1. JSCET model

(1) notation:

gene	G
cell type	K
subject	J (NU in code)
length	$N = \sum_{j=1}^{J} T_j$ (NS in code, # of samples)

$Q = JK$

$N \times 4K$

★ If we consider age effect, $X \in [0,1)$

$$y = X\beta + Au + \varepsilon$$

$N \times G$; $N \times 2K$ ($4K$ for slope) ; $N \times Q$ ($Q = JK$) ; $2K \times G$; $Q \times G$

$$X = \begin{pmatrix} \theta_{111} & \theta_{112} & \dots & \theta_{11K} & z_1\theta_{111} & z_1\theta_{112} & \dots & z_1\theta_{11K} \\ \theta_{121} & \theta_{122} & \dots & \theta_{12K} & z_1\theta_{121} & z_1\theta_{122} & \dots & z_1\theta_{12K} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \theta_{1T_11} & \theta_{1T_12} & \dots & \theta_{1T_1K} & z_1\theta_{1T_11} & z_1\theta_{1T_12} & \dots & z_1\theta_{1T_1K} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \theta_{J11} & \theta_{J12} & \dots & \theta_{J1K} & z_J\theta_{J11} & z_J\theta_{J12} & \dots & z_J\theta_{J1K} \\ \theta_{J21} & \theta_{J22} & \dots & \theta_{J2K} & z_J\theta_{J21} & z_J\theta_{J22} & \dots & z_J\theta_{J2K} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \theta_{JT_J1} & \theta_{JT_J2} & \dots & \theta_{JT_JK} & z_J\theta_{JT_J1} & z_J\theta_{JT_J2} & \dots & z_J\theta_{JT_JK} \end{pmatrix}_{N \times 2K}$$

$$u = \begin{bmatrix} & \text{gene 1} & \dots & \text{gene } G \\ \text{celltype 1 for subject 1} & u_{11} & & \\ \text{celltype 2 for subject 1} & u_{21} & & \\ \vdots & & & \\ \text{celltype } K \text{ for subject } J & u_{JK} & & \end{bmatrix}$$

$$a_{11} = \begin{bmatrix} \theta_{111} \\ \theta_{121} \\ \vdots \\ \theta_{1T_11} \end{bmatrix}$$

$$A = \begin{pmatrix} \mathbf{a}_{11} & 0 & 0 & 0 & \dots & \mathbf{a}_{1K} & 0 & 0 & 0 \\ 0 & \mathbf{a}_{21} & 0 & 0 & \dots & 0 & \mathbf{a}_{2K} & 0 & 0 \\ & & & & & & & & \\ 0 & 0 & \ddots & 0 & \dots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{a}_{J1} & \dots & 0 & 0 & 0 & \mathbf{a}_{JK} \end{pmatrix}_{N \times Q}$$

$N = NS = NU \times T$

(5) 第1个 subject 在第1个 cell type 的 different proportion across time

where $\mathbf{a}_{jk} := (\theta_{j_1k}, \theta_{j_2k}, \cdots, \theta_{jT_jk})'$ is simply a reorganized vector of cell type proportions, to align with random effect $\boldsymbol{u}$.

$Q = NU \times K$

(2) assumption:

① $\varepsilon \sim \text{Normal}(0, \sigma_0^2)$ for every gene g and every subject j

② $u \sim MVN(\vec{0}_{Q\times 1}, \Sigma_u)$ for every gene g,

with $\Sigma_u = Cov(u) = \begin{pmatrix} \sigma_1^2 I_J & & 0 \\ & \sigma_1^2 I_J & \\ 0 & & \ddots \\ & & & \sigma_K^2 I_J \end{pmatrix}$

a_1: $k=1$ (第1个 cell type, 5个时间点)

a_{22}

a_{12}

	V1	V2	V3	V4	...
1	0.02230079	0.00000000	0.00000000	0.00000000	
2	0.01501410	0.00000000	0.00000000	0.00000000	
3	0.01314183	0.00000000	0.00000000	0.00000000	
4	0.04691058	0.00000000	0.00000000	0.00000000	
5	0.04692466	0.00000000	0.00000000	0.00000000	
6	0.00000000	0.07183477	0.00000000	0.00000000	
7	0.00000000	0.01136901	0.00000000	0.00000000	
8	0.00000000	0.06145904	0.00000000	0.00000000	
9	0.00000000	0.02264138	0.00000000	0.00000000	
10	0.00000000	0.04052955	0.00000000	0.00000000	
11	0.00000000	0.00000000	0.03578315	0.00000000	
12	0.00000000	0.00000000	0.08965811	0.00000000	
13	0.00000000	0.00000000	0.05633649	0.00000000	
14	0.00000000	0.00000000	0.08530417	0.00000000	
15	0.00000000	0.00000000	0.01453683	0.00000000	
16	0.00000000	0.00000000	0.00000000	0.05460231	

V50

$u \in \mathbb{R}^{Q\times 1}$

$Q = 50 \times 6 = 300$ in our case

$NS = 250$

$NU = 50$

$T = 5$

③ $y|u \sim N(X\beta + Au, \sigma_0^2 I)$. $W := (y, u) = (W_{observed}, W_{missed})$

$$\Rightarrow \begin{pmatrix} y \\ u \end{pmatrix} \sim N\left(\begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} A\Sigma_u A^T & A\Sigma_u \\ \Sigma_u^T A^T & \Sigma_u \end{bmatrix} \right)$$

let $\eta = (\underbrace{m_1, \cdots, m_K}_{\text{baseline of ctrl group}}, \underbrace{\beta_1, \cdots, \beta_K}_{\text{group effect}}, \sigma_1^2, \cdots, \sigma_K^2, \sigma_0^2)$

If we want to test age,

$\eta = (m_1, \cdots, m_K, \beta_1, \cdots, \beta_K, \underbrace{\alpha_1, \cdots, \alpha_K}_{\text{age effect (across all sample)}}, \underbrace{\gamma_1, \cdots, \gamma_K}_{\text{age * group effect}}, \sigma_1^2, \cdots, \sigma_K^2, \sigma_0^2)$

Note: In islet code, the output will be a G×K matrix for each subject j. (since slope are current unavailable), with each element m_{gk} for ctrl group. $\beta_{gk} + m_{gk}$ for case group for gene g, cell type k.

joint llk $l_0(\eta; y, u) = \log[f(y|u; X, A, \beta, \sigma) \cdot f(u; X, A, \beta, \sigma)]$
(log-likelihood)

$$\propto \frac{N}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}(y-X\beta-Au)^T(y-X\beta-Au) - \frac{J}{2}\sum_{k=1}^{K}\log(\sigma_k^2) - \frac{1}{2}\sum_{k=1}^{K}\frac{1}{\sigma_k^2}u_k^T u_k$$

Details: Check ISLET Additional file 1. Appendix

parameters of interest: $\sigma_0^2, \sigma_1^2, \sigma_2^2, \cdots, \sigma_K^2$, β, u random effect
solved by EM algorithm fixed effect

The EM algorithm calculation will then follow naturally.
E-step:

$$E[\boldsymbol{u}|\boldsymbol{w}_{obs} = \boldsymbol{y}] = \boldsymbol{\Sigma}_u \boldsymbol{A}' \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$V = A\Sigma_u A^T + \sigma_0^2 I_N$

$$E[\boldsymbol{s}'\boldsymbol{s}|\boldsymbol{w}_{obs} = \boldsymbol{y}] = tr(\boldsymbol{A}\boldsymbol{\Sigma}_p\boldsymbol{A}') + (\boldsymbol{A}\boldsymbol{\mu}_p + \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})'(\boldsymbol{A}\boldsymbol{\mu}_p + \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})$$

$$E[\boldsymbol{u}_k'\boldsymbol{u}_k|\boldsymbol{w}_{obs} = \boldsymbol{y}] = tr(\boldsymbol{\Sigma}_{p_k}) + \boldsymbol{\mu}_{p_k}'\boldsymbol{\mu}_{p_k}$$

Here, $\boldsymbol{s} = \boldsymbol{A}\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}$, $\boldsymbol{V} := \boldsymbol{A}\boldsymbol{\Sigma}_u\boldsymbol{A}' + \sigma_0^2\boldsymbol{I}$, $\boldsymbol{\Sigma}_{p_k}$ is the kth diagonal block of matrix $\boldsymbol{\Sigma}_p$, and $\boldsymbol{\mu}_{p_k}$ is the kth sub-vector in $\boldsymbol{\mu}_p$.
M-step:
For the $(t+1)^{th}$ iteration given the t^{th} iteration:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{A}E_{\eta^{(t)}}(\boldsymbol{u}^{(t)}))$$

$$\hat{\sigma}_0^{2(t+1)} = \frac{E_{\eta^{(t)}}[\boldsymbol{s}'\boldsymbol{s}|\boldsymbol{w}_{obs} = \boldsymbol{y}]}{N}$$

$$\hat{\sigma}_k^{2(t+1)} = \frac{E_{\eta^{(t)}}[\boldsymbol{u}_k'\boldsymbol{u}_k|\boldsymbol{w}_{obs} = \boldsymbol{y}]}{J}$$

Lemma 2.1. *If* $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, *and* $\boldsymbol{X} \sim N[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}]$, *then* $[\boldsymbol{X}_1|\boldsymbol{X}_2] \sim N(\boldsymbol{\mu}_{1|2}, \Sigma_{1|2})$, *where* $\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)$ *and* $\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Then we have

$$[\boldsymbol{w}_{mis}|\boldsymbol{w}_{obs}] = [\boldsymbol{u}|\boldsymbol{y}] \sim N_Q(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

and

$$\boldsymbol{\mu}_p = \boldsymbol{\Sigma}_u\boldsymbol{A}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad \boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_u - \boldsymbol{\Sigma}_u\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_u$$

Now we let $\boldsymbol{s} = \boldsymbol{A}\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}$ and $\boldsymbol{V} := \boldsymbol{A}\boldsymbol{\Sigma}_u\boldsymbol{A}' + \sigma_0^2\boldsymbol{I}_N$. To apply Expectation-Maximization (EM) algorithm, we need to evaluate $E_{\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\eta}}[l(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{u})]$, and consequently $E[\boldsymbol{s}'\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\eta}]$, $E[\boldsymbol{u}_k'\boldsymbol{u}_k|Y, \boldsymbol{\eta}]$.

模型现存的问题：(1) 不收敛. source code 只设置了14次迭代.
(2) u=0 的情况解不出来 (但这种情况 real life 少见)
simulation 中，u=0 时. 变为简单的解矩阵问题：$y = X\beta + \varepsilon$
with known X, y, want to solve β.

```
B_sum_val= 2046.932
B_change_val= 6.555662
B_change_prop= 0.3202677 %
Random effect diff2= 2.938226e-06
B_sum_val= 2045.422
B_change_val= 6.16733
B_change_prop= 0.3015187 %
Random effect diff2= 2.80272e-06
B_sum_val= 2044.006
B_change_val= 5.819443
B_change_prop= 0.2847077 %
Random effect diff2= 2.679876e-06
```

考虑 update:
(1) 迭代. 代码中到出了2个可能作为 stopping criteria 的量 B_change_val 和 diff 2.

$$B\_change\_val := sum(|\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}|) / (length(\hat{\beta}^{(t)}))$$

$$diff\ 2 := sum(|\bar{u}^{(t+1)} - \bar{u}^{(t)}|) / ((length(\bar{u}^{(t)}) \cdot \bar{y}^2)$$

$\bar{y} = mean(colMean(y))$

(2) 代码效率：尤其涉及 likelihood ratio test. 代码慢.
(3) 解出 time-wise 的 individual reference panel (4) 其他算法？

2. dispersion - estimation

对于 negative - binomial GLM:

$y \sim NB(\mu, \phi)$

① mean, variance, dispersion 之间的关系?

real life 中 variance 和 dispersion 的取值?

DESeq2:

dispersion max: 24

mean (raw count): 80

< mean (log scale): 3

variance (raw count): = 6×10^8

(with mean = 3000)

$$Y_i|\lambda_i \sim Poisson(\lambda_i)$$
$$\lambda_i \sim Gamma(\mu_i, \phi)$$

It is straightforward to show, under the hierarchical model, that

$$E[Y_i] = \mu_i, \quad var[Y_i] = \mu_i + \phi\mu_i^2$$

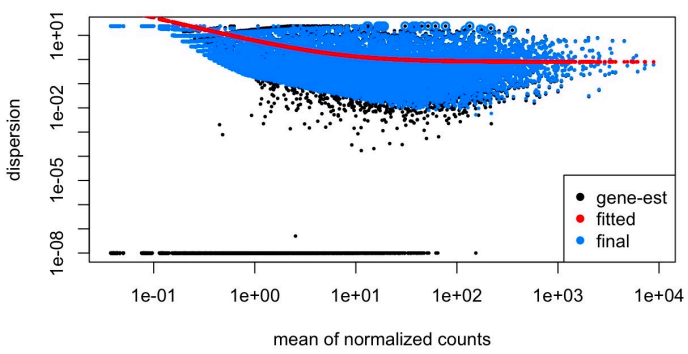
where the variance contains an overdisperion term $\phi\mu_i^2$. The larger ϕ, the greater the overdispersion. The PMF of this NB is:

$$P(y_i; \mu_i, \alpha) = \frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)\Gamma(\alpha)}(\frac{\mu_i}{\mu_i + \alpha})^{y_i}(1 - \frac{\mu_i}{\mu_i + \alpha})^{\alpha}.$$

where $\alpha = 1/\phi$ and and $\Gamma()$ is the gamma function, so that $var[Y_i] = \mu_i + \mu_i^2/\alpha$.

Negative binomial GLMs give larger standard errors than the corresponding Poisson GLMs, depending on the size of $k = 1/\psi$. On the other hand, the coefficient estimates $\hat{\beta}_j$ from a negative binomial GLM may be similar to those produced from the corresponding Poisson GLM. The negative binomial GLM gives less weight to observations with large μ_i than does the Poisson GLM, and relatively more weight to observations with small μ_i, so the coefficients

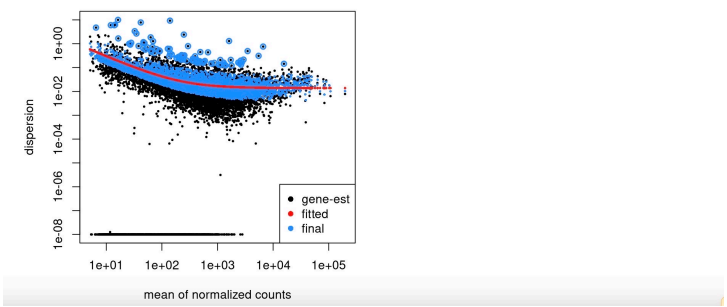
② dispersion shrinkage?



Dispersion plot and fitting alternatives

Plotting the dispersion estimates is a useful diagnostic. The dispersion plot below is typical, with the final estimates shrunk from the gene-wise estimates towards the fitted estimates. Some gene-wise estimates are flagged as outliers and not shrunk towards the fitted value, (this outlier detection is described in the manual page for *estimateDispersionsMAP*). The amount of shrinkage can be more or less than seen here, depending on the sample size, the number of coefficients, the row mean and the variability of the gene-wise estimates.

```
plotDispEsts(dds)
```



写writing sample. 11月URA续期+提交申请(帮我做)

后读之研吗?