Estimating cell counts via linear programming [BPP82-628]

October 8, 2021

Abstract

We describe the linear program of `opti.py` attached to the Jira ticket BBPP82-628. This program estimates the cell counts of various inhibitory neurons in the mouse brain.

This document is a companion for the implementation of the module

`inhibitory_neuron_density_optimization.py`

in `atlas-building-tools`. The linear program of Section 2 is the program currently implemented. It stems out of the original implementation `opti.py` attached to the ticket.

1 Input data and decision variables

We are given a finite set R of brain regions and a finite set M of gene markers. For every region $r \in R$ and every marker $m \in M$, we want to estimate the count $x_{r,m}$ of the cells of r reacting to m. We denote by N the subset of M consisting of the brain regions n which aren't leaves of the brain hierarchy tree. The hierarchy tree is provided by the Allen Institute for Brain Science and the set R is made of unique string identifiers, i.e., the region names. The set of gene markers is for the concrete instance of BBPP82-628:

$$M = \{\mathrm{PV}, \mathrm{SST}, \mathrm{VIP}, \mathrm{GAD67}\}.$$

We single out the marker $m_0 := \mathrm{GAD67}$, as it plays a special rôle. The cells reacting to m_0 are all the inhibitory neurons, which includes the cells reacting to any of the other markers in M.

We are given for every $(r, m) \in R \times M$, an estimate $\tilde{x}_{r,m}$ or $x_{r,m}$ and a non-negative *confidence* number $\sigma_{r,m}$. We are also given an upper bound $b_{m,r}$ for each cell count. These upper bounds correspond to the estimates of the overall neuron counts of the regions r. We want to solve the following linear program: Minimize

$$f((\delta_{r,m})) = \sum_{r,m} \frac{\delta_{r,m}}{\sigma_{r,m}} \quad (\sigma_{r,m} > 0)$$

subject to the inequalities:

$$\begin{cases} 0 \leq \delta_{r,m}, & \text{(1a)} \\ 0 \leq y_{r,m} \leq b'_{r,m}, \text{if } r \text{ is not a leaf}, & \text{(1b)} \\ 0 \leq x_{r,m} \leq b_{r,m}, & \text{(1c)} \\ x_{r,m} - \delta_{r,m} \leq \tilde{x}_{r,m}, & \text{(1d)} \\ -x_{r,m} - \delta_{r,m} \leq -\tilde{x}_{r,m}, & \text{(1e)} \\ \sum_{m \in M \setminus \{m_0\}} x_{r,m} - x_{r,m_0} \leq 0. & \text{(1f)} \\ \sum_{m \in M \setminus \{m_0\}} y_{r,m} - y_{r,m_0} \leq 0, \text{if } r \text{ is not a leaf}. & \text{(1g)} \end{cases}$$

where each pair (r, m) is such that $\sigma_{r,m} > 0$, and to the equalities

$$\begin{cases} x_{r,m} = \tilde{x}_{r,m}, \text{ if } \sigma_{r,m} = 0, & \text{(2a)} \\ y_{r,m} + \sum_{r' \in \text{children(r)}} x_{r',m} - x_{r,m} = 0, \text{ if } \sigma_{r,m} > 0 \text{ and } r \text{ is not a leaf}. & \text{(2b)} \end{cases}$$

The conjunction of the inequalities (3c) and (3d) is equivalent to

$$|x_{r,m} - \tilde{x}_{r,m}| \leq \delta_{r,m}$$

so that minimizing the objective function f above consists in minimizing the (weighted sum of) differences between the region cell counts $x_{r,m}$ and their respective estimates $\tilde{x}_{r,m}$ while enforcing the consistency of cell count sums with respect to the region hierarchy tree.

The variable $y_{r,m}$, for $r \in R$ and r not a leaf, represents the cell count of the (nameless) 3D region whose voxels do not belong to any of the child regions of

r. This variable accounts for the imprecision of the annotated volume where a non-negligible number of voxels could not be assigned to descendant regions within the hierarchy tree. Similarly to $b_{r,m}$, the upper bound $b'_{r,m}$ consists in an known overall neuron count estimate for the 3D region associated to $y_{r,m}$.

In the file `opti.py`, the implemented linear program reduces the number of variables by removing the equalities (2a) and by injecting the values $\tilde{x}_{r,m}$ for which $\sigma_{r,m} = 0$ as constants in the right-hand sides of the inequalities (3e) and equalities (2b). In this linear program the variables are:

- $x_{r,m}, \delta_{r,m}$ for every $(r, m) \in R \times M$ such that $\sigma_{r,m} > 0$.
- $y_{r,m}$ for every $(r, m) \in N \times M$ such that $\sigma_{r,m} > 0$.

The total number of variables is at most $|M|(2|R| + |N|)$. Let F be the set of pairs $(r, m) \in R \times M$ such that $\sigma_{r,m} = 0$. Then the number of variables is at most $|M|(2(|R| - |F|) + |N|)$.

2 An equivalent program with less variables

We aim at simplifying the former program in providing an equivalent program with less decision variables. The previous notation stays in effect and we now emphasize the difference between a region $r \in R$ of the hierarchy tree and a 3D region annotated by a label i. If the volume of a region r is annotated by one or several labels, we call these labels the *atoms* of r.

For i a label of a 3D (atomic) region, we denote by $x_{i,m}$ the number of cells the region labeled by i which react to the gene marker m. Each variable $x_{i,m}$ is assigned an upper bound $b_{i,m}$, a *confidence value* $0 \leq \sigma_{i,m} \leq \infty$, and a known estimate $\tilde{x}_{i,m}$ whenever $\sigma_{i,m} < \infty$. If $\sigma_{i,m} > 0$, the linear program attempts to minimize $\sum_{i \in \text{atoms}(r)} |x_{i,m} - \tilde{x}_{r,m}|$ by introducing the variable $\delta_{r,m}$ in the objective function that represents this distance. If $\sigma_{i,m} = 0$, the estimate $\tilde{x}_{i,m}$ is taken as the actual value of $x_{i,m}$ so that the variable can be removed from the linear program.

The values of the constants $\sigma_{i,m}$ are derived from the values of the constants $\sigma_{r,m}$ defined earlier. If $\sigma_{r,m} = 0$ then $\sigma_{i,m} = 0$ for every atom i of r. If $\sigma_{r,m} > 0$, r is not a leaf region and i is the label of voxels *lying at the root of* r (see our definition of $y_{r,m}$ above), then $\sigma_{i,m} = \infty$, i.e., we have no available estimate to optimize against. We want to solve the following linear program:

Minimize

$$f((\delta_{r,m})) = \sum_{r,m} \frac{\delta_{r,m}}{\sigma_{r,m}} \quad (\sigma_{r,m} > 0)$$

subject to the inequalities:

$$\begin{cases} 0 \leq \delta_{r,m}, & \text{(3a)} \\ 0 \leq x_{i,m} \leq b_{i,m}, & \text{(3b)} \\ \sum_{i \in \text{atoms}(r)} x_{i,m} - \delta_{r,m} \leq \tilde{x}_{r,m}, & \text{(3c)} \\ -\sum_{i \in \text{atoms}(r)} x_{i,m} - \delta_{r,m} \leq -\tilde{x}_{r,m}, & \text{(3d)} \\ \sum_{m \in M \setminus \{m_0\}} x_{i,m} - x_{i,m_0} \leq 0. & \text{(3e)} \end{cases}$$

where the pairs $(r, m), (i, m)$ satisfy $\sigma_{r,m} > 0, \sigma_{i,m} > 0$. The variables $x_{i,m}$ for which $\sigma_{i,m} = 0$ evaluate to the constants $\tilde{x}_{i,m}$ and are subsequently moved to the right hand side of the inequalities. Such resolved variables are therefore removed from the linear program.

In this linear program, the decision variables are:

- $x_{i,m}$ for every $(i, m) \in \text{Labels} \times M$ such that $\sigma_{i,m} > 0$.
- $\delta_{r,m}$ for every $(r, m) \in R \times M$ such that $\sigma_{r,m} > 0$.

As $|\,\text{Labels}\,| = |R|$ in our specific use case, the number of variables is now at most $2|M||R|$ and the number of inequality constraints is at most $4|M||R|+|R|$.