روش تخصیص دیریکله پنهان ترکیبی (compound latent Dirichlet allocation)

در مدل LDA قبلی، تنها داده‌های یک نوع سرطان به مدل داده می‌شد که برای تجزیه و تحلیل مقایسه‌ای بین انواع دیگر سرطان‌ها دارای نقص است. مجموعه داده از یک نوع سرطان با دیگر نوع‌های سرطان دارای امضاهای جهش مشترک است، اما نسبت امضاهای جهش بین نوع‌های سرطان به صورت نسبی متفاوت است. در مدل cLDA فرض بر این است که برای ژنوم‌های هر نوع سرطان ترکیبی از امضاهای جهش پنهان مشترک با دیگر انواع سرطان وجود دارد. با تعیین این امضاهای جهش می‌توان گفت هر امضای جهش به چه میزان در هر ژنوم فردی تاثیر گذاشته است. نسبت امضاهای جهش برای مجموعه ژنوم‌های یک نوع سرطان یکسان است.

تعاریف:

واژگان (vocabulary) v: V نوع جهش در دیکشنری جهش

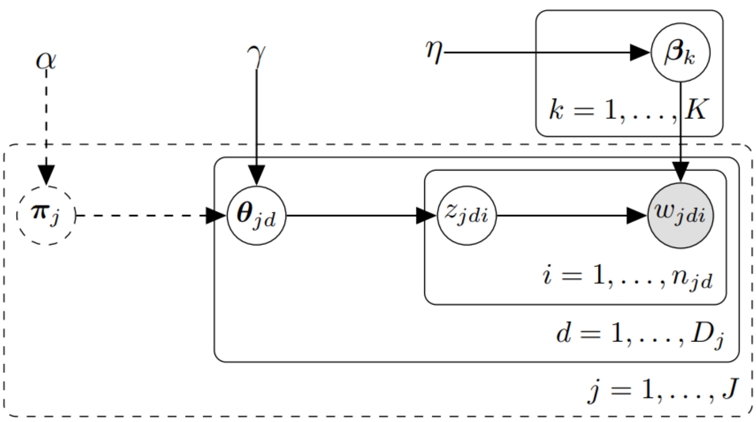
امضاهای جهش (mutational signatures) $\beta_{K\times V}$: K امضا جهش، هر کدام یک توزیع بر روی واژگان v

نوع سرطان j : دارای D_j ژنوم سرطانی

ژنوم سرطانی d با نوع سرطان j : دارای n_{jd} جهش

جهش : index واژگان متناظر با نوع جهش را دارد

$h = (\eta, \alpha, \gamma) \in (0, \infty)^3$:hyperparameters



$$
\begin{aligned}
\boldsymbol{\beta}_k &\overset{\text{iid}}{\sim} \text{Dir}_V(\eta, \ldots, \eta), \text{ for topic } k = 1, \ldots, K \\
\boldsymbol{\pi}_j &\overset{\text{iid}}{\sim} \text{Dir}_K(\alpha, \ldots, \alpha), \text{ for collection } j = 1, \ldots, J \\
\boldsymbol{\theta}_{jd} &\overset{\text{iid}}{\sim} \text{Dir}_K(\gamma\pi_{j1}, \ldots, \gamma\pi_{jK}), \text{ for document } jd \\
z_{jdi} &\overset{\text{iid}}{\sim} \text{Mult}_K(\boldsymbol{\theta}_{jd}), \text{ for each word } w_{jdi} \\
w_{jdi} &\overset{\text{ind}}{\sim} \text{Mult}_V(\boldsymbol{\beta}_{z_{jdi}})
\end{aligned}
$$

$$\beta = \begin{bmatrix} 0 \le & \le 1 \end{bmatrix}_{K \times V} \qquad \pi = \begin{bmatrix} 0 \le & \le 1 \end{bmatrix}_{J \times K} \qquad \theta = \begin{bmatrix} 0 \le & \le 1 \end{bmatrix}_{D_j \times K}$$

$$\alpha = \begin{bmatrix} \quad \end{bmatrix}_{K} \qquad \eta = \begin{bmatrix} \quad \end{bmatrix}_{V} \qquad z = \begin{bmatrix} 1 \le & \le K \end{bmatrix}_{D_j \times n_{jd}} \qquad w = \begin{bmatrix} 1 \le & \le V \end{bmatrix}_{D_j \times n_{jd}}$$

توزیع پیشین:

$$p_h(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\beta}) p_h(\boldsymbol{\theta} \mid \boldsymbol{\pi}) p_h(\boldsymbol{\pi}) p_h(\boldsymbol{\beta}),$$

$$= \prod_{j=1}^{J} \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \pi_{jk}^{\alpha-1} \right) \left[\prod_{j=1}^{J} \prod_{d=1}^{D_j} \prod_{k=1}^{K} \theta_{jdk}^{n_{jdk}} \right]$$
$$\left[\prod_{j=1}^{J} \prod_{d=1}^{D_j} \left(\frac{\Gamma(\gamma)}{\prod_{k=1}^{K} \Gamma(\gamma \pi_{jk})} \prod_{k=1}^{K} \theta_{jdk}^{\gamma \pi_{jk}-1} \right) \right] \qquad \left[\prod_{k=1}^{K} \left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V} \prod_{v=1}^{V} \beta_{kv}^{\eta-1} \right) \right].$$

$n_{jdk} = \sum_{i=1}^{n_{jd}} z_{jdik}$ تعداد جهش از ژنوم d با نوع سرطان j که به امضای جهش k اختصاص یافته است.

$n_{j.k} = \sum_{d=1}^{D_j} \sum_{i=1}^{n_{jd}} z_{jdik}$ تعداد جهش در همه ژنوم‌هایی با نوع سرطان j که به امضای جهش k تخصیص یافته است.

تابع درست‌نمایی:

$$p(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\beta}) = \prod_{j=1}^{J} \prod_{d=1}^{D_j} \prod_{i=1}^{n_{jd}} \prod_{k: z_{jdik}=1} \prod_{v=1}^{V} \beta_{kv}^{w_{jdiv}}$$
$$= \prod_{j=1}^{J} \prod_{d=1}^{D_j} \prod_{k=1}^{K} \prod_{v=1}^{V} \prod_{i \in S_{jdk}} \beta_{kv}^{w_{jdiv}}$$
$$= \prod_{j=1}^{J} \prod_{d=1}^{D_j} \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\sum_{i \in S_{jdk}} w_{jdiv}}$$
$$= \prod_{j=1}^{J} \prod_{d=1}^{D_j} \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{m_{jdkv}},$$

$S_{jdk}=\{i:1\le i\le n_{jd}\ and\ \ z_{jdik}=1\}$ شماره جهش از ژنوم d با نوع سرطان j که امضای جهش آن k است.

$m_{jdkv}=\sum_{i\in S_{jdk}} w_{jdiv}=\sum_{i=1}^{n_{jd}} z_{jdik}w_{jdiv}$ تعداد جهش‌هایی از ژنوم d با نوع سرطان j که امضای جهش آن k است و نوع جهش آن v است.

توزیع پسین:

$$p(z,\theta,\pi,\beta\mid w)=$$

$$\left[\prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{kv}^{\sum_{j=1}^{J}\sum_{d=1}^{D_j} m_{jdkv}+\eta-1}\right]$$
$$\left[\prod_{j=1}^{J}\prod_{d=1}^{D_j}\frac{\prod_{k=1}^{K}\theta_{jdk}^{n_{jdk}+\gamma\pi_{jk}-1}}{\prod_{k=1}^{K}\Gamma(\gamma\pi_{jk})}\right]\left[\prod_{j=1}^{J}\prod_{k=1}^{K}\pi_{jk}^{\alpha-1}\right]$$

استنتاج تغییراتی:

در روش‌های تغییراتی، خانواده‌ای از کران‌های پایین برای توزیع پسین غیر قابل محاسبه در نظر می‌گیریم و از بین آنها نزدیک‌ترین کران را پیدا می‌کنیم. یک روش برای در نظر گرفتن خانواده‌ای کران‌های پایین پایین برای تقریب توزیع‌ها، استفاده از توزیع‌های پارامتری است. معمولا این توزیع پارامتری با فرض استقلال بین متغیرهای پنهان ساده‌تر از توزیع پسین می‌شود. درنتیجه هدف شناسایی پارامترهایی است که کوچکترین کران پایین بدهد.

لگاریتم توزیع حاشیه‌ای:

$$\log p(w\mid\alpha,\eta,\gamma)=L(\lambda,\tau,\rho,\phi;\alpha,\eta,\gamma)+KL(q(\beta,\pi,\theta,z\mid\lambda,\tau,\rho,\phi)\,\|\,p(\beta,\pi,\theta,z\mid w,\alpha,\eta,\gamma))$$

که درآن

$$L(\lambda,\tau,\rho,\phi;\alpha,\eta,\gamma)=\int\sum_{z}q(\beta,\pi,\theta,z)\log\left\{\frac{p(\beta,\pi,\theta,z,w\mid\alpha,\eta,\gamma)}{q(\beta,\pi,\theta,z)}\right\}d\beta d\pi d\theta$$

و

$$KL(q,p) = -\int \sum_{z} q(\beta,\pi,\theta,z) \log\left\{\frac{p(\beta,\pi,\theta,z,w \mid \alpha,\eta,\gamma)}{q(\beta,\pi,\theta,z)}\right\} d\beta d\pi d\theta$$

توزیع فاکتورسازی شده تغییراتی با پارامترهای تغییراتی λ,τ,ρ,ϕ :

$$q(\beta,\pi,\theta,z \mid \lambda,\tau,\rho,\phi) = [\prod_{k=1}^{K} q(\beta_k \mid \lambda_k)][\prod_{j=1}^{J} q(\pi_j \mid \tau_j)[\prod_{d=1}^{D_j} q(\theta_{j,d} \mid \rho_{j,d})[\prod_{i=1}^{n_{j,d}} q(z_{j,d,i} \mid \phi_{j,d,i})]]]$$

توزیع مولفه‌های مستقل آن به صورت:

$$\begin{aligned}
&\beta_k \sim Dir_V(\lambda_k) \\
&\pi_j \sim Dir_K(\tau_j) \\
&\theta_{jd} \sim Dir_K(\rho_{jd}) \\
&z_{jdi} \sim Mult_K(\phi_{jdi})
\end{aligned}$$

Algorithm 3: Variational expectation maximization (VEM)

Data: Observed words $\boldsymbol{w}$ and document metadata

Result: Optimal variational parameters $(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\phi})$

initialize $(\boldsymbol{\lambda}^{(0)}, \boldsymbol{\tau}^{(0)})$;
while *not converged* **do**
    `// Step 1: Expectation`
    for *document* $d = 1, \ldots, D_j$, $j = 1, \ldots, J$ **do**
        initialize $\rho_{jdk}^{(0)} = \frac{\gamma \tau_{jk}}{\tau_{j\cdot}} + \frac{n_{jd}}{K}$, $k = 1, \ldots, K$;
        `// Variational updates for each document`
        while *not converged* **do**
            for *word* w_{jdi}, $i = 1, \ldots, n_{jd}$ **do**
                variational Multinomial update for $\boldsymbol{\phi}_{jdi}$ via (79);
            variational Dirichlet update for $\boldsymbol{\rho}_{jd}$ via (82);
    `// Variational updates for each topic`
    variational Dirichlet update for $\boldsymbol{\lambda}_k$, $k = 1, \ldots, K$ via (85);
    `// Step 2: Maximization`
    `// Constraint Newton updates for collection-level topic mixtures`
    for *collection* $j = 1, \ldots, J$ **do**
        initialize $(a_j^{(0)}, \boldsymbol{\omega}_j^{(0)})$ based on the current $\boldsymbol{\tau}_j$;
        while *not converged* **do**
            constraint Newton update for $\boldsymbol{\omega}_j$ via (72);
            Newton update for a_j via (76);
        set $\boldsymbol{\tau}_j = a_j^{(\text{final})} * \boldsymbol{\omega}_j^{(\text{final})}$
    optimize hyperparameter $h = (\alpha, \gamma, \eta)$;

روش تخمین پارامترهای تغییراتی λ,τ,ρ,ϕ با بیشینه کردن کران پایین $L(\lambda,\tau,\rho,\phi;\alpha,\eta,\gamma)$ که با کمینه کردن معیار واگرایی کولبک-لیبلر بین توزیع تغییراتی q و توزیع پسین p معادل است، صورت می‌گیرد.

کران پایین $L(\lambda,\tau,\rho,\phi;\alpha,\eta,\gamma)$:

$$\begin{aligned}
\mathcal{L}(q,p_h) = &\sum_{k=1}^{K}\mathbb{E}_{q_k}[\log p_\eta(\boldsymbol{\beta}_k)] + \sum_{j=1}^{J}\mathbb{E}_{q_j}[\log p_\alpha(\boldsymbol{\pi})] + \sum_{j=1}^{J}\sum_{d=1}^{D_j}\mathbb{E}_{q_{jd}}[\log p_\gamma(\boldsymbol{\theta}_{jd}\,|\,\boldsymbol{\pi}_j)] \\
&+ \sum_{j=1}^{J}\sum_{d=1}^{D_j}\sum_{i=1}^{n_{dj}}\mathbb{E}_{q_{jdi}}[\log p(z_{jdi}\,|\,\boldsymbol{\theta}_{jd})] + \sum_{j=1}^{J}\sum_{d=1}^{D_j}\sum_{i=1}^{n_{dj}}\mathbb{E}_{q_{jdi}}[\log p(w_{jdi}\,|\,z_{jdi},\boldsymbol{\beta})] \\
&- \sum_{k=1}^{K}\mathbb{E}_{q_k}[\log q(\boldsymbol{\beta}_k\,|\,\boldsymbol{\lambda}_k)] - \sum_{j=1}^{J}\mathbb{E}_{q_j}[\log q(\boldsymbol{\pi}_j\,|\,\boldsymbol{\tau}_j)] \\
&- \sum_{j=1}^{J}\sum_{d=1}^{D_j}\mathbb{E}_{q_{jd}}[\log q(\boldsymbol{\theta}_{jd}\,|\,\boldsymbol{\rho}_{jd})] - \sum_{j=1}^{J}\sum_{d=1}^{D_j}\sum_{i=1}^{n_{dj}}\mathbb{E}_{q_{jdi}}[\log q(\boldsymbol{z}_{jdi}\,|\,\boldsymbol{\phi}_{jdi})]
\end{aligned} \tag{63}$$

$$\rho_{jd.} = \sum_{k=1}^{K}\rho_{jdk} \;،\; \tau_{j.} = \sum_{k=1}^{K}\tau_{jk} \;،\; \lambda_{k.} = \sum_{v=1}^{V}\lambda_{kv}$$

میانگین مولفه‌های احتمالی به صورت زیر است.

$$\begin{aligned}
\beta_k &\sim Dir_V(\lambda_k) \\
\pi_j &\sim Dir_K(\tau_j) \\
\theta_{jd} &\sim Dir_K(\rho_{jd})
\end{aligned}
\quad\Longrightarrow\quad
\begin{aligned}
E_q[\beta_{kv}] &= \lambda_{kv} / \lambda_{k.} \\
E_q[\pi_{jk}] &= \tau_{jk} / \tau_{j.} \\
E_q[\theta_{jdk}] &= \rho_{jdk} / \rho_{jd.}
\end{aligned}$$

میانگین لگاریتم مولفه‌های احتمالی به صورت زیر است.

$$\begin{aligned}
\beta_k &\sim Dir_V(\lambda_k) \\
\pi_j &\sim Dir_K(\tau_j) \\
\theta_{jd} &\sim Dir_K(\rho_{jd})
\end{aligned}
\quad\Longrightarrow\quad
\begin{aligned}
\mathbb{E}_q[\log\beta_{kv}\,|\,\lambda_{kv}] &= \Psi(\lambda_{kv}) - \Psi(\lambda_{k.}) \\
\mathbb{E}_q[\log\pi_{jk}\,|\,\tau_{jk}] &= \Psi(\tau_{jk}) - \Psi(\tau_{j.}) \\
\mathbb{E}_q[\log\theta_{jdk}\,|\,\rho_{jdk}] &= \Psi(\rho_{jdk}) - \Psi(\rho_{jd.})
\end{aligned}$$

اگر $\theta \sim Dir_K(\rho)$، $E[\theta_k] = \rho_k / \rho_.$ و $\alpha \in [0,\infty)$ در نظر بگیریم، عبارت غیر قابل محاسبه $E[\log\Gamma(\alpha\theta_k)]$ با توجه به مقاله (Kim et al., 2013, Theorem 3.1) به صورت زیر قابل تخمین است:

$$\mathbb{E}[\log\Gamma(\alpha\theta_k)] \le \log\Gamma(\alpha\mathbb{E}[\theta_k]) + \frac{\alpha}{\rho_.}(1-\mathbb{E}[\theta_k]) + (1-\alpha\mathbb{E}[\theta_k])\left[\log\mathbb{E}[\theta_k] + \Psi(\rho_.) - \Psi(\rho_k)\right] \tag{65}$$

در نهایت هر میانگین از عبارت کران پایین به صورت زیر محاسبه میشود:

$$
\begin{aligned}
\mathbb{E}_{q_k}[\log p_\eta(\boldsymbol{\beta}_k)] &= \log\Gamma(V\eta) - V\log\Gamma(\eta) + \sum_{v=1}^{V}(\eta-1)\big[\Psi(\lambda_{kv}) - \Psi(\lambda_{k.})\big] \\
\mathbb{E}_{q_j}[\log p_\alpha(\boldsymbol{\pi})] &= \log\Gamma(K\alpha) - K\log\Gamma(\alpha) + \sum_{k=1}^{K}(\alpha-1)\big[\Psi(\tau_{jk}) - \Psi(\tau_{j.})\big] \\
\mathbb{E}_{q_{jd}}[\log p_\gamma(\boldsymbol{\theta}_{jd}\,|\,\boldsymbol{\pi}_j)] &= \mathbb{E}_q[\log\Gamma(\gamma)] - \sum_{k-1}^{K}\mathbb{E}_q[\log\Gamma(\gamma\pi_{jk})] + \sum_{k=1}^{K}\mathbb{E}_q[(\gamma\pi_{jk}-1)\log\theta_{jdk}] \\
&\geq \log\Gamma(\gamma) - \sum_{k=1}^{K}\Big[\log\Gamma(\gamma\mathbb{E}_q[\pi_{jk}]) + \frac{\gamma}{\tau_{j.}}(1-\mathbb{E}_q[\pi_{jk}]) \\
&\qquad\qquad + (1-\gamma\mathbb{E}_q[\pi_{jk}])\big[\log\mathbb{E}_q[\pi_{jk}] + \Psi(\tau_{j.}) - \Psi(\tau_{jk})\big]\Big] \\
&\qquad + \sum_{k=1}^{K}\Big[\gamma\mathbb{E}_q[\pi_{jk}]\mathbb{E}_q[\log\theta_{jdk}] - \mathbb{E}_q[\log\theta_{jdk}]\Big] \\
&\geq \log\Gamma(\gamma) - \frac{\gamma}{\tau_{j.}}(K-1) - (\gamma-K)\Big[\log\tau_{j.} - \Psi(\tau_{j.}) + \Psi(\rho_{jd.})\Big] \\
&\qquad - \sum_{k=1}^{K}\Big[\log\Gamma(\frac{\gamma\tau_{jk}}{\tau_{j.}}) + (1-\frac{\gamma\tau_{jk}}{\tau_{j.}})\big[\log(\tau_{jk}) - \Psi(\tau_{jk}) + \Psi(\rho_{jdk})\big]\Big] \\
\mathbb{E}_{q_{jdi}}[\log p(z_{jdi}\,|\,\boldsymbol{\theta}_{jd})] &= \sum_{k=1}^{K}\phi_{jdik}\big[\Psi(\rho_{jdk}) - \Psi(\rho_{jd.})\big] \\
\mathbb{E}_{q_{jdi}}[\log p(w_{jdi}\,|\,z_{jdi},\boldsymbol{\beta})] &= \sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{jdik}w_{jdiv}\big[\Psi(\lambda_{kv}) - \Psi(\lambda_{k.})\big] \\
\mathbb{E}_{q_k}[\log q(\boldsymbol{\beta}_k\,|\,\boldsymbol{\lambda}_k)] &= \log\Gamma(\lambda_{k.}) - \sum_{v=1}^{V}\log\Gamma(\lambda_{kv}) + \sum_{v=1}^{V}(\lambda_{kv}-1)\big[\Psi(\lambda_{kv}) - \Psi(\lambda_{k.})\big] \\
\mathbb{E}_{q_j}[\log q(\boldsymbol{\pi}_j\,|\,\boldsymbol{\tau}_j)] &= \log\Gamma(\tau_{j.}) - \sum_{k=1}^{K}\log\Gamma(\tau_{jk}) + \sum_{k=1}^{K}(\tau_{jk}-1)\big[\Psi(\tau_{jk}) - \Psi(\tau_{j.})\big] \\
\mathbb{E}_{q_{jd}}[\log q(\boldsymbol{\theta}_{jd}\,|\,\boldsymbol{\rho}_{jd})] &= \log\Gamma(\rho_{jd.}) - \sum_{k=1}^{K}\log\Gamma(\rho_{jdk}) + \sum_{k=1}^{K}(\rho_{jdk}-1)\big[\Psi(\rho_{jdk}) - \Psi(\rho_{jd.})\big] \\
\mathbb{E}_{q_{jdi}}[\log q(\boldsymbol{z}_{jdi}\,|\,\boldsymbol{\phi}_{jdi})] &= \sum_{k=1}^{K}\phi_{jdik}\log\phi_{jdik}
\end{aligned}
\tag{66}
$$

به روز رسانی دیریکله تغییراتی برای انواع سرطان

عبارات شامل τ_j را از کران پایین L جدا می‌کنیم:

$$\mathcal{L}_{[\tau_j]} = \mathbb{E}_{q_j}[\log p_\alpha(\boldsymbol{\pi})] - \mathbb{E}_{q_j}[\log q(\boldsymbol{\pi}_j \mid \boldsymbol{\tau}_j)] + \sum_{d=1}^{D_j} \mathbb{E}_{q_{jd}}[\log p_\gamma(\boldsymbol{\theta}_{jd} \mid \boldsymbol{\pi}_j)]$$

این کران پایین یک فرم بسته برای به روز رسانی τ_j ایجاد نمی‌کند. Kim et al. (2013) پیشنهاد می‌دهد که از به روز رسانی نیوتن با شرط‌های یکسان استفاده شود. برای اینکار لازم است τ_j را به یک پارامتر مقیاس a_j و یک اندازه‌گیری پایه (base measure) w_j با شرط اینکه $\sum_{k=1}^{K} w_{jk} = 1$ تجزیه کنیم. در نتیجه توزیع تغییراتی متناظر برای π_j به صورت $\pi_j = Dir_K(a_j w_j)$ تعریف می‌شود. این تجزیه این امکان را می‌دهد تا از به روز رسانی نیوتن استفاده شود. کران پایین شامل عبارات τ_j با a_j و w_j به صورت زیر باتعریف می‌شود:

$$\begin{aligned}\mathcal{L}_{[a_j\boldsymbol{\omega}_j]} = &\sum_{k=1}^{K}\Big[(\alpha - a_j\omega_{jk})\big[\Psi(a_j\omega_{jk}) - \Psi(a_j)\big] + \log\Gamma(a_j\omega_{jk})\Big] - \log\Gamma(a_j)\\ &- \sum_{d=1}^{D_j}\Big[\frac{\gamma}{a_j}(K-1) + (\gamma - K)\big[\log a_j - \Psi(a_j) + \Psi(\rho_{jd.})\big]\Big]\\ &- \sum_{d=1}^{D_j}\sum_{k=1}^{K}\Big[\log\Gamma(\gamma\omega_{jk}) + (1-\gamma\omega_{jk})\big[\log(a_j\omega_{jk}) - \Psi(a_j\omega_{jk}) + \Psi(\rho_{jdk})\big]\Big]\end{aligned} \tag{67}$$

برای بیشینه کردن $L_{[a_j w_j]}$ براساس w_{jk}، ابتدا عبارات شامل w_{jk} را جدا می‌کنیم:

$$\begin{aligned}\mathcal{L}_{[\omega_{jk}]} = &\Psi(a_j\omega_{jk})\Big[\alpha + D_j - a_j\omega_{jk} - \gamma D_j\omega_{jk}\Big] + \gamma\omega_{jk}\sum_{d=1}^{D_j}\Psi(\rho_{jdk})\\ &+ \log\Gamma(a_j\omega_{jk}) - D_j\Big[\log\Gamma(\gamma\omega_{jk}) + (1-\gamma\omega_{jk})\log(a_j\omega_{jk})\Big]\end{aligned} \tag{68}$$

مشتق اول g_{jk} و مشتق دوم h_{jk} آن:

$$
\begin{aligned}
\frac{\partial}{\partial \omega_{jk}} \mathcal{L}_{[\omega_{jk}]} &= a_j \Psi'(a_j\omega_{jk})\Big[\alpha + D_j - a_j\omega_{jk} - \gamma D_j \omega_{jk}\Big] - \gamma D_j \Psi(a_j\omega_{jk}) + \gamma \sum_{d=1}^{D_j} \Psi(\rho_{jdk}) \\
&\quad - D_j\Big[\gamma \Psi(\gamma\omega_{jk}) + \frac{1}{\omega_{jk}} - \gamma - \gamma \log(a_j\omega_{jk})\Big] \qquad (69) \\
\frac{\partial^2}{\partial \omega_{jk}} \mathcal{L}_{[\omega_{jk}]} &= a_j^2 \Psi''(a_j\omega_{jk})\Big[\alpha + D_j - a_j\omega_{jk} - \gamma D_j \omega_{jk}\Big] - a_j \Psi'(a_j\omega_{jk})\Big[a_j + 2\gamma D_j\Big] \\
&\quad - D_j\Big[\gamma^2 \Psi'(\gamma\omega_{jk}) - \frac{1}{\omega_{jk}^2} - \frac{\gamma}{\omega_{jk}}\Big] \qquad (70)
\end{aligned}
$$

ماتریس هسیان داده شده در (۷۰) قطری است. برای به دست آوردن مقدار به روز رسانی Δw_{jk} در هر مرحله نیوتن مجموعه معادلات خطی زیر را حل می‌کنیم:

$$
\begin{bmatrix} \operatorname{diag}(\boldsymbol{h}) & \mathbf{1} \\ \mathbf{1}^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} \Delta\omega_{jk} \\ u \end{bmatrix} = \begin{bmatrix} -\boldsymbol{g} \\ 0 \end{bmatrix}, \qquad (71)
$$

در نتیجه:

$$
\Delta\omega_{jk} = \left\{ \frac{\sum_{k=1}^{K} \frac{g_{jk}}{h_{jk}}}{\sum_{k=1}^{K} \frac{1}{h_{jk}}} \right\} \begin{bmatrix} \frac{1}{h_{j1}} \\ \cdots \\ \frac{1}{h_{jK}} \end{bmatrix} - \begin{bmatrix} \frac{g_{j1}}{h_{j1}} \\ \cdots \\ \frac{g_{jK}}{h_{jK}} \end{bmatrix} \qquad (72)
$$

شرط $\sum_{k=1}^{K} \Delta w_{jk} = 0$ همچنان برقرار است.

برای بیشینه کردن $L_{[a_j w_j]}$ براساس a_j، ابتدا عبارات شامل a_j را جدا می‌کنیم:

$$
\begin{aligned}
\mathcal{L}_{[a_j]} = \sum_{k=1}^{K} [\Psi(a_j\omega_{jk}) - \Psi(a_j)]\Big(\alpha + D_j - a_j\omega_{jk} - \gamma D_j\omega_{jk}\Big) + \sum_{k=1}^{K} \log\Gamma(a_j\omega_{jk}) \\
- \log\Gamma(a_j) - \frac{\gamma D_j (K-1)}{a_j}
\end{aligned} \qquad (73)
$$

مشتق اول g_j و مشتق دوم h_j آن:

$$
\begin{aligned}
\frac{\partial}{\partial a_j}\mathcal{L}_{[a_j]} &= \sum_{k=1}^{K}\left[\omega_{jk}\Psi'(a_j\omega_{jk})-\Psi'(a_j)\right]\Big(\alpha+D_j-a_j\omega_{jk}-\gamma D_j\omega_{jk}\Big) \\
&\quad +(K-1)\gamma D_j a_j^{-2} \qquad (74)\\
\frac{\partial^2}{\partial a_j}\mathcal{L}_{[a_j]} &= \sum_{k=1}^{K}\left[\omega_{jk}^2\Psi''(a_j\omega_{jk})-\Psi''(a_j)\right]\Big(\alpha+D_j-a_j\omega_{jk}-\gamma D_j\omega_{jk}\Big) \\
&\quad -\sum_{k=1}^{K}\omega_{jk}\left[\omega_{jk}\Psi'(a_j\omega_{jk})-\Psi'(a_j)\right]-2(K-1)\gamma D_j a_j^{-3} \qquad (75)
\end{aligned}
$$

مقدار به روز رسانی a_j در هر مرحله نیوتن:

$$
\Delta a_j = -h_j^{-1} g_j. \qquad (76)
$$

به روز رسانی دیریکله تغییراتی برای جهش‌ها

برای به روز رسانی پارامتر تغییراتی ϕ_{jdik} (احتمال اینکه کلمه jdi ام توسط امضای جهش k ایجاد شده باشد) با بیشینه کردن کران پایین L با شرط اینکه $\sum_{k=1}^{K}\phi_{jdik}=1$ ، تمام عبارات شامل ϕ_{jdik} را از L جدا می‌کنیم:

$$
\begin{aligned}
\mathcal{L}_{[\phi_{jdik}]} &= \phi_{jdik}\Big[[\Psi(\rho_{jdk})-\Psi(\rho_{jd.})]+[\Psi(\lambda_{kv})-\Psi(\lambda_{k.})]-\log\phi_{jdik}\Big] \\
&\quad +\mu_{jdi}\Big[\sum_{k=1}^{K}\phi_{jdik}-1\Big] \qquad (77)
\end{aligned}
$$

از عبارت بالا نسبت به ϕ_{jdik} مشتق می‌گیریم:

$$
\frac{\partial}{\partial\phi_{jdik}}\mathcal{L}_{[\phi_{jdik}]} = [\Psi(\rho_{jdk})-\Psi(\rho_{jd.})]+[\Psi(\lambda_{kv})-\Psi(\lambda_{k.})]-\log\phi_{jdik}-1+\mu_{jdi} \qquad (78)
$$

مشتق بالا را مساوی صفر قرار داده و مقدار بیشینه ϕ_{jdik} را بدست می‌آوریم:

$$
\phi_{jdik} \propto \exp\Big([\Psi(\rho_{jdk})-\Psi(\rho_{jd.})]+[\Psi(\lambda_{kv})-\Psi(\lambda_{k.})]\Big) \qquad (79)
$$

به روز رسانی دیریکله تغییراتی برای ژنوم‌ها

کران پایین L براساس ρ_{jdk} بیشینه می‌کنیم. عبارات شامل ρ_{jdk} در L:

$$
\begin{aligned}
\mathcal{L}_{[\rho_{jd}]} &= \sum_{k=1}^{K}\Big(\frac{\gamma\tau_{jk}}{\tau_{j.}} + \sum_{i=1}^{n_{jd}} \phi_{jdik} - \rho_{jdk}\Big)\left[\Psi(\rho_{jdk}) - \Psi(\rho_{jd.})\right] \\
&\quad - \log\Gamma(\rho_{jd.}) + \sum_{k=1}^{K}\log\Gamma(\rho_{jdk})
\end{aligned}
\tag{80}
$$

از عبارت بالا نسبت به ρ_{jdk} مشتق می‌گیریم:

$$
\frac{\partial}{\partial\rho_{jdk}}\mathcal{L}_{[\rho_{jd}]} = \left[\Psi'(\rho_{jdk}) - \Psi'(\rho_{jd.})\right]\Big(\frac{\gamma\tau_{jk}}{\tau_{j.}} + \sum_{i=1}^{n_{jd}}\phi_{jdik} - \rho_{jdk}\Big) \tag{81}
$$

مشتق بالا را مساوی صفر قرار داده و مقدار بیشینه ρ_{jdk} را بدست می‌آوریم:

$$
\rho_{jdk} = \frac{\gamma\tau_{jk}}{\tau_{j.}} + \sum_{i=1}^{n_{jd}}\phi_{jdik} \tag{82}
$$

به روز رسانی دیریکله تغییراتی برای امضاهای جهش

کران پایین L براساس λ_{kv} بیشینه می‌کنیم. عبارات شامل λ_{kv} در L:

$$
\begin{aligned}
\mathcal{L}_{[\boldsymbol{\lambda}_k]} &= \sum_{k=1}^{K}\sum_{v=1}^{V}\Big(\eta - \lambda_{kv} + \sum_{j=1}^{J}\sum_{d=1}^{D_j}\sum_{i=1}^{n_{jd}}\phi_{jdik}w_{jdiv}\Big)\left[\Psi(\lambda_{kv}) - \Psi(\lambda_{k.})\right] \\
&\quad - \sum_{k=1}^{K}\Big[\log\Gamma(\lambda_{k.}) - \sum_{v=1}^{V}\log\Gamma(\lambda_{kv})\Big]
\end{aligned}
\tag{83}
$$

از عبارت بالا نسبت به λ_{kv} مشتق می‌گیریم:

$$
\frac{\partial}{\partial\lambda_{kv}}\mathcal{L}_{[\boldsymbol{\lambda}_k]} = \left[\Psi'(\lambda_{kv}) - \Psi'(\lambda_{k.})\right]\Big(\eta - \lambda_{kv} + \sum_{j=1}^{J}\sum_{d=1}^{D_j}\sum_{i=1}^{n_{jd}}\phi_{jdik}\Big) \tag{84}
$$

مشتق بالا را مساوی صفر قرار داده و مقدار بیشینه λ_{kv} را بدست می‌آوریم:

$$\lambda_{kv} = \eta + \sum_{j=1}^{J} \sum_{d=1}^{D_j} \sum_{i=1}^{n_{jd}} \phi_{jdik} \tag{85}$$

بهینه‌سازی هایپرپارامترها η, α, γ

مانند الگوریتم تغییراتی EM برای LDA (Blei et al., 2003) ، E-step در cLDA پارامترهای تغییراتی براساس عبارت گفته شده در بالا به روزرسانی می‌شوند. از کران پایین بهینه شده $L(q^*, p)$ به عنوان تقریب قابل محاسبه برای لگاریتم درست‌نمایی حاشیه‌ای $\log p(w)$ می‌توانیم استفاده کنیم. در M-step می‌توانیم هایپر پارامترهای η, α, γ را با بیشینه کردن کران پایین بهینه شده براساس η, α, γ به روز رسانی کنیم.

عبارات شامل هر هایپرپارامتر را از کران پایین بهینه شده جدا می‌کنیم:

$$\mathcal{L}_{[\alpha]} = J \log \Gamma(K\alpha) - JK \log \Gamma(\alpha) + \sum_{j=1}^{J} \sum_{k=1}^{K} \alpha \big[\Psi(\tau_{jk}) - \Psi(\tau_{j.})\big] \tag{86}$$

$$\mathcal{L}_{[\eta]} = K \log \Gamma(V\eta) - KV \log \Gamma(\eta) + \sum_{k=1}^{K} \sum_{v=1}^{V} \eta \big[\Psi(\lambda_{kv}) - \Psi(\lambda_{k.})\big] \tag{87}$$

$$\begin{aligned}\mathcal{L}_{[\gamma]} &= \sum_{j=1}^{J} D_j [\log \Gamma(\gamma) - \frac{\gamma}{\tau_{j.}}(K-1)] - \gamma \sum_{j=1}^{J} \sum_{d=1}^{D_j} \Big[\log \tau_{j.} - \Psi(\tau_{j.}) + \Psi(\rho_{jd.})\Big] \\ &\quad - \sum_{j=1}^{J} \sum_{d=1}^{D_j} \sum_{k=1}^{K} \Big[\log \Gamma(\frac{\gamma \tau_{jk}}{\tau_{j.}}) - \frac{\gamma \tau_{jk}}{\tau_{j.}} \big[\log(\tau_{jk}) - \Psi(\tau_{jk}) + \Psi(\rho_{jdk})\big]\Big]\end{aligned} \tag{88}$$

مشتق اول و دوم هر کدام را محاسبه می‌کنیم:

$$\frac{\partial}{\partial \alpha}\mathcal{L}_{[\alpha]} = JK\big[\Psi(K\alpha)-\Psi(\alpha)\big]+\sum_{j=1}^{J}\sum_{k=1}^{K}\big[\Psi(\tau_{jk})-\Psi(\tau_{j.})\big] \quad (89)$$

$$\frac{\partial^2}{\partial \alpha}\mathcal{L}_{[\alpha]} = JK^2\Psi'(K\alpha)-JK\Psi'(\alpha) \quad (90)$$

$$\frac{\partial}{\partial \eta}\mathcal{L}_{[\eta]} = KV\big[\Psi(V\eta)-\Psi(\eta)\big]+\sum_{k=1}^{K}\sum_{v=1}^{V}\big[\Psi(\lambda_{kv})-\Psi(\lambda_{k.})\big] \quad (91)$$

$$\frac{\partial^2}{\partial \eta}\mathcal{L}_{[\eta]} = KV^2\Psi'(V\eta)-KV\Psi'(\eta) \quad (92)$$

$$\begin{aligned}\frac{\partial}{\partial \gamma}\mathcal{L}_{[\gamma]} &= \sum_{j=1}^{J}D_j[\Psi(\gamma)-\frac{1}{\tau_{j.}}(K-1)]-\sum_{j=1}^{J}\sum_{d=1}^{D_j}\Big[\log\tau_{j.}-\Psi(\tau_{j.})+\Psi(\rho_{jd.})\Big]\\ &\quad -\sum_{j=1}^{J}\sum_{d=1}^{D_j}\sum_{k=1}^{K}\frac{\tau_{jk}}{\tau_{j.}}\Big[\Psi(\frac{\gamma\tau_{jk}}{\tau_{j.}})-\big[\log(\tau_{jk})-\Psi(\tau_{jk})+\Psi(\rho_{jdk})\big]\Big] \quad (93)\end{aligned}$$

$$\frac{\partial^2}{\partial \gamma}\mathcal{L}_{[\gamma]} = \sum_{j=1}^{J}D_j\Psi'(\gamma)-\sum_{j=1}^{J}\sum_{d=1}^{D_j}\sum_{k=1}^{K}\frac{\tau_{jk}^2}{\tau_{j.}^2}\Psi'(\frac{\gamma\tau_{jk}}{\tau_{j.}}) \quad (94)$$

با استفاده از مشتق‌های بدست آمده می‌توانیم η, α, γ با روش نیوتن به روز رسانی کنیم (Blei et al., 2003; Minka, 2000):

روش‌های Newton-Raphson برای یک هسیان با ساختار خاص (Blei et al., 2003):

این روش برای برای تخمین حداکثر درست‌نمایی توزیع دیریکله مورد استفاده قرار می‌گیرد. این تکنیک یک نقطه ثابت از یک تابع را با تکرار عبارت زیر پیدا می‌کند:

$$\alpha_{new} = \alpha_{old} - H(\alpha_{old})^{-1}g(\alpha_{old})$$

که در آن $g(\alpha)$ و $H(\alpha)$ به ترتیب گرادیان و ماتریس هسیان در نقطه α است. این الگوریتم به دلیل محاسبه ماتریس معکوس به زمان $O(N^3)$ نیاز دارد.

اگر ماتریس هسیان فرم زیر را داشته باشد:

$$H = \mathrm{diag}(h) + \mathbf{1}z\mathbf{1}^{\mathrm{T}},$$

ماتریس معکوس به صورت زیر قابل محاسبه است:

$$H^{-1} = \mathrm{diag}(h)^{-1} - \frac{\mathrm{diag}(h)^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathrm{diag}(h)^{-1}}{z^{-1} + \sum_{j=1}^{k} h_j^{-1}}$$

با ضرب گرادیان در آن:

$$(H^{-1}g)_i = \frac{g_i - c}{h_i} \qquad c = \frac{\sum_{j=1}^{k} g_j / h_j}{z^{-1} + \sum_{j=1}^{k} h_j^{-1}}.$$

این عبارت تنها به 2k مقدار g_i و h_i بستگی دارد و در نتیجه پیچیدگی زمانی خطی دارد.

نکات پیاده‌سازی:

مقدار اولیه λ و τ در مقاله (George et al., 2019) به صورت زیر قرار گرفته است:

$$\lambda_{kv} = uniform(0,1) + 1/V$$

$$\tau_{jk} = \alpha + n_j / K$$

α به صورت ورودی گرفته می‌شود. n_j تعداد جهش‌های همه ژنوم‌هایی که نوع سرطان j را دارند. مقدار τ_{jk} نرمال سازی می‌شود تا جمع برابر ۱ شود.

مقادیر اولیه در مقاله (Matsutani et al., 2019) به صورت rnd- قرار گرفته است و نرمال‌سازی می‌شوند.

من در پیاده سازی مقادیر اولیه را به صورت rnd- دادم و نرمال‌سازی کردم.

مقدار اولیه a_j و w_{jk} به صورت زیر قرار گرفته است:

$$a_j = \tau_{j.}$$

$$w_{jk} = \tau_{jk} / \tau_{j.}$$

برای چک کردن همگرایی کل الگوریتم کل مولفه‌های کران پایین L (66) را در نظر می‌گیریم و اگر مجموع آنها از یک threshold کمتر شود، تکرار را متوقف می‌کنیم.

برای چک کردن همگرایی ρ_{jdk} تمام مولفه‌های کران پایین مرتبط با ژنوم‌ها شامل مولفه ۱،۶، ۲،۷ در (66) را در نظر می‌گیریم و اگر مجموع آنها از یک threshold کمتر شود، تکرار را متوقف می‌کنیم.

برای چک کردن همگرایی τ_{jk} تمام مولفه‌های کران پایین مرتبط با انواع سرطان شامل مولفه ۳، ۲،۷ در (66) را در نظر می‌گیریم و $a_{jk}w_j$ را با τ_{jk} جابه‌جا می‌کنیم و اگر مجموع آنها از یک threshold کمتر شود، تکرار را متوقف می‌کنیم.

برای به روز رسانی هایپرپارامترها مقاله (George et al., 2019) از عبارات زیر استفاده می‌کند:

$$\alpha^{new} = \exp(\log(\alpha^{old}) - \frac{f'(\alpha)}{f''(\alpha)\times\alpha + f'(\alpha)})$$

$$\eta^{new} = \exp(\log(\eta^{old}) - \frac{f'(\eta)}{f''(\eta)\times\eta + f'(\eta)})$$

و مقدار γ به روزرسانی نشده است.

برای به روز رسانی هایپرپارامترها مقاله (Matsutani et al., 2019) از عبارات زیر استفاده می‌کند:

$$\alpha_k = \frac{\sum_{s=1}^{S}[\Psi(E[n_{s,k}+\hat{\alpha}_k]) - \Psi(\hat{\alpha}_k)]\hat{\alpha}_k}{\sum_{s=1}^{S}\left[\Psi(n_s + \sum_{k=1}^{K}\hat{\alpha}_k) - \Psi(\sum_{k=1}^{K}\hat{\alpha}_k)\right]}$$

$$\beta_v = \frac{\sum_{k=1}^{K}[\Psi(E[n_{k,v}+\hat{\beta}_v]) - \Psi(\hat{\beta}_v)]\hat{\beta}_v}{\sum_{k=1}^{K}\left[\Psi(\sum_{v=1}^{V}E[n_{k,v}] + \hat{\beta}_v) - \Psi(\sum_{v=1}^{V}\hat{\beta}_v)\right]}$$

که به این عبارات در مقاله (Minka, 2000) در قسمت توزیع چندجمله‌ای- دیریکله اشاره شده است.

من برای به روز رسانی η, α, γ از روش Newton-Raphson که در بالا گفته شد استفاده کردم. این روش در مقاله (Blei et al., 2003) و (Minka, 2000) در قسمت توزیع دیریکله اشاره شده است.

مراجع:

1. Clint P. George , Wei Xia , and George Michailidis. Analyses of Multi-collection Corpora via Compound Topic Modeling, Machine Learning, Optimization, and Data Science. LOD 2019. Lecture Notes in Computer Science, vol 11943. Springer, Cham, 2020
(https://arxiv.org/pdf/1907.01636.pdf)
(https://github.com/clintpgeorge/clda)

2. Taro Matsutani, Yuki Ueno, Tsukasa Fukunaga, and Michiaki Hamada. Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference, Bioinformatics, 35(22), 4543–4552, 2019
3. Thomas Minka. Estimating a dirichlet distribution, 2000a.
4. Thomas P. Minka. Beyond Newton's method. Technical report, Microsoft, 2000b
5. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.