Generalized Adaptive Shrinkage Methods and Applications in Genomics Studies

Mengyin Lu

1 General adaptive shrinkage

1.1 Introduction

From the previous sections, we see that the adaptive shrinkage (*ash*) methods can be extended to deal with data from various distributions and hence used in gene expression analysis. The normal *ash* started from observations with normal likelihood. For variance shrinkage problems, we developed models for gamma (chi-squared) distributed observed variances. In RNA-seq applications, the student t likelihood was used in *ash* to deal with the small sample size issues. The key idea of *ash* and the above extensions is the use of unimodal prior assumption (UA), which is highly adaptive to data and sensible in many contexts (not limited to genomics studies). Hence, apart from data with normal, gamma and t likelihood, it is natural for us to explore potential adaptive shrinkage approaches for generic data.

In practice, we typically use a finite mixture of uniforms to approximate any

unimodal prior. Fortunately, the convolution between a unimodal distribution and general likelihood is generally straightforward using existing software. Here we exploit this to develop a general ash framework that can be applied to many commonly encountered likelihoods (binomial, Poisson, etc.).

1.2 Methods

1.2.1 Models

Suppose we observe Y_j $(j = 1, ..., J)$, which is a random variable with likelihood $\phi_j(\theta_j) := p(Y_j|\theta_j)$, and the parameter of our interest is θ_j. Our goal is to make inference (hypothesis testing, estimation) on θ_j. Under the *ash* [3] framework, we use a Bayesian model to borrow information across the observations and use the posterior distribution to estimate or test θ_j.

We assume that after some transformation $h(\cdot)$, the true parameters θ_j come from a common unimodal prior $g(\cdot)$:

$$h(\theta_j) \sim g(\cdot), \tag{1}$$

where $h(\cdot)$ is the link function. We assume h is a strictly monotone increasing function.

1.2.2 Estimate prior distribution g

As in Stephens [3], we use a mixture of uniform distributions to approximate g:

$$g \sim \sum_{k=1}^{K} Z_k U[a_k, b_k], \tag{2}$$

where $(Z_1, ..., Z_K)$ are latent binary indicators for mixture components following multinomial distribution with $n = 1$ and $P(Z_k = 1) = \pi_k$, $U[a_k, b_k]$ denotes a uniform random variable on $[a_k, b_k]$. Given a fixed grid of a_k and b_k, we use the empirical Bayes method to estimate the mixture proportion π. To do this we first compute the matrix $L = (L_{jk})$ where each entry L_{jk} is the likelihood of θ_j for the k'th prior component:

$$\begin{aligned} L_{jk} &:= p(Y_j|Z_{jk} = 1) && (3)\\ &= \int p(Y_j|\theta_j)p(\theta_j|Z_{jk} = 1)d\theta_j && (4)\\ &= \frac{1}{b_k - a_k}\int_{h^{-1}(a_k)}^{h^{-1}(b_k)} \phi_j(\theta_j)|h'(\theta_j)|d\theta_j, && (5) \end{aligned}$$

where h' is the derivative of h. If the mixture component is a point mass, i.e. $a_k = b_k$, the likelihood is simply given by the probability density $L_{jk} = \phi_j(h^{-1}(a_k))|h'(h^{-1}(a_k))|$.

Then the mixture proportions π are estimated by maximizing the log-likelihood:

$$l(\pi) = \sum_j \log\left(\sum_k \pi_k L_{jk}\right), \quad (6)$$

$$\hat{\pi} = \arg\max_{\pi} l(\pi). \quad (7)$$

This can be done using the same methods as in the normal case [3].

The integral of $\phi_j(\theta_j)|h'(\theta_j)|$ in (5) does not necessarily have analytical form. However, if the link function is identity link $h(x) = x$, we have

$$L_{jk} = \frac{\Phi_j(b_k) - \Phi_j(a_k)}{b_k - a_k}, \quad (8)$$

where $\Phi_j(x) := \int_{-\infty}^{x} \phi_j(y)dy$.

Now we define ψ_j as the density of the distribution proportional to $\phi_j(x)|h'(x)|$:

$$\psi_j(x) := \frac{\phi_j(x)|h'(x)|}{\int_{-\infty}^{\infty} \phi_j(x)|h'(x)|dx}, \tag{9}$$

and $\Psi(x) := \int_{-\infty}^{x} \psi_j(y)dy$ is the corresponding cdf.

If $\phi_j(\cdot;\theta)$ belongs to the exponential family and h is its natural link, $\psi_j(x)$ would be a distribution in the conjugate distribution family of ϕ_j, and its cdf can be used to compute the integral in (5). Some examples are illustrated in Section 1.2.4.

In practice, $\psi_j(\cdot)$ and $\Psi_j(\cdot)$ should be provided to compute the likelihood matrix L in order to fit the prior distribution. Otherwise, we can use numerical integral to calculate (5), but the computational stability might not be guaranteed.

1.2.3 Posterior distribution $p(\theta_j|Y_j, \hat{\pi})$

For any distribution (density) $f(x)$, we denote $f^{\text{trunc}}(x; a, b)$ as its truncated distribution on interval $[a, b]$:

$$f^{\text{trunc}}(x; a, b) := \frac{f(x)}{\int_a^b f(y)dy}, \tag{10}$$

and denote $M^f(a, b)$ as the mean of $f^{\text{trunc}}(x; a, b)$:

$$M^f(a, b) := \int_{-\infty}^{\infty} x f^{\text{trunc}}(x; a, b)dx. \tag{11}$$

For the corner case $a = b$, $f^{\text{trunc}}(x; a, a) := \delta_a$ which is the point mass on a, and $M^f(a, a) = a$.

The posterior distribution of θ_j given observation Y_j and fitted prior mixture proportions $\hat{\pi}$ is given by:

$$p(\theta_j|Y_j, \hat{\pi}) = \frac{p(Y_j|\theta_j)p(\theta_j)}{\int p(Y_j|\theta_j)p(\theta_j)d\theta_j} \tag{12}$$

$$= \frac{\phi_j(\theta_j)g(h(\theta_j))|h'(\theta_j)|}{\sum_k \hat{\pi}_k L_{jk}} \tag{13}$$

$$= \sum_k \tilde{\pi}_{jk} \psi_j^{\text{trunc}}(\theta_j; \tilde{a}_k, \tilde{b}_k), \tag{14}$$

where:

$$\psi_j(x) = \frac{\phi_j(x)|h'(x)|}{\int_{-\infty}^{\infty} \phi_j(x)|h'(x)|dx}, \tag{15}$$

$$\tilde{\pi}_{jk} = \frac{\hat{\pi}_k L_{jk}}{\sum_{k'} \hat{\pi}_{k'} L_{jk'}}, \tag{16}$$

$$\tilde{a}_k = h^{-1}(a_k), \tag{17}$$

$$\tilde{b}_k = h^{-1}(b_k). \tag{18}$$

In other words, the posterior distribution of θ_j is a mixture of truncated ψ_j distribution, truncated on $(\tilde{a}_k, \tilde{b}_k)$, with mixture proportions $\tilde{\pi}_{jk}$:

$$\theta_j|Y_j, \hat{\pi} \sim \sum_{k=1}^{K} \tilde{Z}_{jk} \psi_j^{\text{trunc}}(\theta_j; \tilde{a}_k, \tilde{b}_k), \tag{19}$$

where $(\tilde{Z}_{j1}, ..., \tilde{Z}_{jK})$ are latent binary indicators for mixture components, following

multinomial distribution with $n = 1$ and probability $P(\tilde{Z}_{jk} = 1) = \tilde{\pi}_{jk}$.

Following the posterior distribution, we can calculate other quantities to estimate or test θ_j:

- Posterior mean:

$$\mathrm{E}(\theta_j | Y_j, \hat{\pi}) = \sum_k \tilde{\pi}_k M^{\psi_j}(\tilde{a}_k, \tilde{b}_k), \tag{20}$$

  which can be used as a shrinkage estimator for θ_j.

- Local false discovery rate (lfdr): if the prior includes a mixture component corresponding to the null hypothesis $\theta_j = 0$, i.e. $h^{-1}(a_k) = h^{-1}(b_k) = 0$, then lfdr for θ_j is given by

$$\mathrm{lfdr}_j = P(\theta_j = 0 | Y_j, \hat{\pi}), \tag{21}$$

  which is the posterior mixture proportion for that null component.

- Local false sign rate (lfsr) as defined in [3]:

$$\mathrm{lfsr}_j = P(\theta_j = 0 | Y_j, \hat{\pi}) + \min(P(\theta_j > 0 | Y_j, \hat{\pi}), P(\theta_j < 0 | Y_j, \hat{\pi})), \tag{22}$$

  where

$$P(\theta_j < 0 | Y_j, \hat{\pi}) = \sum_k \frac{\tilde{\pi}_k \Psi_j(0)}{\Psi_j(\tilde{b}_k) - \Psi_j(\tilde{a}_k)}, \tag{23}$$

  and $P(\theta_j > 0 | Y_j, \hat{\pi}) = 1 - P(\theta_j = 0 | Y_j, \hat{\pi}) - P(\theta_j < 0 | Y_j, \hat{\pi})$.

1.2.4 Estimate unknown mode

In previous sections we assume that for the unimodal prior g, the uniform mixture components $\{a_k, b_k\}$ are fixed. However in some cases, the mode is unknown and we would like to estimate the prior using the empirical Bayes method:

$$\hat{g} = \arg \max_{g \text{ unimodal}} l(g), \tag{24}$$

hence $\hat{g}$ optimizes the log-likelihood.

In practice, we solve this optimization problem as follows: for each given mode c, we construct a grid $\{a_k, b_k\}$ which is anchored at mode c and covers a sufficient wide range, and estimate the mixture proportions $\hat{\pi}$ which achieves the maximum log-likelihood (denote by l_c). Thereby, l_c itself is a function of the mode c. We use the numerical optimization function `stats::optimize` in R to search for the optimizer $\hat{c} = \arg\max_c l_c$.

1.2.5 Special cases

Table 1 lists some special cases of general *ash*, where the likelihood ϕ_j is a commonly used distribution. We will discuss *fash*, Poisson *ash* and Binomial *ash* in detail in Section 1.3. Here we define the non-standard log-F distribution $\log F(\cdot; \mu, \nu_1, \nu_2)$ as follows: if for a random variable X, we have $\exp(X - \mu) \sim F(\nu_1, \nu_2)$, then we say X follows the distribution $\log F(X; \mu, \nu_1, \nu_2)$.

1.3 Applications

1.3.1 Adaptive shrinkage of F statistics (*fash*)

A special case for the *ash* methods with general likelihood would be the adaptive shrinkage of F statistics. F statistics are normally used for testing equality of two variances, or multiple-comparison ANOVA problems. In genomic contexts, pooling information across genes may help improve the statistical power of gene-specific F tests. Smyth [2] suggested using the moderated error variance estimates to adjust F statistics, assuming that the gene-specific variances come from a common inverse-gamma prior. Nevertheless, we can directly work on the gene-specific F-statistics and fit their prior more adaptively by a unimodal distribution.

Suppose we have the expression matrix Y for G genes and N samples from $M(>=2)$ conditions. Consider the following two problems related to F test: variability comparison and variance decompositions.

Variability comparison Suppose we would like to compare the expression variability within condition A and the variability within condition B. The statistical

Table 1: Special cases of general *ash*

	Model		Posterior		
Case	ϕ_j	$h(x)$	ψ_j	$\tilde{a}_k$	$\tilde{b}_k$
ash	$N(Y_j;\theta_j,s_j^2)$	x	$N(\theta_j;Y_j,s_j^2)$	a_k	b_k
fash	$\log F(Y_j;\theta_j,\nu_1,\nu_2)$	x	$\log F(\theta_j;Y_j,\nu_2,\nu_1)$	a_k	b_k
Poisson *ash*	$\text{Poisson}(Y_j;c_j\theta_j)$	x	$\text{Gamma}(\theta_j;Y_j+1,c_j)$	a_k	b_k
Poisson *ash*	$\text{Poisson}(Y_j;c_j\theta_j)$	$\log(x)$	$\text{Gamma}(\theta_j;Y_j,c_j)$	e^{a_k}	e^{b_k}
Binomial *ash*	$\text{Bin}(Y_j;n_j,\theta_j)$	x	$\text{Beta}(\theta_j;Y_j+1,n_j-Y_j+1)$	a_k	b_k
Binomial *ash*	$\text{Bin}(Y_j;n_j,\theta_j)$	$\text{logit}(x)$	$\text{Beta}(\theta_j;Y_j,n_j-Y_j)$	$\frac{1}{1+e^{-a_k}}$	$\frac{1}{1+e^{-b_k}}$

model is defined by:

$$Y_{gi} = \mu_g + \beta_{g,c(i)} + e_{gi}, \tag{25}$$

$$e_{gi} \sim N(0, \sigma^2_{g,c(i)}), \tag{26}$$

where g is the index for gene, i is the index for samples and $c(i)$ is the condition indicator, either A or B. Suppose there are N_A and N_B samples in group A and B respectively. All observations are independent with each other.

A straightforward way to estimate the true variance ratio $\frac{\sigma^2_{gA}}{\sigma^2_{gB}}$ (denoted by α_g) is using the ratio of sample variances $\frac{\hat{\sigma}^2_{gA}}{\hat{\sigma}^2_{gB}}$ (denoted by F_g). Its sampling distribution is given by

$$F_g \sim \alpha_g \times F, \tag{27}$$

where F is a F-distributed random variable with degrees of freedom $N_A - 1$ and $N_B - 1$. Let the null hypothesis be: the two conditions have same expression variability ($H_0 : \sigma_{gA} = \sigma_{gB}$), then under the null $\alpha_g = 1$.

Transforming (26), we have

$$\log(F_g) - \log(\alpha_g) | \log(\alpha_g) \sim \log F, \tag{28}$$

where $\log F$ is the logarithm of F-distributed random variable with d.f. $N_A - 1$ and $N_B - 1$. Note that (27) meets the form of general *ash* problem, where $\log(\alpha_g)$ is our parameters of interest with log-F likelihood.

Analogous to (1), assuming $\log(\alpha_g)$ come from a common unimodal prior, the

general *ash* framework can be further used to improve estimates of $\log(\alpha_g)$. According to Table 1, the posterior distribution of $\log(\alpha_g)$ is given by a mixture of truncated $\log F(\cdot; \log(F_g), N_B - 1, N_A - 1)$ distribution (with different truncation limits for different mixture components). By pooling information across genes, the posterior estimates of $\log(\alpha_g)$ are presumably more accurate than the raw noisy estimates $\log(F_g)$.

Variance decomposition Suppose we would like to compare the expression variability explained by conditions to the variability due to noise (or the total variability). The statistical model is defined by:

$$Y_{gi} = \mu_g + \beta_{g,c(i)} + e_{gi}, \tag{29}$$

$$e_{gi} \sim N(0, \sigma^2_{ge}), \tag{30}$$

$$\beta_{g,c(i)} \sim N(0, \sigma^2_{gc}), \tag{31}$$

where $c(i)$ is the condition level of sample i, $\beta_{g,c(i)}$ is the random condition effect. Suppose the design is balanced so we have equal number of samples for each condition (M conditions in total).

Our goal is to compare σ^2_{gc} (variance of condition effect) against σ^2_{ge} (variance among replicates). The ANOVA F-statistic F_g can be used for variance decomposi-

tion, and its sampling distribution is given by

$$F_g = \frac{\mathrm{SST}_g/(M-1)}{\mathrm{SSE}_g/(N-M)} \tag{32}$$

$$= \frac{(\sum_i (\bar{Y}_{g,c(i)} - \bar{Y}_g)^2)/(M-1)}{(\sum_i (Y_{g,c(i)} - \bar{Y}_{g,c(i)})^2)/(N-M)} \tag{33}$$

$$\sim (1 + M\sigma^2_{gc}/\sigma^2_{ge}) \times F \tag{34}$$

$$\sim \alpha_g \times F \quad (\alpha_g := 1 + M\sigma^2_{gc}/\sigma^2_{ge}) \tag{35}$$

where F is a F-distributed random variable with d.f $M-1$ and $N-M$, $\bar{Y}_{g,c(i)}$ is the condition $c(i)$'s expression mean (average of all $Y_{g,c(i)}$'s with condition $c(i)$), and $\bar{Y}_g$ is the overall expression mean for gene g. Let the null hypothesis be: there are no condition effects ($H_0 : \sigma_{gc} = 0$), then under the null $\alpha_g = 1$.

Similarly, we can use the general *ash* framework to fit a unimodal prior for $\log(\alpha_g)$ and use posterior means to estimate $\log(\alpha_g)$. Then the ratio of condition variance and error variance $\sigma^2_{gc}/\sigma^2_{ge}$ can be estimated by transforming the estimate of $\log(\alpha_g)$. Table 1 shows the analytical form of the posterior of $\log(\alpha_g)$: a mixture of truncated $\log F(\cdot; \log(F_g), M-1, N-M)$ distribution (with different truncation limits for different mixture components).

Note that this method can only apply to balanced dataset with equal number of samples for each condition, since (33) does not hold for unbalanced dataset.

Example: variance decomposition for stem cell expression data We have the microarray gene expression data from Burrows et al. [1]. The dataset has four individuals. Each individual has four samples types - Fibroblast, LCL, F-iPSC, L-

iPSC, where L-iPSC refers to iPSCs derived from LCLs, F-iPSC refers to iPSCs derived from Fibroblasts. The L-iPSC type has three replicates A, B and C, and the other three types only have one replicate, so there are 6 samples for each individual.

Burrows et al. [1] were interested in the proportion of expression variance explained by cell type of origin versus that explained by individual in the iPSCs. They performed a linear mixed model with a fixed effect for cell type of origin (i.e. L-iPSC vs F-iPSC) and a random effect for individual. This model did no use the LCLs or the Fibroblasts from these individuals.

We use the naive ANOVA F-test and *fash* to analyze the proportion of variation explained by cell-type or individual. Specifically, we assume that gene expression y_{gij} comes from the following model:

$$y_{gij} = \mu_g + \beta_{gi} + \gamma_{gj} + e_{gij}, \tag{36}$$

where g, i, j are the indices for gene, individual and cell type respectively. β and γ are random effects for individuals and cell-types respectively. Suppose $\beta_{gi} \sim N(0, v_g^{(\text{ind})})$, $\gamma_{gj} \sim N(0, v_g^{(\text{ct})})$ and $e_{gij} \sim N(0, v_g^{(\text{err})})$, we are interested in estimating the "PVE" (proportion of variance explained) by individual or cell-type defined as follows:

$$\text{PVE}_g^{(\text{ind})} := \frac{v_g^{(\text{ind})}}{v_g^{(\text{ind})} + v_g^{(\text{ct})} + v_g^{(\text{err})}}, \tag{37}$$

$$\text{PVE}_g^{(\text{ct})} := \frac{v_g^{(\text{ct})}}{v_g^{(\text{ind})} + v_g^{(\text{ct})} + v_g^{(\text{err})}}. \tag{38}$$

Note that this dataset has unbalanced design (three L-iPSC replicates but just

one F-iPSC sample for each individual), and it is infeasible to use *fash* on unbalanced dataset for PVE analysis as we discussed before. Hence we choose an ad-hoc way: each time we simply use one of the three L-iPSC replicates to form a balanced dataset, and compare the results of three trials. Fortunately the three trials give very similar results. Figure 1 shows the posterior mean of gene-specific PVEs of cell-type or individual estimated by F-test and *fash* for the subsets using each of the L-iPSC replicate.

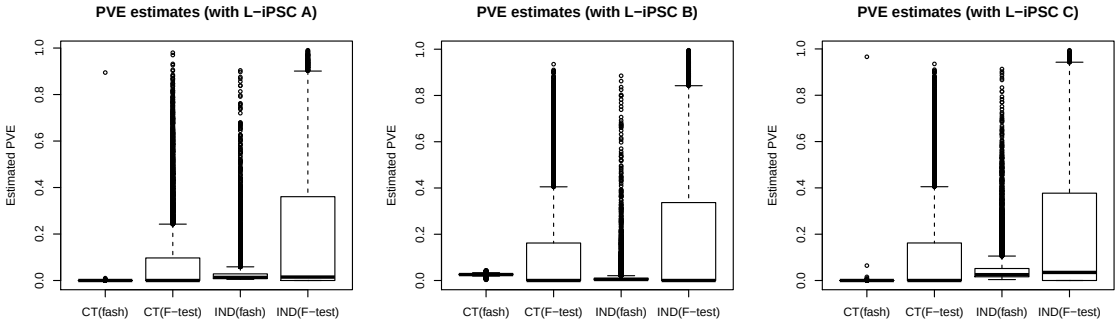

Figure 1: Gene-specific PVE estimates of cell-type (CT) or individual (IND), estimated by F-test and *fash* on Burrows data. Each time we only use one of the three L-iPSC replicates to form a balanced dataset.

Burrows et al. [1] performed the *limma* DE analysis for each pair of cell types and found that $INPP5F$ is the most common DE gene. They also used simple ANOVA R^2 to record the variance explained by cell-types or individuals. The ANOVA R^2 is defined as $R^2 := \text{SST}/(\text{SST}+\text{SSE})$, where SST and SSE are the same as in (31). Note that R^2 is different from our defined "PVE" in (37). They conclude that "individual genetic background captures a much larger proportion of gene regulatory variation than cell type of origin".

The *fash* results are generally consistent with Burrows et al. [1]: all genes have almost zero PVE for cell-types, except for gene $INPP5F$ (ENSG00000198825). Compared to the raw F-test PVE estimates (which are substantially noisy), *fash* tends to shrink them towards 0.

1.3.2 Adaptive shrinkage on binomial data (Binomial *ash*)

Table 1 gives us the analytical forms of Binomial *ash*, where we have binomial observations $Y_j \sim \text{Binomial}(n_j, p_j)$ $(j = 1, ..., J)$ and n_j's are known. The unknown success probability parameter p_j is of our interest. Binomial *ash* allows us to borrow information across the observations and use the posterior distribution to estimate p_j.

Example: comparison between bulk RNA-seq and scRNA-seq data In recent years, single cell RNA-seq (scRNA-seq) methods have been more and more frequently used in gene expression analysis. While bulk RNA-seq data mostly extract gene expression features from millions of cells which have been pooled together, scRNA-seq can capture expression profile of individual cells. The scRNA-seq technologies would allow us to fetch more information about the heterogeneity of gene expression across cells. The comparison between bulk RNA-seq and scRNA-seq could thus be interesting. If we have both scRNA-seq data and corresponding bulk RNA-seq data on the same sample, we might want to quantify the concordance as well as difference between them. Presumably, the difference between bulk RNA-seq and scRNA-seq data may rise from various possible sources: effects due to the dynamics of cell transcription, technical differences in sequencing protocols, etc. Hence, investigating the genes with significant difference might help us better understand the

biological mechanism of certain genes as well as the underlying technical features of scRNA-seq data.

Suppose we have both scRNA-seq and bulk RNA-seq data on the same sample. Let X_{jg}^s denote the observed counts of gene g in single cell j. And let X_g^b denote the counts of gene g in the bulk. We first pool the single cell data into a single count, and define $X_g^s := \sum_j X_{jg}^s$. Now we might want to identify the genes that show the most "significant" deviations between X_g^b and X_g^s, and quantify those deviations.

Suppose the bulk and single cell data are independent, then:

$$X_g^b | C_g \sim \text{Binomial}(C_g, p_g), \tag{39}$$

where $C_g := X_g^b + X_g^s$ is the total count, p_g is the fraction of all reads that come from bulk at gene g. If the single cell data and bulk data are generally concordant, then the bulk RNA-seq expression level should be roughly proportional to scRNA-seq expression level (the ratio relies on sequencing depths). As a result, condition on the total counts C_g, the bulk fraction p_g is supposedly similar across genes. The "outlier" genes where p_g is particularly small or large might be suspicious.

Note that the gene-specific sample bulk fraction $\hat{p}_g := X_g^b / C_g$ is the raw maximum likelihood estimate (MLE) of p_g. We also use the binomial *ash* to estimate p_g, assuming that p_g comes from a unimodal prior with some unknown mode (to be estimated). The posterior mean of p_g (denoted by $\tilde{p}_g$) is thus a shrinkage estimator of p_g, borrowing information across genes. The prior as well as its mode are estimated by the empirical Bayes approach, which makes them adaptive to data.

Tung et al. [4] provide both scRNA-seq and bulk RNA-seq data for same samples

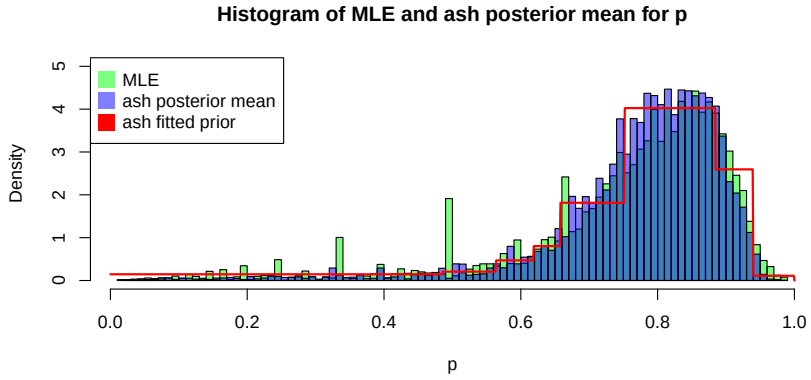

Figure 2: Distribution of sample bulk reads fraction $\hat{p}_g = X_g^b/C_g$ and Binomial *ash* posterior estimates on Tung data (NA19091.r1). The red line is the *ash* fitted prior of p_g.

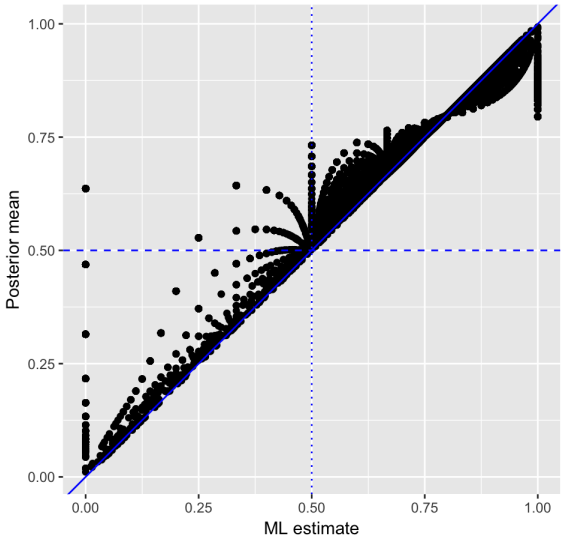

Figure 3: Binomial *ash* posterior estimates $\tilde{p}_g$ versus the ML estimates $\hat{p}_g$ on Tung data (NA19091.r1).

(three individuals and three replicates for each individual). We compare the single cell and bulk data for one replicate NA19091.r1. Genes with both non-zero X_g^b and non-zero X_g^s are selected for our analysis.

Figure 2 shows the distribution of sample bulk reads fraction $\hat{p}_g = X_g^b/C_g$ and the binomial *ash* posterior estimates $\tilde{p}_g$. Although most sample fractions $\hat{p}_g$ are over 0.6, there are some extremely small outliers around 0. The binomial *ash* fitted prior is unimodal with mode around 0.8, and the left tail keeps flat from 0 to near 0.5. Figure 3 plots the posterior estimates $\tilde{p}_g$ from binomial *ash* versus the sample fraction $\hat{p}_g$. Both figures show that on the left side, quite a few small $\hat{p}_g$'s are pushed higher by binomial *ash*. These genes are further examined and turn out to be low expressed genes, and their bulk reads fractions are highly variable due to the small total count C_g. Thereby, binomial *ash* shrinks these posterior means towards the prior mean, after accounting for the lack of informativeness in low expressed genes.

Table 2 lists the genes where the posterior bulk fraction $\tilde{p}_g$ is extremely small or large. We might want to further inspect these genes to investigate the cause of difference between bulk RNA-seq and scRNA-seq expression.

Table 2: Genes with extremely small or large $\tilde{p}_g$ on Tung data (NA19091.r1).

Gene name	Ensemble ID	X_g^s	X_g^b	$\tilde{p}_g$
TSHZ2	ENSG00000182463	64	1	0.030
HIST1H4L	ENSG00000198558	144	4	0.033
MTRNR2L6	ENSG00000270672	433	16	0.038
BCKDHA	ENSG00000248098	23	1547	0.985
RAB19	ENSG00000146955	3	241	0.978
TUBB3	ENSG00000258947	110	4264	0.975

1.3.3 Poisson data

Table 1 provides us the analytical forms of Poisson *ash*, where we have Poisson observations $Y_j \sim \text{Poisson}(c_j\lambda_j)$ $(j = 1, ..., J)$ and c_j's are known scaling factors. The unknown intensity parameter λ_j is of our interest. Poisson *ash* allows us to borrow information across the observations and use the posterior distribution to estimate λ_j.

Example While normal distribution based models have been widely used on classical gene expression data (microarray, bulk RNA-seq), they have non-negligible limitations when handling single cell RNA-seq data, which typically have zero inflation and low count level issues. Thereby count distribution based models are generally preferred for scRNA-seq analysis. In next Chapter, we will discuss the usage of Poisson *ash* on scRNA-seq data and compare with some existing methods.

References

[1] Burrows, C. K., N. E. Banovich, B. J. Pavlovic, K. Patterson, I. G. Romero, J. K. Pritchard, and Y. Gilad (2016). Genetic variation, not cell type of origin, underlies the majority of identifiable regulatory differences in ipscs. *PLoS genetics 12*(1), e1005793. 11, 13

[2] Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology 3*(1), 3. 7

[3] Stephens, M. (2016). False discovery rates: a new deal. *Biostatistics*, kxw041. 2, 3, 6

[4] Tung, P.-Y., J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific reports 7*, 39921. 15