

Bayesian Shrinkage Estimation of the Relative Abundance of mRNA Transcripts using SAGE

Jeffrey S. Morris, Keith Baggerly, and Kevin Coombes ¹

SUMMARY

Serial Analysis of Gene Expression (SAGE) is a technology for quantifying gene expression in biological tissue. SAGE experiments yield count data that can be modeled by a multinomial distribution with two characteristics: skewness in the relative frequencies and small sample size relative to the dimension. As a result of these characteristics, a given SAGE sample will fail to capture a large number of expressed mRNA species present in the tissue. Empirical estimators of mRNA species' relative abundance effectively ignore these missing species, and as a result tend to also overestimate the abundance of the scarce observed species, which make up a vast majority of the total. We have developed a new Bayesian estimation procedure that attempts to quantify our prior information about the population's characteristics, and as a result give better estimation of the relative abundance profiles, given an estimate of the number of unique mRNA transcripts in the tissue. Our method uses a type of Mixture Dirichlet prior, which involves stochastically partitioning the species into abundant and scarce classes, with each class modeled with its own multivariate prior, a scalar multiple of a Symmetric Dirichlet. The resulting estimators have nonlinear shrinkage profiles. We conduct simulation studies which show that our estimator has lower integrated mean squared error than the MLE for the SAGE scenarios simulated, and resulted in relative abundance profiles closer in Euclidean distance to the truth for 100% of the samples simulated. We apply our method to a SAGE library of normal colon tissue, and discuss its implications for assessing differential expression.

Key Words and Phrases: Bayesian Methods; Bayesian Model Averaging; Bioinformatics; Mixture Distributions; Multinomial Distribution; SAGE; Shrinkage Estimators.

Short Title: Shrinkage estimation in SAGE

¹Jeffrey S. Morris, Keith Baggerly, and Kevin Coombes are with the Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Box 447, Houston, TX 77030-4009. This work was partially supported by NIH-NCI Grant 1U19 CA84978-1A1. The authors would like to thank Natalie Blades for introducing us to the log-log plots for clearly displaying the distribution of true frequencies for SAGE tags. We also thank the associate editor and two referees, whose insightful comments have led to a considerable improvement of this paper.

1 Introduction

Serial analysis of gene expression (SAGE) is an experimental method for assessing gene expression in a collection of cells. In a typical SAGE experiment, the cells of interest are used to prepare a sample containing large quantities of RNA (about 500 μg of total RNA, or about 10^{18} mRNA molecules). A prespecified number n of mRNA molecules are selected from the sample, and short sequences (or *tags*, each 10 nucleotides long) from a unique position within the molecules are sequenced. The number sampled typically lies in the range of $n = 10,000 - 100,000$ tags. The results of the experiment are recorded as a list of distinct tags with observed frequency counts. See Polyak and Riggins (2001) for a survey article giving more details on SAGE.

To illustrate the properties of SAGE data and frame our discussion, we consider a SAGE library of size $n = 49,610$ taken from an healthy individual's colon epithelial tissue. This library, referred to as *nc1*, is taken from Zhang et al. (1997), and is publically available on the world wide web (<http://www.sagenet.org/SAGEData/NC1.htm>). In this sample, 17,703 distinct tags were seen, with observed frequencies ranging from one up to a maximum of 1286. The actual number of unique mRNA transcripts expressed in this tissue is likely larger than 17,703, since there are probably numerous mRNA species present in the tissue but not observed in our sample. Stollberg and colleagues (2000) attribute this to sampling error, which they list as one of the four major sources of quantitative error in SAGE. Stollberg and colleagues also developed a method to estimate the number of unique mRNA transcripts expressed, which they find to be $k = 25,336$ in this normal colon tissue. Note that the sample size n is of the same order of magnitude as the number of unique tags k ($n/k \approx 2$ here), which is nearly always the case in SAGE experiments.

Upon plotting a histogram of these SAGE counts, it becomes apparent that the distribution of relative frequencies is heavily skewed right (see Figure 1). Of the 17,703 distinct

tags observed, 13,426 (75.8%) appeared only once in the sample, 1872 (10.6 %) twice, 714 (4.0%) three times. Only 3.6% of the tags had observed frequencies of 9 or greater, yet these accounted for 50% of the total mRNA mass. Skewness of this magnitude is seen in all published SAGE data, and suggests it is a characterization of gene expression data to have many "scarce" mRNA species with low expression levels, and a small number of "abundant" species with much higher expression levels.

Since SAGE is an open system, one can characterize the transcriptome, yielding estimates of the number of unique mRNA transcripts expressed in a given tissue as well as their relative abundance (e.g. see Velculescu et al. 1999). Estimation of the number of unique tags in the tissue involves "estimating the number of species" from incomplete multinomial sampling, a very challenging statistical problem (see Bunge and Fitzpatrick 1993 for review). Not being our focus here, we assume this quantity is known, and demonstrate that our results are not sensitive to its misspecification.

The relative abundance of transcripts is quantified by the relative frequencies of the unique tags. Empirical estimates are easily computed by taking the ratio of the counts for each unique tag and n . These estimators ignore the "missing tags" as well as the skewness characteristic of gene expression data, giving them some undesirable properties (see Section 2.1). Our primary focus in this paper is to develop a new statistical method that appropriately considers these factors, yielding nonlinear shrinkage estimators that have efficiency advantages over the MLEs. Our method is based on a fully specified coherent probability model, so can be used to perform any desired inferences and is potentially expandable to account for other aspects of SAGE data not modeled in this paper.

Frequently, the underlying goal of SAGE experiments is not simply estimation of these true expression levels, but rather the identification of tags differentially expressed between two tissue types. We note that estimation of true expression levels necessarily underlies the

problem of differential expression, so one may reasonably conjecture that assessments based on more efficient estimators may also have improved sensitivity/specificity for detecting differentially expressed tags.

One problem with SAGE data is that sequencing errors can occur, which tends to introduce a number of false small count tags that often differ from the true sequence by one or two bases. Before using our method, we assume that a method for detecting and adjusting for these errors (e.g. Blades 2002) has been applied to the raw data.

An outline of the remainder of the paper is as follows. In Section 2, we propose the multinomial model for SAGE data and demonstrate the weaknesses of some standard estimators. In Section 3, we describe our method, a Bayesian method incorporating a Mixture Dirichlet prior. We explain its shrinkage properties and outline how to fit the resulting model using MCMC. In Section 4, we apply our method to the *nc1*. In Section 5, we describe a simulation study. Section 6 contains discussion of the implications of our results.

2 Multinomial Model for SAGE Data

Let X_i be the number of occurrences of tag i in our SAGE sample of size n , for $i = 1, \dots, k$. Recall k is the total number of expressed tags in the tissue, assumed known. If we assume the mRNA transcripts have been selected randomly and ignore the possibility of sequencing errors, it is natural to model the vector of counts $\underline{X} = (X_1, \dots, X_k)$ as a draw from a multinomial distribution with parameters n and $\underline{\pi}_k = (\pi_1, \pi_2, \dots, \pi_k)^T$. The joint pmf for multinomial vector \underline{X} can be written $f(\underline{X}|\underline{\pi}_k, n) = n! \prod_{i=1}^k (\pi_i^{X_i}/X_i!)$, with $\sum_{i=1}^k X_i = n$, $0 < \pi_i < 1 \forall i$ and $\sum_{i=1}^k \pi_i = 1$.

The vector of relative frequencies $\underline{\pi}_k$ represents the relative abundance of each tag in the tissue, and is the quantity we are interested in estimating in this paper. Since it is estimated that there are 300,000 total mRNA transcripts in a single cell (Hastie and Bishop, 1976), as in Zhang, et al. (1997), we report relative frequencies in SAGE as a fractions over 300,000,

roughly corresponding to the average number of copies per cell.

2.1 Maximum Likelihood Estimation

The empirical estimators discussed in Section 1 are routinely used in analysis of SAGE data. They can be shown to be the maximum likelihood estimators for the $\underline{\pi}_k$, and are unbiased and asymptotically efficient, so have nice classical frequentist properties. Under our conditions, however, we have what is effectively a very small sample, since although n may be large, k is of the same order of magnitude, and many of the multinomial classes are scarce. The MLE performs well for the relatively few abundant species, but has some undesirable properties for the more scarce species comprising a vast majority of the total unique tags. For a given SAGE sample, it automatically underestimates the relative frequencies of all missing species, and as a result tends to overestimate the relative frequencies of the scarce observed species.

Whenever $X_i = 0$, the corresponding empirical estimator is $\hat{\pi}_{i,\text{MLE}} = 0$, which is outside the parameter space. Thus, for the missing tags, we know that $\hat{\pi}_{i,\text{MLE}} < \pi_i$, so we could say that tag i is *underrepresented* in the given sample. From the relative frequency constraint $\sum_{i=1}^k \pi_i = 1$, it follows that $\sum_{\{i: X_i > 0\}} \hat{\pi}_{i,\text{MLE}} > \sum_{\{i: X_i > 0\}} \pi_i$, which implies that, on average, the observed tags are *overrepresented* in the given sample. Considering basic multinomial sampling properties, the more scarce tags are most likely to be overestimated.

To illustrate this point, following is a simple example. Consider a multinomial population with 51 tags, one abundant tag with a relative frequency of $\pi_0 = 0.50$ and the others all scarce with frequencies $\pi_i = 0.01$ for $i = 1, \dots, 50$. Consider a hypothetical sample of size 20 from this population. On average, about 40 of the 50 scarce tags will be missing, with the remaining 10 occurring once. Thus, for the scarce tags, we either have an estimator that is out of the parameter space (0) or at best much larger than the true value (0.05). While the MLE is MVUE, by the sampling characteristics of the problem, the estimator is limited as to how well it can estimate the relative frequencies of the scarce tags for a given sample.

In SAGE, the population is more complex, but these principles transfer. The sampling characteristics of SAGE data cause the MLE to have undesirable properties for estimating the relative frequencies of the scarce tags comprising a vast majority of the total tags in a given tissue. An estimator that yields positive estimates for the missing tags and shrinks the estimates corresponding to the others may have efficiency advantages over the MLE.

It has been shown to be possible in various multivariate settings to construct shrinkage estimators that uniformly dominate MLEs with respect to IMSE, the mean square error summed over the multivariate parameters (e.g. James and Stein 1961, George 1986, Gruber 1998). These estimators have an inherent Bayesian flavor to them, and effectively work by taking the MLEs and shrinking them towards a specified prior mean. The prior structure can be chosen to obtain significant efficiency gains in regions of the parameter space that are of particular interest to the investigators. Much of the work involving shrinkage estimation has been done in the setting of multivariate normal, and involves linear shrinkage. In our multinomial setting, shrinkage estimators can be constructed using a Bayesian approach, requiring the specification of a prior distribution on the $\underline{\pi}_k$.

2.2 Naive Bayesian estimation

A standard Bayesian approach for estimation of multinomial probabilities is to assume a conjugate Dirichlet prior for the set of probabilities $\underline{\pi}_k$, then compute their posterior distribution conditional on the multinomial sample. The joint density for a Dirichlet random variable $\underline{\pi}_k$ of dimension k is given by $f(\underline{\pi}_k|\theta_1, \theta_2, \dots, \theta_k) = \{\Gamma(\sum_{i=1}^k \theta_i) / \prod_{i=1}^k \Gamma(\theta_i)\} \prod_{i=1}^k \pi_i^{\theta_i-1}$, with $\sum_{i=1}^k \pi_i = 1$. Note that this prior satisfies the relative frequency constraints, and ensures that the posterior will, as well.

When there is no prior knowledge on which tags are more likely than others, common practice is to set all Dirichlet parameters to be equal a priori, i.e. $\theta_i \equiv \theta$, which we refer to as a *Symmetric Dirichlet*, or $\text{SymmDir}(\theta)$. Under this prior, the posterior mean estimator

for each relative frequency is $\hat{\pi}_{i,\text{DIR}} = (X_i + \theta)/(n + k\theta)$. A common (and in this case naive) choice for the hyperparameter is $\theta = 1$, described by Jeffreys (1948, Section 3.23) and corresponding to a k -dimensional generalization of the uniform distribution.

This Bayesian estimator can be viewed as taking the MLEs and shrinking them linearly towards the prior mean, which is k^{-1} . The shrinkage is stronger for larger θ , with the estimator being closer to the MLE as θ approaches 0. This results in nonzero estimates for all tags, and shrunken estimates for all tags with observed relative frequencies greater than $1/k$. Applied to our simple example above, a SymmDir(1) prior would yield posterior means of $\hat{\pi}_{i,\text{DIR}} = 1/71 \approx 0.014$ whenever $X_i = 0$, and $\hat{\pi}_{i,\text{DIR}} = 2/71 \approx 0.028$ when $X_i = 1$. These estimates are both closer to the true value 0.01 than the corresponding MLEs. In fact, straightforward calculations show that this estimator has uniformly smaller squared error loss than the MLE for *all* sparse tags in any and every multinomial sample of size $n \leq k = 51$ taken from the population of our example.

The linear shrinkage inherent in this prior, however, will cause it to perform much worse than the MLE for the abundant tags. In the example above, if the abundant tag is observed $X_0 = 10$ times in a given sample, the posterior mean estimator would be $\hat{\pi}_{i,\text{DIR}} \approx 0.15$, versus an MLE of $\hat{\pi}_{i,\text{MLE}} = 0.50$. The linear shrinkage to the mean causes values further from k^{-1} to be shrunken the most, inducing extremely large biases that cause it to perform very poorly for more abundant tags.

The method performs poorly because the prior inaccurately represents the characteristics of the population under consideration – it assumes that *a priori* we think that all unique tags are exchangeable, with relative frequencies of $1/k$, while in reality they are very heterogeneous. There is an unfortunate connotation of "non-informativeness" that goes with the uniform distribution, which is inaccurate since its symmetry implies prior information of homogeneity, which in the Symmetric Dirichlet case is effectively a strong prior belief when

k is large relative to n , and θ is not too small.

Thus, we see from our simple example that the MLEs fail to incorporate important prior information about the problem and, as a result, their performance for a given SAGE sample is less than ideal for the scarce tags, and Bayesian methods based on a "uniform" Symmetric Dirichlet prior make inaccurate prior assumptions that lead to poor performance for the abundant tags. These results carry over to SAGE data, as we demonstrate in Section 5. We would like to find an alternative method that obtains improved estimators for the scarce tags without sacrificing so much efficiency on the abundant ones, which requires nonlinear shrinkage. This method should also ensure the relative frequency constraint is satisfied, and appropriately take into account known prior information.

3 Bayesian Estimation Using Mixture Dirichlet Prior

Recall that by the characteristics of gene expression, we expect that there are a large number of scarce tags and a small number of abundant ones. However, we often do not know *a priori* which tags will be abundant. We quantify this prior knowledge using a type of mixture Dirichlet prior for the $\underline{\pi}_k$. With this prior, the k unique tags are stochastically partitioned between two discrete classes representing "abundant" and "scarce" tags, each with their own separate multivariate distributions, scalar multiples of Symmetric Dirichlets.

3.1 Prior structure

In defining our prior, we first reparameterize the set of parameters corresponding to the true relative frequencies $\underline{\pi}_k$ to the set of parameters $\{\underline{\lambda}, \pi^*, \underline{q}\}$. Each unique tag $i = 1, \dots, k$ is assumed to belong to one of two classes, either "abundant" or "scarce", with λ_i being the indicator of whether tag i belongs to the abundant class. $\underline{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_k)^T$ is the vector of such indicators for all tags. Assuming we don't know the identity of the abundant tags *a priori*, we assume the λ_i are independent and identically distributed Bernoulli(P) random variables, where P is the expected proportion of unique tags belonging to the abundant class.

It is useful to represent the partitioning of the k tags into the two classes by introducing the following notation. The indices of those unique tags belonging to the abundant and scarce classes are given by the sets $\mathcal{A} = \{i : \lambda_i = 1\}$ and $\mathcal{S} = \{i : \lambda_i = 0\}$, respectively, which are of length $k_A = \sum_{i=1}^k \lambda_i$ and $k_S = \sum_{i=1}^k (1 - \lambda_i)$. $\pi^* = \sum_{i \in \mathcal{A}} \pi_i$ represents the "abundant mass", or the total mass of the tags belonging to the abundant class. Likewise $1 - \pi^*$ represents the "scarce mass". π^* is given a $\text{Beta}(\alpha_{\pi^*}, \beta_{\pi^*})$ prior.

The vector \underline{q} of dimension k contains the relative frequencies for each tag *within its class*. Given the class membership parameters $\underline{\lambda}$, \underline{q} is partitioned into $\underline{q}_A = \underline{\pi}_A / \pi^*$ and $\underline{q}_S = \underline{\pi}_S / (1 - \pi^*)$, which are given separate Symmetric Dirichlet priors with parameters θ_A and θ_S , respectively. Note that the interpretation of q_i depends on the class to which tag i belongs. If abundant ($\lambda_i = 1$), then q_i represents the proportion of abundant mass attributable to tag i , while if scarce ($\lambda_i = 0$), then q_i is the proportion of scarce mass attributable to tag i . This construct allows the two classes to have separate multivariate distributions, yet retains the relative frequency constraint $\sum_{i=1}^k \pi_i = 1$.

Following is a summary of our Mixture Dirichlet prior structure. $\underline{q}_A | \underline{\lambda} = \text{Symmetric Dirichlet}(\theta_A)$, $\underline{q}_S | \underline{\lambda} = \text{Symmetric Dirichlet}(\theta_S)$, $\lambda_i = \text{i.i.d. Bernoulli}(P)$, $\pi^* = \text{Beta}(\alpha_{\pi^*}, \beta_{\pi^*})$. The relative frequencies for the unique tags are easily constructed using these quantities as follows: $\pi_i = \{q_i \pi^*\}^{\lambda_i} \{q_i (1 - \pi^*)\}^{(1 - \lambda_i)}$.

3.2 The rich, the poor, and Robin Hood

We now present an analogy to further illuminate the heuristics behind this problem, and why we believe nonlinear shrinkage estimators are appropriate. For tags with zero counts, the standard empirical estimates (MLEs) give zero relative frequency estimates, which are outside the parameter space. Given we know they exist, we would like to give some positive probability to these tags, but because of the relative frequency constraint that the total probability mass must sum to one, this requires decreasing some the estimates for the other

observed tags. There’s no free lunch – in order to “pay” the zero count tags, we need to “steal” some probability mass from other tags.

A Bayesian estimator will do this. The posterior mean can be written as a weighted average of the MLE and prior mean, and thus it effectively shrinks the MLEs towards the prior mean. In this way, we obtain positive estimates for the zero counts and subsequently shrink the estimates for the observed tags towards their prior mean. The choice of prior determines the shrinkage rule, i.e. who we “steal from” in order to “pay the zeros”.

By performing linear shrinkage, the simple Symmetric Dirichlet prior is what we could call a “Robin Hood” prior. That is, it steals the most mass from the “richest” or most abundant tags. This makes sense if we actually believe that the tags are exchangeable, since then large counts are taken to be a aberrations for which the most reasonable course of action is to shrink the most. However, in settings where the exchangeable hypothesis is not true, like SAGE, this type of shrinkage is undesirable, leading to estimators with poor properties.

What we would like in our setting is a “reverse Robin Hood” prior. We would like to “steal” from the most “poor” tags (with low counts) and leave the “rich” alone, since the sampling properties of the problem suggest the poorer tags are those holding on to the mass rightfully belonging to the zero tags. In other words, we would like an estimator that shrinks the MLEs nonlinearly, whereas tags with large counts are left largely unaffected, but those with small counts shrunken. This can be accomplished using our Mixture Dirichlet Prior.

3.3 Nonlinear shrinkage and the Mixture Dirichlet prior

We have heuristically motivated the utility of nonlinear shrinkage estimation in SAGE, and we describe how our Mixture Dirichlet prior yields nonlinear shrinkage profile.

For the following illustration, suppose we know the values of P , π^* , and $\underline{\lambda}$. In this case, the posterior mean relative frequency estimators for scarce tag i and abundant tag i' can be

written in closed form as

$$\widehat{\pi}_{i,STRAT} = (1 - \pi^*) \left(\frac{X_i + \theta_S}{n_S + k_S \theta_S} \right), \quad (1)$$

$$\widehat{\pi}_{i',STRAT} = \pi^* \left(\frac{X_{i'} + \theta_A}{n_A + k_A \theta_A} \right), \quad (2)$$

where k_A and k_S are the number of abundant and scarce tags, respectively, and $n_A = \sum_{i \in \mathcal{A}} X_i$ and $n_S = \sum_{i \in \mathcal{S}} X_i$ are the total observed counts for tags in the abundant and scarce classes.

Making use of the approximations $n_S/n \approx 1 - \pi^*$ and $k_S/k \approx 1 - P$, we can simplify (1) to be written as a linear combination of the MLE, $\widehat{\pi}_{i,MLE} = X_i/n$, and the prior mean for a scarce tag, $\pi_S = \{(1 - \pi^*)/(1 - P)\} * (1/k)$, with the weight placed on the MLE being $\gamma_S = n/\{n + (1 - \pi^*)^{-1}(1 - P)k\theta_S\}$, which is typically < 0.5 under the conditions of SAGE, with $n \sim k$, π^* large, and P small. For example, if $\pi^* = 0.50$, $P = 0.01$, $\theta_S = 1$, and $n/k = 1$, we get $\gamma_S = 0.33$. As a result, scarce tags are shrunken strongly towards their prior mean π_S , which is very close to zero. This shrinkage towards the mean causes species with zero counts to have positive relative frequency estimators, and causes tags with counts greater than $n * \pi_S$ to be shrunken.

Similarly, we can simplify (2) to be approximately written as a linear combination of the MLE and the prior mean for an abundant tag, $\pi_A = \{\pi^*/P\}(1/k)$, with the weight on the MLE being $\gamma_A = n/\{n + (\pi^*)^{-1}(P)k\theta_A\}$, which is ≈ 1 under the conditions of our problem. Using the example above, with $\theta_A = 0.5$ we get $\gamma_A = 0.99$. As a result, abundant tags are shrunken towards their prior mean π_A , but since $\gamma_A \approx 1$, the magnitude of this shrinkage is very weak, thus resulting in estimates close to the MLEs, as desired.

We see that zero count tags get positive estimates, the estimates for scarce tags are shrunken towards a small value, while the estimates for more abundant tags are very close to their MLEs. The estimates given above assume knowledge of $\underline{\lambda}$ and π^* , which we of course do not have, but these approximate results hold for tags whose membership in either the

scarce or abundant class given its observed count is unquestionable, i.e. whose posterior probability of $\lambda_i = 1$ is close to zero or one. For tags in between, the estimator is effectively a combination of (1) and (2), with the weight placed on each depending on the posterior probability of $\lambda_i = 1$. This can be viewed as Bayesian model averaging over models with tag i belonging to the abundant or scarce class, and yields a smooth nonlinear shrinkage profile that shrinks less for larger observed counts. Figure 2 plots the Symmetric and Mixture Dirichlet estimators versus the observed counts, with the solid line indicating the unshrunk MLEs, clearly demonstrating the linear and nonlinear shrinkage profiles.

3.4 Selection of prior hyperparameters

The prior hyperparameters θ_A , θ_S , and P work together to determine the shape of the shrinkage curve. We now discuss their effect, and give recommendations for their selection.

Generally, larger values for θ_A and θ_S tend to result in stronger shrinkage towards the prior mean within the respective classes. Since we would like to have little shrinkage for the abundant tags, a natural thought would be to make θ_A very close to zero. However, choosing θ_A too small relative to θ_S will make the boundary between scarce and abundant species very sharp, resulting in a nonsmooth shrinkage profile, which increases the MSE for tags near the boundary of the scarce and abundant classes (the intermediate tags described in Section 5). Thus, our general recommendation is to choose $\theta_A < \theta_S$ to reduce the amount of shrinkage in the abundant tags, but with θ_A not too close to zero.

The hyperparameter P represents the expected relative size of the abundant class. The model is sensitive to this parameter, with a degenerate mixture with linear shrinkage profiles resulting if P is chosen too small or too large (see Section 5). We recommend choosing P by trial and error, ensuring that the resulting shrinkage profile is nondegenerate. Given P , we have found that a Uniform(0,1) prior works well for π^* .

3.5 Fitting the Model

The posterior distribution of the $\underline{\pi}_k$ under the Stratified Dirichlet prior structure is not available in closed form as long as $\underline{\lambda}$ is unknown, so we use a Gibbs sampler for estimation.

Following are the steps of the Gibbs sampler, in order.

1. Sample $\underline{\lambda}$ from $f(\underline{\lambda}|\underline{X}, P)$, described below. Based on this sample, redefine the indices $\mathcal{A} = \{i : \lambda_i = 1\}$ and $\mathcal{S} = \{i : \lambda_i = 0\}$.
2. Sample π^* from its full conditional, which is $\text{Beta}(\alpha_{\pi^*} + n_A, \beta_{\pi^*} + n_S)$.
3. Sample $\underline{q}_{\mathcal{A}}$ from its full conditional, which is a Dirichlet of dimension k_A with parameters $\{X_i + \theta_A, i \in \mathcal{A}\}$.
4. Sample $\underline{q}_{\mathcal{S}}$ from its full conditional, a Dirichlet of dimension k_S with parameters $\{X_i + \theta_S, i \in \mathcal{S}\}$.

Recall k_A and k_S are the number of abundant and scarce tags, respectively, and $n_A = \sum_{i \in \mathcal{A}} X_i$ and $n_S = \sum_{i \in \mathcal{S}} X_i$ the total counts within each class.

In step 1, the λ_i are updated one at a time by drawing $u \sim \text{Uniform}(0, 1)$, and setting $\lambda_i = 1$ if $u < \alpha_i$, with the $\alpha_i = \text{Pr}(\lambda_i = 1 | \underline{\lambda}_{(-i)}, \underline{X}, P)$, where $\underline{\lambda}_{(-i)}$ is the set of all $\underline{\lambda}$ except λ_i . The expression for α_i can be written as $O_i / (O_i + 1)$, where O_i is the conditional posterior odds that tag i is abundant, which is the product of the prior odds $P / (1 - P)$ and the conditional Bayes Factor BF_i , given by

$$\begin{aligned}
 BF_i = & \left[\frac{\Gamma(n_{A(-i)} + \alpha_{\pi^*}) \Gamma(n_{S(-i)} + \beta_{\pi^*})}{\Gamma(n + \alpha_{\pi^*} + \beta_{\pi^*})} \right] * \left[\frac{\Gamma(\theta_A + X_i) \Gamma(\theta_S)}{\Gamma(\theta_A) \Gamma(\theta_S + X_i)} \right] \\
 & * \left[\frac{(n_{A(-i)} + X_i)! n_{S(-i)}!}{n_{A(-i)}! (n_{S(-i)} + X_i)!} \right] \\
 & * \left[\frac{\Gamma\{\theta_A k_{A(-i)} + n_{A(-i)}\} \Gamma\{\theta_A k_{A(-i)} + \theta_A\}}{\Gamma\{\theta_A k_{A(-i)}\} \Gamma\{\theta_A k_{A(-i)} + \theta_A + n_{A(-i)} + X_i\}} \right] \\
 & * \left[\frac{\Gamma\{\theta_S k_{S(-i)} + \theta_S + n_{S(-i)} + X_i\} \Gamma\{\theta_S k_{S(-i)}\}}{\Gamma\{\theta_S k_{S(-i)} + \theta_S\} \Gamma\{\theta_S k_{S(-i)} + n_{S(-i)}\}} \right]. \tag{3}
 \end{aligned}$$

$\Gamma(x) = \int_0^\infty \exp(-u) u^{x-1} du$ is the Gamma function, $k_{A(-i)} = \sum_{j \neq i} \lambda_j$ and $k_{S(-i)} = \sum_{j \neq i} (1 - \lambda_j)$ are the number of abundant and scarce tags, leaving out transcript i , and $n_{A(-i)} =$

$\sum_{j \neq i} \lambda_j X_j$ and $n_{S(-i)} = \sum_{j \neq i} (1 - \lambda_j) X_j$ are the total abundant and scarce counts, again leaving out transcript i . If one is willing to impose monotonicity (if $\lambda_i = 1$ then $\lambda_{i'} = 1$ for all $i' : X_{i'} > X_i$), then the parameter k_S is sufficient for the $\underline{\lambda}$, and samples from its posterior can be obtained using a single Metropolis step.

For illustration, with $n = 10,000$, $\theta_S = 1$ and $\theta_A = 0.5$, $P \approx 0.01$ and $\pi^* \approx 0.40$, when $X_i = 1, 2$ and 3 , we have $\alpha_i \approx 0.007, 0.03$, and 0.12 , respectively. When $X_i > 9$, we see $\alpha_i > 0.995$. Whenever $X_i = 0$, we set $\lambda_i = 0$ with probability 1, which considerably saves computational time. Various computational tricks were used for efficient updating of the λ_i , which is by far the most costly computational step in the procedure. The C++ code for performing the MCMC and the derivations for all the conditional distributions given above is available by request from the first author.

4 Example: Application to *nc1* Data Set

We now apply our method to *nc1* to estimate the relative abundance of the expressed transcripts in this individual's normal colonic tissue. We start by assuming the number of unique expressed mRNA transcripts is $k = 25,336$, as estimated for this tissue by Stollberg, et al. (2000). Our simulation studies in the next section suggest that our method's performance is robust to k , so this estimate seems sufficient. Before fitting our model, we correct for likely sequencing errors using Blades (2002), yielding 17,264 unique tags, with a maximum revised count of 1314. The Dirichlet parameters used were $\theta_S = 1.0$ and $\theta_A = 0.2$, and π^* was given a uniform prior. P was chosen to be 0.005. The results presented here are from a single chain of 2000 Gibbs samples after a burn-in of 500.

Our method gives joint posterior samples for the relative frequencies of all unique tags, but for this paper we simply demonstrate how they compare with other estimators by plotting the shrinkage curve in Figure 3. The relative frequency estimates from our method are generally smaller than the MLE for tags with observed counts less than 7, and very close to the MLE

for more abundant tags. The arguments made in Section 2 suggest that our estimators should have efficiency benefits over the other two methods, which we now investigate using a simulation study.

5 Simulation Study

The "true" set of relative frequencies $\underline{\pi}_k$ for simulation were obtained by pooling together the observed counts from six SAGE libraries from breast cancer tissue in a study at M.D. Anderson Cancer Center. These samples consist of 495,947 sequenced tags, with $k = 44,984$ unique tags. Of the unique tags, 684 (1.5%) have observed relative frequencies of greater than 50 copies per cell (i.e. $\pi_i > 50/300,000$), accounting for 41% of the total mRNA mass. Given the large number of sequenced tags, the observed relative frequencies in this pooled sample should give a reasonable approximation to the distribution of true relative frequencies for unique tags in a biological tissue sample (see Velculescu, et al. 1999).

We performed simulations based on sample sizes of $n = 10,000$ and $n = 50,000$, currently among the most prevalent sample sizes seen in the SAGE literature. In each case, we randomly generated 100 samples of size n from a multinomial population with relative frequencies $\underline{\pi}_k$. For each sample, we obtained estimates using 3 methods, maximum likelihood, Bayesian with Symmetric Dirichlet(1) prior, and Bayesian with our Mixture Dirichlet prior. For the mixture method, we used $\theta_S = 1.00$ and $\theta_A = 0.50$ for the Dirichlet parameters and a uniform prior for π^* . The parameter P was chosen to yield a nondegenerate shrinkage profile. To assess sensitivity to estimation of k , the $n = 10,000$ simulations were run with estimates of k smaller (25,000) and larger (65,000) than the true value (44,984), and various values of P (0.01, 0.03, 0.0425, 0.06) were chosen to assess sensitivity to P . Our results for each simulated data set come from a single chain of Gibbs samples, 2000 for $n = 10,000$ simulations and 500 for $n = 50,000$ simulations, after a burn-in of 100.

The squared error loss for the three estimators was calculated for each of the 44,984

tags in each dataset. The squared error loss for an estimator for tag i in dataset j is $SE_{ij} = (\hat{\pi}_{ij} - \pi_i)^2$. From this, the mean square error for each tag $MSE_i = 100^{-1} \sum_{j=1}^{100} SE_{ij}$ was computed, as was the relative efficiency (RE_i) to the MLE. Summing MSE_i over tags, we also computed the integrated mean squared error, $IMSE = \sum_{i=1}^k MSE_i$, for each estimator. In order to compare estimators on a data set by data set basis, we computed the integrated squared error for each sample j , given by $ISE_j = \sum_{i=1}^k SE_{ij}$.

Figure 4 contains plots of the relative efficiency of the Symmetric and Mixture Dirichlet estimators as a function of the true relative frequencies in the two simulations. First, consider the performance of the Symmetric Dirichlet. For both sample sizes, the estimator was more efficient than the MLE for scarce tags, but performed increasingly poorly for more abundant tags, with the relative efficiency close to zero for the most abundant ones. The IMSE for $n = 10,000$ and $n = 50,000$ were 4489 and 1546, respectively versus 995 and 601 for the MLE. This was what we expected based on our discussion in Section 2.2.

For $n = 10,000$, the Mixture Dirichlet estimator showed efficiency improvements of more than 35% over the MLE based on IMSE (IMSE=995 for MLE vs. 643 for Mixture Dirichlet). Efficiency gains of this order were seen for every one of the 100 simulated data sets, as measured by ISE. It is important in this setting that we do not limit ourselves to aggregate performance measures, but also examine performance across the various regions of the parameter space, since a loss function treating all tags equally may not match the inferential goals of a given investigation. For the scarce tags (0-50 copies per cell), the mixture method was more efficient with RE of up to 12. These scarce tags account for 98.4% of the total number of unique tags. In the region of 200-1000 copies per cell (0.37% of total tags), its performance was essentially equivalent to the MLE. For an intermediate range (50-200 copies per cell, 1.2% of tags) and for the most abundant tags (> 1000 copies per cell, 0.03% of tags), the mixture method was outperformed by the MLE, with minimum RE

near 0.60 in the intermediate range and 0.50 for the most abundant tag in the population.

We have found that the mixture is sensitive to choice of the parameter P . Figure 5 contains the relative efficiency and shrinkage plots for various choices of P . If chosen too small, the model converges to a state where all tags are scarce, i.e. the Symmetric Dirichlet case. If chosen too large, it converges to a state where all observed tags are abundant. In both of these degenerate cases, the method has a linear shrinkage profile and performs poorly. In between these two extremes lie values of P that yield nondegenerate mixtures. Within this range, there is a tradeoff in performance between the intermediate and most abundant regions. If P is smaller, there is less shrinkage for the most abundant tags but a less smooth shrinkage profile, trading off more performance in the intermediate range. For larger P , there is more shrinkage and thus worse performance for the most abundant tags, but the shrinkage profile is smoother, improving performance in the intermediate range.

When the value of k was misspecified, the mixture method showed similar efficiency gains over the MLE as when it was correctly specified (IMSE=687 for $k = 25,000$ and IMSE=651 for $k=65,000$), with improvements and trade-offs at the same regions of the parameter space (see Figure 6). This suggests that our method is robust to selection of k , as long as P is selected to yield nondegenerate shrinkage profiles.

For $n = 50,000$, the Mixture Dirichlet again had smaller IMSE than the MLE (173 vs. 201), outperforming the MLE for scarce tags (0-15 copies per cell, 93.5% of total), with RE of up to 12, with less efficiency than the MLE for an intermediate range (15-50 copies per cell, 4.9% of total), and equivalent for the abundant genes. The magnitude of improvement for scarce tags was again larger than the efficiency loss in the intermediate range (RE > 0.50).

6 Discussion

Maximum likelihood estimators for the relative abundances of SAGE tags fail to make use of known information about the data. They ignore prior knowledge that the population contains

many scarce tags and few abundant ones, and that the sampling properties of the problem suggest that there many unique tags missed by the sampling, and many tags with small counts whose empirical relative frequencies are larger than their true relative frequencies. An estimator that takes this information into consideration and performs nonlinear shrinkage has considerable efficiency gains over the MLE for the scarce tags making up a vast majority of the total, without sacrificing too much efficiency for the abundant ones. The Mixture Dirichlet prior we have introduced is one way to quantify this prior information and obtain nonlinear shrinkage estimators.

The key benefit of our method comes from the fact that the nonlinear shrinkage profile imposed by our prior helps correct for some of the sampling limitations in the data. It is possible that other methods could be constructed to give nonlinear shrinkage profiles, and would likely experience similar types of efficiency gains (and tradeoffs) as our method.

For example, Good(1953) discusses the inadequacy of the MLE in incomplete multinomial sampling contexts, and proposes an empirical Bayes estimator he attributes to A. M. Turing. In simulations, the performance of their estimator was similar to ours. Their IMSE was similar to ours for $n = 10,000$ (643 vs. 643) and $n = 50,000$ (160 vs. 173). Their estimator had slightly extended regions of improved efficiency (0-60 copies per cell for $n = 10,000$ and 0-20 for $n = 50,000$) and reduced efficiency (60-500 copies per cell for $n = 10,000$ and 20-150 for $n = 50,000$) within the parameter space. Their method requires a somewhat *ad hoc* step of plugging in a smoothed version of the histogram of the observed counts for the histogram of true counts, and is not fully model based like ours.

The fact that our method flows naturally from a fully specified coherent probability model gives it several inferential advantages over other more *ad hoc* methods. First, we obtain posterior samples from the joint distribution of all tags' expression levels, from which any inference, univariate or multivariate, can be obtained. For example, estimates, posterior

intervals, density estimates, and Bayesian hypothesis tests are available for any quantities derivable from the relative frequencies for any set of tags. Since our model is multivariate, we can also estimate and make probability statements on inter-tag summaries, e.g. correlations and clustering, which are not available from tag-by-tag modeling or *ad hoc* approaches.

Second, the model can be extended, if desired, to incorporate other features of SAGE data. For example, in principle the model can be supplemented to simultaneously estimate k , the true number of unique tags in the tissue. Denoising could be incorporated by adding an element onto the model that detects likely sequencing errors. Further, it is possible to add an additional hierarchical level to the model accommodating multiple libraries, so that both the library to library and multinomial sources of variability are appropriately taken into account. These extensions require additional work beyond the scope of this paper, but are made possible by our model-based approach.

As previously mentioned, the goal of many SAGE experiments is the identification of tags differentially expressed between two types of tissue. There are a number of methods in the current literature for testing for differential expression of SAGE tags between two libraries, all which are applied on a gene-by-gene basis and implicitly involve estimation of π_1 and π_2 , the true relative frequencies for a given gene in the two libraries, given the observed counts A and B . Zhang, et al. (1997) use a permutation test, Madden, et al. (1997) perform a hypothesis test on $H_0 : \pi_1 - \pi_2 = 0$ using the Normal approximation to the Poisson, Audic and Claverie (1997) compute the predictive distribution of $B|A$ assuming a Poisson-Gamma model. All these methods implicitly use the maximum likelihood estimator for π_1 and π_2 . Chen, et al. (1998) have a Bayesian method whereby they place a Symmetric Beta prior on the quantity $\phi = \pi_1/(\pi_1 + \pi_2)$ and compute the posterior distribution of ϕ . Their method effectively uses estimates of π_1 and π_2 that are shrunken towards each other by their prior.

Given two libraries, our method can be applied separately to the two libraries and the

outputted posterior samples used to assess differential expression. For example, posterior credible intervals can be constructed on the fold change for each unique tag, i.e. the ratio of the relative abundances, or the posterior probability of a greater than two-fold change can be computed, e.g.. If a zero count is observed for one of the two libraries, we can use the posterior samples from one of the anonymous zero count tags. This is reasonable when we are willing to believe that the corresponding mRNA transcript is one of those present in a small quantity in the other tissue, but missed by SAGE due to sampling error. This is an advantage for our method, since other methods must resort to *ad hoc* adjustments such as substituting a count of one when zero counts are encountered. The estimates of fold change difference using the MLE or our nonlinear shrinkage estimator differ considerably in some cases. For example, a tag with observed counts of 1 and 6 in two libraries of size 10,000 will give a fold change estimate of 6 using the MLE but 16.5 using our method.

Our simulation results in Section 5 suggest that the Mixture Dirichlet estimator has large efficiency benefits over the MLE for the majority of tags that are scarce, and tradeoffs of a smaller magnitude for tags in a particular intermediate range and for the most abundant tags. Efficiency gains in estimation of the relative abundances should transfer over to more sensitivity and specificity in assessing differential expression. Thus, we expect to see gains in comparisons that involve at least one scarce tag, and a degree of tradeoff when a tag is in the specific regions where the RE is less than one.

A key question to consider with any shrinkage estimator is whether the gains of efficiency in one part of the parameter space justify the tradeoffs in other regions. The answer to this question depends on the investigator, but an argument could be made that the benefits of our method outweigh the costs. First, the region of benefit contains a vast majority of the tags. In most experiments, not all tags are equally important, but frequently the tags that appear to be candidates for differential expression contain a scarce tag in one of the two

libraries. Second, the loss of efficiency in the intermediate and most abundant regions is small in magnitude compared to the gain for the scarce tags. Third, we have found that for more abundant tags, the library-to-library variability dominates sampling variability, so a marginal loss of sampling efficiency has a small practical consequence for multi-library inference, while by contrast the sampling variability dominates inference for the scarce tags.

We have discovered that our Mixture Dirichlet model is sensitive to choice of P . Trial and error on this parameter is necessary to obtain nondegenerate shrinkage curves, which is required to effectively incorporate our method. If interest is mainly in point estimation, one easily implemented alternative to fitting our full model is to apply a suitably nonlinear shrinkage curve to the raw counts to obtain revised counts. Our method must work hand-in-hand with a method for dealing with sequencing errors and estimating the number of unique tags k , and in multilibrary settings, it is important to also allow the modeling of library-to-library heterogeneity in the relative frequencies. There is clearly more work to be done, but we believe the ideas introduced in this paper can serve as an important starting point for a rigorous probability model-based approach for the analysis of SAGE data.

7 References

- Audic, S. and Claverie, J. M. (1997). The significance of digital gene expression profiles. *Genome Research* **7**, 986–995.
- Blades N. (2002). Noise and Shadows, Statistical Issues with SAGE Data. Ph.D. Dissertation, Johns Hopkins University.
- Bunge J. and Fitzpatrick M. (1993). Estimating the Number of Species: A Review. *Journal of the American Statistical Association* **88**, 364–373.
- Chen, H., Centola, M., Altschul, S. F., and Metzger, H. (1998). Characterization of gene expression in resting and activated mast cells. *Journal of Experimental Medicine* **188**(9), 1657–1668.
- George, E. I. (1986). Minimax Multiple Shrinkage Estimation. *The Annals of Statistics* **14**,

188–205.

- Gruber, M. H. J. (1998). *Improving Efficiency by Shrinkage* New York: Marcel Dekker.
- Good, I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* **40**, 237–264.
- Hastie, N. D. and Bishop, J. O. (1976). The Expression of Three Abundance Classes of Messenger RNA In Mouse Tissues. *Cell* **9**, 761-774.
- James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol 1, Berkeley, University of California Press, 361–379.
- Jeffreys, H. (1948). *Theory of Probability*. Oxford: Clarendon Press.
- Johnson, N. J., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: John Wiley and Sons.
- Madden, S. L., Galella, E. A., Zhu, J., Bertelsen, A. H., and Beaudry, G. A. (1997). SAGE transcript profiles for p53-dependent growth regulation. *Oncogene* **15**, 1079–1085.
- Polyak, K. and Riggins, G. J. (2001). Gene Discovery Using the Serial Analysis of Gene Expression Technique: Implications for Cancer Research. *Journal of Clinical Oncology* **19**, 2948–2958.
- Stollberg J., Urschitz J., Urban Z., and Boyd C. D. (2000). A Quantitative Evaluation of SAGE. *Genome Research* **10**, 1241–1248.
- Velculescu V. E., Madden S. L., Zhang L., Lash A. E., Yu J., Rago C., Lal A., Wang C. J., Beaudry G. A., Ciriello K. M., Cook B. P., Dufault M. R., Ferguson A. T., Gao Y., He T. C., Hermeking H., Hiraldo S. K., Hwang P. M., Lopez M. A., Luderer H. F., Mathews B., Petroziello J. M., Polyak K., Zawel L., Zhang W., Zhang X., Zhou W., Haluska F. G., Jen J., Sukumar S., Landes G. M., Riggins G. J., Vogelstein B., and Kinser K. W. (1999). Analysis of Human Transcriptomes. *Nature Genetics* **23**, 387–388.
- Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., Kinzler, K. W. (1997). Gene Expression Profiles in Normal and Cancer Cells. *Science* **276**, 1268–1272.

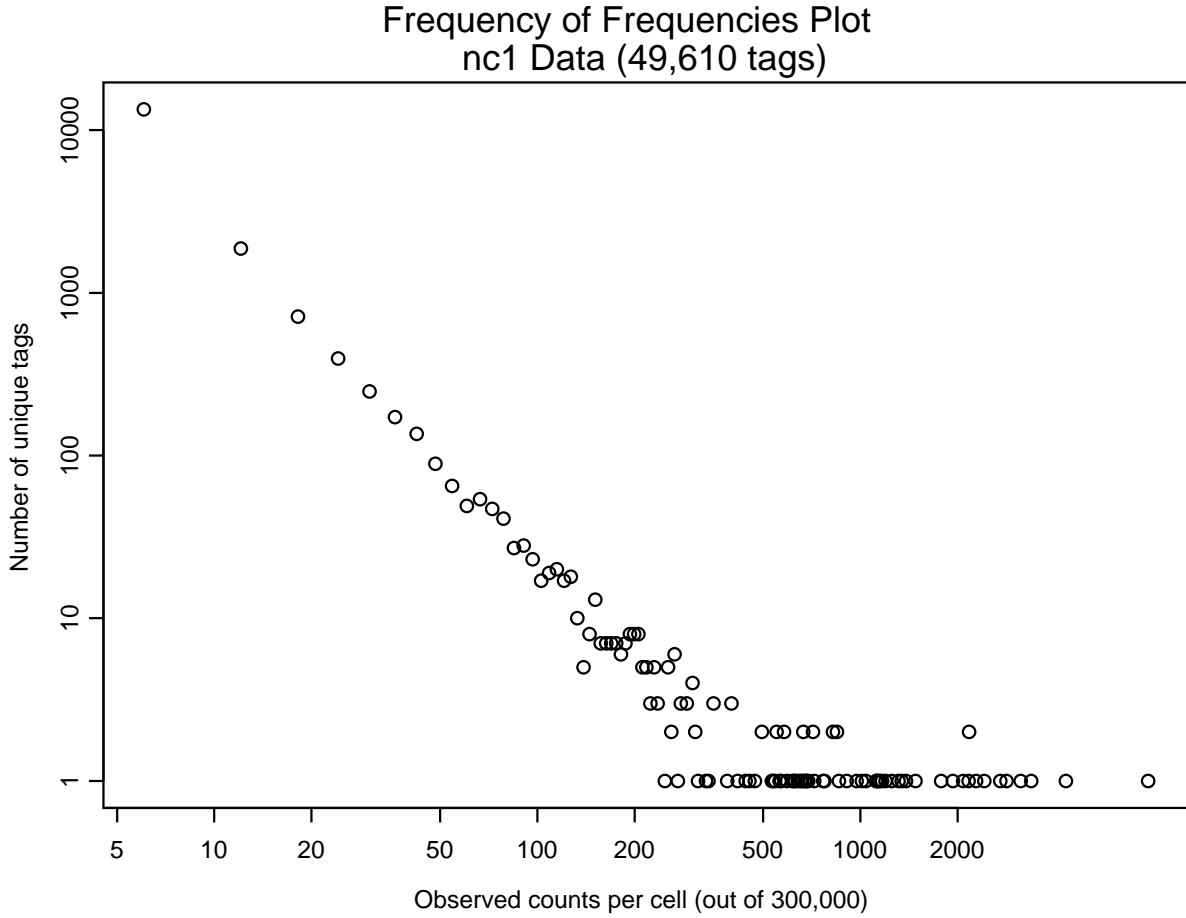


Figure 1: Number of unique tags in SAGE library *nc1* with certain observed relative frequencies. Note the extreme skewness in the distribution of the true frequencies. This linear shape in this log-log scale is characteristic of SAGE data. Both axes are in the log scale to make the plot more readable.

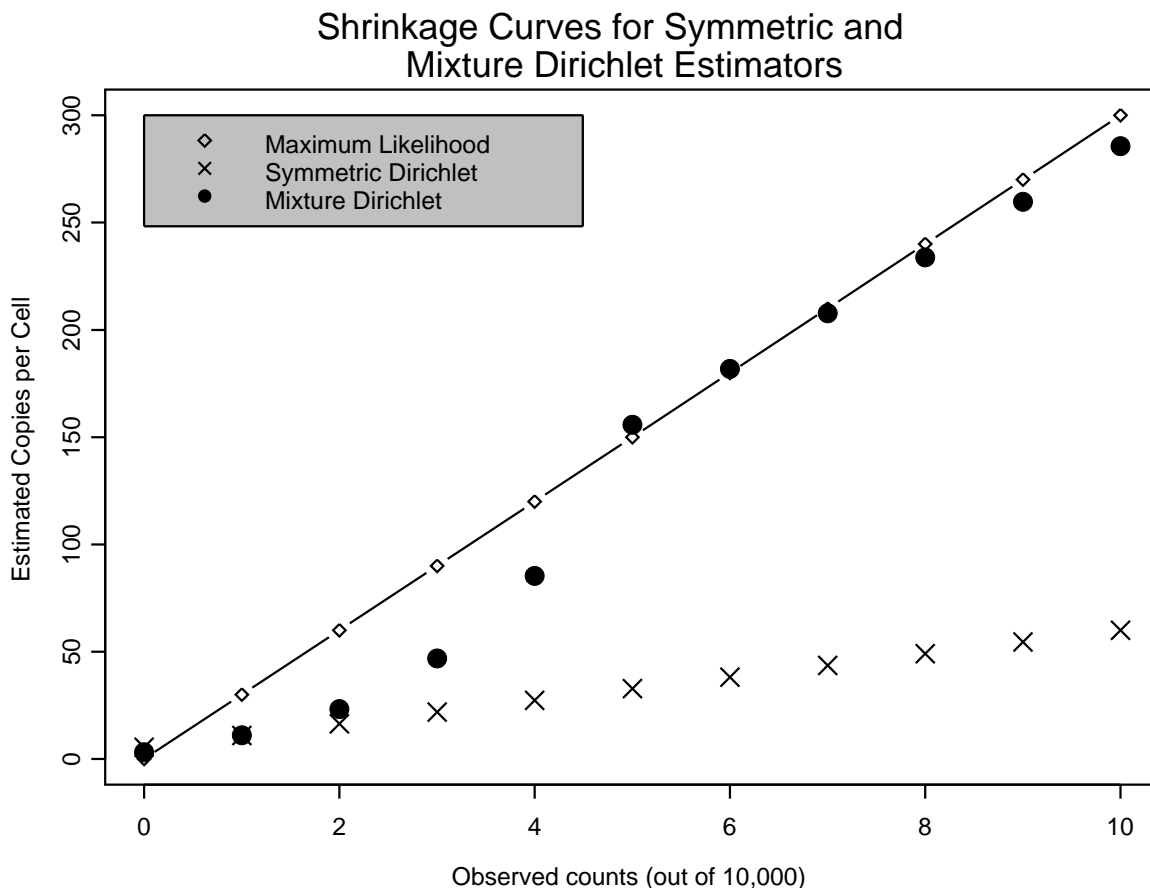


Figure 2: Shrinkage curves for Bayesian estimators using our mixture Dirichlet prior and a Symmetric Dirichlet prior for SAGE data with $n = 10,000$ and $k = 44,984$. The shrinkage curves plot the Bayesian estimates versus the observed counts to demonstrate their shrinkage profiles relative to the MLE, indicated by the solid line. The hyperparameters of the Mixture Dirichlet are $\theta_S = 1.0$ and $\theta_A = 0.5$, with $P = 0.0425$ and a uniform priors for π^* . The hyperparameter for the Symmetric Dirichlet distribution is $\theta = 1.0$. Note the nonlinear shrinkage profile for the Mixture Dirichlet contrasted with the linear profile for the simple Symmetric Dirichlet. This plot only gives the shrinkage plot for observed counts from 0 to 10. If it were extended to include larger counts, the Mixture Dirichlet would remain relatively close to the MLE line, while the Symmetric Dirichlet would continue on its linear course, moving further and further away from the MLE.

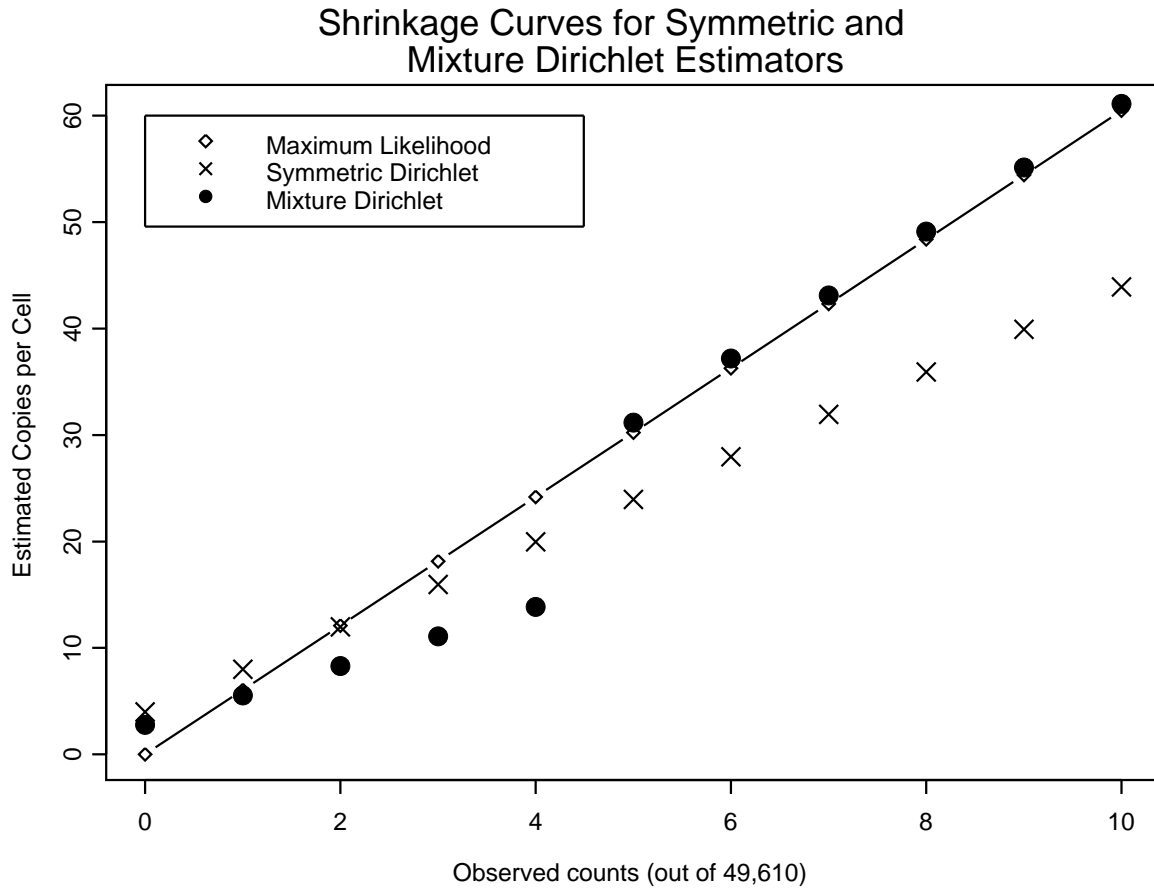


Figure 3: Shrinkage curves for Mixture Dirichlet estimator applied to *nc1* data set, assuming $k = 25,536$ unique transcripts. The reference line corresponds to the MLEs. The hyperparameters for the mixture Dirichlet are $\theta_S = 1.0$ and $\theta_A = 0.2$, with $P = 0.005$ and uniform prior on π^* .

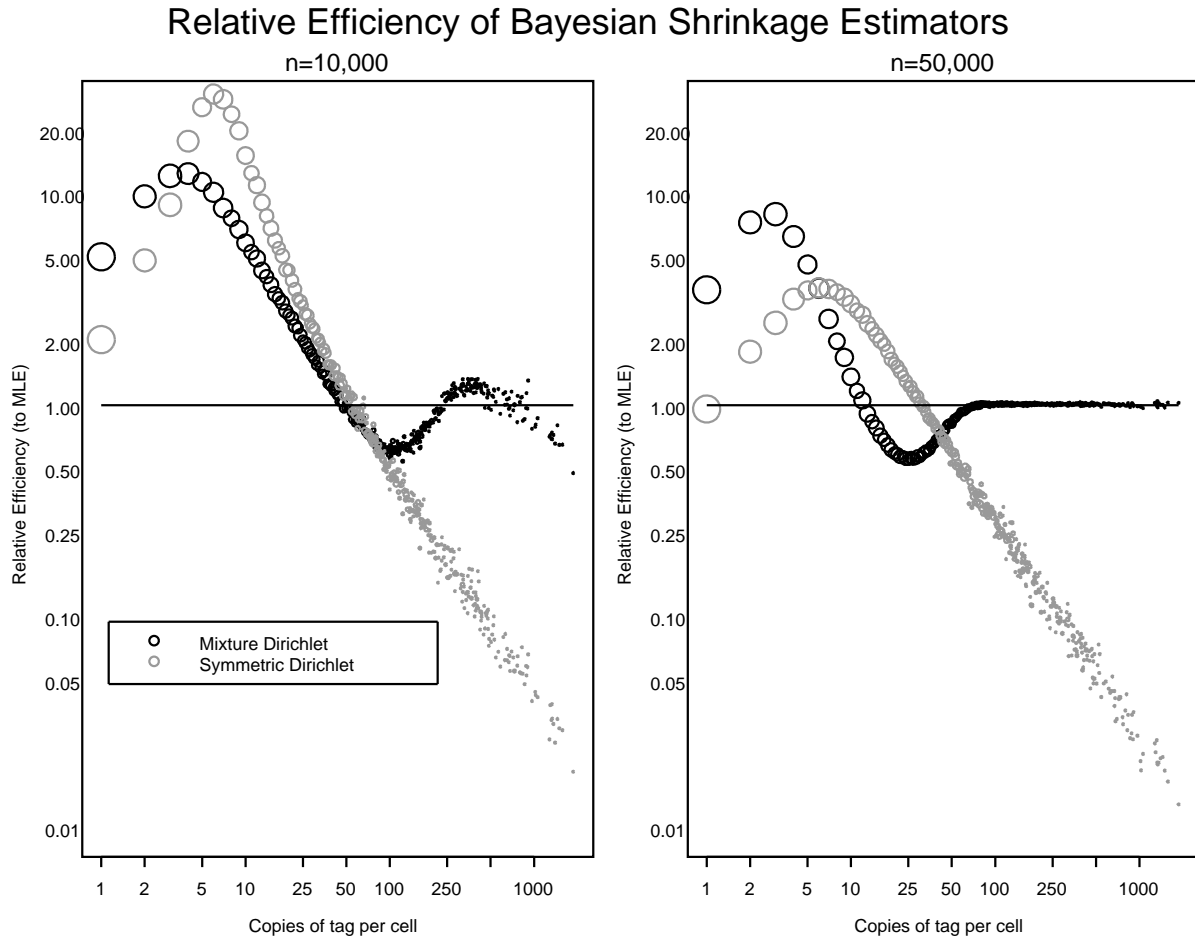


Figure 4: Relative Efficiency of estimators from a simulation of 100 multinomial samples of size 10,000 and 50,000 taken from a SAGE-like population. The horizontal axis consists of true relative frequencies multiplied by 300,000 to represent number of copies per cell containing 300,000 total mRNA transcripts. To aide presentation, the results for unique tags with like true relative frequencies have been combined, and the size of each plotted circle made proportional to the $\log(\text{number of unique tags})$ with that true relative frequency.

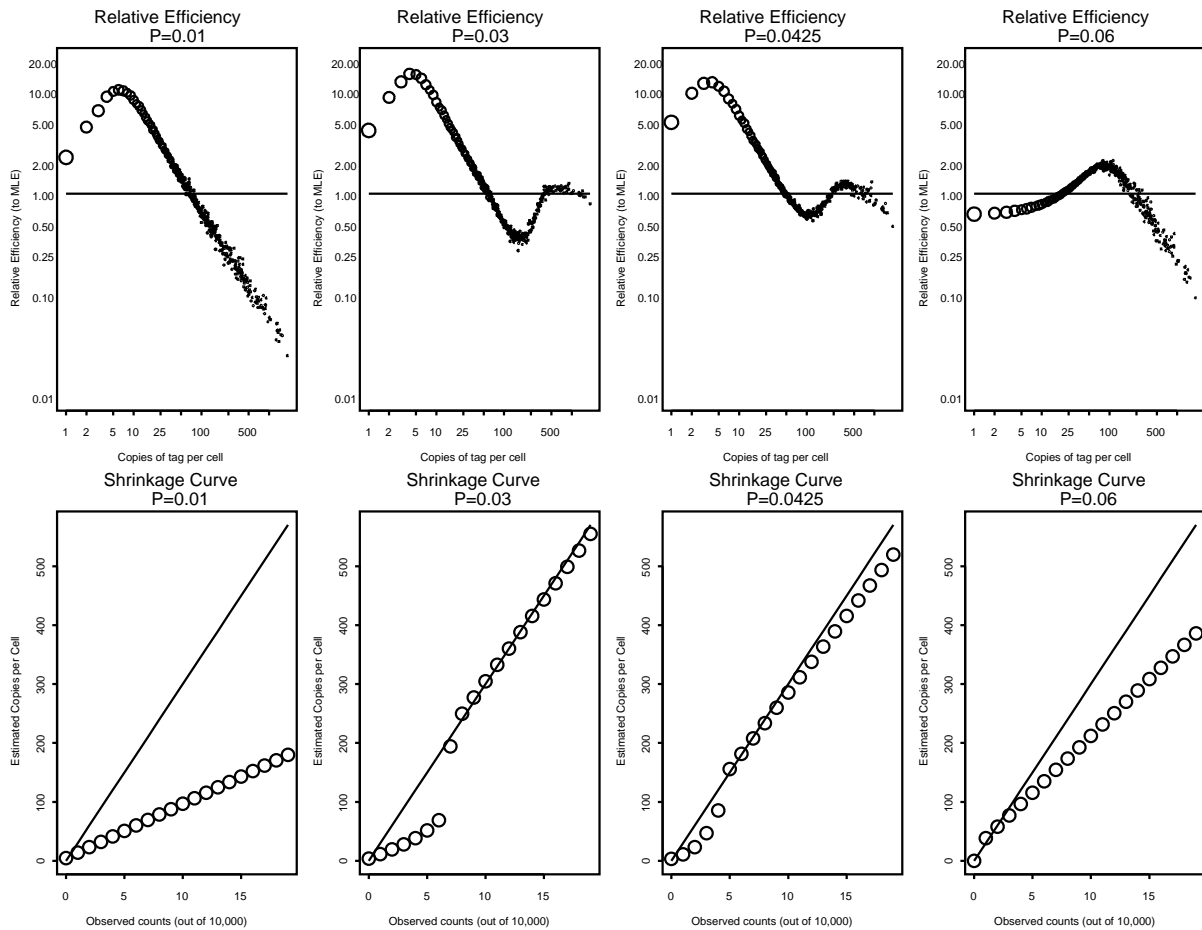


Figure 5: Relative Efficiency and Shrinkage plots for mixture Dirichlet estimators for various choices of P from a simulation of 100 multinomial samples of size 10,000 taken from a SAGE-like population, with $\theta_S = 1.0$ and $\theta_A = 0.5$. Note that the values of P giving a nondegenerate mixture lie within an intermediate range.

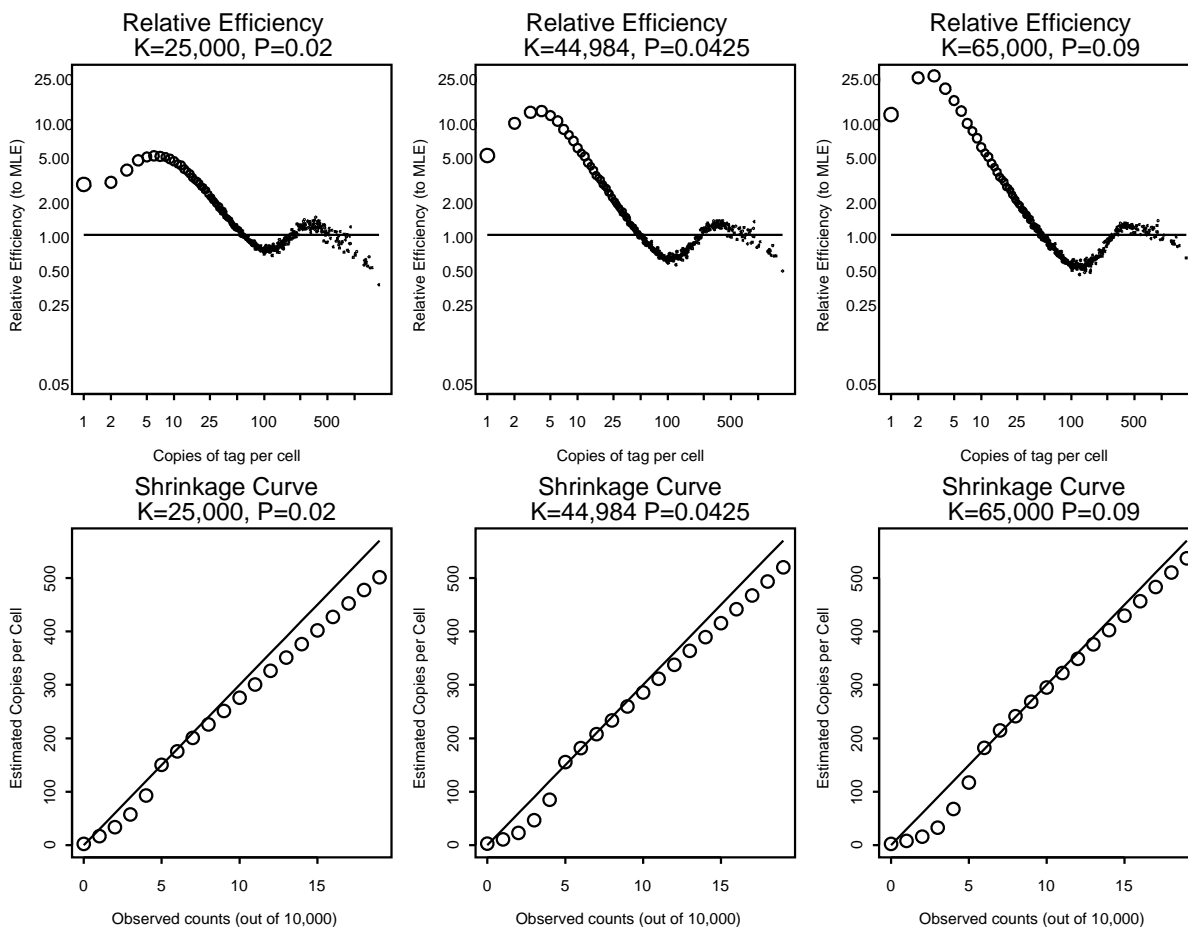


Figure 6: Relative Efficiency and Shrinkage plots for mixture Dirichlet estimators for various choices of k and P from a simulation of 100 multinomial samples of size 10,000 taken from a SAGE-like population. Note that the performance of the method is robust to misspecification of k , given P is chosen to give a nondegenerate mixture.