Age Distribution Inference of Blood cells

Aodong Li

al5350@nyu.edu

1 Progress - May 2

To distinguish a uniform distribution assumption and a finite exponential distribution assumption, we can take advantage of the Riemann sum approximation to the integral 6.

First notice that the shape of the cdf of a uniform distribution is linear whereas the shape of the cdf of an exponential distribution is non-linear. Hence it is fair to expect that if we increase the input, i.e., G, by a step and keep it fixed starting from a certain point, the response difference should decrease linearly for a while for the uniform distribution assumption whereas exponentially for the exponential distribution assumption.

For example, a sequence of G is changing like this:

... 100 100 100 150 150 150 ...

For the uniform distribution, the response difference changes linearly in theory, like

... 0 0 0 0.1 0.08 0.06 ...

until all the cells are renewed by the new glucose concentration. But for the exponential distribution, the response difference changes faster since the difference day by day is not constant.

1.1 Simulation

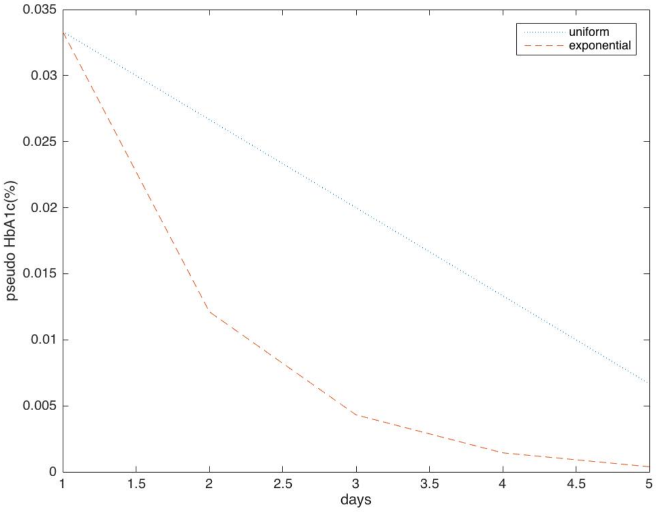
The simulation shows an age distribution with a support of 5 days. The same setting is used as in the previous simulation. The pdf of the uniform distribution is

$$f_{RBC_{age}}(x) = \frac{1}{5}, \quad x \in [0, 5]$$

and the pdf of the finite exponential distribution is

$$g_{RBC_{age}}(x) = \frac{e^{-x}}{1 - e^{-5}}, \quad x \in [0, 5]$$

The glucose concentration is increased from 100mg/dL to 150mg/dL and kept constant. The day-by-day response difference plot is as follows.



It can be seen that the response difference of the uniform distribution is decreasing linearly but the response difference of the finite exponential distribution is decreasing exponentially.

2 Progress - May 1

For the previous uniform distribution assumption, generate a sequence of glucose measurements G by

```
G = 100+10*randn(20, 1)
```

Then the predicted H measurement is 5.1, and the most probable result is $M \approx 76$, which renders a pdf

$$f_{RBC_{age}}(t) = \frac{1}{152} \qquad 0 \le t \le 152,$$

which greatly exceeds the longest age of cells in practice. So there might be something on assumptions going wrong.

The work given by Nathan et al. presents that there is a linear relationship between estimated G and $HbA1c$ measurements

$$\mathrm{AG}_{\mathrm{mg/dl}} = 28.7 \times \mathrm{HbA1c}_{\%} - 46.7, R^2 = 0.84, P < 0.0001$$

$$\mathrm{HbA1c}_{\%} = \frac{\mathrm{AG}_{\mathrm{mg/dl}} + 46.7}{28.7} \tag{1}$$

This empirical result shed a light on the assumptions we use and make them more accurate. It turns out that our assumption should make the simulation results accorded with the empirical results.

It is noted that there is a bias in the equation, which is 1.627. That means that in the extreme case even when no glucose in the blood exist, $HbA1c$ measurements still give a result of 1.627%. It can be seen as a solid bias. So the computation in the equation 6 should take this into consideration. In practice, we subtract 1.627 from the $HbA1c$ measurements and then conduct the computation.

It turns out that after taking this into consideration, for diabetes I, $M \approx 49$ and for non-diabetes, $M \approx 51$. The results are much more consistent.

This is under the assumption that the age is distributed uniformly.

3 Notes on Fourier transform and its applications - April 24

3.1 Some fundamental questions and answers

- In which form do we think of the complex number better?

  Vectors. For example, better to think of $w = x + iy$ as a vector (x, y) in the complex space $\mathbf{C}$.

- What does a complex exponential $e^{2\pi int}$ mean for signals?

  A vector runs along a circle counter-clockwise with frequency n. That vector is $(\cos(2\pi nt), \sin(2\pi nt))$, where $e^{2\pi int} = \cos(2\pi nt) + i\sin(2\pi nt))$.

- What does negative frequencies mean?

  A vector runs along a circle clockwise similarly as the above.

- Why and how do we use complex exponentials to represent real periodic signals?

  Simplicity! $c_{-n}e^{-2\pi int} + c_n e^{2\pi int}$ is able to represent $a_n \cos(2\pi nt) + b_n \sin(2\pi nt)$ where $c_{-n}, c_n \in \mathbf{C}$ and $a_n, b_n \in \mathbf{R}$. This fact comes from any function f can be decomposed into even component $\frac{f(t)+f(-t)}{2}$ and odd component $\frac{f(t)-f(-t)}{2}$.

- What if I am not familiar with the complex representations?

  Get over it!

3.2 The whole picture of Fourier analysis

The whole journey is starting with periodic cases with period T, which leads to Fourier series. Then we adopt a transition to non-periodic cases by taking $T \to \infty$, which gives rise to Fourier transform.

The spectrum of a periodic function is a discrete set of frequencies, possibly an infinite set (when there's corner in the space or time domain). By contrast, the Fourier transform of a non-periodic signal produces a continuous spectrum, or a continuum of frequencies. This can be thought of as $T \to \infty$, the space interval $1/T$ between spectrums decreases all the way to continuity.

However, the modern world is full of discrete signals rather than ideal continuous signals because the recording machine can only generate discrete digital signals. Thus an approximation to the continuous signals by discrete ones is necessary, which lays the building stone for modern applications.

The transition from the Fourier Transform to the Discrete Fourier Transform takes advantage of the delta function δ, because of the equivalent nature between the sampling of continuous signals by δ and discrete signals. Furthermore, by this definition, N points in time or space domain always result in N points in frequency domain and automatically exert **periodicity** N on time and frequency domain.

For the convolution operation now, since we have only finite N points, and by the definition of convolution

$$h[m] = \sum_{k=0}^{N-1} f[k]g[m-k], m = 0, ..., N-1$$

the periodicity of g has to be used in defining h, because the index on g will be negative for $m < k$. Also note that h is periodic.

Now we have two (dual) representations for the same object, in time domain and frequency domain. This duality gives us easier manipulations of the signals in one domain while it may be difficult in another domain, for example, the convolution and filtering.

3.3 Ring a bell

If the water is the same murkiness throughout, meaning, for example, uniform density of stuff floating around in it, then it is natural to assume that the intensity of light decreases by the same percent amount per length of path traveled.

Constant percent change characterizes exponential growth, or decay, so the attenuation of the intensity of light passing through a homogeneous medium is modeled by

$$I = I_0 e^{-\mu x},$$

more generally,

$$I = I_0 e^{-\int_L \mu(x)dx},$$

where L is the line the light travels along. It is common to call the number

$$p = \int_L \mu(x)dx = -\ln(\frac{I}{I_0})$$

the attenuation coefficient.

4 Recent Progress - April 17

4.1 Specify the linear system

New signal comes in continuously at any time. So if we want to get the correct output sequence, we have to eliminate the previous signal effects, through which we may be able to do the deconvolution pointwise to the signal.

To simplify the notation, let $H(t)$ be $HbA1c(t)$ and $F(t)$ be $1-\tilde{F}_{RBC_{age}}(t;\theta)$ where $\tilde{F}_{RBC_{age}}(t;\theta)$ is the cdf of the age of blood cells and controlled by parameter θ. At this moment, we assume that θ does not change over time and is an unknown constant.

Now suppose we can measure multiple Hs and Gs. The best case would be one measurement of G and one measurement of H at the same time.

The following equations hold up to a constant and is an approximation of formula 6. (We leave out of the constant glycation rate.)

$$\begin{aligned}
H(0) &= G(0)F(0) + G(-1)F(1) + ... + G(-n)F(n) \\
H(1) &= G(1)F(0) + G(0)F(1) + ... + G(-(n-1))F(n) \\
&... \\
H(n) &= G(n)F(0) + G(n-1)F(1) + ... + G(0)F(n)
\end{aligned}$$

which is essentially a linear system and is equivalent to the matrix form,

$$\begin{bmatrix} & & \\ & G & \\ & & \end{bmatrix} \begin{bmatrix} \\ F \\ \\ \end{bmatrix} = \begin{bmatrix} \\ H \\ \\ \end{bmatrix}$$

where $1 = F(0) \geq F(1) \geq ... \geq F(n) \geq 0$ as a constraint.

The only problem left is to specify the support n of the age distribution. So we can construct a linear programming (LP) with the objective function to minimize the distance between $F(n)$ and 0.

4.2 Specify the linear programming problem

The constraint $1 = F(0) \geq F(1) \geq ... \geq F(n) \geq 0$ can be rewritten as

$$\begin{aligned}
F(0) &= 1 \\
F(0) - F(1) &\geq 0 \\
F(1) - F(2) &\geq 0 \\
&... \\
F(n-1) - F(n) &\geq 0 \\
F(n) &\geq 0
\end{aligned}$$

which is equivalent to the form $AF \geq 0$ and $F(0) = 1$ where

$$A = \begin{bmatrix} 1 & -1 & 0 & ... & 0 \\ 0 & 1 & -1 & 0 & ... \\ ... & & & & \\ 0 & 0 & ... & 0 & 1 \end{bmatrix}.$$

The final constrained problem is

$$\begin{aligned} &\min_F (c^\top F)^2 \\ \text{subject to } &GF = H \\ &AF \geq 0 \\ &F(0) = 1 \end{aligned}$$

where $c = (0, ..., 0, 1)^\top \in \mathbf{R}^n$, $A \in \mathbf{R}^{n \times n}$, $G \in \mathbf{R}^{n \times n}$, $F \in \mathbf{R}^n$, and $H \in \mathbf{R}^n$.

4.3 Simulation

The simulation is constructed using CVX[1]. We can try many different parameter n and choose the one with the smallest objective value as the solution.

All the points are generated randomly by the commands

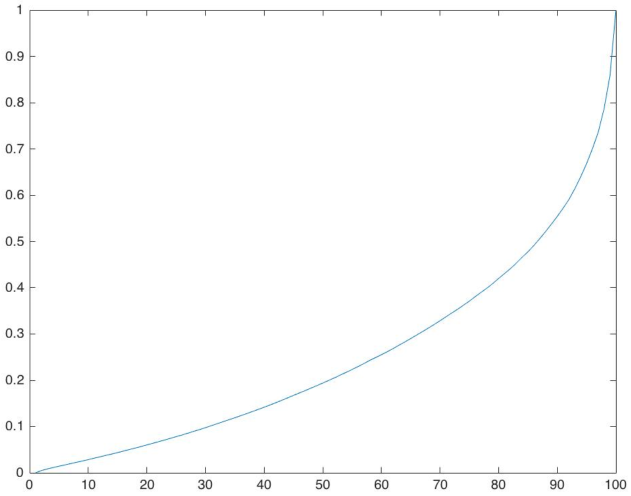
```
G_seq = 100+10*randn(2*n, 1); % G sequences
H = 5+0.1*randn(n,1); % H sequences
```

But maybe due to the random values, the linear system is not consistent. CVX solver gives infeasible solution.

If we relax the constraints by reducing rows of A, i.e., expanding the null space of A, the programming might be solvable but it's hard to tell whether the solution is correct. We still do not know the exact support of the distribution. In fact, we can only assume a support and compute the corresponding distribution.

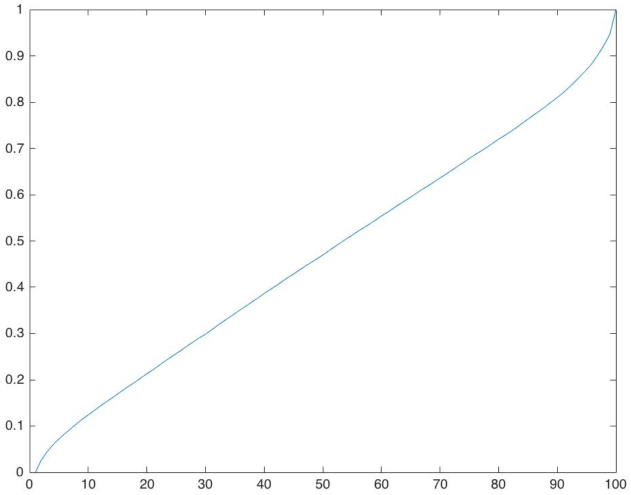
When reducing the number of constraints to 1, the resulting cdf looks like the following,

[1] http://cvxr.com/cvx/doc/basics.html



It is interesting to notice that the cdf is not linear but rather approximately an exponential. But I'm not sure if this suggests anything. If this is true, the assumption of uniform distribution may be invalid.

After doing the correction, meaning that we subtract 1.627 from the $HbA1c$ measurements, the cdf looks very different and varies from different support size n.



5 Recent Progress - April 2

5.1 Redefine the problem

From my point of view, the original formula 6 does not give a direct computation of $HbA1c$, but involves the related factors, i.e., $G(t)$ and $f_{RBC_{age}}(t)$, and how they affect the measurement $HbA1c$ together.

Based on the formula 6 and further assumptions, we try to define a direct computation formula of $HbA1c$.

First define a new factor $v(t) = f(G(t))$, the glycation rate – the added percentage of glycation per day. Specifically, one of its valid units can be %/d.

Thus the new formula that can directly compute $HbA1c$ is

$$HbA1c(t) = \int_0^\infty v(t-a)\left[1 - \int_0^a f_{RBC_{age}}(\tau)\mathrm{d}\tau\right]\mathrm{d}a. \tag{2}$$

5.2 Specify the glycation rate

Assumptions:

1. Suppose the glucose concentration G contributes linearly to the glycation rate v.
2. Suppose the varying glucose concentration does not affect the age of the cells, so that we can measure the age distribution accurately.
3. Suppose the oldest cells' life is 120 days.
4. Suppose the oldest cells reach a 20% level of glycation in 250 mg/dL glucose concentration, which is the inner environment of a person with type 1 diabetes.

The glycation rate can be specified as

$$v(t) = \frac{G(t)20\%}{250\text{mg/dL} \times 120\text{d}}. \tag{3}$$

So the final formula of $HbA1c$ is

$$HbA1c(t) = \int_0^\infty \frac{G(t-a)20\%}{250\text{mg/dL} \times 120\text{d}}\left[1 - \int_0^a f_{RBC_{age}}(\tau)\mathrm{d}\tau\right]\mathrm{d}a. \tag{4}$$

5.3 Simulations

Suppose the $G(t)$ is circulant, i.e., we extend the measurement sequence periodically due to the practical limitation. The simulation sequence is collected from PhysioNet[2].

Here we use $G[n]$ =[135, 140, 169, 215, 224, 201, 265, 252, 332, 325, 296, 240, 285, 294, 276, 273, 296, 286, 349], where the last entry is the latest measurement. And suppose the measurement interval is 1 day. The $HbA1c$ is computed as 10.5%[3] by prediction from the mean of glucose concentration $\mathbb{E}[G] \approx 255.42$mg/dL.

5.3.1 Uniform distribution

Suppose $f_{RBC_{age}}(t)$ follows from uniform distribution with mean M_{RBC},

$$f_{RBC_{age}}(t) = \frac{1}{2M_{RBC}} \qquad 0 \leq t \leq 2M_{RBC}.$$

For approximation, through binning the domain of $f_{RBC_{age}}(t)$, we have

$$\sum_{n=0}^{M_{RBC}-1} \tilde{v}(n) \times 1 \times \frac{2M_{RBC} - n}{2M_{RBC}} = HbA1c \tag{5}$$

where $\tilde{v}(n) = \frac{\tilde{G}(n)20\%}{250\text{mg/dL}\times 120\text{d}}$ and $\tilde{G}(n) = G(-n)$.

This results in $M_{RBC} \approx 58$ days, thus showing

$$f_{RBC_{age}}(t) = \frac{1}{116} \qquad 0 \leq t \leq 116.$$

Because we have the similar assumed value with G, the result is no wonder related to the oldest age of cells.

NOTE: under this distribution assumption, there is only one parameter that needs to be determined, so only one measurement of $HbA1c$ is needed.

[2] https://physionet.org/physiobank/database/mimic2cdb-ps/s20794/#234

[3] https://www.uptodate.com/contents/calculator-glycemic-assessment-using-conventional-or-si-units-for-hemoglobin-a1c

6 Questions - March 30

1. Should the G term in the formula be $G(t-a)$ instead of $G(-a)$ in the original word text?

$$HbA1c(t) = \int_0^\infty G(t-a)\left[1 - \int_0^a f_{RBC_{age}}(\tau)\mathrm{d}\tau\right]\mathrm{d}a \qquad (6)$$

2. Is this formula equal to the practical measurement up to a constant? If yes, so how do we deal with this constant? (Suppose G is a constant, and M is the mean of a uniform distribution, then H = GM where the true mean $M_{age} = \frac{M}{\alpha}$. Ideally set the maximum of the support to be 120 days and then do a scale).

3. Can we use monte carlo integration to solve this problem? Is there any relationship between monte carlo integration with sampling theorem?

4. How do we define the measurement at t=0, i.e., HbA1c(0)?

5. Is this formula equal to the practical measurement up to a constant?

6. When we mention to "measure multiple HbA1c", are we referring to multiple measurements on the same blood sample or multiple measurements on different blood samples?

7. At this time, if we take G(t) as a constant and assume the shape of f, it seems solvable for uniform assumption and finite exponential assumption. But in practice, because both G and f are changing, and every time there are new signals entering, I can not see the process of taking deconvolution method, can you give me some more hints on this method? Or do we need to take other assumptions?