Theoretical Analysis of Gene Expression Reconstruction

1 Single Transition Case

1.1 Exact Bounds for Dense Sampling

Let $p(s_d = t_i|s_g, \sigma^2)$ be the probability of selecting the i'th time point conditioned on the actual step time and the noise in the measured data, and let $L(t_i)$ denote the likelihood of the observed data for a specific time point t_i. In order to select t_i as the step point, we need the likelihood defined by this point to be higher than any other point. Thus;

$$p\big(s_d = t_i|s_q, \sigma^2\big) = p\big(L_i > L_j,\, i \neq j\big) \tag{1}$$

Let $\hat{L}_i = \log(L_i)$, and This can also be interpreted as follows. Let $M = L(i)$, then this can be written as follows:

$$p\big(s_d = t_i|s_q, \sigma^2\big) = \int_{-\infty}^{\infty} p(\hat{L}_i = m) p\big(\hat{L}_j \leq m,\, i \neq j\big) dm \tag{2}$$

where $p(\hat{L}_i = m)$ is chi-squared distribution, and $p\big(\hat{L}_j \leq m,\, i \neq j\big)$ is the probability likelihoods of all other time points is smaller than m. Let $S_i = \{t_1, t_2, \ldots, t_{i-1}\}$ be the set of sorted time points that are smaller than t_i, and $M_i = \{t_{i+1}, t_{i+2}, \ldots, t_T\}$ be the set of time points larger than t_i. For $t_j \in S_i$, $p(\hat{L_i} \geq \hat{L_j})$ is:

$$p(\hat{L_i} \geq \hat{L_j}) = \frac{1}{2\sigma^2}\left(\sum_{m=j}^{i-1} -(d_m)^2 - \sum_{m=j}^{i-1} -(d_m - 1)^2\right) = \sum_{m=j}^{i-1} d_m \leq \frac{s}{2} \tag{3}$$

due to gaussian assumption where s is the number of points between t_j and t_{i-1} including both time points. Similar inequalities for all points in S_i return the

following dependent equations:

$$d_{i-1} \leq 0.5 \tag{4}$$
$$d_{i-1} + d_{i-2} \leq 1 \tag{5}$$
$$d_{i-1} + d_{i-2} + d_{i-3} \leq 1.5 \tag{6}$$
$$\dots \tag{7}$$
$$d_{i-1} + \ldots + d_1 \leq \frac{i-1}{2} \tag{8}$$

Similar analysis for points in M_i return the following dependent equations:

$$d_i \geq 0.5 \tag{9}$$
$$d_i + d_{i+1} \geq 1 \tag{10}$$
$$d_i + d_{i+1} + d_{i+2} \geq 1.5 \tag{11}$$
$$\dots \tag{12}$$
$$d_i + \ldots + d_{T-1} \geq \frac{T-i}{2} \tag{13}$$

These joint integrals are independent of each other, so overall expression becomes:

$$p\big(s_d = t_i | s_q, \sigma^2\big) = \int_{-\infty}^{\infty} p(\hat{L}_i = m)\, p\big(\hat{L}_j \leq m,\, j \in S_i\big)\, p\big(\hat{L}_j \leq m,\, j \in M_i\big) dm \tag{14}$$

where $p\big(\hat{L}_j \leq m,\, j \in S_i\big)$ and $p\big(\hat{L}_j \leq m,\, j \in M_i\big)$ are probabilities of satisfying the equations for S_i and M_i above. $p\big(\hat{L}_j \leq m,\, j \in S_i\big)$ can be expressed by the following nested integral:

$$p\big(\hat{L}_j \leq m,\, j \in S_i\big) = \int_{-\infty}^{0.5} p(x_{i-1}) \int_{-\infty}^{1-x_{i-1}} p(x_{i-2}) \dots \int_{-\infty}^{\frac{i-1}{2} - \sum_{t=2}^{i-1} x_t} p(x_1) d_{x_1} d_{x_{i-2}} d_{x_{i-1}} \tag{15}$$

where $p(x)$ is gaussian probability distribution function. Since $p(x)$ is same for all time points, we can estimate the area by visualization. Let $x = cdf(0.5)$, and $D(n)$ be the area by using only topmost n equations. Then, integral can be estimated recursively by:

$$D(n+1) = D(n)(1 - \frac{1}{n+1}(1-x)) \tag{16}$$

with base case $D(1) = x$, and resulting integral is equal to $D(i-1)$. Similarly, $p\big(\hat{L}_j \leq m,\, j \in M_i\big)$ can be estimated by the following recursive equation:

$$U(n+1) = U(n)(1 - \frac{x}{n+1}) \tag{17}$$

where $U(n)$ is the area by using only topmost n equations and base case is $T(1) = 1 - x$. $p\big(\hat{L}_j \leq m,\, j \in M_i\big) = U(T-i)$.

In main equation 2, $p(m)$ is same for all time points and independent of other parts, so it integrates out to 1. As a result, Eq. 2 becomes:

$$p\big(s_d = t_i | s_q, \sigma^2\big) = U(T - i)\, D(i - 1) \tag{18}$$