

2

Shrinkage Estimation for SAGE Data using a Mixture Dirichlet Prior

Jeffrey S. Morris,
M.D. Anderson Cancer Center

Keith A. Baggerly,
M.D. Anderson Cancer Center

Kevin R. Coombes,
M.D. Anderson Cancer Center

Abstract

Serial Analysis of Gene Expression (SAGE) is a technique for estimating the gene expression profile of a biological sample. Any efficient inference in SAGE must be based upon efficient estimates of these gene expression profiles, which consist of the estimated relative abundances for each mRNA species present in the sample. The data from SAGE experiments are counts for each observed mRNA species, and can be modeled using a multinomial distribution with two characteristics: skewness in the distribution of relative abundances 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. Standard empirical estimates of the relative abundances effectively ignore these missing, unobserved species, and consequently tend to also overestimate the abundance of the scarce observed species comprising a vast majority of the total. In this chapter, we review a new Bayesian procedure that yields improved estimates for the missing and scarce species without trading off much efficiency for the abundant species. The key to the procedure is the mixture Dirichlet prior, which stochastically partitions the mRNA species into abundant and scarce strata, with each stratum modeled with its own multivariate prior, a scalar multiple of a symmetric Dirichlet. Simulation studies demonstrate that the resulting shrinkage estimators have efficiency advantages over the MLE for SAGE scenarios simulated.

2.1 Introduction

Serial analysis of gene expression (SAGE) is a method for estimating the gene expression profile of a biological sample of interest. In this chapter, we review a method introduced in Morris, Baggerly, and Coombes (2003) for obtaining Bayesian shrinkage estimates of these profiles using a fully specified probability model. Bayesian inference using this method is straightforward since the estimators arise from a coherent probability model. An outline of the chapter is as follows.

In Section 2.2, we provide an overview of SAGE and motivate the use of the multinomial likelihood to model the resulting data, then in Section 2.3 describe some standard methods for estimating multinomial relative frequencies. After explaining why we believe these standard approaches are inadequate for SAGE data, we review a method based on the mixture Dirichlet prior. Section 2.4 describes the prior and Section 2.5 provides implementation details for obtaining posterior mean estimates for the mRNA species' relative abundances. In Section 2.6, we present the results of a simulation study demonstrating the benefit of shrinkage estimation based on the mixture Dirichlet prior, and Section ?? contains some conclusions.

2.2 Overview of SAGE

2.2.1 Measuring Gene Expression

The central dogma in genetics is that DNA is used as a template to make mRNA molecules, which then assemble the proteins that perform the biological tasks of a living organism. The "expression level" of a gene refers to the amount of its corresponding mRNA present in a cell, which is taken to be a rough surrogate for how active that gene is in coding its protein. Levels of gene expression are important to study in cancer research, as well as other biological applications. It is often of interest to compare the gene expression profiles of different biological samples, for example to identify genes differentially expressed across biological conditions (e.g. cancer/normal) or clinical outcomes (e.g. response to chemotherapy/no response).

The most common method for measuring gene expression is cDNA hybridization using microarrays, which can simultaneously measure the expression levels of the genes represented on the array, typically numbering in the thousands. In this method, sequences of DNA from the genes of interest are arranged on an array, either by spotting or direct

synthesis. The sample of interest is then labeled and hybridized with the targets on the array. The amount of material hybridized to each target is estimated by measuring the corresponding staining intensity. After normalization, these intensities are used as an estimate of the expression levels of the target genes. Note that microarrays are a "closed system", since they only provide information on the prespecified genes that have been placed on the array.

An alternative method for measuring gene expression is Serial Analysis of Gene Expression (SAGE), introduced by Velculescu et al. (1995). Unlike microarrays, SAGE is an "open system", since one need not pre-specify the genes of interest. It is possible to obtain expression level estimates for any gene expressed in the sample. First, a sample of n mRNA transcripts are selected from the biological sample and complementary cDNA strands are constructed. Commonly, the number of transcripts sampled is between $n = 10,000$ and $100,000$. For each selected transcript, a 10-base region at a specific location within its sequence is isolated, sequenced, and recorded. This sequence is called a tag. Ideally, these tags uniquely identify the source mRNA, and in practice this is roughly true. Thus, the relative expression level of a gene, measured by the relative quantity of the corresponding mRNA species, is approximated by the relative frequency of its corresponding SAGE tag. The data from a SAGE experiment consists of the counts for each unique tag observed in the sample. The collection of these counts for a biological sample is called a "SAGE library."

2.2.2 Characterizing SAGE data

Suppose that from a SAGE sample consisting of n transcripts, we obtain a library containing k^* unique tags, with the counts for each unique tag represented by X_i , $i = 1, \dots, k^*$, with $\sum_{i=1}^{k^*} X_i = n$. In a typical library, the number of unique tags observed is between $k^* = 3000$ and $30,000$. By the nature of the sampling, it is very likely that the SAGE sample has failed to capture a number of mRNA species present in the biological sample. Let k_0 be the number of these "missing species", and $k = k^* + k_0$ the true number of expressed transcripts in the sample. If we knew k_0 or at least had an estimate of it, we could append our dataset with $X_i = 0$ for $i = k^* + 1, \dots, k$, to include the zero counts for the missing species.

Assuming the sampling of mRNA proceeds in a roughly independent fashion, the vector $\underline{X} = (X_1, \dots, X_k)'$ can be modeled as a random draw from a multinomial distribution with parameters n and $\underline{\pi} = (\pi_1, \dots, \pi_k)'$,

Table 2.1. *Skewness in SAGE data (Velculescu, et al. 1999)*

Copies/cell	% of Tags	% of Mass
≤ 5	89.9%	23%
5-50	9.2%	30%
50-500	0.8%	27%
500-5000	0.1%	20%

with $\sum_{i=1}^k \pi_i = 1$. The vector $\underline{\pi}$ characterizes the relative expression profile of the biological sample, with π_i representing the true relative abundance of the mRNA species corresponding to unique tag i . These are the main parameters of interest in SAGE, and efficient inference depends on efficient estimation of these parameters. It is estimated that a typical cell contains roughly 300,000 mRNA transcripts, so frequently the π_i are reported as fractions over 300,000 to represent "mean copies of the transcript per cell."

Table 2.2.2 summarizes the distribution of relative frequencies across tags estimated by pooling together a series of large colon cancer SAGE libraries (Velculescu, et al. 1999). Note that a vast majority of the observed tags have relative expression levels of no more than 5 per cell, i.e. $\pi_i \leq 5/300000$. Although containing almost 90% of all unique tags, this group only accounts for 23% of the total probability mass; the π_i for the unique tags in this group sum to 0.23. There are progressively fewer genes with larger relative frequencies. Only 1/1000 of genes are present at rates of over 500 per cell, but these few abundant genes account for almost as much probability mass as the scarcest ones. The 1% most abundant tags account for almost 50% of the total probability mass of the sample.

Thus, we see that there are a small number of "abundant" genes, and a large number of "scarce" genes. This is an inherent characteristic of SAGE data, and gene expression in general. We could say from this that the distribution of the π_i is very strongly skewed right. Note that this skewness, along with the fact that n is not large relative to k , contributes to the fact that a large number of mRNA species are missing from any given SAGE sample.

2.3 Methods for Estimating Relative Abundances

2.3.1 Maximum Likelihood Estimation

It is typical to estimate the relative abundances for mRNA species π_i by the standard empirical estimators $\hat{\pi}_{i,\text{MLE}} = X_i/n$. These are trivial to compute and are maximum likelihood estimators, so are asymptotically efficient. Thus, for large enough samples, they can be shown to outperform all other estimators. However, since $n \approx k$ and the distribution of π_i is strongly skewed, a SAGE library is essentially a small sample, even with values of n that seem "large". In this setting, the MLE performs well for the relatively few abundant species, but has undesirable properties for the scarce species comprising the vast majority of the total number of unique tags. For a given SAGE sample, it underestimates the relative frequency for all missing tags, and as a result, tends to overestimate the relative frequencies of the scarce observed tags.

The estimator for each missing tag is zero, which we know is less than the true value. As a result of the relative frequency constraint that the π_i 's must sum to 1, this implies that the MLE tends to, on average, overestimate the relative frequencies for the non-missing tags, i.e. $\sum_{i: X_i > 0} \hat{\pi}_{i,\text{MLE}} > \sum_{i: X_i > 0} \pi_i$. The genes with small but nonzero counts will be the ones most likely to be overestimated. Thus, for a given data set, the MLE will underestimate π_i for any species with zero counts, and tend to overestimate π_i for genes with small nonzero counts.

To further illustrate this point, consider the following toy example. Suppose we have a biological sample with 51 expressed genes. One is abundant with relative frequency $\pi_i = 0.50$, and the other fifty are scarce with $\pi_i = 0.01$. Suppose we sample 20 transcripts and record the counts for each. On average, 40 of the 50 scarce tags will be missing, with the other 10 occurring once. Thus, our estimate for a scarce tag will either be much smaller (0) or much larger (≥ 0.05) than the true value. In this effectively small sample setting, these estimators are limited as to how well they can estimate the relative abundances of the scarce tags.

We would like to find an estimator that improves on the MLE. We would like for it to give positive estimates when $X_i = 0$, which would require shrinking the estimates for the genes with counts greater than zero in order to honor the relative frequency constraint $\sum_{i=1}^k \pi_i = 1$. This can be accomplished by specifying a prior distribution for $\underline{\pi}$ and using a Bayesian estimator.

2.3.2 Bayesian Estimation With a Symmetric Dirichlet Prior

The Dirichlet distribution is commonly used as a prior distribution with the multinomial likelihood, since it is conjugate. When there is no prior knowledge on which of the multinomial categories are more likely than the others, it is common practice is to set all Dirichlet parameters to be the same *a priori*, which we refer to as a *symmetric Dirichlet* with parameter θ , or $\text{SymmDir}(\theta)$. A common choice for this hyperparameter is $\theta = 1$, described by Jeffreys (1961, Section 3.23), and corresponding to a k -variate generalization of the Uniform distribution.

In that case, the posterior distribution of $\underline{\pi}$ is Dirichlet with parameters $\theta + X_i$ for $i = 1, \dots, k$. The posterior mean for π_i is $\hat{\pi}_{i,\text{DIR}} = \frac{X_i + \theta}{n + k\theta}$, which can also be written as $\left(\frac{n}{n+k\theta}\right) * X_i/n + \left(\frac{k\theta}{n+k\theta}\right) * 1/k$. Thus, we see that the posterior mean is a weighted average of the MLE and the prior mean, with the weight determined largely by the sample size. As n gets very large, the estimator is approximately X_i/n , the MLE. When n is smaller, there is more weight given to the prior mean, $1/k$.

Now, consider our toy example again, this time using the posterior mean estimator with this Dirichlet prior, assuming $\theta = 1$. When $X_i = 0$, $\hat{\pi}_{i,\text{DIR}} = 1/71 = 0.014$, and when $X_i = 1$, $\hat{\pi}_{i,\text{DIR}} = 2/71 = 0.028$. These are both closer to the true value of 0.01 than the MLE. In fact, it can be shown that, for the scarce genes in this toy example, the Bayesian estimators are always closer to the truth than the MLEs, no matter when X_i is! However, this estimator performs abysmally for the abundant gene. When $X = 10$, we see the posterior mean estimate is 0.15, versus the MLE of 0.50. Recall the true value was 0.50. We see in our example that the Bayesian method with Dirichlet prior results in improved estimation for the scarce species relative to the MLE, but induces a severe bias that results in horrible estimates for the abundant species. This is not just true in this contrived example, but these results would also hold for real SAGE data.

2.3.3 Robin Hood and Nonlinear Shrinkage

Following is an analogy to illustrate the heuristics behind this problem. We know that the MLEs yield zero relative frequency estimates for all species not seen in our SAGE sample. Since we know that there are a large number of true mRNA species out there with positive relative frequencies that were simply missed by our random sampling procedure, we would like to have some positive probability mass for these species.

Because of the relative frequency constraint, this means we need to decrease the estimators for some of the observed classes. There's no free lunch; in order to "pay" the zero count classes, we need to "steal" some probability mass from the other classes.

A Bayesian estimator will do this – because the Bayesian posterior mean takes a weighted average of the prior mean and empirical estimator, it will result in a positive relative frequency estimate for the zero counts, and will shrink the estimates for tags with empirical estimates greater than the prior mean. The form of the prior determines how this shrinkage is done; that is, who do we "steal from" to "pay the zeroes."

The simple symmetric Dirichlet performs linear shrinkage, so steals the most mass from the "richest" or most abundant classes. This prior could be called a "Robin Hood" prior, since it steals from the rich to pay the poor. This is reasonable when we truly believe the multinomial classes are exchangeable, since then it is likely that any very large count is an aberration, so should be shrunken the most. However, in settings like SAGE where we know there is a great deal of heterogeneity in the relative frequencies, this type of shrinkage is undesirable and, as we see from our simulation studies, leads to very poor estimators.

In SAGE, we would like a prior that is the opposite of a Robin Hood prior. It would be best to "steal" from the most "poor" classes, those with low counts, and leave the "rich" alone, since the sampling properties of the problem suggest that the poorest classes are the ones whose empirical probability estimates are holding on to the mass rightfully belonging to the zero classes. In other words, we would like an estimator that shrinks the MLEs in a nonlinear fashion, where classes with large counts are left largely unaffected, but those with small counts are shrunken.

This idea of nonlinear shrinkage has been implemented in other statistical settings. For example, in wavelet regression, noise is removed from a functional signal by shrinking wavelet coefficients towards zero in such a way that the smallest wavelet coefficients, likely to consist mostly of noise, are shrunken the most, while the large coefficients, likely to consist of signal, are left largely unaffected. In that setting, this shrinkage is achieved by using a prior that is a mixture of a Normal and point mass at zero (Vidakovic, 1998), or alternatively a mixture of two Normals, one with a small and the other a large variance (Chipman, Kolaczyk, and McCulloch, 1997). Here we accomplish nonlinear shrinkage via a mixture Dirichlet prior, which we now describe.

2.4 Mixture Dirichlet Distribution

One reason the symmetric Dirichlet prior fails in this context is that it inaccurately represents the population of relative frequencies in SAGE. It assumes that all k classes are *a priori* exchangeable, with relative frequencies of $1/k$, while in reality they are very heterogeneous. As previously mentioned, the characteristics of gene expression suggest that there should be a small number of very abundant tags, and a very large number of scarce tags. However, we typically do not know *a priori* which tags will be abundant. We quantify this prior knowledge through the following mixture Dirichlet prior, introduced by Morris, Baggerly, and Coombes (2003). This distribution stochastically partitions the tags into scarce and abundant strata, each of which has its own multivariate distribution, a scalar multiple of a symmetric Dirichlet.

2.4.1 Prior Specification

In order to specify this prior, we first introduce a new set of parameters $\{\lambda, \pi^*, \underline{q}\}$. Each unique tag is assumed to belong to one of two classes, either "abundant" or "scarce." The parameter $\lambda_i = 1$ indicates that unique tag i belongs the abundant class. We assume *a priori* that $\lambda_i \sim \text{Bernoulli}(P)$, where P represents the prior proportion of unique tags belonging to the abundant class. Let the indices of tags belonging to the scarce and abundant classes be represented by $\mathcal{S} = \{i : \lambda_i = 0\}$ and $\mathcal{A} = \{i : \lambda_i = 1\}$, which are of length $k_S = \sum_{i=1}^k (1 - \lambda_i)$ and $k_A = \sum_{i=1}^k \lambda_i$, respectively. The parameter $\pi^* = \sum_{i=1}^k \lambda_i X_i$ represents the "abundant mass", and is given a $\text{Beta}(\alpha_{\pi^*}, \beta_{\pi^*})$.

The vector $\underline{q} = (q_1, \dots, q_k)'$ contains the relative frequencies for each unique tag *within its class*. Given $\underline{\lambda}$, \underline{q} is partitioned into $\underline{q}_S = \underline{\pi}_S / (1 - \pi^*)$ and $\underline{q}_A = \underline{\pi}_A / \pi^*$. If $\lambda_i = 1$, then the q_i represents the relative proportion of abundant mass attributable to tag i , while if $\lambda_i = 0$, q_i represents the relative proportion of scarce mass attributable to tag i . The vectors \underline{q}_A and \underline{q}_S are assumed to each follow symmetric Dirichlet distributions of dimension k_A and k_S and parameters θ_A and θ_S , respectively. This construct allows the scarce and abundant tags to have their own prior distributions, yet honors the relative frequency constraint $\sum_{i=1}^k \pi_i = 1$. If k_0 is unknown, we give it an improper prior $f(k_0) \propto k_0^{-1}$, which was suggested by Jim Berger as a good reference prior in this context (personal communication).

Following is a summary of our mixture Dirichlet prior structure.

$$\begin{aligned} \underline{q}_A | \underline{\lambda} &= \text{Symmetric Dirichlet}(\theta_A), \\ \underline{q}_S | \underline{\lambda} &= \text{Symmetric Dirichlet}(\theta_S), \\ \lambda_i &\sim \text{Bernoulli}(P), \\ \pi^* &= \text{Beta}(\alpha_{\pi^*}, \beta_{\pi^*}), \\ f(k_0) &\propto k_0^{-1}. \end{aligned}$$

The relative frequencies for the individual multinomial classes are constructed using these quantities as follows: $\pi_i = \{q_i \pi^*\}^{\lambda_i} \{q_i (1 - \pi^*)\}^{(1 - \lambda_i)}$.

This prior yields a nonlinear shrinkage profile, whereby the scarce species are shrunk strongly towards zero while the abundant species are more or less left alone, leaving positive mass for the missing species we know are there. Figure 2.4.1 contains shrinkage curves for a SAGE sample of size $n = 10,000$ for the MLE and Bayesian estimators with symmetric Dirichlet prior with $\theta = 1$ and a mixture Dirichlet prior with $\theta_A = 1$, $\theta_S = 0.5$, and $P = 0.005$. Note the linear and nonlinear shrinkage profiles which characterize the symmetric and mixture Dirichlet priors, respectively.

2.4.2 Selection of Prior Hyperparameters

The hyperparameters θ_A , θ_S , and P largely determine the shape of the shrinkage curve. In general, larger values of θ_A and θ_S lead to stronger shrinkage towards the prior means within the abundant and scarce classes, respectively. Because we would like more shrinkage for scarce tags, we recommend making $\theta_S > \theta_A$. However, making θ_A too small will have the side effect of making the boundary between scarce and abundant species too sharp, which can hinder the efficiency of the shrinkage estimator. We have found that $\theta_S = 1$ and $\theta_A = 0.5$ have worked well in the examples we have tried. The hyperparameter P has a strong influence on the location of the boundary between scarcity and abundance. We have found that values of P between 0.005 and 0.03 seem to work well in practice, with P larger as n is larger.

2.5 Implementation Details

With the mixture Dirichlet prior, there is no closed form expression for the posterior distribution of $\underline{\pi}$ that would allow efficient random variate

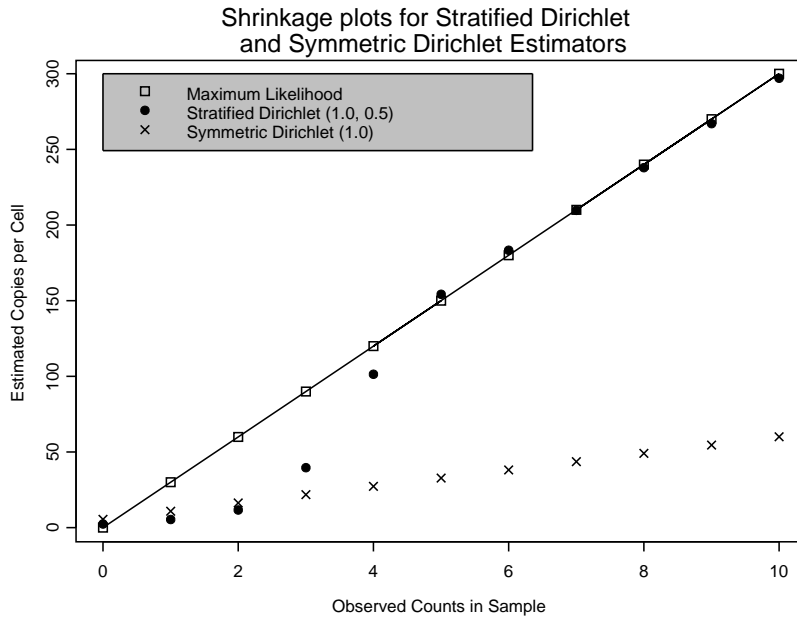


Fig. 2.1. *Shrinkage curves.* Shrinkage curves for Bayesian estimators using symmetric and mixture Dirichlet priors for SAGE data with $n = 10,000$ and $k = 44,984$. The shrinkage plots plot the Bayesian posterior mean estimates versus the observed counts to demonstrate their shrinkage profiles relative to the MLE, given by the line. The hyperparameters of the mixture Dirichlet are $\theta_S = 0.5$ and $\theta_A = 1.0$, with $P = 0.005$ and a uniform prior for π^* . The hyperparameter for the symmetric Dirichlet is $\theta = 1.0$. 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 away from the MLE.

generation. The following Markov Chain Monte Carlo procedure can be used to obtain posterior samples from this distribution.

- (i) For $i = 1, \dots, k$, sample λ_i from $\text{Bernoulli}(\alpha_i)$, with $\alpha_i = Pr(\lambda_i = 1 | \lambda_{(-i)}, \underline{X}, P)$, and $\lambda_{(-i)}$ is the set of all $\lambda_j, j = 1, \dots, k$ except for the i^{th} one. The formula for α_i is given below. Based on this sample, redefine the sets of indices $\mathcal{A} = \{i : \lambda_i = 1\}$ and $\mathcal{S} = \{i : \lambda_i = 0\}$.

- (ii) Sample π^* from its complete conditional distribution, which is $\text{Beta}(\alpha_{\pi^*} + n_A, \beta_{\pi^*} + n_S)$.
- (iii) Sample \underline{q}_A and \underline{q}_S from their complete conditional distributions, which are symmetric Dirichlets of dimension k_A and k_S and parameters $\{X_i + \theta_A, i \in \mathcal{A}\}$ and $\{X_i + \theta_S, i \in \mathcal{S}\}$, respectively.
- (iv) If one wishes to update k_0 , the number of "missing species", then sample k_0 from $f(k_0 | \underline{X}^*, \underline{\lambda}, \theta_S, P)$ using a Metropolis step, where \underline{X}^* is the vector containing all $X_i > 0$ from our sample. Details for this step are given below.

Given these posterior samples, then, posterior samples of π_i are constructed by $\pi_i = \{q_i \pi^*\}^{\lambda_i} \{q_i (1 - \pi^*)\}^{(1 - \lambda_i)}$.

In step (i), the probability $\alpha_i = O_i / (O_i + 1)$, 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

$$\begin{aligned}
 BF_i &= \left\{ \frac{\Gamma(n_{A(-i)} + \alpha_{\pi^*} + X_i) \Gamma(n_{S(-i)} + \beta_{\pi^*})}{\Gamma(n_{A(-i)} + \alpha_{\pi^*}) \Gamma(n_{S(-i)} + \beta_{\pi^*} + X_i)} \right\} & (2.1) \\
 &\times \left\{ \frac{\Gamma(n_{A(-i)} + k_{A(-i)} \theta_A) \Gamma(n_{S(-i)} + k_{S(-i)} \theta_S + X_i + \theta_S)}{\Gamma(n_{A(-i)} + k_{A(-i)} \theta_A + X_i + \theta_A) \Gamma(n_{S(-i)} + k_{S(-i)} \theta_S)} \right\} \\
 &\times \left\{ \frac{\Gamma(k_{A(-i)} \theta_A + \theta_A) \Gamma(k_{S(-i)} \theta_S)}{\Gamma(k_{A(-i)} \theta_A) \Gamma(k_{S(-i)} \theta_S + \theta_S)} \right\} \left\{ \frac{\Gamma(\theta_A + X_i) \Gamma(\theta_S)}{\Gamma(\theta_A) \Gamma(\theta_S + X_i)} \right\}.
 \end{aligned}$$

$\Gamma(x) = \int_0^\infty \exp(-u) u^{x-1} du$ 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 species i , and $n_{A(-i)} = \sum_{j \neq i} X_j \lambda_j$ and $n_{S(-i)} = \sum_{j \neq i} X_j (1 - \lambda_j)$ are the total abundant and scarce counts, again leaving out species i . The first and second factors (in curly braces) arise from two Dirichlet multinomial distributions for the observed tag counts in A and S . The third and fourth factors arise from a Beta-Binomial distribution for k_A and k_S . Note that this expression differs from the expression (3) in Morris, Baggerly and Coombes (2003), which contains typographical errors.

In step (iv), to update k_0 , we generate a proposal value $k'_0 \sim N(k_0, \sigma_{k_0})$ for some chosen proposal standard deviation σ_{k_0} . We accept this new proposal with probability $\min(1, \gamma)$, given by

$$\begin{aligned}
 \gamma &= \frac{\Gamma\{(k_S^* + k'_0) \theta_S\} \Gamma\{n_S + (k_S^* + k'_0) \theta_S\}}{\Gamma\{(k_S^* + k_0) \theta_S\} \Gamma\{n_S + (k_S^* + k_0) \theta_S\}} & (2.2) \\
 &\times \frac{\Gamma(k_S^* + k'_0 + 1) \Gamma(k'_0 + 1) k'_0}{\Gamma(k_S^* + k_0 + 1) \Gamma(k_0 + 1) k_0} (1 - P)^{k'_0 - k_0},
 \end{aligned}$$

where $k_S^* = \sum_{i=1}^{k^*} (1 - \lambda_i)$. The first two terms on the second line come from the fact that the generalized multinomial distribution must be used when the number of zero-class categories are unknown, which induces an extra combinatoric term into the likelihood, see Boender and Kan (1987). In updating k_0 , we assume that all missing species are scarce, i.e. $\lambda_i = 0$ for all $i : X_i = 0$.

2.6 Simulation Study

In Morris, Baggerly, and Coombes (2003), a simulation study was performed to compare the relative performance of four estimators in estimating the relative frequencies of mRNA transcripts in SAGE: the MLE, a Bayesian estimator with symmetric Dirichlet prior, a Bayesian estimator with the mixture Dirichlet prior, and an alternative empirical Bayes-based shrinkage estimator from Good (1953).

2.6.1 Description of Simulation

The simulation was based on a true set of relative frequencies $\underline{\pi}$ obtained by pooling together the observed counts from six SAGE libraries from breast cancer tissue in a study at M.D. Anderson Cancer Center. All totaled, these data consist of 495,947 sequenced tags, with $k = 44,984$ unique tags, assumed to represent different mRNA species. There were 684 (1.5%) of the tags with relative frequencies greater than 50 copies per cell, accounting for 41% of the total mRNA mass. With such a large number of sequenced tags, our hope is that the set of observed relative frequencies from this pooled sample is a reasonable approximation for the distribution of true relative frequencies of mRNA transcripts in a biological tissue sample.

Simulations were performed with SAGE sample sizes of $n = 10,000$ and $n = 50,000$ transcripts. For each, 100 samples of size n were randomly generated from a multinomial population with relative frequencies $\underline{\pi}$. For each sample, $\underline{\pi}$ was estimated using various estimators: (1) maximum likelihood and Bayesian posterior means using the (2) symmetric Dirichlet prior with $\theta = 1$, and (3) the mixture Dirichlet prior with $\theta_S = 0.5$ and $\theta_A = 1.0$, and (4) a shrinkage estimator described by Good (1953), and attributed to Turing. This estimator is empirical Bayes and involves substituting a smoothed histogram of the observed counts for the histogram of true counts. For the mixture Dirichlet, $P = 0.005$ was used for the $n = 10,000$ simulation, and $P = 0.01$ was used for the

$n = 50,000$ simulations. The total number of mRNA transcripts k was assumed known at 44,984, and π^* was given a Uniform prior. The estimates were based on 2000 MCMC iterations obtained after a burn-in of 100.

The squared error loss for each of the three estimators was computed for each of the $k = 44,984$ species in each dataset. The squared error loss for species i in dataset j , SE_{ij} , is given by $(\hat{\pi}_{ij} - \pi_i)^2$. From this, the mean square error for each species i was computed as $MSE_i = 100^{-1} \sum_{j=1}^{100} SE_{ij}$. These measures compared the performance for each individual species over repeated sampling. Overall performance averaging over species was assessed using integrated mean square error, $IMSE = \sum_{i=1}^k MSE_i$. In order to compare estimators for each given data set, the integrated square error for each sample j was computed as $ISE_j = \sum_{i=1}^k SE_{ij}$. All these summary measures were rescaled by a factor of 10^7 for readability.

2.6.2 Results

Figure ?? summarizes the simulation results for the $n = 10,000$ and $n = 50,000$ simulations. For each simulation, the relative efficiency of the three estimators to the MLE for each species is plotted against the true relative abundance. The relative efficiency for species i for estimator l is given by $RE_{il} = MSE_{i,MLE} / MSE_{il}$. To make the plot more readable, the relative efficiencies for species with similar true abundances were averaged together and plotted as a single circle, with the size of the circle made proportional to the log of the number of species averaged at that abundance.

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 201 for the MLE. This was what we expected based on our discussion in Section 2.3.2.

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. For the scarce tags (0-50 copies per cell), the mixture method was more efficient with RE of up to 25. These scarce tags account for

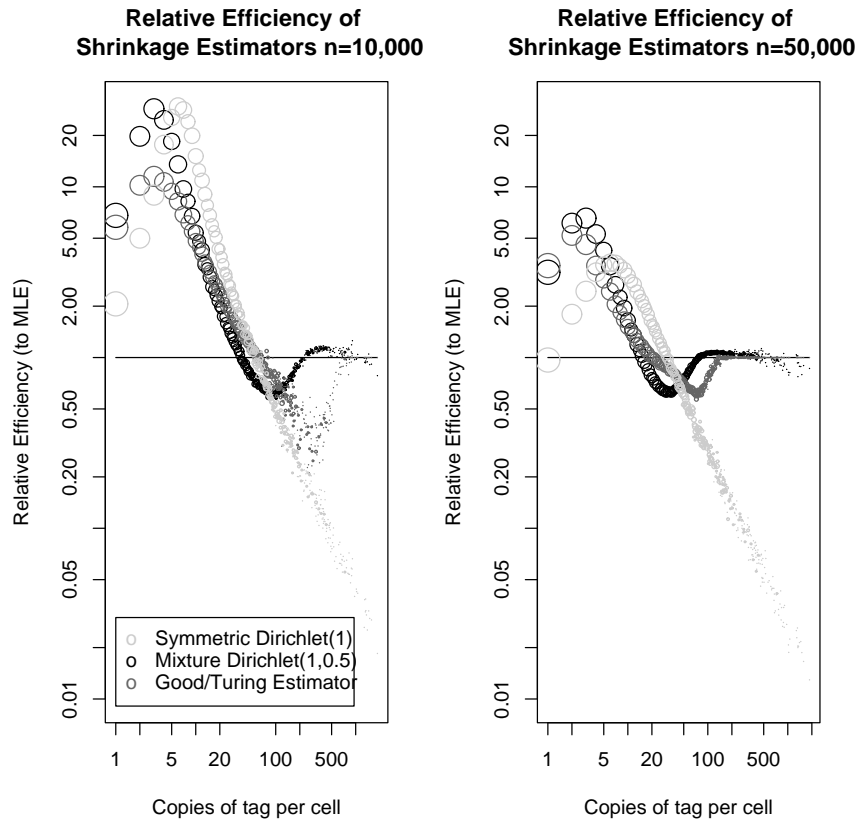


Fig. 2.2. *Simulation Results.* Relative Efficiency of estimators from a simulation of 100 multinomial samples of size 10,000 and 50,000 taken from a SAGE-like population. The horizontal axis consists of true relative frequencies multiplied by 300,000 to represent number of copies per cell containing 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.

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. Turing’s estimator had identical IMSE to the Mixture Dirichlet (643), with slightly extended regions of improved efficiency (0-60 copies per cell) and reduced efficiency (60-500 copies per cell) within the parameter space.

For $n = 50,000$, the mixture Dirichlet again had smaller IMSE than the MLE (166 vs. 201), outperforming the MLE for scarce tags (0-15 copies per cell, 93.5% of total), with RE of up to 6, with less efficiency than the MLE for an intermediate range (15-50 copies per cell, 4.9% of total), and equivalent for the abundant tags. The magnitude of improvement for scarce tags was again larger than the efficiency loss in the intermediate range ($RE > 0.50$). Turing’s estimator had a slightly smaller IMSE (160) than the mixture Dirichlet, and again had extended regions of improved efficiency (0-20 copies per cell) and reduced efficiency (20-150 copies per cell) within the parameter space.

2.7 Conclusion

We have introduced a new method for estimating the relative abundance profiles of SAGE tags that explicitly takes into account the skewed nature of the data and as a result can have efficiency advantages over the MLE. Its key benefit is that the nonlinear shrinkage profile imposed by our prior helps correct for some of the sampling limitations in the data. Other methods could be constructed that yield nonlinear shrinkage profiles, and would likely experience similar types of efficiency gains (and tradeoffs) as our method.

Turing’s estimator cited in Good (1953) had similar performance to ours, although it is not model-based. Blades, et al. (2005) propose a mixture model for SAGE data that partitions the data into scarce and abundant tags based on whether they belong to the linear or noise portions of a $\log(\text{number of tags})$ vs. $\log(\text{frequency of tags})$ plot. The nonlinear shrinkage profile inherent in their method results in more shrinkage for intermediate and abundant tags than ours, with shrinkage for tags with observed counts into the 10’s and 100’s.

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, including fold-change differences. The efficiency advantages seen in estimation should translate to inferential procedures with better properties. For example, fold-change assessments involving species with very small counts in one group should be more accurate using inferential procedures based on the shrinkage estimator introduced here.

Another advantage of this method is that it can give an estimate of k_0 , the number of missing species, and thus can estimate k , the total number of unique mRNA transcripts expressed in the biological sample. In that case, it is important to first apply a method to correct likely sequencing errors before applying the procedure described in this chapter, for example the procedures described by Colinge and Feger (2001) or Blades, Parmigiani, and Velculescu (2005). We believe these estimates of k should be taken with a grain of salt, however, since they are still based on a severely oversimplified model for the distribution of the π_i across species, consisting of a mixture of 2 symmetric Dirichlets. It would be interesting to consider other more flexible prior distributions for $\underline{\pi}$ that do an even better job of accommodating the heterogeneity seen in these expression data. This would likely lead to even better shrinkage estimators for $\underline{\pi}$ as well as improved estimates of the number of mRNA species present in the sample, k .

2.8 References

- Blades N., Parmigiani G., and Velculescu, V. (2005). Error estimation in SAGE libraries. *Genome Biology*, in press.
- Blades N., Kern S., Jones J., and Parmigiani G. (2005). Denoising SAGE libraries: A mixture model approach. *Bioinformatics*, in press.
- Boender C.G.E. and Kan A.H.G.R. (1987). A multinomial Bayesian approach to the estimation of population and vocabulary size. *Biometrika* **74**, 849–856.
- Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, **92**, 1413–1421.
- Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264.

- Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford: Clarendon Press.
- Morris, J. S., Baggerly, K. A., and Coombes, K. R. (2003). Bayesian shrinkage estimators of the relative abundance of mRNA transcripts using SAGE. *Biometrics* **59**, 476–486.
- Velculescu V. E., et al. (1999). Analysis of human transcriptomes. *Nature Genetics* **23**, 387–388.
- Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes Factors. *Journal of the American Statistical Association* **93**, 173–179.