Shrinkage Estimation of Proportion via Logit Penalty

Yoonsuh Jung

Department of Statistics, University of Waikato, Hamilton, New Zealand

Abstract

By releasing the unbiasedness condition, we often obtain more accurate estimators due to the bias-variance tradeoff. In this paper, we propose a class of shrinkage proportion estimators which show improved performance over the sample proportion. We provide the "optimal" amount of shrinkage. The advantage of the proposed estimators is given theoretically as well as explored empirically by simulation studies and real data analyses.

Keywords Biased estimator; Penalization; Sample proportion; Shrinkage proportion estimator.

1 Introduction

One of the simplest analyses is to find an estimator of binomial parameter p. The most popular estimator is the sample proportion X/n that is a minimum variance unbiased estimator where X follows binomial distribution with sample size n and proportion p. Shrinkage method which allows bias in the estimation has been explored under the regression model explicitly (Copas, 1983) or indirectly through

penalization since the work of Tibshirani (1996). Although shrinkage estimators prevail in the literature these days, there are only few shrinkage methods in the estimation of proportion presumably due to its simpleness. Examples are the works of Ahmed and Rohatgi (1996) and Singh et al. (2007) that proposed shrinkage estimators of proportion under randomized response survey where sensitive questions are often asked. One of the well-known shrinkage estimator of the proportion is a posterior mean with uniform prior (Bolstad, 2007, Chapter 9), which performs better than X/n when p is not close to 0 or 1 where X follows binomial distribution with n and p.

We propose a class of shrinkage estimators for proportion in Section 2. In Section 2.1, another aspect of the proposed methods based on the likelihood is discussed. The desirable extent of shrinkage and a simple rule for the amount of shrinkage are addressed in Section 2.2. In Section 2.2, we also establish theoretic advantage from the suggested shrinkage estimator, which supports improved performance across most of p values. In Section 3 and Section 4, we conduct simulation studies and real data analyses, respectively, to compare the proposed methods with sample proportion. We conclude the paper with discussions in Section 5.

2 Shrinkage Estimation of Proportion

Let X follows binomial distribution with sample size n and parameter p where $0 < p < 1$. We consider shrinking the sample proportion by the shrinkage parameter λ. That is,

$$\hat{p}(\lambda) = (1-\lambda)X/n = (1-\lambda)\tilde{p}, \tag{1}$$

where $\tilde{p}$ is a sample proportion. Since it is desirable that $0 \leq \hat{p}(\lambda) \leq 1$, the value of λ is restricted to $0 \leq \lambda \leq 1$. Then, mean squared error (MSE) of $\hat{p}(\lambda)$ is given as,

$$MSE(\hat{p}(\lambda)) = (1-\lambda)^2 p(1-p)/n + \lambda^2 p^2. \tag{2}$$

The MSE value of $\tilde{p}$, $MSE(\tilde{p})$, is simply its variance, $p(1-p)/n$. Then, we want (2) to be smaller than $p(1-p)/n$. Let $Q(\lambda) = MSE(\tilde{p}) - MSE(\hat{p}(\lambda))$. Now, it can be easily checked that $Q(\lambda) > 0$ is equivalent to

$$\lambda\{\lambda - \frac{2(1-p)}{np+1-p}\} < 0. \tag{3}$$

Noticing that $2(1-p)/(np+1-p)$ can be larger than 1, the range of λ which produce smaller MSE for $\hat{p}(\lambda)$ is

$$\begin{cases} 0 < \lambda < \frac{2(1-p)}{np+1-p} & \text{if } p > 1/(n+1) \\ 0 < \lambda \leq 1 & \text{if } p \leq 1/(n+1). \end{cases} \tag{4}$$

Any value of λ between 0 and 1 will produce smaller MSE when p is smaller than $1/(n+1)$. This is because $\tilde{p}$ is likely to be small when $p < 1/(n+1)$, thus any value of $\lambda \in (0,1)$ would bring small bias in $\hat{p}(\lambda) = (1-\lambda)\tilde{p}$. Further, we seek to find the value of λ which maximizes $Q(\lambda)$, the difference in MSE. By setting the derivative of $Q(\lambda)$ to be zero, the maximizer can be easily obtained as

$$\lambda_{opt} = \frac{1-p}{np+1-p}. \tag{5}$$

Note that $0 < \lambda_{opt} < 1$, and $\lambda_{opt} = O(1/n)$. Thus, the amount of shrinkage disappears as sample size increases and $\hat{p}(\lambda_{opt})$ is asymptotically unbiased, while

achieves smaller MSE in finite sample. Relative efficiency of $\hat{p}(\lambda_{opt})$ with respect to $\tilde{p}$ can be calculated as

$$\frac{MSE(\hat{p}(\lambda_{opt}))}{MSE(\tilde{p})} = \frac{np + 1 - p}{np} > 1. \tag{6}$$

The relative efficiency is greater than 1, and approaches to 1 as p moves to 1. Since p is unknown, λ_{opt} has to be estimated, which will be discussed later.

2.1 Perspective of Likelihood

X/n is a maximum likelihood estimator (MLE) whose log likelihood is given as $X \log p + (n - X) \log (1 - p) \equiv l_n$. In terms of likelihood, $\hat{p}(\lambda)$ in (1) can be understood as a penalized MLE, which maximizes the penalized likelihood of

$$pl_n = l_n - \lambda X \log p/(1 - p). \tag{7}$$

Then, differentiating (7) with respect to p results in

$$\frac{\partial}{\partial p} pl_n = X/p - (n - X)/(1 - p) - \lambda X\{1/p + 1/(1 - p)\}.$$

Now, we set the above derivative equal to zero and solve the equation to obtain the penalized MLE, $\hat{p}(\lambda) = X(1 - \lambda)/n$, which is equivalent to (1).

The $logit$ function $\log p/(1 - p)$ is monotone increasing (or, $-X \log p/(1 - p)$ is monotone decreasing). Further, its absolute value is symmetric about $p = 1/2$ and attains 0 when $p = 1/2$. If λ were allowed to have a negative value when $p > 1/2$, the proposed $logit$-type penalty has an effect of encouraging estimation towards 0 when $p < 1/2$, while discouraging towards 0 (or, encouraging estima-

tion towards 1) when $p > 1/2$. However, we restricted the value of λ within $[0, 1]$ in (1), thus resulted in $\lambda_{opt} \in (0, 1)$. Since we did not allow negative λ, (or, allow shrinkage only towards 0), we always shrink the proportion estimate toward 0. Then, it is not desirable to shrink the sample proportion to zero when, in fact, the true p is close to one. However, the amount of shrinkage we want by λ_{opt} is close to 0 when p is large as shown in Figure 1. In other words, when p is close to 0, λ_{opt} shrink the sample proportion towards 0 severely, whereas shrink the sample proportion towards 0 very slightly when p is close to 1. This is because it is not desirable to shrink the sample proportion to zero when, in fact, the true p is close to 1.

2.2 Estimation of λ_{opt}

We observed that wide range of λ can be used when p is small in (4). On the contrary, the length of interval for λ in (4), $2(1-p)/(np+1-p)$ decreases as p increases to 1. This implies that the estimation of λ_{opt} should be more careful when p is close to 1. We found that employing X/n for the estimation of λ_{opt} does not yield satisfactory performance. We also found that the over-estimation of λ_{opt} often lead to unacceptably large bias, while we still get benefits from the bias-variance tradeoff with under-estimated λ_{opt} from empirical studies. Figure 1 shows the value of λ_{opt} at various sample size. When p is close to 0, the λ_{opt} value is relatively large, and it decreases to 0 as p goes to 1. Since the under-estimation of λ_{opt} is somewhat beneficial, the estimation of λ_{opt} is quite easy when p is close to 0, but it is much difficult when p is close to 1. This fact leads us to choose λ_{opt} based on p close to 1. Although $\lambda_{opt} = (1-p)/(np+1-p) = O(1/n)$, the numerator $(1-p)$ becomes small for large p. So, $\hat{\lambda}_{opt} = 1/n$ was observed to

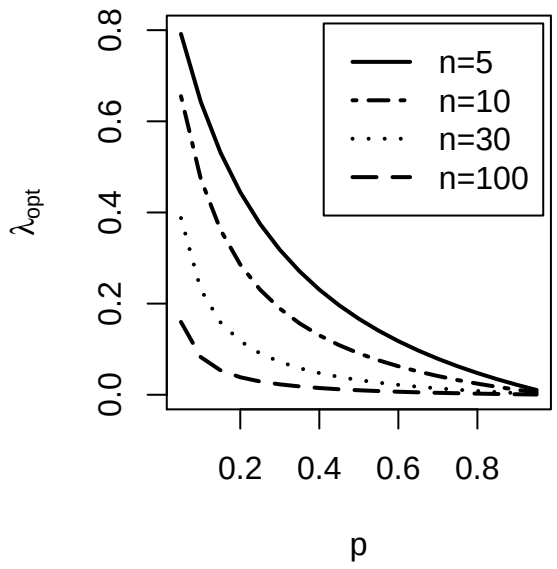

Figure 1: Plot of λ_{opt} against p at various sample size n.

be not small enough when p is large, although works good for small p as will be described in Section 4. Instead, $\hat{\lambda}_{opt} = 1/n^2$ often works satisfactory, thus it is selected as our estimator of λ_{opt}. Then, after some algebra with $\hat{\lambda}_{opt} = 1/n^2$, we have,

$$MSE(\hat{p}(\hat{\lambda}_{opt})) - MSE(\tilde{p}) \leq 0 \qquad \text{for} \quad 0 < p \leq p_{max} = \frac{2n^2 - 1}{2n^2 + n - 1}. \quad (8)$$

It can be easily checked that $p_{max} > 0.95$ for any $n > 9$. Thus, we can get more accurate proportion estimator with the suggested $\hat{\lambda}_{opt}$ compared to the sample proportion in most values of p. When sample size is large, the suggested estimator has asymptotic normality.

Theorem 1 *For any* $0 < p < 1$*, we have,*

$$T = \frac{\hat{p}(\hat{\lambda}_{opt}) - p}{\sqrt{var(\hat{p}(\hat{\lambda}_{opt}))}} \xrightarrow{d} N(0,1), \tag{9}$$

as $n \to \infty$.

Proof. First, re-express T as,

$$\frac{(1-1/n^2)X/n - p}{\sqrt{(n^2-1)^2 p(1-p)/n^5}} = \frac{X/n - p}{\sqrt{p(1-p)/n}}\left(\frac{n^2}{n^2-1}\right) - \frac{X/n}{\sqrt{p(1-p)/n}}\left(\frac{1}{n^2-1}\right). \tag{10}$$

Since $(X/n)/\sqrt{p(1-p)/n}$ converges to $N(p,1)$ in distribution by CLT, and $n^2/(n^2-1) \xrightarrow{p} 1$, the first quantity in the RHS converges to $N(0,1)$, while the second quantity in the RHS is $O_p(1/n^2)$. Applying Slutsky's Theorem completes the proof. □

In the perspective of the penalized likelihood, it might sound better to change the sign of λ according to the sign of $p-1/2$. But, due to the fact that p is unknown, changing the sign of λ appropriately is quite a difficult task. In fact, changing the sign of λ according to the sign of $X/n-1/2$ was far from satisfaction under our experiments. When one consider penalization methods, cross-validation is one of the most popular ways for choosing a proper value for the penalty parameter. However, several versions of cross-validation (such as leave-one-out, 2-fold, 5-fold, and, 10-fold) did not lead to adequate estimation of λ_{opt} at universal p, but showed superior performance only at certain range of p. Another estimation method such as bootstrap (Efron and Tibshirani, 1994) might look promising. So, we tried a two-step estimator such as a procedure of estimating λ_{opt} by bootstrapping in advance, then using it for $\hat{p}(\hat{\lambda}_{opt})$. It certainly yielded better estimation of p when p is close to 0,

but not when p is close to 1. Thus, we stick to $\hat{\lambda}_{opt} = 1/n^2$, and use this rule for the simulations in the next section.

3 Simulation studies

To compare the performance of $\hat{p}(\hat{\lambda}_{opt})$ with X/n, we generate 300,000 Monte Carlo (MC) samples from Binomial distribution with n=10, 15, and 20, and various p from 0.05 to 0.95. MSE is used to measure the discrepancy between the true p and the two estimates. For the estimation of λ, $1/n^2$, which is a rule explored in Section 2.2 is always considered. In addition, we employ λ_{opt} to see the 'achievable' lower bound of MSE by more accurate estimation of λ_{opt}. The mean of MSE and its standard error with selected n and p are given in Table 1.

Table 1: Mean value of MSE and its standard error (in parenthesis) from 300,000 MC samples. '*% dec.*' represents percentage reduction in MSE by switching from X/n to $\hat{p}(1/n^2)$. All values are multiplied by 1000.

p	0.05	0.10	0.25	0.50	0.75	0.90	0.95
				n=10			
X/n	4.76 (0.02)	9.06 (0.03)	18.8 (0.05)	25.1 (0.06)	18.8 (0.05)	9.06 (0.03)	4.76 (0.02)
$\hat{p}(1/n^2)$	4.67 (0.02)	8.88 (0.03)	18.5 (0.05)	24.6 (0.06)	18.5 (0.05)	8.96 (0.03)	4.76 (0.02)
$\hat{p}(\lambda_{opt})$	1.64(0.004)	4.75 (0.01)	14.5 (0.03)	22.8 (0.05)	18.2 (0.05)	8.96 (0.03)	4.74 (0.02)
% dec.	1.98	1.98	1.96	1.88	1.68	1.07	0.11
				n=15			
X/n	3.19 (0.01)	6.03 (0.02)	12.5 (0.03)	16.7 (0.04)	12.5 (0.03)	6.03 (0.02)	3.19 (0.01)
$\hat{p}(1/n^2)$	3.16 (0.01)	5.98 (0.02)	12.4 (0.03)	16.5 (0.04)	12.4 (0.03)	6.00 (0.02)	3.18 (0.01)
$\hat{p}(\lambda_{opt})$	1.40(0.002)	3.75 (0.01)	10.4 (0.02)	15.7 (0.04)	12.3 (0.03)	5.99 (0.02)	3.17 (0.01)
% dec.	0.89	0.89	0.88	0.85	0.79	0.60	0.30
				n=20			
X/n	2.40 (0.007)	4.54 (0.01)	9.40 (0.02)	12.5 (0.03)	9.40 (0.03)	4.54 (0.01)	2.40 (0.007)
$\hat{p}(1/n^2)$	2.39 (0.007)	4.52 (0.01)	9.35 (0.02)	12.5 (0.03)	9.36 (0.02)	4.52 (0.01)	2.39 (0.008)
$\hat{p}(\lambda_{opt})$	1.22 (0.002)	3.11 (0.01)	8.16 (0.02)	11.9 (0.03)	9.25 (0.03)	4.52 (0.01)	2.39 (0.008)
% dec.	0.50	0.50	0.50	0.49	0.46	0.37	0.24

In case of $n = 10$, there are MSE reduction of about 1.5% to 2.0% with the suggested method compared to X/n, and the amount of reduction drops to

near zero when p is close to 1. With the λ_{opt}, MSE reduction is surprisingly low when p is close to zero. Thus, we see that an improved estimation of λ_{opt} is needed, although $\hat{\lambda}_{opt} = 1/n^2$ shows superior performance over the sample proportion in various p. The overall tendency for $n = 20$ is similar to the case of $n = 10$ with the smaller amount of reduction, and this is expected with a minute shrinkage of $\hat{\lambda}_{opt} = 1/20^2$. The results from aforementioned methods such as bootstrapping and cross-validation for the estimation of λ_{opt} do not show universal MSE reduction, thus they are omitted.

4 Applications to real data

Carseats data set from James et al. (2013) contains 400 samples with 11 variables regarding sales information of car seats. The data set is available in R package `ISLR`. One of the variables indicates whether the location of the car seats' stores is U.S. or not. As we do not know the true population proportion of the car seat stores in U.S., we treat the 400 samples as a population. From the data set, the (assumed) population proportion is 0.645, which will become our target. Now, we take random samples of size n = 5, 10, 15, and 20 from the population ($n = 400$) and estimate the proportion by the sample proportion $\tilde{p} = X/n$, and by the suggested method $\hat{p}(1/n^2)$. Then, the squared distance between the true p (0.645) and the estimated values is calculated. We repeat this procedure for 100,000 times and the mean of MSE values from X/n and $\hat{p}(1/n^2)$ are recorded in Table 2. The percentage reduction in MSE (*% dec.*) shows the superior performance of $\hat{p}(1/n^2)$ over X/n. In addition, we present the results from $\hat{p}(1/n^{1.5})$ and $\hat{p}(1/n)$. Overall, the results from $\hat{p}(1/n^{1.5})$ and $\hat{p}(1/n)$ are also acceptable.

Another data set we examine is a well-known *Iris* data set (Fisher, 1936) where the measurements of three species of iris were recorded. The data set is available in R package `datasets`. There are 50 flowers from each of three species of *Iris setosa, versicolor*, and *virginica*. Again, we regard this data set as an (unrealistic) population where 1/3 is the true population proportion of each kind. Then, we select random samples of size n = 5, 10, 15, and 20 from the whole data set ($n = 150$) and repeat the same procedure as we performed with *Carseats* data set. The results are shown in Table 2.

Table 2: Mean value of MSE and its standard error (in parenthesis) from 100,000 MC samples. '*% dec.*' represents percentage reduction in MSE by switching from X/n to $\hat{p}(1/n^2)$. All values are multiplied by 1000.

n	5	10	15	20
		Carseats data	($p = 0.645$)	
X/n	45.60 (0.186)	22.51 (0.097)	14.87 (0.065)	10.97 (0.048)
$\hat{p}(1/n^2)$	42.67 (0.182)	22.10 (0.096)	14.75 (0.064)	10.91 (0.048)
$\hat{p}(1/n^{1.5})$	41.10 (0.183)	21.52 (0.095)	14.48 (0.064)	10.77 (0.048)
$\hat{p}(1/n)$	45.75 (0.201)	22.38 (0.102)	14.79 (0.067)	10.94 (0.050)
% dec.	6.42	1.81	0.84	0.48
		Iris data	($p = 1/3$)	
X/n	43.27 (0.178)	22.01 (0.090)	13.49 (0.059)	9.763 (0.043)
$\hat{p}(1/n^2)$	40.07 (0.161)	20.61 (0.088)	13.37 (0.058)	9.715 (0.043)
$\hat{p}(1/n^{1.5})$	36.79 (0.144)	19.81 (0.084)	13.06 (0.056)	9.560 (0.042)
$\hat{p}(1/n)$	32.18 (0.119)	18.13 (0.074)	12.24 (0.052)	9.090 (0.039)
% dec.	7.41	1.94	0.87	0.49

In Table 2, we can see the MSE values of the proposed methods are significantly lower than those from the sample proportion. The amount of MSE reduction achieved from both the data sets are about 7% when sample size n is 5. The benefits we obtained from the suggested methods gradually decrease as the sam-

ple size increases since the proposed methods are asymptotically equivalent to the sample proportion.

5 Discussion

In this article, we suggest a class of shrinkage estimators for proportion. We provide the desirable amount of shrinkage, and a simple rule which is a reciprocal of the squared sample size. Although the suggested method outperforms the sample proportion, more accurate estimation of the shrinkage parameter will improve the suggested methods.

Another perspective of the shrinkage estimator as a penalized maximum likelihood estimator, which was reviewed in detail by van Houwelingen (2001) provides wider view of shrinkage estimators, thus give us clues to shrink in other ways. For example, the likelihood of X with a penalty term of $\lambda \log p(1-p)$ results in $(X+\lambda)/(n+2\lambda)$, which is another proportion estimator shrunken towards 0.5. With the choice of $\lambda = 1$, it is also a well-known Bayesian posterior mean with uniform prior. Note that X/n produce larger MSE for p around 0.5 and smaller MSE around 0 and 1. In overall, this pattern is maintained when $(X+1)/(n+2)$ is compared with our suggested estimators. For a future research, it may worthwhile to explore desirable amount of shrinkage in $(X+\lambda)/(n+2\lambda)$ as we demonstrated in this article. Extension of the suggested idea to the shrinkage of the regression parameters could be another interesting topic.

References

Ahmed, S., Rohatgi, V., 1996. Shrinkage estimation of the proportion in randomized response. Metrika 43, 17 – 30.

Bolstad, W. M., 2007. Introduction to Bayesian Statistics (2nd Edition). John Wiley & Sons, Inc., Hoboken, New Jersey.

Copas, J. B., 1983. Regression, prediction and shrinkage. Journal of the Royal Statistical Society. Series B (Methodological) 45 (3), 311 – 354.

Efron, B., Tibshirani, R., 1994. An Introduction to the Bootstrap. Chapman & Hall/CRC.

Fisher, R., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7 (Part II), 179 – 188.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. Springer-Verlag New York.

Singh, H. P., Shukla, S., Mathur, N., 2007. Shrinkage estimation of proportion of population possessing stigmatizing character in unrelated question randomized response technique. Journal of Indian Society of Agricultural Statistics 61 (1), 1 – 13.

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

van Houwelingen, J. C., 2001. Shrinkage and penalized likelihood as methods to improve predictive accuracy. Statistica Neerlandica 55 (1), 17 – 34.