STATISICAL MODELS FOR READING COUNT DATA

by

Minh Thu Bui

Submitted in partial fulfillment of the

requirements for Departmental Honors in

the Department of Mathematics

Texas Christian University

Fort Worth, Texas

May 2, 2022

STATISTICAL MODELS FOR READING COUNT DATA

Project Approved:

Supervising Professor: Cornelis Potgieter, Ph.D.

Department of Mathematics

Ken Richardson, Ph.D.

Department of Mathematics

Bingyang Wei, Ph.D.

Department of Computer Science

ABSTRACT

This thesis considers parameter estimation for different statistical models used on count data. The motivating data consists of multiple independent count variables with a moderate sample size per variable. The data were collected during the assessment of oral reading fluency (ORF) in school-aged children. A sample of fourth-grade students were given one of ten available passages to read with these differing in length and difficulty. The observed number of words read incorrectly (WRI) is used to measure ORF. Five models are considered for WRI scores, namely the binomial, the Poisson, the zero-inflated binomial, the zero-inflated Poisson, and the beta-binomial distributions. We aim to efficiently estimate passage difficulty, a quantity expressed as a function of the underlying model parameters. In addition to considering ordinary maximum likelihood, two types of penalty functions are considered for penalized likelihood. The goal of shrinkage is to encourage parameter estimates either closer to zero or closer to one another. A simulation study evaluates the efficacy of the shrinkage estimates using Mean Square Error (MSE) as metric. Big reductions in MSE relative to unpenalized maximum likelihood are observed. The thesis concludes with an analysis of the motivating ORF data.

Statistical Models for Reading Count Data

Minh Thu Bui

April 15, 2022

1 Introduction

1.1 Measuring Oral Reading Fluency

The definition of Oral Reading Fluency (ORF) is "the oral translation of text with speed and accuracy," see for example Fuchs et al. (2001) and Shinn et al. (1992). Reading fluency is a skill developed during childhood that is needed to understand the meaning of texts and literary pieces. There is a strong correlation between reading fluency and reading comprehension, see Allington (1983); Johns and Lunn (1983); Samuels (1988); Schreiber (1991). According to DiSalle and Rasinski (2017), once a student has identified a word and read it correctly, their focus generally shifts from word recognition (attempting to recognize the word) to comprehension (making meaning of the word). This leads to overall understanding of the text. These authors have claimed that the incompetent ORF levels are the cause of up to 90% of reading fluency issues. If a child is not fluent in their reading, their ability to read comprehensively is hindered and they will have trouble in grasping the meaning of texts. Thus, ORF is a method of evaluating whether a child is at their appropriate reading level compared to their peers and provides a quantifiable score to identify at-risk students with poor reading skills.

In this thesis, ORF data collected from a sample of 508 fourth-grade students is analyzed. Each student was given one randomly selected passage (out of ten available) to read and the number of words read incorrectly (WRI) was recorded. This resulted in a maximum of 53 and a minimum of 49 observations per passage. Moreover, their lengths vary from three to five sentences each, with a median of 51 words, and the minimum is 44 words while the maximum is 69 words. Reading sessions were recorded so that observer error in counting the number of words read correctly and incorrectly can be minimized. Strong readers tend to have low WRI scores and weak readers tend to have high WRI scores. However, as the passages may not all be equal in difficulty, it is important to be cautious in using WRI scores obtained from different passages to measure overall ORF levels in a classroom setting.

For the 10 reading passages, $Count_j$ denotes a vector that contains the WRI data for the studets who read passage j, $j = 1, 2, ..., 10$; each data value is the number of the words a specific student got wrong while reading the passage. Furthermore, the integer N_j indicates the number of words j passage and n_j denotes

the number of students who read passage j.

For the initial analysis of ORF data, we considered five different models on the count data. These were the binomial, Poisson, beta-binomial, zero-inflated binomial, and zero-inflated Poisson distributions. The reason why we choose these five models are as follow:

1. The binomial model is easy to interpret as it is expressed directly in terms of the success probability, with a success being represented the participant reading the words *incorrectly*. Furthermore, the binomial distribution naturally conforms to the bounds on the data, i.e. the length of the passage.

2. Poisson model is a great fit to measure probability of rare events, in which the WRI scores tend to be small. However, the Poisson distribution does not preserve the natural range of the data, i.e. the model potentially allows more WRI than the passage length.

3. Beta-binomial distribution generalizes the binomial distribution by allowing the success probability to vary to each word in the passage.

4. Zero-inflated models are designed for datasets where many observations are equal to 0. In the data being considered, it is frequently the case where the passage is read without any errors so the WRI score is 0. We show in the next subsection how zero inflation can be applied to both the binomial and Poisson models.

In the remainder of chapter 1, we give an overview of each of the model mentioned above. In chapter 2, we review the technique of maximum likelihood and measure how well the proposed models fit the WRI data by both visual inspection and the Akaike Information Criterion (AIC). In chapter 3, we consider modified estimation by introducing bias through the use of penalty functions. The introduction of bias has the potential to improve parameter estimation when considering mean squared error (MSE) as a criterion. Next, chapter 4 considers the practical implementation of penalized estimation and then chapter 5 presents some simulation results. Furthermore, chapter 6 illustrates the final data analysis and conclusions following in chapter 7.

1.2 Models' Origins and Probability Mass Function

The first model being considered is the binomial distribution, which was formulated by a Swiss mathematician Jakob Bernoulli in 1713, see Routledge (2018). It became a much more widely used statistical method after Ronald Fisher published his work using the binomial distribution in 1936. Binomial distribution considers data with a fixed number of trials with the probability of either success or failure. In our thesis, the probability of success is interpreted as the expected proportion of Words Read Incorrectly (WRI). The model pmf is

$$f_j(x) = \binom{N_j}{x} p_j^x (1-p_j)^{N_j - x} \text{ with } x = 0, 1, ..., N_j.$$

Note that p_j denotes the success probability (probability of getting a word wrong) and N_j is the length of the j passage.

The Poisson model was developed by the French mathematician Siméon-Denis Poisson in 1830 to describe the number of times a gambler would win a rarely won game of chance in a large number of tries, see Routledge (2020). Poisson distribution is appropriate for counting rare events over a period of time. This is an approximation in present setting since the Poisson distribution does not have an upper bound. However, a majority of students read most words correctly and WRI can be considered a rare event count. The model pmf is

$$f_j(x) = \frac{\lambda_j^x e^{\lambda j}}{x!} \text{ with } x = 0, 1, 2, ...$$

with λ_j denoting the mean number of words read incorrectly for passage j.

We also consider the beta-binomial distribution, a generalization of the binomial distribution that does not assume a fixed success probability. The beta-binomial distribution is relevant to the WRI context as not all words in the passage are equally difficult to read. For more background, see Griffiths (1973). The pmf is given by

$$f_j(x) = \binom{N_j}{x} \frac{B(x + \alpha_j, N_j - x + \beta_j)}{B(\alpha_j, \beta_j)} \quad \text{with } x = 0, 1, ..., N_j$$

where $B(x, y)$ denotes the Beta function. In this model, the success probabilities are assumed to be drawn from a beta distribution, a continuous distribution that takes values on the interval [0,1] and depends on two parameters, $\alpha_j > 0$ and $\beta_j > 0$.

Next, regarding the zero-inflated models, let X denote a count variable (for example Poisson or binomial), and let $f(x)$ denote the pmf of that distribution. The zero-inflated version, denoted Y, behaves like the original distribution X but it has proportionally more zeros than X. There has been much research that deals with data that containing excess zero values. One of the first to use is Cohen (1967). Now, let $\gamma \in [0,1]$ denote the proportion of excess zeros. The pmf of Y can be written in terms of the pmf of X as

$$g_Y(y) = \begin{cases} \gamma + (1 - \gamma) f_X(0), & \text{if } y = 0 \\ (1 - \gamma) f_X(y), & \text{if } y \geq 1 \end{cases}$$

We consider zero-inflated versions of both of binomial and Poisson distributions. Given the pmf of all the distribution models being considered, we now explore how to find the model that best represents our dataset using maximum likelihood estimation in the next section.

2 Data Analysis and Maximum Likelihood Estimators

2.1 Maximum Likelihood Estimation

Maximum likelihood estimation is a commonly used technique that helps us evaluate the parameters that describe our dataset. Our goal is to estimate a parameter (or sets of parameters) called θ, assuming θ is the

parameter(s) of interest. Then, maximum likelihood follows a procedure of (1) constructing the likelihood function $f(x|\theta)$, then (2) evaluating the log-likelihood function, which is the natural log of $f(x|\theta)$, and (3) setting the derivative of $f(x|\theta)$ equal to 0 to solve for the best-fitted parameter(s), $\hat{\theta}$.

Below, the method of maximum likelihood estimation is illustrated using the Poisson distribution as reference.

Example 1. Let x_1, x_2,..., x_n denote a random sample drawn from the population. Construct a log-likelihood function of a Poisson(λ) population. The likelihood function of a Poisson(λ) distribution is

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_j} e^{-\lambda}}{x_i!}.$$

The log-likelihood is defined as

$$\begin{aligned}
l(\lambda) &= \ln\left(\mathcal{L}(\lambda)\right) \\
&= \ln\left(\prod_{i=1}^{n} \frac{\lambda^{x_j} e^{-\lambda}}{x_i!}\right) \\
&= \sum_{i=1}^{n} \ln\left(\frac{\lambda^{x_j} e^{-\lambda}}{x_i!}\right) \\
&= \sum_{i=1}^{n} [\ln(e^{-\lambda}) - \ln(x_i!) + \ln(\lambda^{x_i})] \\
&= \sum_{i=1}^{n} [-\lambda - \ln(x_i!) + x_i \ln(\lambda)] \\
&= -n\lambda - \sum_{i=1}^{n} \ln(x_i!) + \ln\lambda \sum_{i=1}^{n} x_i.
\end{aligned}$$

Thus, the log-likelihood of a Poisson(λ) is $l(\lambda) = -n\lambda - \sum_{i=1}^{n} \ln(x_i!) + \ln\lambda \sum_{i=1}^{n} x_i$. Next, we take a derivative of the log-likelihood functions. In particular, we find the MLE $\hat{\theta}$ by solving the equation $l'(\lambda) = 0$.

$$\begin{aligned}
\frac{dl}{d\lambda} &= (-n\lambda - \sum_{i=1}^{n} ln(x_i!) + ln\lambda \sum_{i=1}^{n} x_i)' \\
&\Rightarrow -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i = 0 \\
&\Leftrightarrow \frac{1}{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} \\
&\Leftrightarrow \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}
\end{aligned}$$

Hence, the MLE of λ is $\hat{\lambda} = \bar{x}$. If we take a closer look at the Poisson model, we can see that it has a closed-form solution. However, this is not always the case for every statistical model. For models without a closed-form solution, numerical methods must be used to find the MLE. We demonstrate the numerical

calculations of the MLE calculation here using the R software.

The log-likelihood function is defined below as a R function that requires two inputs, the model parameters and a data object. The function returns the value of the negative log-likelihood. Below is an example for Poisson distribution:

```
poisson_nllh <- function(data, par) {
                x <- data
                lambda <- par
            llh <- sum(dpois(x, lambda, log = TRUE))
                return(-llh)}
```

In this function, `dpois()` calculates the Poisson pmf and the argument `log = TRUE` indicates that these values should be returned on a natural log scale. The input `data` is an array that contains observations while `par` is the model parameter(s) to be optimized. We use similar functions to evaluate the log-likelihoods of other distributions with the only difference being the command that evaluates the pmf. For example, for the binomial distribution, we use `dbinom()` and for beta-binomial `dbbinom()`. In R, we use `optim()` to calculate the maximum likelihood estimator (MLE) for each distribution. By default, `optim()` uses a Nelder-Mead algorithm to minimize the input function. However, if we specify `method = 'BFGS'` then the function will use the Broyden–Fletcher–Goldfarb–Shanno algorithm and numerically evaluate the gradient. Below is a code snippet illustrating this for the Poisson distribution. Note that `Count.data` is the name of our data

```
MLE_poisson <- optim(1,poisson_nllh, data = Count.data)
```

The numeric value 1 in the above code snippet is the initial value used by the optimization routine. There are four outputs: `par, value, count`, and `convergence`. Output `par` shows the MLEs of the parameters, which is $\hat{\lambda}$ in Poisson distribution. Meanwhile, `value` gives us the value of the negative log-likelihood at the MLEs. Also, `count` is a vector that reports the number of calls to the log-likelihood function and the gradient. Finally, `convergence` returns the value 0 if convergence was achieved, while other values indicate potential numerical difficulties.

2.2 Empirical and Model-based Probabilities for Count Data Records

In this section, we visually inspect the data for $Count_1$. Recall that $Count_1$ represents the WRI scores of the first passage. Our visualization will include both empirical and model probabilities.

Empirical probability represents probability values we observe when performing experiments or surveys. In other words, these values can be obtained from the observed dataset. For example, in $Count_1$, the empirical probability of 0 equals the number of times 0 is observed divided by n_1, which is the total number of observations in the passage. Let f_k be the number of times an value k is observed and n be the total number of trials, then the formula to calculate this is

$$\text{EP}_k = \frac{f_k}{n} \quad \text{with } f_k = \#\{x_j = k\}.$$

Then, we also have the model-based probabilities for $Count_1$. Model-based probabilities are the probability values produced by applying the models on the dataset. Thus, we have five model-based probability graphs relative to five statistical models we mentioned. The general formula for this is $f(x|\hat{\theta})$ where $\hat{\theta}$ represents the maximum likelihood estimators of the models. For example, for the binomial distribution, let $\hat{p}_j$ denote the MLE of the success probability p_j in model. Then

$$f_j(x|\hat{p}) = \binom{N_j}{x}(\hat{p}_j)^x(1-\hat{p}_j)^{N_j-x} \quad \text{with } x = 0, 1, ..., N_j.$$

The figure below shows the empirical probabilities and model-based probabilities for various distributions calculated for the $Count_1$ data.

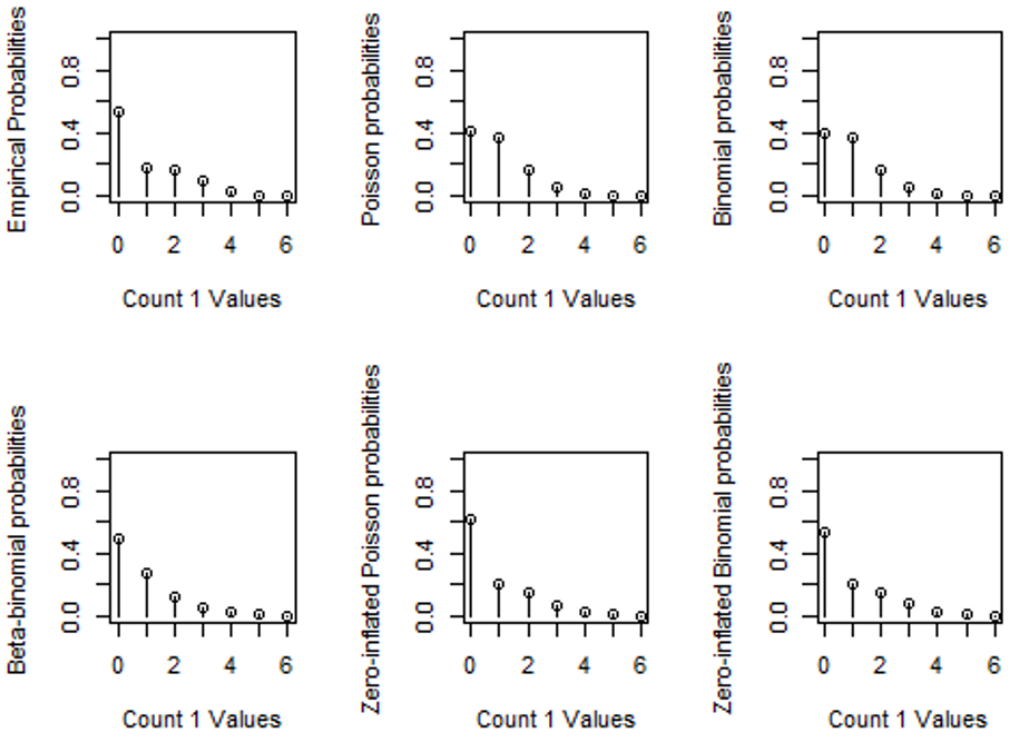

Figure 1: Empirical probabilities of each observation in Count1

Generally, the closer the overall trend of values in the model-based probability models to the empirical graph, the more accurate the statistical distribution is. Here, observe that the zero-inflated and beta-binomial models share a similar trend as the empirical one as there was a significant drop from 0 to 1 and a decreasing trend on the rest of the values. Tentatively, we suspect that the beta-binomial, zero-inflated Poisson, and zero-inflated binomial distribution models may give us the best results, i.e. better parameter estimators. Next, this will be more formally investigated using a tool called the *Akaike Information Criterion* (AIC).

2.3 The Akaike Information Criterion

A formal tool to assess how well the model fits the data we have is to use the Akaike Information Criterion (AIC), which follows the formula of

$$\text{AIC} = 2k - 2\ln\hat{L}$$

with k the number of estimated parameters in the model and $\hat{L}$ the maximum value of the likelihood function for the model. In other words, $\hat{L}$ is the log-likelihood function evaluated at the MLEs. AIC attempts to avoid overfitting by penalizing methods with a large number of parameters. Based on the AIC score, we can determine which distribution is best-suited. A smaller AIC score represents a better model fit.

Table 1 reports the AIC scores when fitting the five statistical distributions to each passage using maximum likelihood. For each passage, the minimum AIC value is printed in bold. AIC should be interpreted in a relative sense. This means the minimum value may not correspond to the true distribution, but indicates which distribution is the most suited out of those considered.

Model Distribution	Count 1	Count 2	Count 3	Count 4	Count 5	Count 6	Count 7	Count 8	Count 9	Count 10
Poisson	134.8	186.9	255.0	145.5	193.3	210.8	215.6	211.3	183.5	246.1
Binomial	135.2	186.2	259.0	143.9	195.4	211.5	218.1	213.3	183.9	250.8
Beta-Binomial	132.2	**175.7**	**201.2**	**143.0**	**170.0**	**187.3**	**167.5**	**180.3**	**159.8**	**189.9**
Zero-inflated Poisson	128.8	180.0	236.1	143.3	191.4	200.6	185.7	197.8	174.5	218.5
Zero-inflated Binomial	**128.6**	180.7	240.9	143.4	195.0	202.5	188.4	200.7	176.1	223.0

Table 1: AIC scores for all five models being considered

With the AIC number we have for each model from table 1 , the beta-binomial distribution model almost always has the smallest AIC scores compared to the other models. In particular, in nine out of ten available passages, the beta-binomial model performs better than others. The one exception is the first passage where the zero-inflated models yield smaller results. However, all other AIC scores are close to one another in $Count_1$, suggesting little difference between the model performances for this passage.

Maximum likelihood estimation relies on the independence of the passages. This means the estimated parameters for one passage are found independently of the parameters of any other passage. However, since all of the passages were created to be similar in the level of difficulty, parameter values should not be too far away from each other. Thus, in order to bring this kind of structure to the parameter estimation process, we use the penalization shrinkage method. Specifically, a penalty function is used to shrink parameter values closer to one another or meet certain "conditions" and the method is called penalized maximum likelihood estimation.

3 Shrinkage through Penalized Likelihood Methods

3.1 Penalized Maximum Likelihood Estimation

Penalized maximum likelihood estimation is a method that allows for the possibility of achieving greater overall accuracy in parameter estimation by allowing for the potential of bias. Intuitively, unbiased estimators are seen as "good". However, if we use MSE as criterion, then we must balance a trade-off between bias and variance.

As we previously noted, the passages in ORF assessment are generally designed to be comparable in difficulty level. Passages are also designed to not be overly challenging for proficient readers. These two passage properties can be incorporated in the estimation of WRI proportions using penalty functions. Specifically, penalty functions are considered that encourage the estimated passage-specific WRI proportions to be close to one another and/or close to zero. The use of parameter shrinkage is further motivated by a small sample size per passage relative to the number of passages. In the next section, we will give a brief overview of some of the topics that have been considered in the literature related to penalized maximum estimation.

3.2 Literature Review

There is, of course, a rich literature on parameter shrinkage in various statistical models. One of the earliest examples is the James-Stein estimator of the mean, see Stein et al. (1956). This estimator is often described as "borrowing" information between variables to obtain a more efficient estimator.

One of the most frequently encountered applications of shrinkage is in regression models with a large number of predictor variables. The lasso, developed by Tibshirani (1996), is one such technique which revolutionized parameter estimation in generalized linear models (GLMs). The lasso shrinks regression parameters towards zero using an L_1 penalty, resulting in predictors being "dropped" from the model by setting the corresponding coefficients equal to zero. The lasso was predated by ridge regression using an L_2 penalty, see Hoerl and Kennard (1970). This approach can result in some regression coefficients being very close to zero, but unlike the lasso does not eliminate potential predictor variables from the model altogether. The monographs Gruber (2017) and Hastie et al. (2019) are very good resources for further exploration of shrinkage in regression models.

In this thesis, the parameters of interest are success proportions (with a success being that a word has been read *incorrectly* during an assessment). Shrinkage as applied to the estimation of proportions has received limited attention in the literature. In the univariate case, (Lemmer, 1981b) considered three different estimators for a binomial success probability, and Lemmer (1981a) proposed estimators of the type $w\hat{p} + (1 - w)p_0$ where p_0 is an *a priori* guess. However, neither of these papers consider likelihood-based methods nor provide guidance on selecting the amount of shrinkage. Hansen (2016) considered three shrinking approaches, namely restricted maximum likelihood, an efficient minimum distance approach, and a projection approach. However, the work of Hansen requires the specification of a "shrinkage direction", which is similar

to the selection of a penalty function.

3.3 Penalty Functions

In this section, we explore the use of penalty functions in a general setting. Assume we observe a sample of $X_1, X_2, ..., X_n$ from a population with a parameter of interest is θ. The parameter θ can have different meanings, such as a population average, population variance, or population proportion, etc. Let $\hat{\theta}$ be an estimate of θ. We can decide if $\hat{\theta}$ is a good estimator based on two criteria: (1) unbiasedness and (2) small variance of the estimated values. When given multiple unbiased estimators, it is sensible to choose the ones with smallest variance. However, when comparing an unbiased and biased estimator, the decision is less clear. Take a look at the Mean Squared Errors (MSE) formula,

$$MSE = E[(\hat{\theta} - \theta)^2] = Variance + Bias^2$$

The MSE measures the closeness of the estimated value to the original value. Generally, **bias-variance tradeoff theory** states that a reduction in the variance of an estimator is associated with an increase in the bias and vice versa. In other words, we can sometimes find a biased estimator that has a smaller MSE than an unbiased estimator. In this thesis, the introduction of (possible) bias is done by using penalty functions.

Given observed data $x_1, x_2, x_3, ..., x_J$. Let $l(\theta)$ be the log-likelihood function and $h(\theta)$ the penalty function. Next, we aim to minimize the penalized log-likelihood function

$$D(\theta) = -l(\theta) + \lambda h(\theta)$$

for some optimal value of λ, a constant that determines how aggressively the penalty is enforced.

The minimum of $D(\theta)$ is known as a penalized likelihood estimation. This section gives a couple of examples to demonstrate the idea.

Example 2. Consider using the penalty function of $h(\theta) = \sum_{j=1}^{J} \log(1 - p_j)$ that potentially bring the parameters close to 0. We illustrate here the penalized estimators for a binomial model. Let $x_j \sim Bin(N_j, p_j)$ the likelihood function of binomial distribution.

$$L = \prod_{j=1}^{J} \binom{N_j}{x_j} p_j^{x_j} (1 - p_j)^{N_j - x_j}$$

Then, the log-likelihood function is:

$$l(\mathbf{p}) = \sum_{j=1}^{J} \log \binom{N_j}{x_j} + \sum_{j=1}^{J} x_j \log(p_j) + \sum_{j=1}^{J} (N_j - x_j) \log(1 - p_j).$$

We then have

$$D(\mathbf{p}) = -\sum_{j=1}^{J} \log \binom{N_j}{x_j} - \sum_{j=1}^{J} x_j \log(p_j) - \sum_{j=1}^{J} (N_j - x_j) \log(1 - p_j) + \lambda \sum_{j=1}^{J} \log(1 - p_j).$$

We take the partial derivatives of $D(\mathbf{p})$ and set these equal to 0:

$$\frac{dD(\mathbf{p})}{dp_1} = -\frac{x_1}{p_1} + \frac{N_1 - x_1}{1 - p_1} - \frac{\lambda}{1 - p_1} = 0$$

$$\Leftrightarrow \frac{-x_1(1 - p_1) + p_1(N_1 - x_1) - \lambda p_1}{p_1(1 - p_1)} = 0$$

$$\Leftrightarrow -x_1(1 - p_1) + p_1(N_1 - x_1) - \lambda p_1 = 0$$

$$\Leftrightarrow -x_1 + N_1 p_1 - \lambda p_1 = 0$$

$$\Leftrightarrow -x_1 + p_1(N_1 - \lambda) = 0$$

$$\Leftrightarrow \tilde{p}_1 = \frac{x_1}{N_1 - \lambda}$$

Further, notice that the estimated probability cannot exceed 1, we have

$$\tilde{p}_1 = \begin{cases} \dfrac{x_1}{N_1 - \lambda}, & \text{if } \lambda < N_1 \\ 1, & \text{otherwise} \end{cases}$$

Similar results hold for $\tilde{p}_2$, $\tilde{p}_3$,..., $\tilde{p}_J$.

Example 2 above illustrates shrinkage of the estimated probabilities closer to 1. In example 3 below, we illustrate a different kind of penalty. Specifically, we consider shrinking all probabilities to a common value $\kappa \in (0, 1)$.

Example 3. Penalized Estimators for Binomial Distribution Model with a different penalty function of

$$h(\mathbf{p}) = \sum_{j=1}^{J} (\kappa log(p_j) + (1 - \kappa) \log(1 - p_j)).$$

The log-likelihood function, similar to example 2, is

$$l(\mathbf{p}) = \sum_{j=1}^{J} \log \binom{N_j}{x_j} + \sum_{j=1}^{J} x_j \log(p_j) + \sum_{j=1}^{J} (N_j - x_j) \log(1 - p_j)$$

Combining the log-likelihood and the penalty, we have

$$D(\mathbf{p}) = -\sum_{j=1}^{J} \log \binom{N_j}{x_j} - \sum_{j=1}^{J} x_j \log(p_j) - \sum_{j=1}^{J} (N_j - x_j) \log(1 - p_j) - \lambda \sum_{j=1}^{J} (\kappa \log(p_j) + (1 - \kappa) \log(1 - p_j))$$

Thus, taking derivative of $D(\mathbf{p})$ with respect to p_1, we obtain

$$\begin{aligned}\frac{dD(\mathbf{p})}{dp_1} &= -\frac{x_1}{p_1} + \frac{N_1 - x_1}{1-p_1} - \frac{\lambda K}{p_1} + \frac{\lambda(1-\kappa)}{1-p_1} \\ &= \frac{-x_1 - \kappa\lambda}{p_1} + \frac{(N_1 - x_1) + \lambda(1-\kappa)}{1-p_1}.\end{aligned}$$

Setting this gradient equal to 0, we get

$$\begin{aligned}&\Leftrightarrow -x_1 - \kappa\lambda + x_1 p_1 + \kappa p_1 \lambda + N_1 p_1 - x_1 p_1 + \lambda p_1 - \kappa p_1 \lambda = 0 \\ &\Leftrightarrow -x_1 - \kappa\lambda + N_1 p_1 + \lambda p_1 = 0 \\ &\Leftrightarrow p_1(N_1 + \lambda) = x_1 + \kappa\lambda \\ &\Leftrightarrow \tilde{p}_1 = \frac{x_1 + \kappa\lambda}{N_1 + \lambda} \\ &\Leftrightarrow \tilde{p}_1 = \frac{x_1}{N_1}\frac{N_1}{N_1+\lambda} + \kappa\frac{\lambda}{N_1+\lambda} \\ &\Leftrightarrow \tilde{p}_1 = \hat{p}_1\frac{N_1}{N_1+\lambda} + \kappa\frac{\lambda}{N_1+\lambda}\end{aligned}$$

Similarly, $\tilde{p}_j = \hat{p}_j \frac{N_j}{N_j+\lambda} + \kappa\frac{\lambda}{N_j+\lambda}$ for all j. Let $w_j = \frac{N_j}{N_j+\lambda}$, we can rewrite $\tilde{p}_j$ as $\tilde{p}_j = \hat{p}_j w_1 + \kappa(1-w_1)$. The penalized solution is therefore expressed as a linear combination of the MLE and the value κ.

There is no specific rules on how to choose an appropriate penalty functions. Usually, we can specify multiple penalty functions that implement similar constraints. It is therefore reasonable to choose one that has nice calculus properties since our solutions involve evaluating gradients for easier minimization. However, things get more complicated when we must also select a good value for λ. In chapter 4, we consider data-driven methods to find the optimal λ.

Example 4: Shrinking normal means.

We illustrate another example of shrinkage and consider the Normal means problem. Let $X_j \sim N(\mu_j, \sigma_j^2)$, $j = 1, \ldots, n$ be independent variables with unknown means μ_j and known variances σ_j^2. Assume $\boldsymbol{\mu} = \begin{pmatrix}\mu_1 & \mu_2 & \cdots & \mu_n\end{pmatrix}^\top$. The log-likelihood function (ignoring the constant of proportionality that doesn't depend on the unknown parameters) is

$$\ell(\boldsymbol{\mu}) = -\frac{1}{2}\sum_{j=1}^{n}\left(\frac{X_j - \mu_j}{\sigma_j}\right)^2.$$

Now, say we want to apply shrinkage forcing the means to be "close" to one-another using penalty

$$\text{Pen}(\boldsymbol{\mu}) = \sum_{j=1}^{n}\sum_{k=1}^{n}(\mu_j - \mu_k)^2.$$

The corresponding penalized minimization problem is

$$D(\boldsymbol{\mu}) = -\ell(\boldsymbol{\mu}) + \frac{\lambda}{4}\ \text{Pen}(\boldsymbol{\mu}) = \frac{1}{2}\sum_{j=1}^{n}\left(\frac{X_j - \mu_j}{\sigma_j}\right)^2 + \frac{\lambda}{4}\ \sum_{j=1}^{n}\sum_{k=1}^{n}(\mu_j - \mu_k)^2$$

where $\lambda/4$ is a specified constant (and we divide by 4 for convenience reasons that will become clear in a moment). The score functions are obtained by taking partial derivatives and setting equal to 0. For example, we have

$$\frac{\partial D}{\partial \mu_1} = -\frac{1}{\sigma_1^2}(X_1 - \mu_1) + \lambda\sum_{k=2}^{n}(\mu_1 - \mu_k) = 0$$

which can be written as

$$\left[\frac{1}{\sigma_1^2} + \lambda(n-1)\right]\mu_1 - \lambda\mu_2 - \cdots - \lambda\mu_n = \frac{1}{\sigma_1^2}X_1.$$

and this can be written as

$$\left(\frac{1}{\sigma_1^2} + \lambda(n-1) \quad -\lambda \quad \cdots \quad -\lambda\right)\boldsymbol{\mu} = \frac{1}{\sigma_1^2}X_1.$$

In general, if we define $\boldsymbol{L}_j$ as the row vector with jth element $1/\sigma_j^2 + \lambda(n-1)$ and all other elements $-\lambda$ then for the jth variable we have

$$\frac{\partial D}{\partial \mu_j} = -\frac{1}{\sigma_j^2}(X_j - \mu_j) + \lambda\sum_{k\neq j}(\mu_j - \mu_k) = 0$$

which can be written as

$$\boldsymbol{L}_j\boldsymbol{\mu} = \frac{1}{\sigma_j^2}X_j.$$

If we now define $\boldsymbol{L}$ to be the $n\times n$ matrix with jth row equal to $\boldsymbol{L}_j$, and we define $\tilde{\boldsymbol{X}} = \left(X_1/\sigma_1^2 \quad \cdots \quad X_n/\sigma_n^2\right)^\top$ then combining the n score equations gives

$$\boldsymbol{L}\boldsymbol{\mu} = \tilde{\boldsymbol{X}}$$

so that the penalized solution is for a given value of λ is

$$\hat{\boldsymbol{\mu}}_\lambda = \boldsymbol{L}^{-1}\tilde{\boldsymbol{X}}.$$

Now it is of interest to find a closed-form expression for $\boldsymbol{L}^{-1}$, which will lead to a closed-form expression for $\hat{\boldsymbol{\mu}}_\lambda$. Note that $\boldsymbol{L}$ can be written as

$$\boldsymbol{L} = \begin{pmatrix} \sigma_1^{-2} + n\lambda & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} + n\lambda & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \sigma_n^{-2} + n\lambda \end{pmatrix} - \begin{pmatrix} \lambda & \lambda & \cdots & \lambda \\ \lambda & \lambda & \cdots & \lambda \\ \cdots & \cdots & \ddots & \cdots \\ \lambda & \lambda & \cdots & \lambda \end{pmatrix}.$$

This representation is convenient when recalling the Sherman-Morrison inverse formula which states that

$$(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^\top)^{-1}=\boldsymbol{A}^{-1}-\frac{1}{1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}}(\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{A}^{-1}).$$

with

$$\boldsymbol{A}=\begin{pmatrix}\sigma_1^{-2}+n\lambda & 0 & \cdots & 0\\ 0 & \sigma_2^{-2}+n\lambda & \cdots & 0\\ \cdots & \cdots & \ddots & \cdots\\ 0 & 0 & \cdots & \sigma_n^{-2}+n\lambda\end{pmatrix},$$

and $\boldsymbol{u}=\begin{pmatrix}-\lambda^{1/2} & -\lambda^{1/2} & \cdots & -\lambda^{1/2}\end{pmatrix}^\top$ while $\boldsymbol{v}=-\boldsymbol{u}$. The above formula can be applied directly. $\boldsymbol{A}$ is diagonal, so the inverse is given by

$$\boldsymbol{A}^{-1}=\begin{pmatrix}\sigma_1^2/(1+\sigma_1^2n\lambda) & 0 & \cdots & 0\\ 0 & \sigma_2^2/(1+\sigma_2^2n\lambda) & \cdots & 0\\ \cdots & \cdots & \ddots & \cdots\\ 0 & 0 & \cdots & \sigma_n^2/(1+\sigma_n^2n\lambda)\end{pmatrix}.$$

Now,

$$\begin{aligned}\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u} &= \begin{pmatrix}\lambda^{1/2} & \cdots & \lambda^{1/2}\end{pmatrix}\begin{pmatrix}\sigma_1^2/(1+\sigma_1^2n\lambda) & 0 & \cdots & 0\\ 0 & \sigma_2^2/(1+\sigma_2^2n\lambda) & \cdots & 0\\ \cdots & \cdots & \ddots & \cdots\\ 0 & 0 & \cdots & \sigma_n^2/(1+\sigma_n^2n\lambda)\end{pmatrix}\begin{pmatrix}-\lambda^{1/2}\\ \cdots\\ -\lambda^{1/2}\end{pmatrix}\\ &= -\lambda\sum_{k=1}^n\frac{\sigma_k^2}{\sigma_k^2+n\lambda}.\end{aligned}$$

From this,

$$\frac{1}{1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}}=\left(1-\lambda\sum_{k=1}^n\frac{\sigma_k^2}{\sigma_k^2+n\lambda}\right)^{-1}:=c_\lambda.$$

Furthermore, adopting the shorthand

$$w_k=\frac{\sigma_k^2}{\sigma_k^2+n\lambda},$$

we have

$$\begin{aligned}
\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{A}^{-1} &= \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} \begin{pmatrix} -\lambda & \cdots & -\lambda \\ \cdots & \ddots & \cdots \\ -\lambda & \cdots & -\lambda \end{pmatrix} \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} \\
&= -\lambda \begin{pmatrix} w_1^2 & w_1w_2 & \cdots & w_1w_n \\ w_1w_2 & w_2^2 & \cdots & w_2w_n \\ \cdots & \cdots & \ddots & \cdots \\ w_1w_n & w_2w_n & \cdots & w_n^2 \end{pmatrix}.
\end{aligned}$$

Finally, this means $\boldsymbol{L}^{-1}$ has diagonal elements

$$(\boldsymbol{L}^{-1})_{ii} = w_i(1 + c_\lambda\lambda w_i)$$

and when $i \neq j$,

$$(\boldsymbol{L}^{-1})_{ij} = c_\lambda\lambda w_i w_j.$$

The penalized estimators of the mean values are subsequently given by

$$\begin{aligned}
\tilde{\mu}_i &= w_i(1 + c_\lambda\lambda w_i)X_i/\sigma_i^2 + \sum_{j\neq i} c_\lambda\lambda w_i w_j X_j/\sigma_j^2 \\
&= w_i(1 + c_\lambda\lambda w_i)X_i/\sigma_i^2 + \sum_{j=1}^{n} c_\lambda\lambda w_i w_j X_j/\sigma_j^2 - c_\lambda\lambda w_i^2 X_i/\sigma_i^2 \\
&= \frac{1}{\sigma_i^2 + n\lambda}X_i + c_\lambda\lambda w_i \sum_{j=1}^{n} \frac{1}{\sigma_j^2 + n\lambda}X_j.
\end{aligned}$$

This last expression formulates the solution in terms of X_i and a weighted sum of all the observations.

4 Data-driven Shrinkage

In Section 3.3, examples of different penalized likelihood estimators were illustrated. However, in the presentation, the value of the parameter λ was assumed known. As λ controls the relative importance of the penalty function, it is important to choose a value resulting in parameter estimates with small MSE. Note that MSE cannot be calculated in practice as the true parameter values are unknown. Here, we discuss an alternative selection approach known as V-fold cross-validation (VFCV).

Consider an dataset consisting of I independently sampled variables, with the ith variable consisting of n_i independent observations. Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$ denote the observations corresponding to the ith variable. VFCV proceeds by partitioning the data into V subsets of roughly equal size. For the ith variable,

let $\mathcal{I}_{i,v}$, $v = 1, \ldots, V$ denote a partition of the indices such that $\bigcup_v \mathcal{I}_{i,v} = \{1, \ldots, n_i\}$ and $\mathcal{I}_{i,v_1} \bigcap \mathcal{I}_{i,v_2} = \varnothing$ for all $v_1 \neq v_2$.

VFCV repeatedly creates subsets of the data for model training, in each instance leaving out one of the V subsets per variable. The subsets left out in each iteration are then used for model validation. More specifically, the model building data subsets are used to estimate penalized parameter estimates for various degrees of penalty enforcement, say K possible values of λ satisfying $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_K$. The negative log-likelihood function is then evaluated using penalized estimators corresponding to each possible value of λ and using the validation subsets. This is repeated V times, and the optimal value of λ_{opt} is chosen as where the negative log-likelihood function averaged over the validation subsets is minimized.

Algorithmically, implementation of VFCV proceeds as follows. For each fold $v = 1, \ldots, V$:

- For the i^{th} variable, form a training dataset by excluding the vth fold, $\boldsymbol{x}^{(v)}_{train,i} = \{x_{ij} : \ j \notin \mathcal{I}_{i,v}\}$, and let the vth fold equal to the validation set, $\boldsymbol{x}^{(v)}_{valid,i} = \{x_{ij} : \ j \in \mathcal{I}_{i,v}\}$. Let $n_i^{(v)}$ denote the number of observations in $\boldsymbol{x}^{(v)}_{train,i}$. Also let $\boldsymbol{x}^{(v)}_{train}$ and $\boldsymbol{x}^{(v)}_{valid}$ denote the collection of the training and validation sets for all I variables.

- For each value $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_K$, find the estimators $\tilde{\boldsymbol{\theta}}^{(v)}_{train}(\lambda_k)$ that minimize the penalized negative log-likelihood function

$$D_k(\boldsymbol{\theta}) = -l\left(\boldsymbol{\theta} \left| \boldsymbol{x}^{(v)}_{train}\right.\right) + \lambda_k \bar{n}^{(v)} \mathrm{Pen}(\boldsymbol{\theta})$$

  where $\bar{n}^{(v)} = (1/I)\sum_i n_i^{(v)}$.

- Calculate the validation function by evaluate the negative log-likelihood at this estimator,

$$\tilde{D}^{(v)}(\lambda_k) = -\ell\left(\tilde{\boldsymbol{\theta}}^{(v)}_{train}(\lambda_k) \left| \boldsymbol{x}^{(v)}_{valid}\right.\right).$$

The VFCV score is then defined as

$$\mathrm{CV}_k = \mathrm{CV}(\lambda_k) = \sum_{v=1}^{V} \tilde{D}^{(v)}(\lambda_k), \tag{1}$$

and the optimal shrinkage level is taken to be the λ that minimizes CV_k, i.e. $\lambda_{opt} = \lambda_{k^*}$ where $k^* = \mathrm{argmin}_k \mathrm{CV}_k$. Note that after the optimal penalty level has been chosen using VFCV, penalized estimators are calculated one more time using the full dataset. The penalized likelihood estimator with data-driven shrinkage, denoted $\tilde{\boldsymbol{\theta}}_{pen}$, is the value that minimizes

$$D_{opt}(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}|\boldsymbol{x}) + \lambda_{opt}\bar{n}\mathrm{Pen}(\boldsymbol{\theta})$$

where $\bar{n} = (1/I)\sum_i n_i$. The literature on cross-validation recommends various choices for V, with common

values ranging from $V = 2$ to $V = 10$. The choice $V = n$ is equivalent to leave-one-out cross-validation and can become computationally expensive. As discussed in (Arlot and Celisse, 2010), the size of the validation set has an effect on the bias of the penalized estimator, while the number of folds V controls for the variance of the estimated penalization parameter. These authors also discuss some asymptotic considerations of cross-validation. If n_{train} denotes the size of the training set, then for $n_{train}/n \to 1$, cross-validation is asymptotically equivalent to Mallows' C_p and therefore asymptotically optimal. Furthermore, if $n_{train}/n \to \gamma \in (0, 1)$, then asymptotically the model is equivalent to Mallows' C_p multiplied by (or over-penalized by) a factor $(1 + \gamma)/(2\gamma)$. Asymptotics notwithstanding, throughout the remainder of this paper, $V = 10$ is used. This strikes a balance having larger training sets and reasonable computational costs.

5 Simulation Studies

In this chapter, the performance of shrinkage estimation is considered both for the binomial model as well as two related models, the zero-inflated binomial and the beta-binomial. The reason we consider the binomial model is that our data consists of success and failure counts, i.e. reading *incorrectly* versus *correctly*. Moreover, the zero-inflated binomial and beta-binomial models tackle issues arisen with count data with many 0s, such as unbalanced and skewed results. The simulation studies also consider whether the model estimation is improved by using penalization. In most scenarios investigated here, no closed-form solutions for the penalized estimators are available. Even so, these simulation studies are very useful for investigating the properties of different penalty functions as they impact estimation for the three models. Note that simulations are restricted here to $I = 10$ independent variables, each consisting of $N = 40$ trials and having $n = 50$ independent observations per variable. This choice was made so that the simulations would, at least in part, closely resemble the real data motivating this work.

5.1 The Binomial Model

A sample $\mathcal{X} = \{X_{ij}, i = 1, \ldots, I,\ j = 1, \ldots, n\}$ was generated with independent observations $X_{ij} \sim \text{Bin}(N, p_i)$ with $I = 10$, $N = 40$, and $n = 50$. The binomial success probabilities p_i were sampled from different scaled beta distributions. Specifically, for success probability lower and upper bounds a and b, three shapes of the success probability distribution were considered, namely a skewed distribution $(p_i - a)/(b - a) \sim \text{Beta}(2, 5)$, a very flat distribution $(p_i - a)/(b - a) \sim \text{Beta}(5/4, 5/4)$, and a bell-shaped distribution $(p_i - a)/(b - a) \sim \text{Beta}(10, 10)$. The three success probability distributions are illustrated in Figure 2 below. The λ term controlling how aggressively the penalty gets enforced was chosen using cross-validation using 63 possible values ranging from 0 to $10,000$ spaced approximately equi-distant on a logarithmic scale. VFCV was used to determine the λ for each penalty function under consideration.

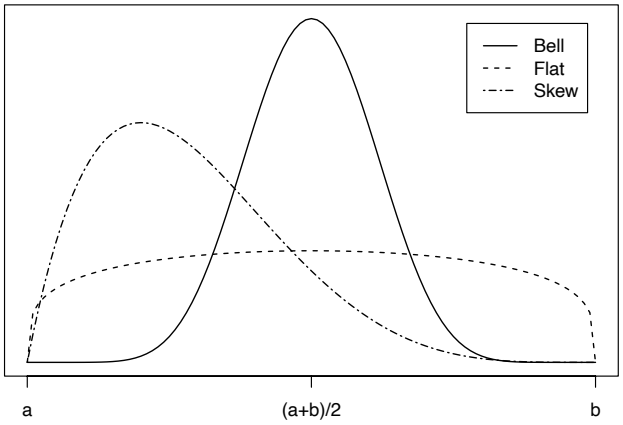

Figure 2: Success probability distributions considered in the simulation study.

In addition to the estimators resulting from the use of different penalty functions, maximum likelihood estimators were also calculated. In total, $K = 500$ samples were generated for each configuration of success probability bounds (a, b) and Beta shape parameters. Summarized in the tables below are the Monte Carlo estimates of the MSE ratios. For the kth sample $\mathcal{X}_k$, let $\boldsymbol{p}_k = (p_{k,1}, \ldots, p_{k,10})$ denote the true success probabilities simulated from a specified scaled Beta distribution. Let $\hat{\boldsymbol{p}}_k$ denote the MLE and let $\tilde{\boldsymbol{p}}_k$ denote a penalized estimator found using VFCV. Define Sum of Squared Deviations $\text{MSD}(\boldsymbol{q}, \boldsymbol{p}) = (1/I) \sum_{i=1}^{I} \left[(q_i - p_i)^2\right]$. Then, the Monte Carlo MSE ratios are defined as

$$\text{MSE}_{\text{Pen}} = \frac{(1/K) \sum_{k=1}^{K} \text{MSD}(\tilde{\boldsymbol{p}}_k, \boldsymbol{p}_k)}{(1/K) \sum_{k=1}^{K} \text{MSD}(\hat{\boldsymbol{p}}_k, \boldsymbol{p}_k)}$$

where the subscript "Pen" emphasizes the specific penalty function used to obtain the estimators. An MSE ratio less than 1 indicates superior performance of the penalized estimator. In Table 2, the results of shrinking success probabilities closer to one another are presented. The penalties $\sum_{i=1}^{I} p_i^2$ and $\sum_i \log p_i$ from Section 2.1 were considered.

		Penalty	
$p_i \in (a, b)$	Shape	L_2	Q_2
$(0.01, 0.05)$	Skew	0.928	0.906
	Flat	0.935	0.942
	Bell	0.705	0.704
$(0.08, 0.20)$	Skew	0.960	0.952
	Flat	0.969	0.973
	Bell	0.854	0.856
$(0.31, 0.35)$	Skew	0.411	0.411
	Flat	0.652	0.652
	Bell	0.292	0.293

Table 2: MSE ratios comparing penalized parameter estimates to maximum likelihood when shrinking estimators closer to one another.

In Table 2, the performance of the L_2 and Q_2 penalties is nearly indistinguishable. When shrinking parameters closer to one another, large gains in efficiency are sometimes realized. This is especially notable when the Beta shape from which the success probabilities are generated is bell-shaped, i.e. the p_i are close to one another. In all instances, VFCV results in penalized estimators with performance superior to maximum likelihood. Altogether, these simulations illustrate that both the average success probability and the spacing of the p_i relative to that average are important in determining the reduction in MSE. For penalties shrinking the p_i closer to one another, an MSE ratios below 0.3 was realized, showing dramatic improvement due to shrinkage.

5.2 The Beta-binomial Model

The probability mass function of the beta-binomial distribution is given by

$$f(x|N,\alpha,\beta) = \binom{N}{x} \frac{B(x+\alpha, N-x+\beta)}{B(\alpha,\beta)},\ x = 0,1,...,N$$

where $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$ is the Beta function and N is the number of trials. The parameters $\alpha > 0$ and $\beta > 0$ control the mean and variance of the model. Specifically, for $p = \alpha/(\alpha+\beta) \in (0,1)$ and $\nu = (\alpha+\beta+N)/(\alpha+\beta+1) \in (1,N)$, the mean and variance can be written as $E[X] = Np$ and $Var[X] = Np(1-p)\nu$. In this parameterization, p and ν denote, respectively, the expected proportion of successes and the the over-dispersion of the model relative to the binomial.

Samples $\mathcal{X} = \{X_{ij}, i = 1,\ldots,I,\ j = 1,\ldots,n\}$ were generated with independent Beta-Binomial variables, $X_{ij} \sim \text{BetaBin}(N,\alpha_i,\beta_i)$, with $I = 10$, $N = 40$, and $n = 50$. The overall expected success proportions p_i and the overdispersion ν_i were sampled from scaled beta distributions as per Figure 2 with the specific bounds (a,b) listed in the table below. In this case, the BetaBin simulation considered three penalty functions. Letting $p_i = \alpha_i/(\alpha_i+\beta_i)$, $i = 1,\ldots,I$, these were: $\text{Pen}_2(\boldsymbol{p}) = \sum_i p_i^2$, $\text{Pen}_{L_2}(\boldsymbol{p}) = \sum_i \sum_j (p_i - p_j)^2$, and $\text{Pen}_{full}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_i \sum_j (\alpha_i - \alpha_j)^2 + \sum_i \sum_j (\beta_i - \beta_j)^2$. These are again termed, respectively, *zero shrinkage*, *mean shrinkage*, and *full shrinkage*. Again, after choosing the λ shrinkage levels using VFCV, one final estimator, termed *minCV*, was calculated by selecting among the three penalized estimators the one with the smallest CV score.

For each parameter configuration, a total of $M = 500$ samples were generated. The MSE ratios are reported in Table 3.

			Penalty			
$p_i \in (a_1, b_1)$	$\nu_i \in (a_2, b_2)$	Shape	Zero	Mean	Full	minCV
$(0.05, 0.10)$	$(4, 6)$	Skew	0.917	0.474	0.428	0.429
		Flat	0.928	0.702	0.591	0.604
		Bell	0.921	0.290	0.270	0.271
$(0.12, 0.22)$	$(2, 5)$	Skew	0.974	0.722	0.726	0.708
		Flat	0.977	0.903	0.948	0.889
		Bell	0.973	0.476	0.466	0.463
$(0.17, 0.22)$	$(3, 8)$	Skew	0.971	0.301	0.400	0.331
		Flat	0.971	0.445	0.762	0.481
		Bell	0.968	0.217	0.242	0.227
$(0.05, 0.06)$	$(2, 10)$	Skew	0.905	0.170	0.469	0.211
		Flat	0.891	0.188	0.733	0.213
		Bell	0.893	0.155	0.187	0.175

Table 3: MSE ratios for Beta-Binomial success proportions $\boldsymbol{p} = (p_1, \ldots, p_{10})$ comparing penalized parameter estimates to maximum likelihood for different penalization approaches.

When looking at the results in Table 3, it comes as no surprise that *zero shrinkage* is the least effective approach here, even while still being more effective than maximum likelihood. For most of the simulation configurations, MSE ratios under *mean* and *full shrinkage* are comparable. Here, the *minCV* approach is also very impressive, in most instances nearly matching the best-performing method. This reaffirms that VFCV can be effectively used to choose both the level of shrinkage for a specific penalty function, but then also choose from among competing penalty functions.

5.3 The Zero-inflated Binomial Model

The pmf of the zero-inflated binomial (ZIB) distribution is

$$f(x|N, q, \gamma) = \begin{cases} \gamma + (1-\gamma)(1-q)^N & \text{when } x = 0 \\ (1-\gamma)\binom{N}{x} p^x (1-q)^{N-x} & \text{when } x = 1, \ldots, N \end{cases}$$

where γ represents the excess zero probability, and p and N are the binomial success probabilities and number of trials. For $X \sim \text{ZIB}(N, q, \gamma)$, it follows that $E[X] = Nq(1-\gamma)$. Consequently, the parameter $p = E[X]/N = q(1-\gamma)$ is the expected proportion of successes in a ZIB model. The parameter p is of primary interest when considering possible penalty functions, especially under the assumption that the different ZIB distributions are "similar" to one another.

In the simulation study, a sample $\mathcal{X} = \{X_{ij}, i = 1, \ldots, I,\ j = 1, \ldots, n\}$ was generated with independent ZIB variables, $X_{ij} \sim \text{ZIB}(N, q_i, \gamma_i)$. As with the binomial model simulation, $I = 10$, $N = 40$, and $n = 50$. The overall expected success proportions p_i and the excess zero probabilities γ_i were sampled from scaled beta distributions as per Figure 2 with the specific bounds (a, b) listed in the table below. Letting $p_i = (1-\gamma_i)q_i$, $i = 1, \ldots, I$, the ZIB simulation considered three penalty functions: $\text{Pen}_2(\boldsymbol{p}) = \sum_i p_i^2$,

$\text{Pen}_{L_2}(\boldsymbol{p}) = \sum_i \sum_j (p_i - p_j)^2$, and $\text{Pen}_{full}(\boldsymbol{\gamma}, \boldsymbol{q}) = \sum_i \sum_j (\gamma_i - \gamma_j)^2 + \sum_i \sum_j (q_i - q_j)^2$. The first of these, termed *zero shrinkage*, results in success proportions closer to 0. The second, termed *mean shrinkage*, results in mean success proportions p_i closer to one another. The third, termed *full shrinkage*, shrinks all γ_i closer to one another and all q_i closer to one another.

In addition to using VFCV to select the level of shrinkage for the above three penalities, a combined estimator, termed *minCV*, was calculated by selecting the model parameters associated with the penalty function having minimum VFCV score. The same set of 63 λ values ranging from 0 to 10,000 were used. The Monte Carlo MSE ratios for the success proportions $\boldsymbol{p}$ are in Table 4.

			Penalty			
$p_i \in (a_1, b_1)$	$\gamma_i \in (a_2, b_2)$	Shape	Zero	Mean	Full	minCV
$(0.01, 0.05)$	$(0.10, 0.14)$	Skew	0.957	0.888	0.981	0.958
		Flat	0.977	0.942	0.979	0.983
		Bell	0.964	0.668	0.836	0.755
$(0.04, 0.06)$	$(0.20, 0.30)$	Skew	0.968	0.364	0.368	0.356
		Flat	0.971	0.562	0.526	0.523
		Bell	0.968	0.258	0.246	0.239
$(0.15, 0.30)$	$(0.04, 0.06)$	Skew	1.006	0.969	0.860	0.885
		Flat	1.010	1.005	0.808	0.819
		Bell	1.009	0.821	0.873	0.899
$(0.05, 0.06)$	$(0.20, 0.70)$	Skew	0.963	0.203	0.635	0.273
		Flat	0.955	0.223	0.934	0.259
		Bell	0.951	0.183	0.372	0.245

Table 4: MSE ratios for ZIB success proportions $\boldsymbol{p} = (p_1, \ldots, p_{10})$ comparing penalized parameter estimates to maximum likelihood for different penalization approaches.

These MSE ratios in Table 4 are based on $M = 500$ simulated datasets for each possible simulation configuration. While the *zero shrinkage* penalty does result in some efficiency gains in most scenarios, overall MSE ratios close to 1 suggest little improvement from from using this penalty. On the other hand, both *mean* and *full shrinkage* result in large decreases in the MSE ratios. Overall, it cannot be said that either *mean* and *full* shrinkage performs best. This makes sense, as it depends on the configuration of all parameters and not just the mean parameters. Finally, while *minCV* does not always have the smallest MSE ratio, it is generally close to the minimum. This suggests that data-driven selection of the level of shrinkage as well as the penalty function leads to good performance for the model.

6 Data Analysis and Findings

The work presented in this paper was motivated by the reading data collection a sample of 508 elementary-school aged children. Each child read one randomly selected passage out of ten possible passages. This resulted in roughly 50 Words Read Incorrectly (WRI) scores per passage. The passage lengths varied from 44 to 69 words with an average length of 51 words. Of interest is to accurately and efficiently estimate

passage difficulty as measured by the average proportion of words read incorrectly. Note that higher WRI proportions indicate that a passage is more difficult. Figure 3 provides information about the passage-specific WRI proportions. The solid dot in each violin plot represents the mean WRI proportion, a typical estimate of passage difficulty.

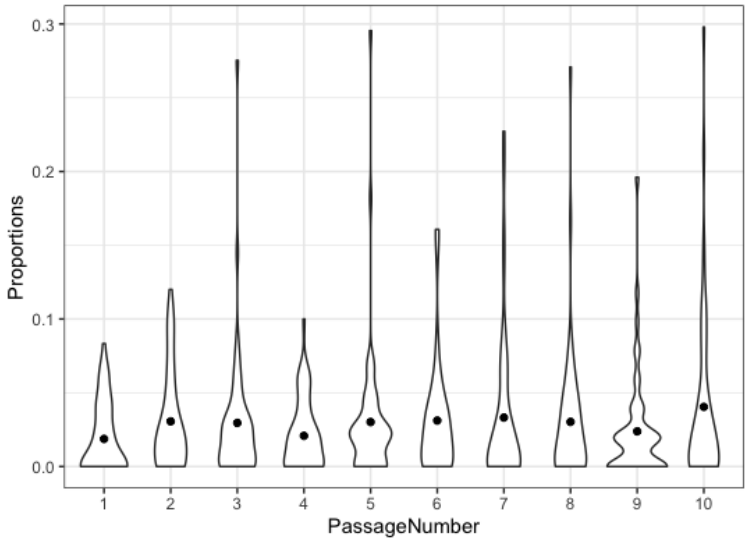

Figure 3: Violin plot comparison of passage WRI proportions

Note that the mean WRI proportions in Figure 3 appear fairly close to one another, meaning passages are crafted to be within a narrow range of difficulty to reinforce fairness in grading and evaluation of students. Thus, as one expects the WRI proportions to be close to one another, appropriate shrinkage may result in improved estimates.

Three models and three types of shrinkage were considered for the data at hand. In each instance, the same set of partitions were used to select a smoothing parameter with 10-fold cross validation. Table 5 reports the cross-validation scores as defined in (1).

Distribution	Penalty	CV Score	$\log(\lambda_{opt}+1)$
Binomial	None	1025.5	–
	Zero	1024.9	3.56
	Mean	1017.1	4.36
ZIB	None	964.7	–
	Zero	964.3	2.78
	Mean	959.6	3.96
	Full	950.4	3.56
BetaBin	None	869.7	–
	Zero	869.5	2.41
	Mean	866.3	3.56
	Full	851.9	0.04

Table 5: 10-fold CV scores and optimal λ values for the three distributions considered.

	Maximum Likelihood			Mean Shrinkage			Full Shrinkage		
Passage	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{p}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{p}$
P1	1.28	66.34	0.019	1.20	52.67	0.022	0.70	27.45	0.025
P2	1.51	45.15	0.032	1.50	47.47	0.031	0.91	27.44	0.032
P3	0.84	19.85	0.040	0.80	22.81	0.034	0.96	27.44	0.034
P4	2.47	160.0	0.015	2.25	123.5	0.018	0.67	27.45	0.024
P5	1.17	42.54	0.027	1.17	41.65	0.027	0.83	27.45	0.030
P6	1.18	29.10	0.039	1.13	32.52	0.034	0.97	27.44	0.034
P7	0.53	19.48	0.026	0.53	18.75	0.027	0.74	27.44	0.026
P8	0.87	25.37	0.033	0.86	27.20	0.031	0.88	27.44	0.031
P9	0.85	32.25	0.026	0.84	30.74	0.027	0.79	27.45	0.028
P10	0.65	17.03	0.037	0.63	19.14	0.032	0.89	27.44	0.031

Table 6: Beta-binomial parameter estimates for the WRI data.

It is evident from Table 5 that all variations of the beta-binomial model have much lower cross-validation scores than either the binomial or zero-inflated binomial models. With regards to penalty type, full shrinkage works best for this model with mean shrinkage being a distant second choice. Table 6 shows the beta-binomial parameter estimates obtained using maximum likelihood as well as penalized likelihood with mean shrinkage and full shrinkage. It is interesting to note that in the full shrinkage solution, the $\tilde{\beta}_i$ values have all been shrunk to within 0.01 of a common value, but the $\tilde{\alpha}_i$ still exhibit a fair spread of values. In the maximum likelihood sense, the estimated success proportions range from 0.019 to 0.04, while the full shrinkage values range from 0.024 to 0.034. The latter shows much more adherence to the idea that the passages are similar in terms of difficulty.

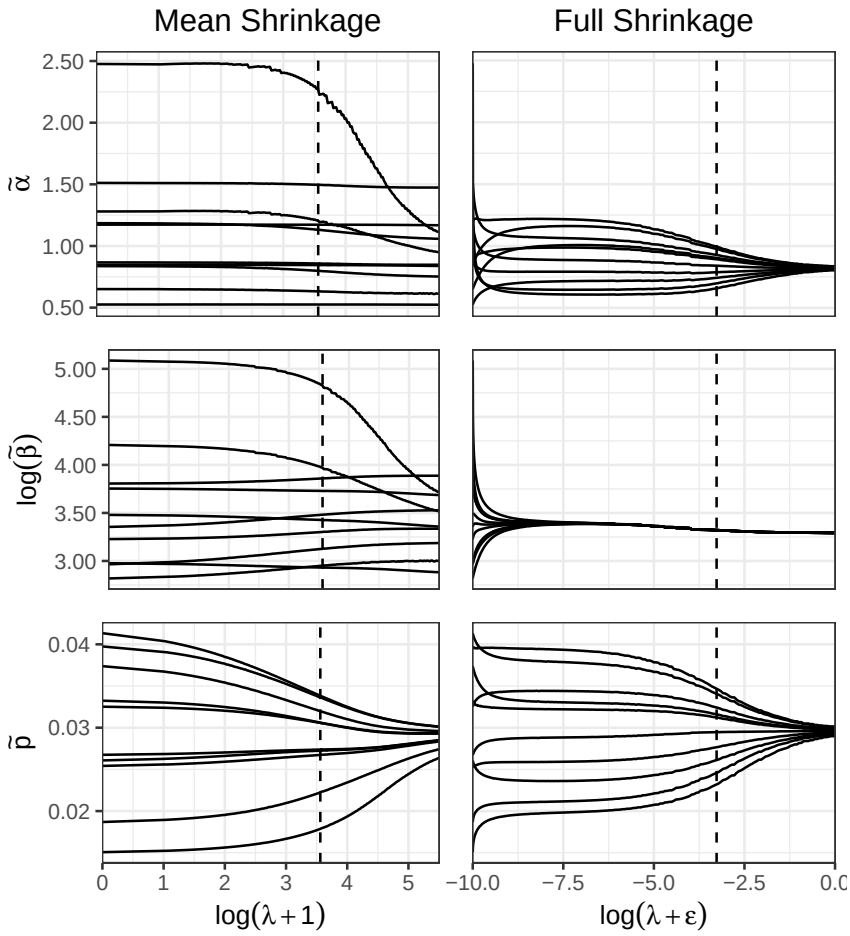

Figure 4: Beta-binomial parameter estimates under mean shrinkage and full shrinkage. Dashed line indicates optimal shrinkage. Scale value to improve full shrinkage plot readability is $\varepsilon = e^{-10}$.

For the interested reader, Figure 4 shows the penalized likelihood estimate trajectories for mean shrinkage and full shrinkage as a function of λ. The estimates of $\tilde{\beta}$ are presented on a logarithmic scale. For mean shrinkage, the horizontal scale is $\log(\lambda + 1)$ and for full shrinkage it is $\log(\lambda + \varepsilon)$ with $\varepsilon = 10^{-10}$. These adjustments were all made to improve readability of the plots. Dashed vertical lines indicate the optimal shrinkage solutions as determined by VFCV.

Under mean shrinkage, the passage-specific $\tilde{\alpha}_i$ and $\tilde{\beta}_i$ still exhibit a large spread even when the success proportions $\tilde{p}_i = \tilde{\alpha}_i/(\tilde{\alpha}_i + \tilde{\beta}_i)$ are close to one another. Under full shrinkage, the $\tilde{\beta}_i$ values are very quickly shrunk to a nearly common value while the $\tilde{\alpha}_i$ still exhibit some spread.

7 Conclusions

The goal of this project is to consider various statistical models for WRI data. After evaluating 5 potential models in chapter 2 using maximum likelihood, the beta-binomial and zero-inflated binomial appears to

have the best performance when considering AIC. However, standard maximum likelihood does not allow for sharing of information across passages. Because we believe that passages are similar in difficulty, we next consider penalized estimation of multiple independent success proportions from the observed multivariable count data. In chapter 3, a few examples of shrinkage are given. In chapter 4, we demonstrate how cross-validation can be used.

The application of interest considered WRI scores realized by students during an ORF assessment. The simulation results in chapter 5 show that across the scenarios considered, large decreases in MSE are often achieved. There is also very little risk in using penalized likelihood, as using V-fold cross validation never resulted in a large increase in MSE. When applying the methodology to the data of interest in chapter 6, it is seen that the resulting penalized estimators of the success proportions have a much tighter spread. This affirms the expectation that the passages are very similar in difficulty, with estimated difficulty scores ranging from 2.4% to 3.4% of words expected to be read incorrectly. Even so, this does highlight one important avenue for future research. If students are reading different passages to assess ORF, it is desirable to have a method that standardizes WRI scores to be independent of passage difficulty. Also, in practice students typically read multiple passages, so exploring methods accounting for correlated WRI scores need to be considered.

References

Allington, R. L. (1983). Fluency: The neglected reading goal. *The reading teacher*, 36(6):556–561.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Cohen, A. C. (1967). Estimation in mixtures of two normal distributions. *Technometrics*, 9(1):15–28.

DiSalle, K. and Rasinski, T. (2017). Impact of short-term intense fluency instruction on students' reading achievement: A classroom-based, teacher-initiated research study. *Journal of Teacher Action Research*, 3(2):1–13.

Fuchs, L. S., Fuchs, D., Hosp, M. K., and Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific studies of reading*, 5(3):239–256.

Griffiths, D. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, pages 637–648.

Gruber, M. H. (2017). *Improving efficiency by shrinkage: the James-Stein and ridge regression estimators*. Routledge.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132.

Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Johns, J. L. and Lunn, M. K. (1983). The informal reading inventory: 1910–1980. *Literacy Research and Instruction*, 23(1):8–19.

Lemmer, H. (1981a). From ordinary to bayesian shrinkage estimators. *South African Statistical Journal*, 15(1):57–72.

Lemmer, H. (1981b). Note on shrinkage estimators for the binomial distribution. *Communications in statistics-theory and methods*, 10(10):1017–1027.

Routledge, R. (2018). binomial distribution. `https://www.britannica.com/science/binomial-distribution`. Accessed: 04-08-2022.

Routledge, R. (2020). Poisson distribution. `https://www.britannica.com/science/Poisson-distribution`. Accessed: 04-08-2022.

Samuels, S. J. (1988). Decoding and automaticity: Helping poor readers become automatic at word recognition. *The reading teacher*, 41(8):756–760.

Schreiber, P. A. (1991). Understanding prosody's role in reading acquisition. *Theory into practice*, 30(3):158–164.

Shinn, M. R., Knutson, N., Good III, R. H., Tilly III, W. D., and Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, 21(3):459–479.

Stein, C. et al. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.