Supplementary Material for 'Generalized Hierarchical Sparse Model for Arbitrary-Order Interactive Antigenic Sites Identification in Flu Virus Data'

Proof of Theorem 1

Proof. We prove the theorem by contradiction. Suppose the optimal solution of problem (10) is $\{\dot{\boldsymbol{\theta}}^{(1)}, \cdots, \dot{\boldsymbol{\theta}}^{(K)}\}$ and there exists some element $\dot{\theta}^{(k)}_{\langle i_1,\cdots,i_k\rangle}$ that its sign differs from the the element $v^{(k)}_{\langle i_1,\cdots,i_k\rangle}$, i.e. $sign\left(\dot{\theta}^{(k)}_{\langle i_1,\cdots,i_k\rangle}\right) = -sign\left(v^{(k)}_{\langle i_1,\cdots,i_k\rangle}\right)$. Now, let $\ddot{\theta}^{(k)}_{\langle i_1,\cdots,i_k\rangle} = -\dot{\theta}^{(k)}_{\langle i_1,\cdots,i_k\rangle}$, and replace $\dot{\theta}^{(k)}_{\langle i_1,\cdots,i_k\rangle}$ with $\ddot{\theta}^{(k)}_{\langle i_1,\cdots,i_k\rangle}$ in the optimal solution to get $\{\ddot{\boldsymbol{\theta}}^{(1)}, \cdots, \ddot{\boldsymbol{\theta}}^{(K)}\}$. It is easy to see that $\{\ddot{\boldsymbol{\theta}}^{(1)}, \cdots, \ddot{\boldsymbol{\theta}}^{(K)}\}$ still satisfies the hierarchical chain constraint in problem (10) and the value of the second ℓ_1 term in the objective function remains the same under the new solution. Now since $\frac{\tau}{2}\sum_{k=1}^{K}\|\ddot{\boldsymbol{\theta}}^{(k)} - \mathbf{v}^{(k)}\|_2^2 < \frac{\tau}{2}\sum_{k=1}^{K}\|\dot{\boldsymbol{\theta}}^{(k)} - \mathbf{v}^{(k)}\|_2^2$, we conclude that $\{\dot{\boldsymbol{\theta}}^{(1)}, \cdots, \dot{\boldsymbol{\theta}}^{(K)}\}$ is not the optimal solution, which makes a contradiction.

Now, we know that the signs of the elements in the optimal solution $\{\boldsymbol{\theta}^{*(1)}, \cdots, \boldsymbol{\theta}^{*(K)}\}$ must be the same with the signs of the corresponding elements in $\{\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(K)}\}$. Then, by letting $\bar{\boldsymbol{\theta}}^{(k)} = |\boldsymbol{\theta}^{(k)}|$ for $k \in \mathbb{N}_K$, we can directly obtain the conclusion in Theorem 1. □

Proof of Lemma 1

Proof. The first statement is obvious. We now adopt the induction technique to prove the second statement. In the following analysis, a sequence $(\dot{s}_1, \cdots, \dot{s}_K)$ is said to be better (worse) than another sequence $(\ddot{s}_1, \cdots, \ddot{s}_K)$ for problem (19) if both the sequences are feasible for problem (19), i.e., satisfying the hierarchical chain constraints and the objective value of problem (19) at $(\dot{s}_1, \cdots, \dot{s}_K)$ is smaller (larger) than that at $(\ddot{s}_1, \cdots, \ddot{s}_K)$.

When $H = 2$ and $u_1 \le u_2$, we assume the optimal solution is (s_1^*, s_2^*), where $s_1^* \ge s_2^*$. If $s_1^* > s_2^*$, there must exist a $\check{s}$ that $s_1^* > \check{s} > s_2^*$ and $u_1 \le \check{s} \le u_2$. Otherwise, if $u_1 \le u_2 < \check{s} < s_1^*$, we can immediately get $s_2^* = u_2$, and then (s_2^*, s_2^*) is better than (s_1^*, s_2^*), which contradicts the fact that (s_1^*, s_2^*) is the optimal solution. The case that $s_2^* < \check{s} < u_1 \le u_2$ can be proved similarly.

Now we have $s_1^* > \check{s} > s_2^*$ and $u_1 \le \check{s} \le u_2$. Assume $\check{s} = \frac{\omega_1 u_1 + \omega_2 u_2}{\omega_1 + \omega_2}$, we can immediately obtain that $(\check{s}, \check{s})$ is better than (s_1^*, s_2^*), which again contradicts the fact that (s_1^*, s_2^*) is the optimal solution. Therefore, we must have $s_1^* = s_2^* = \frac{\omega_1 u_1 + \omega u_2}{\omega_1 + \omega_2}$.

Then we assume that the statement holds for any $k \le K - 1$. We will show that when $k = K$, the statement also holds. Actually, given $k = K$ and $u_1 \le \cdots \le u_K$, the optimal solution must have the form $(\check{s}, \cdots, \check{s})|_{K-1} \bowtie s_K^*$, i.e. $(\check{s}, \cdots, \check{s}, s_K^*)$, where $\check{s} \ge \bar{s}_K^*$. Otherwise, suppose the optimal solution is denoted by $(\dot{s}_1, \dot{s}_2, \cdots, \dot{s}_K)$ with at least one equality dissatisfied in inequalities $\dot{s}_1 \ge \dot{s}_2 \ge \cdots \ge \dot{s}_{K-1}$. Then we can immediately obtain a contradiction that the sequence $(\check{s}^*, \cdots, \check{s}^*)|_{K-1} \bowtie \dot{s}_K$ is better than $(\dot{s}_1, \dot{s}_2, \cdots, \dot{s}_K)$ where $(\check{s}^*, \cdots, \check{s}^*)|_{K-1}$ is the optimal solution of the problem of size $k = K - 1$ corresponding to the sequence $(u_1, \ldots, u_{K-1})$. Similarly, we can get that the optimal solution have the form $s_1^* \bowtie (\check{s}, \cdots, \check{s})|_{K-1}$, i.e. $(s_1^*, \check{s}, \cdots, \check{s})$, where $s_1^* \ge \check{s}$. Combing those results we complete the proof. □

A Useful Lemma

Lemma 3. *For any input $(u_1, \cdots, u_K)$ and $(\omega_1, \cdots, \omega_K)$, if the optimal solution of problem (19) is $(s^*, \cdots, s^*)|_K$, then for any $\check{s}$ and $\dot{s}_1 \ge \cdots \ge \dot{s}_K$ such that $s^* \ge \check{s} \ge \dot{s}_1$ or $\dot{s}_K \ge \check{s} \ge s^*$, the sequence $(\dot{s}_1, \cdots, \dot{s}_K)$ is not better than $(\check{s}, \cdots, \check{s})|_K$.*

Proof. We first prove it when $s^* \ge \check{s} \ge \dot{s}_1$. Given any K and the sequence $(u_1, \cdots, u_K)$, we consider the feasible sequence $(\dot{s}_1, \cdots, \dot{s}_K)$ for problem (19), where $\dot{s}_1 \ge \cdots \ge \dot{s}_K$. Then, we can obtain that the sequence $(\dot{s}_1, \cdots, \dot{s}_1)|_K$ is not worse than $(\dot{s}_1, \cdots, \dot{s}_K)$, because if $(\dot{s}_1, \cdots, \dot{s}_1)|_K$ is worse, there must exist a sequence $(\ddot{s}_2, \cdots, \ddot{s}_K)$, where $s^* > \ddot{s}_2 \ge \cdots \ge \ddot{s}_K$, such that the optimal solution for the sub-sequence $(s_2, \cdots, s_K)$ is $(\ddot{s}_2, \cdots, \ddot{s}_K)$, and in that case $(s^*, \ddot{s}_2, \cdots, \ddot{s}_K)$ is better than $(s^*, \cdots, s^*)|_K$, which contradicts with the fact that $(s^*, \cdots, s^*)|_n$ is the optimal solution. Therefore $(\dot{s}_1, \cdots, \dot{s}_1)|_n$ is better than $(\dot{s}_1, \cdots, \dot{s}_K)$. Furthermore, since $s^* \ge \check{s} \ge \dot{s}_1$, it is easy to see that $(\check{s}, \cdots, \check{s})|_K$ is not worse than $(\dot{s}_1, \cdots, \dot{s}_1)|_K$ due to the convexity of the function $f(x) = \sum_k \omega_k (x - u_k)^2$ for positive weights ω_k. So we complete the proof when $s^* \ge \check{s} \ge \dot{s}_1$. The case that $\dot{s}_K \le \check{s} \le s^*$ can be proved similarly and we finish the proof. □

Proof of Lemma 2

Proof. The case that $\dot{u}^* \ge \ddot{u}^*$ is obvious. Then we prove the case that $\dot{u}^* < \ddot{u}^*$. In this case, we denote the optimal solution for the concatenated sequence by $(s_1^*, \cdots, s_l^*, s_{l+1}^*, \cdots, s_n^*)$, where $s_1^* \ge \cdots \ge s_l^* \ge s_{l+1}^* \ge \cdots \ge s_n^*$. Then it is easy to show that $s_l^* \ge \dot{u}^*$, because if $s_l^* < \dot{u}^*$, substituting the sub-sequence $(s_1^*, \cdots, s_l^*)$ with $(\dot{u}^*, \cdots, \dot{u}^*)|_l$ in $(s_1^*, \cdots, s_l^*, s_{l+1}^*, \cdots, s_n^*)$ will lead to a better feasible solution, which makes a contradiction. Similarly, we can show that $s_{l+1}^* \le \ddot{u}^*$. Then based on Lemma 3, substituting the two sub-sequences $(s_1^*, \cdots, s_l^*)$ and $(s_{l+1}^*, \cdots, s_n^*)$ with $(s_l^*, \cdots, s_l^*)|_l$ and $(s_{l+1}^*, \cdots, s_{l+1}^*)|_{n-l}$ respectively will generate a new solution that is not worse than the previous one. Note that $s_l^* \ge \dot{u}^*$, $s_{l+1}^* \le \ddot{u}^*$, $\dot{u}^* < \ddot{u}^*$ and $s_l^* \ge s_{l+1}^*$. Then the optimal solution is achieved when $s_l^* = s_{l+1}^*$ due to the convexity of the objective function, making the optimal solution have the form $(s^*, \cdots, s^*)|_n$. Plugging the form into problem (19), we get $s^* = \frac{\sum_{k=1}^{n}\omega_k u_k}{\sum_{k=1}^{n}\omega_k}$, in which we reach the conclusion. □

Proof of Theorem 2

Proof. In Algorithm 3, step 1 splits the initial sequence $(u_1, \cdots, u_K)$ into several non-decreasing sub-sequences. According to Lemma 1, the solutions for those sub-sequences take the form that the entries in the solution are identical. Then, steps 2-14 concatenate the solutions of these sub-sequences according to Lemma 2 iteratively. According to Lemma 2, the global optimality can be guaranteed for any concatenation operation. So Algorithm 3 can find the optimal solution in step 15 for problem (19). □