转录组学混合样本解卷积：回归分析方法比较与前沿进展

张逸凡

1 问题背景

在生物学实验与临床医学研究中，批量 RNA 测序（Bulk RNA-seq）是获取组织或器官基因表达水平的常规方法。然而，该方法仅能测定细胞群体的平均表达量，无法实现单细胞分辨率[1]。这一局限性在临床应用中尤为显著：若需解析样本中不同细胞类型的比例及其特异性基因表达特征（例如用于空间域检测[2]），则需借助更精细的计算方法。

假设某样本的基因表达量（$Gene_{exp}$）可表示为 k 种细胞类型表达量的线性组合：

$$Gene_{exp} = \sum_{i=1}^{k} c_i Gene_{exp}^{i} \tag{1}$$

其中，c_i 为第 i 类细胞的占比。该模型与信号处理中的卷积运算形式相似[3]。已知 $Gene_{exp}$ 时，通过求解此方程反演细胞占比（c_i）以及单细胞基因表达量（$Gene_{exp}^{i}$）的过程称为解卷积（deconvolution）；若假定我们已知单细胞基因表达量（$Gene_{exp}^{i}$），通过求解上式反演细胞占比（c_i）的过程称为部分解卷积（partial-deconvolustion）。本文主要探讨后者。由于式 (1) 本质上是一个线性模型，线性回归分析方法自然成为转录组混合样本解卷积的核心工具。

本综述将系统回顾回归分析在该领域的应用，对比不同解卷积方法的优劣，并复现部分前沿技术的计算结果，以评估其实际性能。

2 数学建模

设 n 为（我们所能测量到或所关心的）基因数，p 为样本数，q 为细胞类型数。$M \in R^{n\times p}$ 为混合矩阵（Mixture Matrix），可以分块为 $(m_1, \cdots, m_p)$，其第 i, j 元素表示第 i 个基因在第 j 个样本上的平均表达量。

$H \in R^{n\times r}$ 表示原始细胞类型基因表达的参考特征矩阵（reference expression matrix）。其 n 行对应 n 个不同基因，而 r 列可以分组为 q 组：$H = (h_1, \cdots, h_q)$，h_i 有 r_i 列，代表对第 i 个细胞类型进行 r_i 次独立的单细胞 RNA 测序（single cell RNA sequencing, scRNA-seq[4]），$\sum_{i=1}^{q} r_i = r$。$G \in R^{n\times q}$ 为参考矩阵（注意区分其与 H 的区别），其第 i, j 元素由将 h_j 的第 i 行元素进行平均得到，表示对细胞类型 j 进行多次单细胞 RNA 测序后基因 i 的平均表达量。$C \in R^{q\times p}$ 为比例矩阵（proportion matrix），可分块为 $C = (c_i, \cdots, c_p)$，其第 j 列表示第 j 个样本 q 个细胞的占比。

若我们承认 (1) 式，那么我们就有 $M \approx GC$。于是，解卷积可以正式定义为寻找 G, C 的估计 $\hat{G}, \hat{C}$，使得

$$argmin_{\hat{G},\hat{C}\geq 0}\delta(\hat{G}\hat{C}, M)$$

这里 $\delta(\cdot,\cdot)$ 是一个损失函数。

若我们仅已知 M，对 G, C 做估计，那么所考虑的问题为解卷积。若我们已知 M 以及 G 和 C 中的一个，并对另外一个做估计，那么所考虑的问题为部分解卷积。一般来讲，由 G 来推断 C 的问题是超定的，因为基因数 n 一般远大于细胞类型数 q；而由 C 来推断 G 的问题是欠定的，因为样本数 p 一般会小于细胞类型数 q。

一般来讲，同时推断 G 和 C 的问题可以由迭代的求解两种部分解卷积问题来得到，这种方法称为交替非负最小二乘法（alternating nonnegative least squares, ANLS）：

$$\begin{cases} \hat{C} \leftarrow argmin_{C\geq 0}\big(\delta(\hat{G}\hat{C} - M)\big) \\ \hat{G} \leftarrow argmin_{G\geq 0}\big(\delta(\hat{C}^T\hat{G}^T - M^T)\big) \end{cases}$$

由于在实际应用中，参考矩阵 G 一般可以通过单细胞 RNA 测序获得，且由 M, G 推断 C 的问题较为常见，本综述报告主要探讨这类问题。由矩阵的分块，我们容易说明这实际上等价于独立的考虑 p 个样本的回归问题：

$$argmin_{\hat{c_i}\geq 0}\delta(G\hat{c_i}, m_i), i = 1, 2, \cdots, p \tag{2}$$

这样，所考虑的问题和我们课程上涉及的多元回归问题形式上就相同了。然而，在简化问题的同时，认为 p 个样本是独立的假设也许也会忽略一部分信息，尤其是我们已知这 p 个样本的空间分布的时候（此时，我们运用空间转录组学的方法测量空间中位点的平均基因表达量，每个空间位点的基因表达量为许多细胞基因表达量的平均值。可以想象，相邻位点的基因表达量应该相似）。一些最前沿的计算生物学技术运用了这部分信息，我们将在综述报告的最后进行介绍，并给出结果复现。

由 (2) 式，我们已经将所考虑的问题转化为课程中涉及的回归问题的形式（注意这里 $\hat{c}_i, m_i$ 都是 n 维向量）。于是，我们仅需考虑这独立的 p 个回归问题。我们沿用课程上的记号：给定样本 $\{(\mathbf{x_i}, y_i)\}_{i=1}^n$，这里 $\mathbf{x_i}$ 为 q 维向量，$X = (\mathbf{x_1}, \cdots, \mathbf{x_n})^T$ 为设计矩阵，$\mathbf{y} = (y_1, \cdots, y_n)^T$。我们试图拟合函数 f，使得残差 $r_i = y_i - f(\mathbf{x_i})$ 之和尽可能小。由于我们假设基因表达是线性的，因此 f 被限定为线性函数。

不仅如此，在课程中我们学习过，为解决变量之间的多重共线性或进行变量选择，我们经常会对目标函数增加一个惩罚项（正则项）$\mathcal{R}$。于是，整个优化问题可以写为

$$argmin_{\mathbf{w}\in R^q} \sum_{i=1}^{m} \mathcal{L}(y_i - \mathbf{x_i}^T\mathbf{w}) + \lambda\mathcal{R}(\mathbf{w}) \tag{3}$$

这里 $\mathcal{L}$ 为某种损失函数，λ 为调参参数，用于控制惩罚项强度。其中，损失函数若选为 $\mathcal{L}_2$ 惩罚，那么优化问题的形式就与我们课程所涉及的最小二乘线性回归相同。此外，本报告将会涉及的损失函数还有 $\mathcal{L}_1$ 损失，Huber 损失[5]

$$\mathcal{L}_{Huber}^{(M)}(r_i) = \begin{cases} r_i^2, & \text{if } |r_i| \le M \\ M(2|r_i| - M), & \text{else} \end{cases}$$

和在支持向量回归（support vector regression[6]）中，用到的损失函数 Hinge 损失（Hinge Loss）：

$$\mathcal{L}_{\epsilon}^{(\epsilon)}(r_i) = max(0, |r_i| - \epsilon)$$

其中，Huber 损失中的 M 和 Hinge 损失中的 ϵ 均为调参参数。以上所有的四个损失函数可以由图 1 直观刻画：

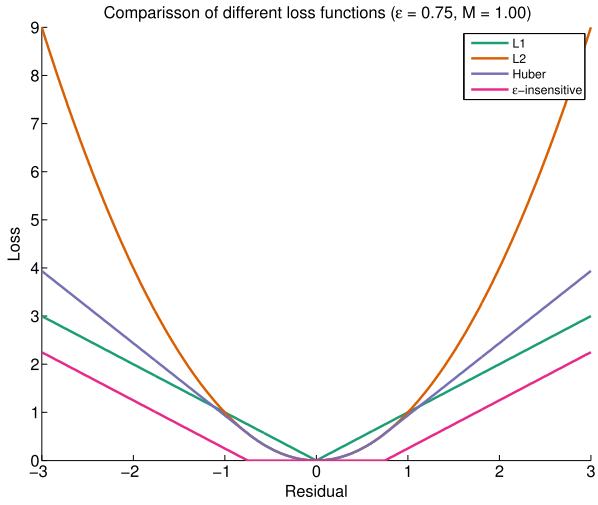

图 1: 四种损失函数比较

对于惩罚项 $\mathbf{R}$，我们可以对 $\mathbf{w}$ 施加 $\mathcal{L}_2$ 惩罚（岭回归）[7]，$\mathcal{L}_1$ 惩罚（LASSO）[8]，以及二者的加权平均（Elastic Net）[9]。

3 转录组学解卷积的经典方法

2009 年 Abbas 等人[10] 首次提出运用经典最小二乘法（Ordinary Least Squares, OLS）进行转录组学解卷积，具体应用于血液样本。我们知道，对于 OLS 我们可以得到解析解

$$\hat{\mathbf{w}} = (X^T X)^{-1} X^T \mathbf{y}$$

然而，由生物学背景，$\mathbf{w}$ 表示样本中每类细胞的占比，因此其元素应非负，且满足归一化条件 $\sum_{i=1}^{q} w_i = 1$。于是，Abbas 等人在得到解析解 $\hat{\mathbf{w}}$ 后，强制将其中所有小于 0 的元素赋值为 0，并进行标准化：

$$w_i \leftarrow \frac{w_i}{\sum_{j=1}^{q} w_j}, i = 1, \cdots, q$$

然而，虽然这样求解很方便，但是求解得到的 $\hat{w}$ 与真值偏差较大，且没有很好的理论性质。因此，Qiao 等人[11] 提出了使用非负最小二乘法（Non-negative Least Squares, NNLS）求解回归问题，具体表述为：

$$argmin_{\mathbf{w} \geq 0} ||\mathbf{y} - X\mathbf{w}||_2^2$$

尽管我们在课程上学习过求解带约束的最小二乘问题（可以给出解析解），但课程上主要涉及的是等式约束，而我们所考虑的是不等式约束，求解解析解较为困难。在实际应用中，Qiao 等人运用 Hanson-Lawson 算法[12]求解 NNLS 问题，它将 $\mathbf{w}$ 的 q 个下标分为被动集（所有大于零元素的下标组成的集合）$\mathcal{P} = \{i | w_i > 0\}$ 和主动集（所有等于零元素的下标组成的集合）$\mathcal{Z} = \{i | w_i = 0\}$，每个循环从主动集向被动集移入负梯度最大的元素下标允许优化。具体算法可写为：

Algorithm 1 Lawson-Hanson 非负最小二乘 (NNLS) 算法

Require: 设计矩阵 $X \in \mathbb{R}^{n \times q}$, 响应向量 $\mathbf{y} \in \mathbb{R}^n$

Ensure: 非负权重向量 $\mathbf{w} \geq 0$，最小化 $\|\mathbf{y} - X\mathbf{w}\|_2$

初始化被动集 $\mathcal{P} = \emptyset$，主动集 $\mathcal{Z} = \{1, 2, \ldots, q\}$
初始化 $\mathbf{w} = \mathbf{0}$，计算负梯度 $\mathbf{v} = X^T(\mathbf{y} - X\mathbf{w})$
while $\mathcal{Z} \neq \emptyset$ 且 $\max_{j \in \mathcal{Z}}(v_j) >$ 容差 **do**
    选择 $j^* = \arg\max_{j \in \mathcal{Z}} v_j$ ▷ 选择违反约束最严重的权重
    将 j^* 从 $\mathcal{Z}$ 移到 $\mathcal{P}$
    repeat
        求解 $\arg\min_{\mathbf{w}_\mathcal{P}} \|\mathbf{y} - X_\mathcal{P}\mathbf{w}_\mathcal{P}\|_2$ ▷ 无约束最小二乘问题
        设置 $\mathbf{w}_\mathcal{Z} = \mathbf{0}$
        if $\mathbf{w}_\mathcal{P}$ 的所有分量 ≥ 0 **then**
            退出循环 ▷ 当前解可行，继续迭代
        else ▷ 处理 $\mathbf{w}_\mathcal{P}$ 不在可行域中的情况
            计算步长 $\alpha = \min_{i \in \mathcal{P}, w_i \leq 0} \left(\frac{w_i}{w_i - \hat{w}_i} \right)$
            更新 $\mathbf{w} \leftarrow \mathbf{w} + \alpha(\hat{\mathbf{w}} - \mathbf{w})$
            将所有 $w_j = 0$ 的索引从 $\mathcal{P}$ 移回 $\mathcal{Z}$
        end if
    until $\mathbf{w}$ 的所有分量非负
    更新对偶变量 $\mathbf{v} = X^T(\mathbf{y} - X\mathbf{w})$
end while
return 非负权重解 $\mathbf{w}$

然而，此方法没有考虑归一化条件 $\sum_{j=1}^{q} w_j = 1$，仍需在得到 $\hat{\mathbf{w}}$ 后进行标准化处理。

Gong 等人[13]提出了在非负约束以及和为一 (sum-to-one, STO) 约束下

求解最小二乘问题，并应用在了血液样本上。他们应用 Lawson 和 Hanson 所提议的二次规划方法求解该问题。[12] 具体来讲，将目标函数改写为

$$\|X\mathbf{w}-\mathbf{y}\|_2^2=(X\mathbf{w}-\mathbf{y})^T(X\mathbf{w}-\mathbf{y})=\mathbf{w}^TX^TX\mathbf{w}-2\mathbf{y}^TX\mathbf{w}+\mathbf{y}^T\mathbf{y}$$

于是优化问题可以等价写为二次规划的标准形式：

$$argmin_{\mathbf{w}}\frac{1}{2}\mathbf{w}^TQ\mathbf{w}+\mathbf{c}^T\mathbf{w},\quad \text{st.}\quad \mathbf{w}\geq 0, A\mathbf{w}=b$$

这里 $Q=X^TX$ 为半正定矩阵，$\mathbf{c}=-2X^T\mathbf{y}, A=\mathbf{1}^T, b=1$。于是，我们可以调用求解二次规划的经典算法，如内点法[14] 或信赖域反射法[15]。

为解决组成细胞之间的多重共线性问题，以及考虑到参考矩阵中可能存在样本中不存在的细胞类型（因此需要进行变量选择），Altboum 等人[16] 提出在上述两个限制条件下，对损失函数施加 Elastic Net 惩罚，用 R 语言的 glmnet 宏包进行求解（主要利用坐标下降法[17]）。

以上所有方法的损失函数全都为 $\mathcal{L}_2$ 损失，它对异常值较为敏感。并且，与 Hinge Loss 相比，$\mathcal{L}_2$ 损失对于低噪声的样本也会施加一定的惩罚。于是，Newman 等人提出 CIBERSORT[18]，这是一种基于支持向量回归（support vector regression, SVM[19]）的方法，比当时其他的方法性能上占有优势，一段时间内被广为应用。

支持向量回归的思想来源于支持向量机(support vector machine, SVM[20])。直观的来讲，它试图寻找一个超平面，在避免过拟合的条件下（由 $\mathcal{L}_2$ 惩罚控制）使得尽可能多的样本点落入超平面的 ϵ-带中（图 2）：

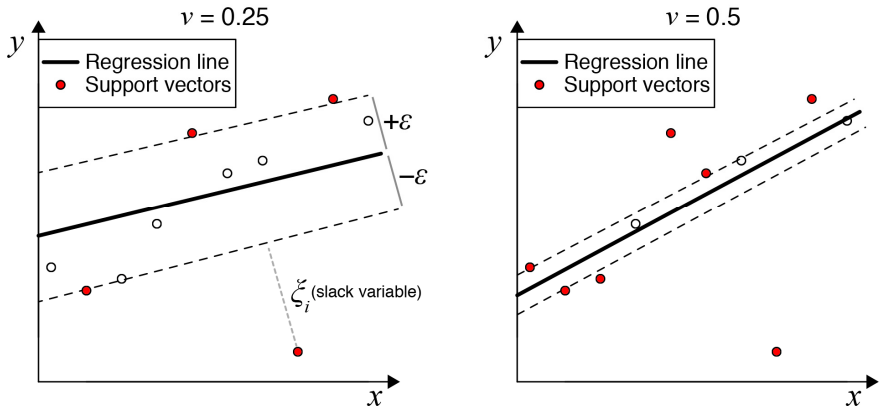

图 2: 支持向量回归

其数学表达式可以写为：

$$\min_{\mathbf{w},\xi_i^+,\xi_i^-}\left\{\frac{1}{2}\|\mathbf{w}\|_2^2+C\sum_{i=1}^{m}(\xi_i^+ +\xi_i^-)\right\}$$

Subject to:

$$\begin{cases} y_i-\mathbf{w}\cdot x_i\le\epsilon+\xi_i^+ \\ -(\epsilon+\xi_i^-)\le y_i-\mathbf{w}\cdot x_i \\ 0\le\xi_i^+,\xi_i^- \end{cases}$$

这里 ϵ 即为带宽，C 为控制损失和惩罚的超参数。为了与本文 (3) 式形式相同，我们把它等价的写为目标函数 + 惩罚项的形式：

$$\min_{\mathbf{w}}\left\{\sum_{i=1}^{m}\mathcal{L}_\epsilon^{(\epsilon)}(y_i-\mathbf{w}^T x_i)+\lambda\mathcal{R}_2(\mathbf{w})\right\} \tag{4}$$

运用支持向量回归进行建模的好处是它对噪声较小的点不敏感，其回归超平面仅由那些位于 ϵ-带之外的点（支持向量）决定，从而选择那些较为关键的基因。

Newman 等人运用 R 语言中的 svm 宏包求解 (4) 式（通过转化为对偶问题求解）。得到解后，再对 $\hat{\mathbf{w}}$ 中所有负值元素强行赋值为 0，并进行归一化，保证满足 STO 条件。

4 转录组学解卷积经典方法性能比较

从上一小节中我们可以发现，不同的转录组学解卷积经典方法，本质上来讲就是不同的损失函数、约束条件和惩罚项的组合。在这一小节，我们根据 Mohammadi 等人在论文 “A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues[21]” 中给出的综述性结果，讨论不同的组合的实际性能表现。

4.1 不同损失函数和约束条件的性能比较

在这一小节中，我们讨论前文提到过的 $\mathcal{L}_1$ 损失、$\mathcal{L}_2$ 损失、Huber 损失、Hinge 损失四种损失函数以及是否施加非负约束、是否是加 STO 约束对回归性能的影响。

首先，在七个不同的数据集中，不同损失函数在是否施加约束条件下的平均计算时间由图 3 刻画。可以看出，$\mathcal{L}_2$ 损失所需要的平均计算时间最短，Huber 损失所需要的平均计算时间最长，而其他两种损失所需的平均计算时间介于二者之间。

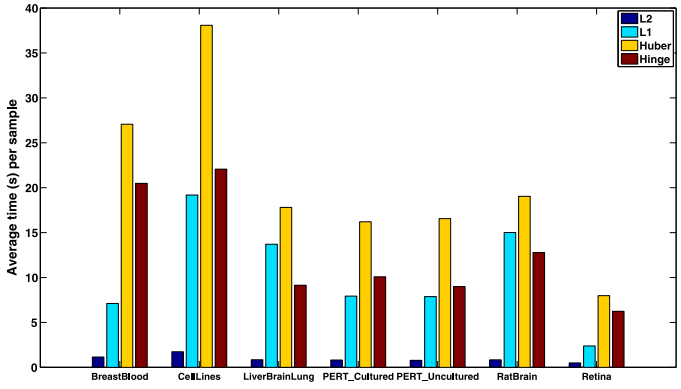

图 3: 不同损失函数平均计算时间

不仅如此，Mohammadi 等人运用平均绝对差异（Mean Absolute Difference, mAD）作为衡量 C 与其估计值 $\hat{C}$ 的差异度量：

$$mAD = \frac{100}{p \times q} \sum_{i=1}^{q} \sum_{j=1}^{p} |c_{ij} - \hat{c}_{ij}|$$

比较了四种损失函数以及是否施加非负约束、STO 约束的回归问题在七个不同的数据集上的比较 (图 4)：

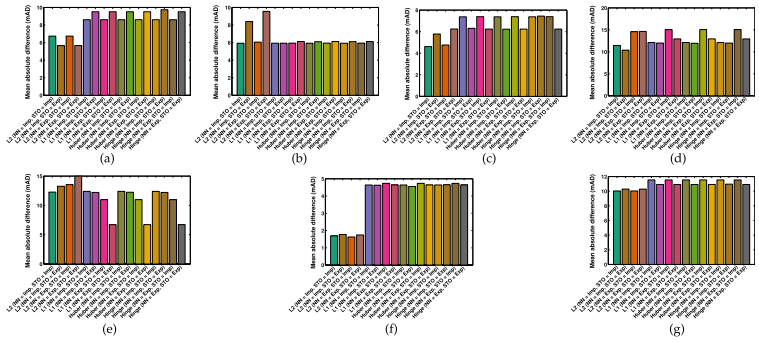

图 4: 不同损失函数/约束条件的 mAD

可以发现，总体来说 $\mathcal{L}_2$ 损失的表现强于其他所有损失函数。结合其计算时长极快的特点，在实际生物学数据上（异常点较少），最简单的 $\mathcal{L}_2$ 损失比起 $\mathcal{L}_1$ 损失、Huber 损失和 Hinge 损失更有优势。另一个令人惊讶的结论是，在多数情况，施加 STO 条件后，回归问题的解的 mAD 反而增加了。Mohammadi 等人认为，这是由于混合样本的 RNA 测序技术和 scRNA-seq 的基因测序技术之间的技术差异导致的。换言之，对于某些基因，可能混合样本的 RNA 测序技术会测出较高的值，而 scRNA-seq 测序技术会测出较低的值；对其他基因则相反。因此，此时施加 STO 限制就没有意义了。

4.2 施加惩罚项的影响

在本小节中，我们探讨惩罚项（$\mathcal{L}_2$ 惩罚以及 $\mathcal{L}_1$ 惩罚）对回归表现（mAD）的影响。

首先，需要说明的是 $\mathcal{L}_1$ 惩罚主要用于变量选择（选择与混合样本相关的细胞种类），而 Mohammadi 等人所应用的参考矩阵已经由生物学背景知识挑选了相关的细胞种类。因此，施加 $\mathcal{L}_1$ 惩罚不会对回归表现有本质提升。而施加 $\mathcal{L}_2$ 惩罚可以解决细胞之间的多重共线性问题。Mohammadi 等人探究了添加 $\mathcal{L}_2$ 损失后 mAD 的下降值，结果如图 5 所示。由图五可见，对于大部分数据集，添加 $\mathcal{L}_2$ 惩罚可以改善回归性能。

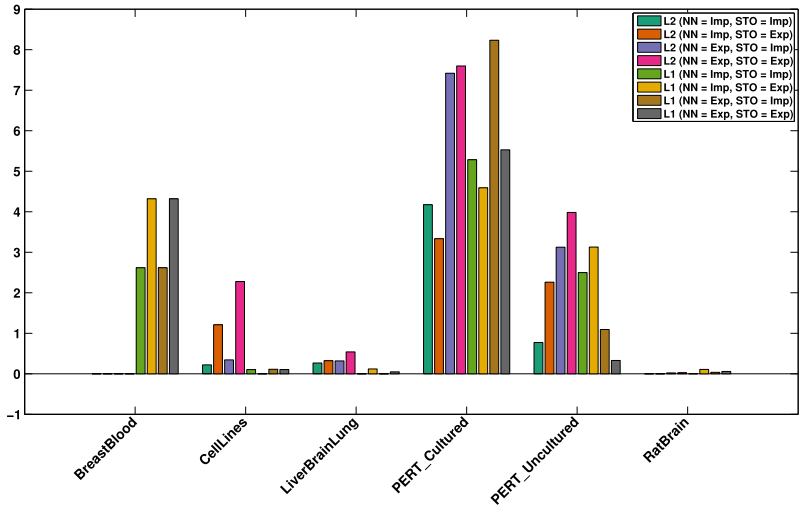

图 5: $\mathcal{L}_2$ 惩罚对回归表现的影响

至此，我们可以对经典的转录组学解卷积方法加以总结：运用岭回归的方法（$\mathcal{L}_2$ 损失 $+\mathcal{L}_2$ 惩罚，并且不施加约束）是最经济实惠且有效的解卷积方法。在下一小节中，我将总结几个近年来较为先进、广为使用的转录组学解卷积方法。

5 前沿转录组学解卷积方法

前面小节中，所提到的转录组学解卷积方法大多都是经典回归方法的直接套用，并没有过多考虑到解卷积问题的生物学意义，且没有利用不同样本之间的相关性信息，因此在准确率、鲁棒性、可拓展性方面均表现欠佳。近年来，许多基因组学的学者基于以上经典方法，提出了一些前沿的转录组学解卷积方法，我将在此小节汇总介绍。

5.1 Bisque

Bisque[22] 是 Jew 等人在 Nature Communications 上首次提出的批量 RNA 解卷积方法，其主要创新点在于解决了 scRNA-seq 和批量 RNA 测序之间技术差异所造成的解卷积性能下降的问题，这也是造成上一小节中对回归方程施加 STO 约束后，回归性能反而下降的主要原因。

Bisque 对 scRNA-seq 的数据要求较高（但随着 scRNA-seq 技术的发展，这样的数据实际上很容易获得）：假定我们已经获得 p' 个样本中每个样本中每个细胞的细胞类型以及在 n 个基因上的表达量。如果我们对 scRNA-seq 数据中每种细胞类型的基因表达量进行平均，仍能得到参考矩阵 $G \in R^{n\times q}$。此外，我们还能得到在 scRNA-seq 中观察到的比例矩阵 $C' \in R^{q\times p'}$。二者相乘，即可得到对于 scRNA-seq 的伪批量（pseudo bulk）矩阵 $M' = GC' \in R^{n\times p'}$。

对于任意基因 j，我们希望消除 scRNA-seq 技术与批量 RNA 测序技术之间的差异。也就是说，对于 $m'_j \in R^{p'}, m_j \in R^p$，我们希望其均值和样本方差尽量类似。由于 scRNA-seq 所测得的样本量 p' 一般较少，其样本方差的方差可能会很大，因此我们不使用其无偏估计，而是使用有偏估计

$$\hat{\sigma_j} = \frac{1}{p'+1}\sum_{i=1}^{p'}(m'_{i,j} - \frac{1}{n}\sum_{k=1}^{n} m'_{k,j})^2$$

对于 m_j，我们使用其样本方差作为其方差的估计

$$\sigma_{X_j} = \frac{1}{p-1}\sum_{i=1}^{p}(m_{i,j} - \frac{1}{n}\sum_{k=1}^{n} m_{k,j})^2$$

Bisque 在做回归解卷积之前，先对混合矩阵 M 的每一行做一个线性变换

$$m_{j,transformed} = \frac{m_j - \mathbf{1_p}\frac{1}{n}\sum_{k=1}^{n} m_{k,j}}{\sigma_{X_j}}\hat{\sigma}_j + \mathbf{1_p}\frac{1}{n}\sum_{k=1}^{n} m'_{k,j}$$

从而得到转换过的混合矩阵 $M_{transformed}$，再利用这个矩阵以及参考矩阵 G 做带有非负约束和 STO 约束的最小二乘解卷积

$$argmin_C\|GC - M_{transformed}\|_2 \quad st. \quad \sum_{i=1}^{q} c_{i,j} = 1, \forall j = 1, \cdots, p, \quad C \geq 0$$

Jew 等人说明，这种先对混合矩阵的每一行做一次线性变换再进行解卷积的方法极大提升了经典解卷积方法的性能。

5.2 SCDC

SCDC[23] 是 Dong 等人在 Bioinformatics 杂志上提出的解卷积方法，旨在融合多个参考矩阵。Dong 等人发现，如果只是简单的将所有参考矩阵对每个细胞类型的基因表达取平均，解卷积的效果不会很好，原因是因为不同的 scRNA-seq 数据集拥有批次效应。于是，Dong 等人提议用 R 个不同的 scRNA-seq 数据集所生成的参考矩阵独立进行解卷积，从而得到 R 个不同的参考矩阵 $\hat{G}_1, \hat{G}_2, \cdots, \hat{G}_R$ 以及 R 个不同的比例矩阵 $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_R$。

Dong 等人提议用 $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_R$ 的加权平均作为最终比例矩阵 C 的估计：

$$\hat{C} = \sum_{i=1}^{R} w_i \hat{C}_i$$

直觉上来讲，我们希望 $\hat{C}$ 与 C 之差越小越好：

$$argmin_{w_1,\cdots,w_R}\delta(C, \sum_{i=1}^{R} w_i \hat{C}_i)$$

这里，$\delta(\cdot,\cdot)$ 为某种损失函数。然而，上式无法优化，因为我们不知道 C 的真值。于是，Dong 等人将上式替换为

$$argmin_{w_1,\cdots,w_R}\delta(M,\sum_{i=1}^{R}w_i\hat{M}_i) \tag{5}$$

这里 $\hat{M}_i=\hat{G}_i\hat{C}_i$ 为预测的混合样本基因表达矩阵。Dong 等人说明，从经验角度来讲，对于某参考矩阵，若使用该参考矩阵解卷积得到的 $\hat{C}_i$ 距离真值 C 偏差较小，那么其回归残差 $M-M_i$ 也一般较小。直观上来讲，回归残差越小的参考矩阵所解出的细胞占比矩阵在最终的细胞占比矩阵估计中的权重越大。

对于损失函数 $\delta(\cdot,\cdot)$ 的选取，SCDC 默认使用 Spearman 相关系数[24]，并使用网格搜索的方法优化 (5) 式。

5.3 DWLS

阻尼加权最小二乘法（Dampened Weighted Least Squares, DWLS[25]）是 Tsoucas 等人发表在 Nature Communications 上的转录组学解卷积方法，旨在解决在 OLS 情境下，某些平均表达量较低的基因或比例较小的细胞类型对总的 $\mathcal{L}_2$ 误差贡献较小，从而导致偏差和信息利用不全面的问题。

Tsoucas 等人提议，将原有的最小二乘形式 $argmin_{\mathbf{w}}\sum_{i=1}^{n}\left(y_i-(X\mathbf{w})_i\right)^2$ 修改为

$$argmin_{\mathbf{w}}\sum_{i=1}^{n}w_i\left(y_i-(X\mathbf{w})_i\right)^2$$

这样，对于平均表达量较小的基因 j，我们可以调高其权重 w_j，从而利用到所有基因的信息。Tsoucas 等人通过数学推导，确定了最佳的 w_i 的形式：

$$w_i=\frac{1}{(X\mathbf{w})_i^2}$$

并且他们还证明，如此形式的最小二乘问题也可以解决细胞比例不平均的问题。为了防止上式的 w_i 趋于无穷大，他们给上式的分母加了一个阻尼常数，并在非负条件下进行优化。

注意到 w_i 的表达式与加权最小二乘法的解有关，因此 Tsoucas 等人给出了一个迭代的求解方法：首先求解无权重非负最小二乘问题，然后迭代地赋值权重 w_i 并求解加权非负最小二乘问题，直至收敛。

5.4 MuSiC

多个体单细胞解卷积（MUlti-Subject SIngle Cell deconvolution, MuSiC[26]）是 Wang 等人发表在 Nature Communications 上的转录组学解卷积方法，旨在自动化标记基因（Marker Genes）的选取并解决细胞类型之间的多重共线性问题。

一般来讲，参考矩阵 G 不能用所有 scRNA-seq 中含有的基因来构建，而是需要预处理。首先，我们需要过滤掉所有在混合样本中不存在的基因。其次，我们需要只保留标记基因，这类基因通常满足两个条件：

- 在不同样本（或个体）之间表达量接近（方差较小）
- 在同一细胞类型的不同细胞之间表达量接近（方差较小）

第一点确保用于构建 scRNA-seq 参考矩阵的样本和我们需要解卷积的混合 RNA 测序样本之间，标记基因的表达量类似，这样解卷积不会引入过多偏差。第二点确保同一类细胞类型内细胞基因表达量噪声较小，也可降低解卷积偏差。

然而，之前的方法必须依赖对于标记基因的预筛选步骤，多数方法会筛选那些跨样本方差、跨细胞方差小于一定阈值的基因作为标记基因。这样的方法需要人为对数据进行预处理，且与所选取的阈值有关。而 MuSiC 则采用加权最小二乘的方法，对基因加以其跨样本方差的倒数作为权重。这样，我们可以容纳所有的基因所带来的信息，并且省略掉了认为构建参考矩阵的步骤。

不仅如此，MuSiC 还解决了细胞之间多重共线性的问题。Wang 等人没有像前人那样利用岭回归来解决这个问题，而是先对所有的细胞类型利用 scRNA-seq 数据进行系统聚类，得到一个树状结构。用户可以人为选择聚类的精细程度。然后，MuSiC 使用在该大类细胞间表达差异不大的基因类型构建参考矩阵，利用加权最小二乘法估计每一细胞大类在每个样本内的占比。接着，对于每一大类，我们可以重复调用 MuSiC，进行进一步的细胞类型细分（图 6）。

与 DWLS 相同，MuSiC 使用一个迭代算法求解在非负条件和 STO 条件下的加权最小二乘问题。

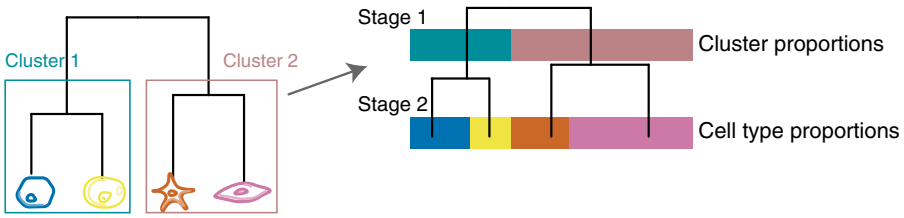

图 6: MuSiC 用系统聚类解决细胞间多重共线性

6 回归分析在前沿空间转录组学解卷积的应用拓展

前文中，我们考虑的问题中涉及的所有“样本”都几乎可以看作相互独立进行解卷积的，因为我们没有什么理由预设它们之间的相关性。然而，随着 RNA 测序技术的发展，出现了空间转录组学（Spatial Transcriptomics, ST）测序的技术[27]，它可以同时测量许多位点（spot）的二维空间坐标以及其平均基因表达量（每个位点由许多细胞组成）。由此可见，ST 技术中每个位点相当于前文讨论的样本，我们需要对 ST 技术产生的数据进行解卷积。由于我们可以获知所有位点的二维坐标信息，将所有位点独立解卷积就无法利用这部分空间坐标的信息。因此，Ma 等人提出基于条件自回归的解卷积方法（Conditional Autoregressive-based Deconvolution, CARD[28]），并发表在 Nature Biotechnology 期刊上，该模型能够同时应用 ST 测序数据的基因表达信息以及空间坐标信息。作为最前沿的解卷积技术，我们还会在报告最后给出 CARD 的结果复现。

6.1 模型描述

与前面的模型相同，Ma 等人将混合矩阵 $M \in R^{n\times p}$ 建模为参考矩阵 $G \in R^{n\times q}$ 与比例矩阵 $C \in R^{q\times p}$ 的乘积和残差项 $E \in R^{n\times p}$ 之和：

$$M = GC + E \tag{6}$$

不同的是，Ma 等人给予残差矩阵 E 正态假设：$e_{i,j} \quad i.i.d \sim N(0, \sigma_e^2)$，这里 σ_e^2 为待估参数。

至此，我们还没有用到 ST 数据的空间信息。于是，Ma 等人考虑了一

个条件自回归模型[29]：

$$c_{k,i} = b_k + \phi \sum_{j=1, j\neq i}^{n} w_{i,j}(c_{k,j} - b_k) + \epsilon_{i,k} \tag{7}$$

这里 b_k 为第 k 个细胞类型在整个 ST 数据的平均占比（未知），ϕ 为自回归的强度参数，$\epsilon_{i,k} \sim N(0, \sigma_{i,k}^2) = \frac{\lambda_k}{\sum_{i=1}^{n} w_{i,j}}$，$\lambda_k$ 为调参参数，$w_{i,j}$ 为由第 j 个位点推测第 i 个位点的权重，Ma 等人使用高斯核函数：

$$K_G(s_i, s_j) = exp(-\frac{\|s_i - s_j\|_2^2}{2\sigma^2})$$

其中带宽 σ 选取为 0.1。

直观的来讲，(6) 式要求 GC 与 M 的残差 E 尽量小，包含了每个位点独立的基因表达信息；而 (7) 式则通过要求位点与其相邻位点的细胞组成类似，将所有空间中的位点联系起来。Ma 等人联合 (6) 式和 (7) 式，并使用极大似然方法联合估计所有待确定的参数 $b_k, \lambda_k, \phi, \sigma_e^2, C$。Ma 等人给出了一个约束下的优化算法，用于计算所有待定参数的数值解。

6.2 结果复现

CARD 已经实施为一个 R 的宏包，我们直接调用其中的解卷积方程即可。我们利用论文提供的一个 428 个位点、25753 个基因、20 个细胞种类的数据做解卷积。CARD 的运行速度很快，上述数据只需要几秒钟即可运行结束。我们将运行结果可视化：

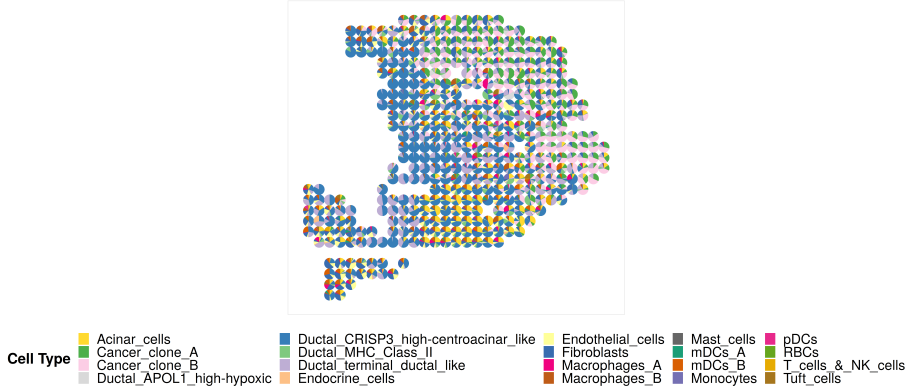

图 7: CARD 解卷积结果复现

7 小结与感想

本综述性报告回顾了从 2009 年到 2021 年这 12 年应用回归方法进行转录组学解卷积的 10 篇主要论文。对于其经典方法，过往文献主要涉及常规最小二乘法（OLS）、带非负约束或 STO 约束的最小二乘法、增加 Elastic Net 惩罚的最小二乘法和支持向量回归。本报告对这些经典转录组学解卷积方法进行了系统性的性能比较。此外，本报告还展示了近年来发表在 Nature Communications, Nature Biotechnology 以及 Bioinformatics 等顶刊的前沿转录组学解卷积方法 Bisque，SCDC，DWLS，MuSiC 和 CARD，阐述其算法及优势，并对最新一篇空间转录组学解卷积工具 CARD 做了结果复现。

基因表达的线性性假设使得线性回归分析在转录组学解卷积任务中得以施展拳脚。实际上，也有一些学者提出了回归方法以外的解卷积方法[30]，大多数都依赖极大似然估计或贝叶斯模型，但不可否认的是，回归方法仍是这一领域的主流方法。从文献回顾中也能看出，许多学者为了把回归方法应用到具体的生物学问题中，对回归方法做了许多改进，产生了许多变体。

线性回归这一概念，大概我们在中学就已经有所耳闻，甚至可能认为它没有其他更加复杂的模型那样耐人寻味。但实际上，它在最前沿的基因组学科技中却应用广泛。也希望这篇综述性报告能够成为一个起点，有朝一日我也能在我的基因组学科研中，应用到回归方法。

参考文献

[1] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[2] Ying Ma and Xiang Zhou. Accurate and efficient integrative reference-informed spatial domain detection for spatial transcriptomics. *Nature Methods*, 21(7):1231–1244, 2024.

[3] Sylvestre François Lacroix. *TRAITÉ DU CALCUL DIFFÉRENTIEL ET DU CALCUL INTÉGRAL.: TOME PREMIER*, volume 1. Chez JBM DUPRAT, Libraire pour les Mathématiques, quai des Augustins, 1797.

[4] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[5] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[6] Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9, 1996.

[7] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[8] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[9] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

[10] Alexander R Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS one*, 4(7):e6098, 2009.

[11] Wenlian Qiao, Gerald Quon, Elizabeth Csaszar, Mei Yu, Quaid Morris, and Peter W Zandstra. Pert: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS computational biology*, 8(12):e1002838, 2012.

[12] Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995.

[13] Ting Gong, Nicole Hartmann, Isaac S Kohane, Volker Brinkmann, Frank Staedtler, Martin Letzkus, Sandrine Bongiovanni, and Joseph D Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PloS one*, 6(11):e27156, 2011.

[14] Pault T Boggs, Paul D Domich, and Janet E Rogers. An interior point method for general large-scale quadratic programming problems. *Annals of Operations Research*, 62:419–437, 1996.

[15] Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

[16] Zeev Altboum, Yael Steuerman, Eyal David, Zohar Barnett-Itzhaki, Liran Valadarsky, Hadas Keren-Shaul, Tal Meningher, Ella Mendelson, Michal Mandelboim, Irit Gat-Viks, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular systems biology*, 10(2):720, 2014.

[17] Stephen J Wright. Coordinate descent algorithms. *Mathematical programming*, 151(1):3–34, 2015.

[18] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

[19] Vladimir Vapnik. Support vector method for function estimation, September 7 1999. US Patent 5,950,146.

[20] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[21] Shahin Mohammadi, Neta Zuckerman, Andrea Goldsmith, and Ananth Grama. A critical survey of deconvolution methods for separating cell

types in complex tissues. *Proceedings of the IEEE*, 105(2):340–366, 2016.

[22] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M Garske, Jae Hoon Sul, Kirsi H Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications*, 11(1):1971, 2020.

[23] Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M Perou, Fei Zou, and Yuchao Jiang. Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Briefings in Bioinformatics*, 22(1):416–427, 01 2020.

[24] Charles Spearman. The proof and measurement of association between two things. In James J. Jenkins and Donald G. Paterson, editors, *Studies in Individual Differences: The Search for Intelligence*, pages 45–58. Appleton-Century-Crofts, 1961.

[25] Daphne Tsoucas, Rui Dong, Huidong Chen, et al. Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, 10:2975, 2019.

[26] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R. Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10:380, Jan 2019.

[27] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, Jul 2016.

[28] Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*, 40:1349–1359, Sep 2022.

[29] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

[30] Tinyi Chu, Zhong Wang, Dana Pe' er, and Charles G Danko. Cell type and gene expression deconvolution with bayesprism enables bayesian integrative analysis across bulk and single-cell rna sequencing in oncology. *Nature cancer*, 3(4):505–517, 2022.