiet`	If `TRUE`, suppress output of JAGS (or rjags) when updating models (default `FALSE`)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2022 doi:10.20944/preprints202203.0182.v1

Table A2. `ordinalbayes` parameters for each penalized ordinal Bayesian model.

model	Parameters in ordinalbayes call to specify	Description
lasso	`a, b`	The penalty parameter λ ~Gamma(`a, b`) (default `a = 0.1, b = 0.1`)
normalss	`sigma2.0`	The variance for the spike (set to some small positive value, e.g. 0.01)
	`sigma2.1`	The variance for the slab (set to some large positive value, e.g. 10)
	`gamma.ind="fixed", pi.fixed`	Use a constant prior for π_j of `pi.fixed` (default 0.05)
	`gamma.ind="random", c.gamma, d.gamma`	Use a random prior for π_j ~Beta(`c.gamma, d.gamma`), for example, `c.gamma = 0.01, d.gamma = 0.19`.
dess	`a, b`	The penalty parameter λ ~Gamma(`a, b`) (default `a = 0.1, b = 0.1`)
	`lambda0`	The parameter value for the spike, e.g. `lambda0 = 20`
	`gamma.ind="fixed", pi.fixed`	Use a constant prior for π_j of `pi.fixed` (default 0.05)
	`gamma.ind="random", c.gamma, d.gamma`	Use a random prior for π_j ~Beta(`c.gamma, d.gamma`), for example, `c.gamma = 0.01, d.gamma = 0.19`.
regressvi	`a, b`	The penalty parameter λ ~Gamma(`a, b`) (default `a = 0.1, b = 0.1`)
	`gamma.ind="fixed", pi.fixed`	Use a constant prior for π_j of `pi.fixed` (default 0.05)
	`gamma.ind="random", c.gamma, d.gamma`	Use a random prior for π_j ~Beta(`c.gamma, d.gamma`), for example, `c.gamma = 0.01, d.gamma = 0.19`.

References

1. Vu, M.; Yu, J.; Awolude, O.A.; Chuang, L. Cervical cancer worldwide. *Current Problems in Cancer* **2018**, *42*, 457–465.
2. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **2021**, *71*, 209–249.
3. Prat, J. Ovarian, fallopian tube and peritoneal cancer staging: rationale and explanation of new FIGO staging 2013. *Best Practice & Research Clinical Obstetrics & Gynaecology* **2015** *29*, 858–869.
4. Cohen, P.A.; Jhingran, A.; Oaknin, A; Denny, L. Cervical cancer. *The Lancet* **2019**, *393*, 169–182.
5. Colombo, N.; Sessa, C.; du Bois, A.; Ledermann, J.; McCluggage, W.G.; McNeish, I.; Morice, P.; Pignata, S.; Ray-Coquard, I.; Vergote, I.; Baert, T.; Belaroussi, I.; Dashora, A.; Olbrecht, S.; Planchamp, F.; Querleu, D. ESMO-ESGO consensus conference recommendations on ovarian cancer: pathology and molecular biology, early and advanced stages, borderline tumours and recurrent disease. *Annals of Oncology* **2019**, *30*, 672–705.
6. Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodology)* **1996**, *58*, 267–288.
7. Zhu, J.; Hastie, T. Classification of gene microarrays by penalized logistic regression. *Biostatistics* **2004**, *5*, 427–443.
8. Archer, K.J.; Williams, A.A.A. L_1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine* **2012**, *31*, 1464–1474.
9. Archer, K.J.; Hou, J.; Zhou, Q.; Ferber, K.; Layne, J.G.; Gentry, A.E. ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer Informatics* **2014**, *13*, 187–195.
10. Wurm, M.J.; Rathouz, P.J.; Hanlon, B.M. Regularized ordinal regression and the ordinalNet R package. *Journal of Statistical Software* **2021**, *99*, 1–42.

11. Yi, N.; Xu, S. Bayesian LASSO for quantitative trait loci mapping. *Genetics* **2008**, *179*, 1045–1055.
12. Hans, C. Bayesian lasso regression. *Biometrika* **2009**, *96*, 835–845.
13. Li, J.; Das, K.; Fu, G.; Li, R.; Wu, R. The Bayesian lasso for genome-wide association studies. *Bioinformatics* **2011**, *27*, 516–523.
14. Lykou, A.; Ntzoufras, I. On Bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing* **2013**, *23*, 361–390.
15. Biswas, S.; Xia, S.; Lin, S. Detecting rare haplotype-environment interaction with logistic Bayesian LASSO. *Genetic Epidemiology* **2013**, *38*, 31–41.
16. Biswas, S.; Papachristou, C. Evaluation of logistic Bayesian LASSO for identifying association with rare haplotypes. *BMC Proceedings* **2014**, *8*, 554.
17. Zhang, Y.; Hofmann, J.N.; Purdue, M.P.; Lin, S.; Biswas, S. Logistic Bayesian LASSO for genetic association analysis of data from complex sampling designs. *Journal of Human Genetics* **2017**, *62*, 819–829.
18. Zhang, Y.; Archer, K.J. Bayesian variable selection for high-dimensional data with an ordinal response: identifying genes associated with prognostic risk group in acute myeloid leukemia. *BMC Bioinformatics* **2021**, *22*, 539.
19. Zhang, Y.; Archer, K.J. Bayesian penalized cumulative logit model for high-dimensional data with an ordinal response. *Statistics in Medicine* **2021**, *40*, 1453–1481.
20. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **2014**, *15*, 550.
21. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; Ceccarelli, M.; Bontempi, G.; Noushmehr, H. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **2016**, *44*, e71.
22. The Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature* **2017**, *543*, 378–384.
23. Lee, S-J.; Kim, B-G.; Choi, Y-L.; Lee, J-W. Increased expression of calpain 6 during the progression of uterine cervical neoplasia: immunohistochemical analysis. *Oncology Reports* **2008**, *19*, 859–863.
24. Liu, J.; Li, D.; Zhang, X.; Li, Y.; Ou, J. Histone Demethylase KDM3A Promotes Cervical Cancer Malignancy Through the ETS1/KIF14/Hedgehog Axis. *OncoTargets and Therapy* **2020**, *13*, 11957–11973.
25. Shang, C.; Huang, J.; Guo, H. Identification of an Metabolic Related Risk Signature Predicts Prognosis in Cervical Cancer and Correlates With Immune Infiltration. *Frontiers in Cell and Developmental Biology* **2021**, *9*, 677831.
26. Shao, L.; Cui, Y.; Li, H.; Liu, Y.; Zhao, H.; Wang, Y.; Zhang, Y.; Ng, K.M.; Han, W.; Ma, D.; Tao, Q. CMTM5 exhibits tumor suppressor activities and is frequently silenced by methylation in carcinoma cell lines. *Clinical Cancer Research* **2007**, *13*, 5756–5762.
27. Shao, L.; Guo, X.; Plate, M.; Li, T.; Wang, Y.; Ma, D.; Han, W. CMTM5-v1 induces apoptosis in cervical carcinoma cells. *Biochemical and Biophysical Research Communications* **2009**, *379*, 866–871.
28. Bhat, S.; Kabekkodu, S.P.; Varghese, V.K.; Chakrabarty, S.; Mallya, S.P.; Rotti, H.; Pandey, D.; Kushtagi, P.; Satyamoorthy, K. Aberrant gene-specific DNA methylation signature analysis in cervical cancer. *Tumor Biology* **2017**, *39*, 1010428317694573.
29. Prokopczyk, B.; Sinha, I.; Trushin, N.; Freeman, W.M.; El-Bayoumy, K. Gene expression profiles in HPV-immortalized human cervical cells treated with the nicotine-derived carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. *Chemico-Biological Interactions* **2009**, *177*, 173–180.
30. Hu, Y.; Zheng, M.; Wang, S.; Gao, L.; Gou, R.; Liu, O.; Dong, H.; Li, X.; Lin, B.. Identification of a five-gene signature of the RGS gene family with prognostic value in ovarian cancer. *Genomics* **2021**, *113*, 2134–2144.
31. Belotti, Y.; Lim, E.H.; Lim, C.T. The Role of the Extracellular Matrix and Tumor-Infiltrating Immune Cells in the Prognostication of High-Grade Serous Ovarian Cancer. *Cancers* **2022**, *14*, 404.
32. Xu, Y.; Wang, W.; Chen, J.; Mao, H.; Liu, Y.; Gu, S.; Liu, Q.; Xi, Q.; Shi, W. High neuropilin and tolloid-like 1 expression associated with metastasis and poor survival in epithelial ovarian cancer via regulation of actin cytoskeleton. *Journal of Cellular and Molecular Medicine* **2020**, *24*, 9114–9124.
33. Ralph, S.; Brenchley, P.E.C.; Summers, A.; Rosa, D.D.; Swindell, R.; Jayson, G.C. Heparanase gene haplotype (CGC) is associated with stage of disease in patients with ovarian carcinoma. *Cancer Science* **2007**, *98*, 844–849.
34. Wang, L.; Chen, Y-J.; Hou, J.; Wang, Y-Y.; Tang, W-Q.; Shen, X-Z.; Tu, R-Q. Expression and clinical significance of BIRC6 in human epithelial ovarian cancer. *Tumor Biology* **2014**, *35*, 4891–4896.
35. Paudel, I.; Herrnandez, S.M.; Portalatin, G.M.; Chambers, T.P.; Chambers, J.W. Sab concentrations indicate chemotherapeutic susceptibility in ovarian cancer cell lines. *Biochemical Journal* **2018**, *475*, 3471–3492.
36. Park, J.S.; Choi, S.B.; Kim, H.J.; Cho, N.H.; Kim, S.W.; Kim, Y.T.; Nam, E.J.; Chung, J.W.; Kim, D.W. Intraoperative Diagnosis Support Tool for Serous Ovarian Tumors Based on Microarray Data Using Multicategory Machine Learning. *International Journal of Gynecological Cancer* **2016**, *26*, 104–113.
37. Xu, W.; Presnell, S.R.; Parrish-Novak, J.; Kindsvogel, W.; Jaspers, S.; Chen, Z.; Dillon, S.R.; Gao, Z.; Gilbert, T.; Madden, K.; Schlutsmeyer, S.; Yao, L.; Whitmore, T.E.; Chandrasekher, Y.; Grant, F.J.; Maurer, M.; Jelinek, L.; Storey H.; Brender, T.; Hammond, A.; Topouzis, S.; Clegg, C.H.; Foster, D.C. A soluble class II cytokine receptor, IL-22RA2, is a naturally occurring IL-22 antagonist. *Proceedings of the National Academy of Sciences U.S.A.* **2001**, *98*, 9511–9516.
38. Wu, L.; Hu, X.; Dai, H.; Chen, K.; Liu, B. Identification of an m6A Regulators-Mediated Prognosis Signature For Survival Prediction and Its Relevance to Immune Infiltration in Melanoma. *Frontiers in Cell and Developmental Biology* **2021**, *9*, 718912.

 doi:10.20944/preprints202203.0182.v1

39. Denys, A.; Allain, F. The Emerging Roles of Heparan Sulfate 3- O-Sulfotransferases in Cancer. *Frontiers in Oncology* **2019**, *9*, 507.
40. Luo, Y.; Sun, F.; Peng, X.; Dong, D.; Ou, W.; Xie, Y.; Luo, Y. Integrated Bioinformatics Analysis to Identify Abnormal Methylated Differentially Expressed Genes for Predicting Prognosis of Human Colon Cancer. *International Journal of General Medicine* **2021**, *14*, 4745–4756.