A linear time algorithm for VAFPP projection

Henri Schmidt

July 18, 2023

1 Model for tumor deconvolution

Definition 1. *A rooted tree $\mathcal{T}$ on n vertices is an n-clonal tree for a mutation set $[n] = \{1, \ldots, n\}$ if each edge is labeled by exactly one mutation in $[n]$.*

Let F be an n-by-m matrix of frequencies measured on a set of m mutations across a set of n samples. Given the matrix F, the variant allele frequency projection problem (VAFPP) is to

Problem 1. *Given a frequency matrix F and a clonal matrix B, the* variant allele frequency p-projection problem *(p-VAFPP) is to find a usage matrix U such that*

$$\sum_{i=1}^{m} \|F_i - (UB)_i\|_p \tag{1}$$

is minimized.

First, notice that it suffices to consider the case where there is only a single sample, since the objective is separable with respect to the samples. That is, we can assume F is a row vector, which we denote as f^T, and the goal is to find a usage vector u^T such that $\|f^T - u^T B\|_p$ is minimized.

Here, we will consider the case where $p = 1$, since it has not yet been studied in the literature and is more robust to outliers. The case where $p = 2$ is the well-known case studied by [1], and they derive an efficient $O(mn^2)$ time algorithm to solve the 2-VAFPP problem.

We start by writing out a linear programming formulation of the 1-VAFPP problem. Let f^T be a row vector of frequencies, and let u^T be a row vector of usages. Let B be an n-by-n clonal matrix. Then, the 1-VAFPP problem is equivalent to the following linear program.

$$\begin{aligned}
&\max_{u \geq 0, z \geq 0} -\sum_{i=1}^{n} z_i \\
\text{subject to} \quad & z_i \geq f_i - \sum_{j=1}^{n} u_j B_{ji} \quad \text{for all } i \in [n] && (2)\\
& z_i \geq \sum_{j=1}^{n} u_j B_{ji} - f_i \quad \text{for all } i \in [n] && (3)\\
& 1 \geq \sum_{i=1}^{n} u_i && (4)
\end{aligned}$$

Then, we can write out the dual problem by associating a dual variable α_i with the constraint in (2), a dual variable β_i with the constraint in (3), and a dual variable γ with the constraint in (4). Then, the dual linear program is as follows.

$$
\begin{aligned}
&\min_{\alpha\geq 0,\beta\geq 0,\gamma\geq 0} \gamma + \sum_{i=1}^{n} f_i(\beta_i - \alpha_i) \\
&\text{subject to} \quad \sum_{j=1}^{n} B_{ij}(\beta_j - \alpha_j) + \gamma \geq 0 \quad \text{for all } i \in [n] && (5)\\
&\qquad\qquad \alpha_i + \beta_i \leq 1 \quad \text{for all } i \in [n] && (6)
\end{aligned}
$$

We can perform a change of variables by setting $\lambda_i = \beta_i - \alpha_i$. Since α_i and β_i are non-negative and their sum is bounded by 1, $\lambda_i \in [-1, 1]$. Then, writing the constraints in matrix form and using a slack variable to remove the inequality constraint, we have the following equivalent, dual linear program.

$$
\begin{aligned}
&\min_{\gamma\geq 0,\psi\geq 0} \gamma + f^T\lambda && (7)\\
&\text{subject to} \quad B\lambda = \psi - \gamma\mathbb{1} && (8)\\
&\qquad\qquad \lambda_i \in [-1, 1] \quad \text{for all } i \in [n] && (9)
\end{aligned}
$$

We now make use of the following lemma.

Lemma 1. *Let B be an n-by-n clonal matrix and A the corresponding ancestor-child matrix, where $A_{i,j} = 1$ if j is a parent of i and is otherwise 0. Then,*

$$
B = (I - A)^{-1} \quad \textit{and} \quad [(I - A)v]_i = \begin{cases} v_i - v_{parent(i)} & \textit{if } i \neq \textit{root}, \\ v_i & \textit{otherwise.} \end{cases}
$$

where $parent(i)$ is the parent of vertex i in the tree corresponding to B.

Applying the above lemma and noting that $(\psi_i - \gamma) - (\psi_j - \gamma) = \psi_i - \psi_j$, we obtain:

$$
\lambda_i = \left[(I - A)^{-1}(\psi - \gamma\mathbb{1})\right]_i = \begin{cases} \psi_i - \psi_{\text{parent}(i)} & \text{if } i \neq \text{root}, \\ \psi_i - \gamma & \text{otherwise.} \end{cases}
$$

Finally, noting that $\lambda_i \in [-1, 1]$ and ψ_i, γ non-negative implies $\psi_i, \gamma \in [0, 1]$, we can remove the variable λ. Then, re-writing the objective as a linear function of γ and ψ, we have the following equivalent, dual linear program.

$$
\begin{aligned}
&\min \quad \gamma(1 - f_{\text{root}}) + \sum_{i=1}^{n} \psi_i \left(f_i - \sum_{j\in\text{child}(i)} f_j \right) && (10)\\
&\text{subject to} \quad \psi_i, \gamma \in [0, 1] && (11)
\end{aligned}
$$

Notice that this linear program is trivial to solve, by setting

$$
\gamma = 0 \text{ and } \psi_i = \begin{cases} 0 & \text{if } f_i \geq \sum_{j\in\text{child}(i)} f_j, \\ 1 & \text{otherwise.} \end{cases}
$$

which takes objective value 0 if and only f satisfies the sum condition, providing another proof of the sufficiency of this condition.

Theorem 1. *Given a frequency vector $f \in \mathbb{R}^n$ and a clonal matrix $B \in \mathbb{R}^{n \times n}$, the minimum of*

$$\|f^T - u^T B\|_1$$

over all usage vectors $u \in \mathbb{R}^n$ is equal to

$$\sum_{i=1}^{n} \max\left\{0, \sum_{j \in child(i)} f_j - f_i\right\},$$

where $child(i)$ is the set of children of vertex i in the tree corresponding to B.

References

[1] *Bei Jia, Surjyendu Ray, Sam Safavi, and José Bento. Efficient projection onto the perfect phylogeny model. In* Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.*