Contents

VAE-CME

Xinyi Zhou

2023.6.23

1 Birth Death Model

Consider a simple non-Markovian system where molecules are produced at a rate ρ and are removed from the system (degraded) after a fxed time delay τ:

$$\emptyset \xrightarrow{\rho} N, N \overset{\tau}{\Rightarrow} \emptyset \tag{1}$$

The training set is the distribution from 1×10^4 samples using the SSA.In the experiment, we assume $\rho = 20$, $\tau = 10$ and truncation $N = 271$.

Both encoder and decoder are multilayer perceptron with one hidden layer. The objective function is chosen as the sum of mean-squared-error and KL divergence. For the training we used the standard adaptive moment estimation algorithm (ADAM). The weight of mean-squared-error needs to be increased, and the learning rate needs to be decreased from approximately 0.25 to 0.01 during the training process. The fitting performance after training is shown in Fig. 1.

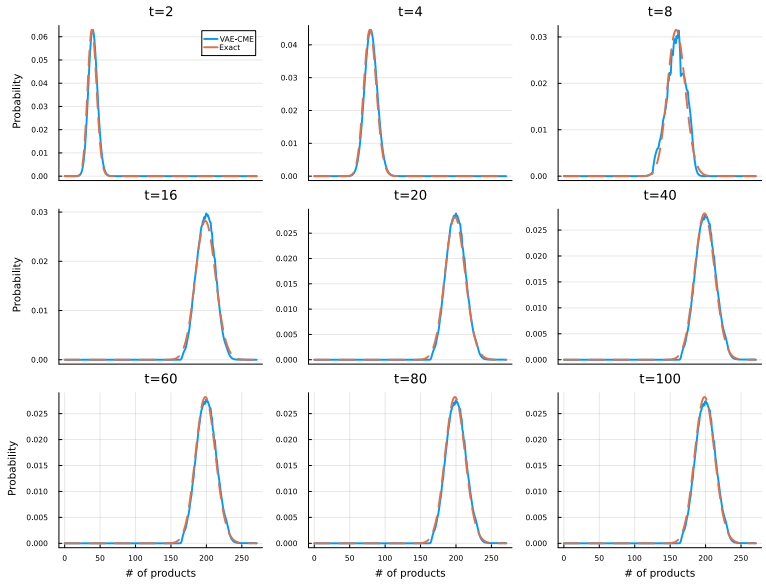

Figure 1: Birth Death Model Fitting

2 On Off Model

We also consider On Off Model wherein the promoter switches between an active and inactive state, RNAP binding occurs only in the active state, which is followed by delayed degradation modelling the RNAP movement along the gene and its detachment; this can be described by the reaction scheme:

$$G^* \xrightarrow{k_{\text{on}}} G, G \xrightarrow{k_{\text{off}}} G^*, G \xrightarrow{\rho} G + N. \tag{2}$$

The same as Birth Death Model, The training set is the distribution from 1×10^4 samples using the SSA.In the experiment, we assume $k_{\text{on}} = 1$, $k_{\text{off}} = 1$, $\rho = 20$, and truncation $N = 45$.

The same objective function, optimizer, and adjustments to weights and learning rate are used for training as in the Birth Death Model. The fitting performance after training is shown in Fig. 2.

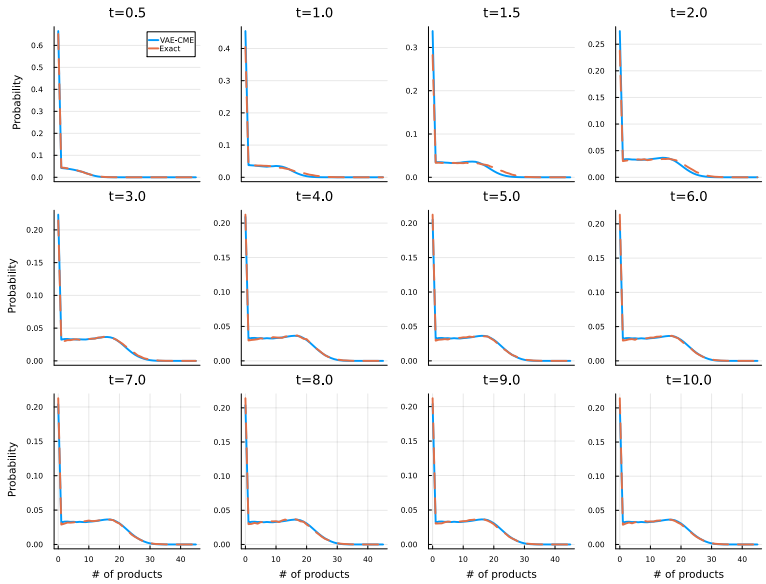

Figure 2: On Off Model fitting

3 Bursty Model

We consider Bursty Model, which is the same as Birth Death Model, except that the binding of RNAPs to the promoter occurs in bursts whose size i is distributed according to the geometric distribution $b^i/(1+b)^{i+1}$; this can be described by the reaction scheme:

$$\begin{aligned} \emptyset &\xrightarrow{\frac{\alpha b^i}{(1+b)^{i+1}}} iN, i = 1, 2, 3, ... \\ N &\overset{\tau}{\Rightarrow} \emptyset \end{aligned} \tag{3}$$

To achieve the best fitting performance, the analytical solution of the Bursty model's probability distribution (See SI in [1]) is used as the training set. In the experiment, we assume $\alpha = 0.0282$, $b = 3.46$, $\tau = 120$ and truncation $N = 64$.

The same as before, we choose the sum of mean-squared-error and KL divergenc as the objective function, ADAM as the optimizer. The weight of mean-squared-error needs to be increased, and the learning rate needs to be decreased. The fitting performance after training is shown in Fig. 3

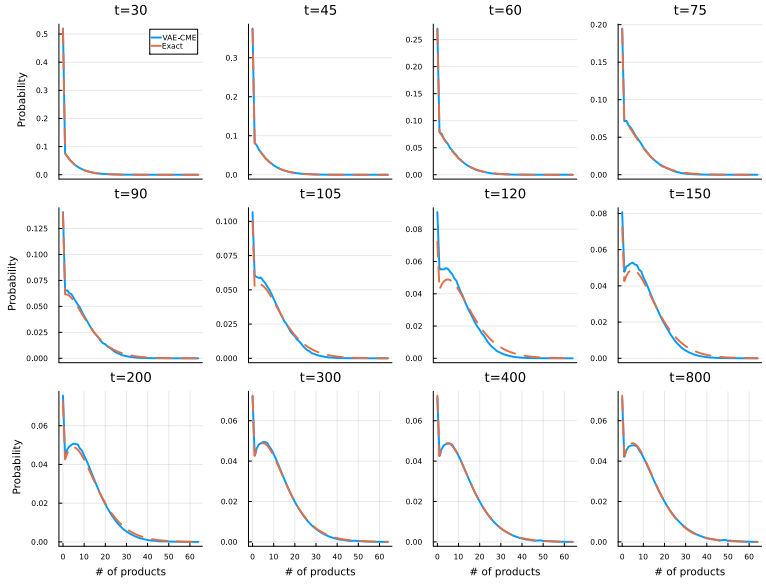

Figure 3: Bursty Model Fitting

3.1 Control Mean of time delay τ

We have designed one encoder and two decoders. The two decoders are designed to reconstruct the probability distributions of $\tau = 30$ and $\tau = 120$, respectively. The two decoders share the same set of parameters. An additional node, which we call Attribute in this paper, is added to the input layer of the decoder (i.e., the latent space). For the decoder corresponding to $\tau = 120$, the value of the additional node is set to 1. For the decoder corresponding to $\tau = 30$, the value of the additional node is set to 0. The objective function ischosen as the sum of mean-squared-error between the outputs of the two decoders and the probability distributions obtained for Bursty Model of $\tau = 30$ and $\tau = 120$ and KL divergence. To achieve the best fitting performance, we choose the analytical solution of the Bursty model's probability distributions as the training set. The fitting performance after training is shown in Figs. 4-5. Inspired by the set of biases in the output layer, which is

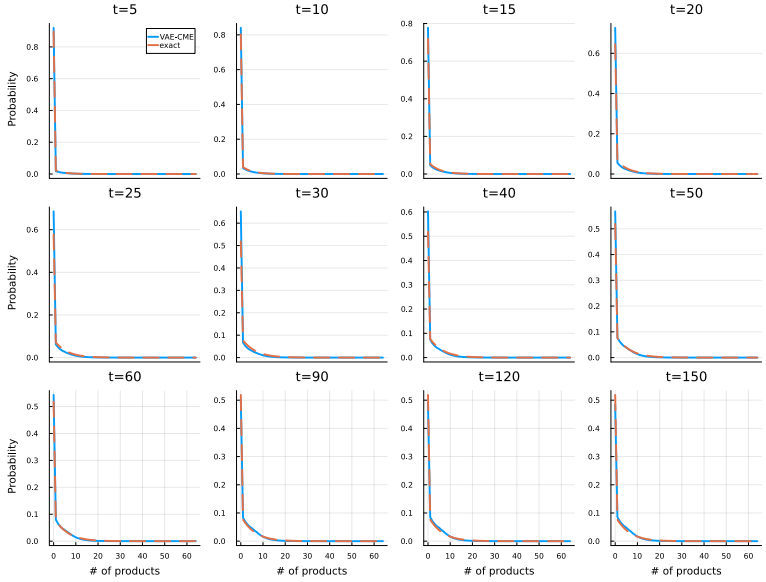

Figure 4: Bursty Model Fitting $\tau = 30$

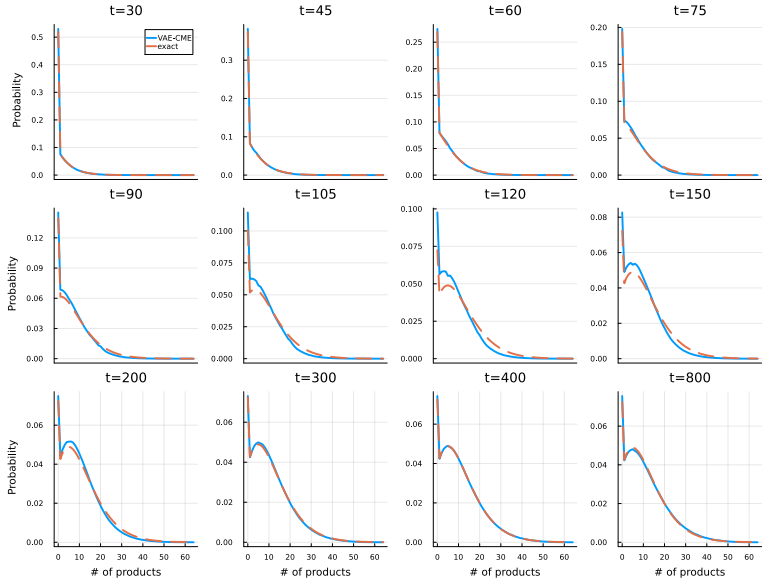

Figure 5: Bursty Model Fitting $\tau = 120$

$r_n = \frac{n}{\tau}$ [1], $n = 1, ..., N$, we found that there is a proportional relationship between the Attribute and $1/\tau$(See Fig. 6). Following this proportional relationship, it is possible to predict the probability distributions obtained by the Bursty Model for any elongation time τ between 30 and 120. The predicting performance is shown in Figs. 7-9

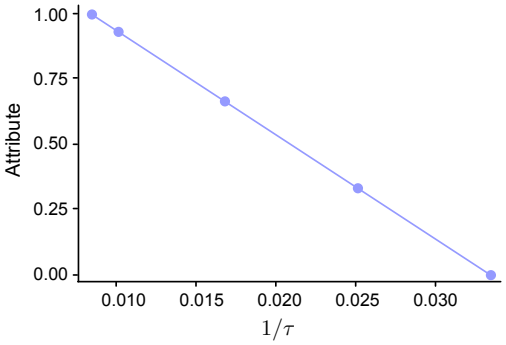

Figure 6: $1/\tau$ Attribute

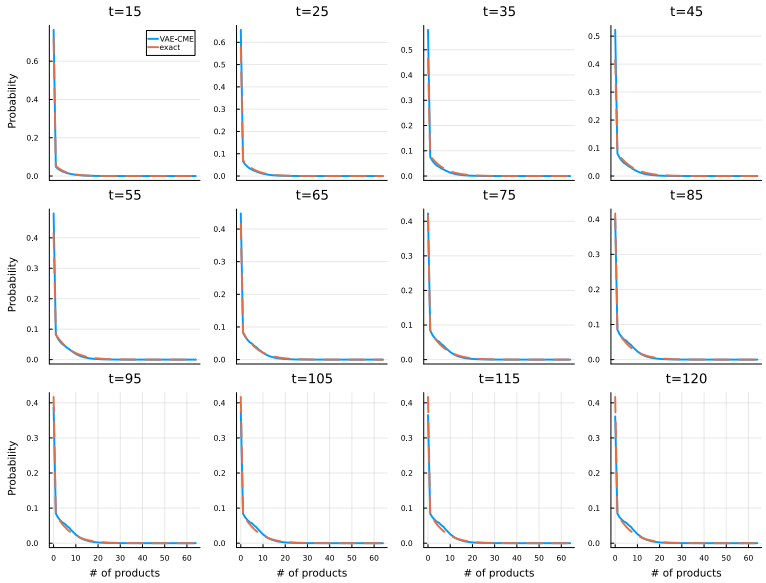

Figure 7: Bursty Model Predicting $\tau = 40$

3.2 Control Variance of time delay τ

When the elongation time τ is a random variable sampled from a distribution, VAE alse works very well. Here we assume τ follows a two-point distribution, it means that it can take only two discrete values,τ_1 and τ_2 with probabilities p and $1-p$, respectively. In the experiment, we assume $\alpha = 0.0282$, $b = 3.46$, $p = 0.6$. We obtain five sets of data by different values of τ (See Table. 1). Note that the mean value of τ is 120 across these five sets of data.

dataset	τ_1	τ_2	Variance
dataset #1	L/3	L	4266
dataset #2	3L/10	1.05L	5400
dataset #3	L/4	1.125L	7350
dataset #4	L/5	1.2L	9600
dataset #5	L/6	1.25L	11266

Table 1: Value of τ, $L = 200$

We use dataset #1 and dataset #5 as the training sets, and the remaining datasets will be used as the prediction sets. The same as before, the two decoders are designed to reconstruct the probability distributions of dataset #1 and dataset #5, the corresponding attribute values are 0 and 1, respectively. Figs 10-11 show the fitting performance of dataset #1 and dataset #5.

We found that there is a proportional relationship between the Attribute and T_1 or T_2 (See Fig. 12). Figs 13-15 show the predicting performance of the remaining datasets.

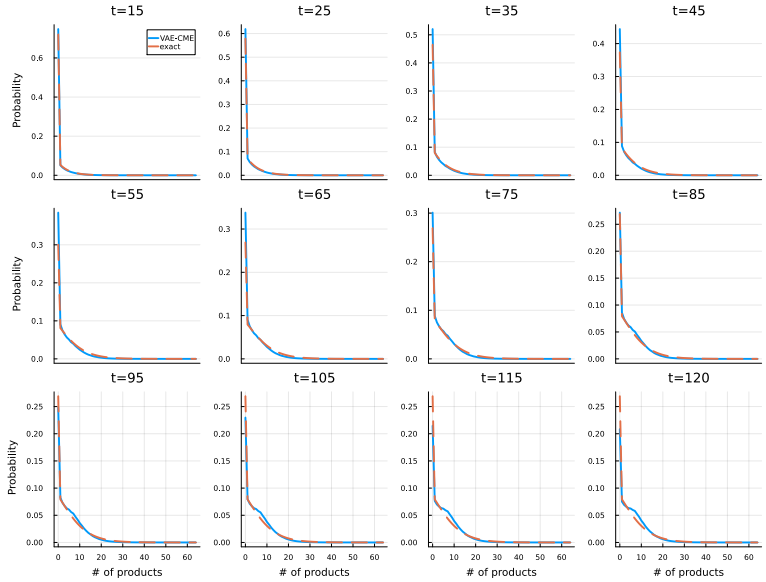

Figure 8: Bursty Model Predicting $\tau = 60$

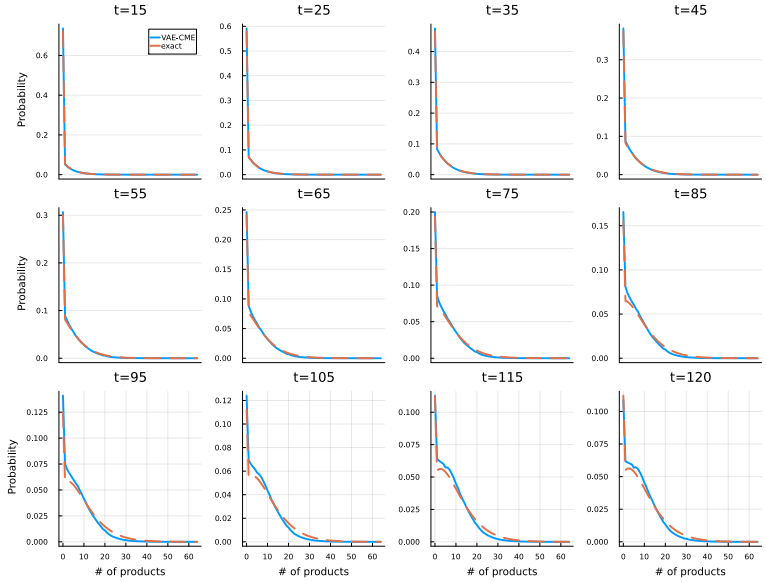

Figure 9: Bursty Model Predicting $\tau = 100$

4 Oscillation Model

Here we consider Oscillation Model, which is a simple genetic negative feedback loop whereby (i) a protein X is transcribed by a promoter, (ii) subsequently after a fixed time delay τ, X turns (via some set of unspeci"ed biochemical processes) into a protein Y and (iii) finally Y binds the promoter and reduces the rate of transcription of X. This can be described by the reaction scheme:

$$\emptyset \xrightarrow{J_1(Y)} X, X \overset{\tau}{\Rightarrow} Y, Y \xrightarrow{J_2(Y)} \emptyset. \tag{4}$$

The function $J_1(Y)$ and $J_2(Y)$ is defined as follows:

$$\begin{aligned} J_1(Y) &= k_1 S \frac{K_d^p}{K_d^p + Y^p}, \\ J_2(Y) &= k_2 E_T \frac{Y}{K_m + Y}. \end{aligned} \tag{5}$$

In this example, we assume $k_1 = k_2 = S = E_T = K_d = K_m = 1, p = 2$ for simplicity. Unlike Models considered earlier, the delay master equation corresponding to this model has no known analytical solution.

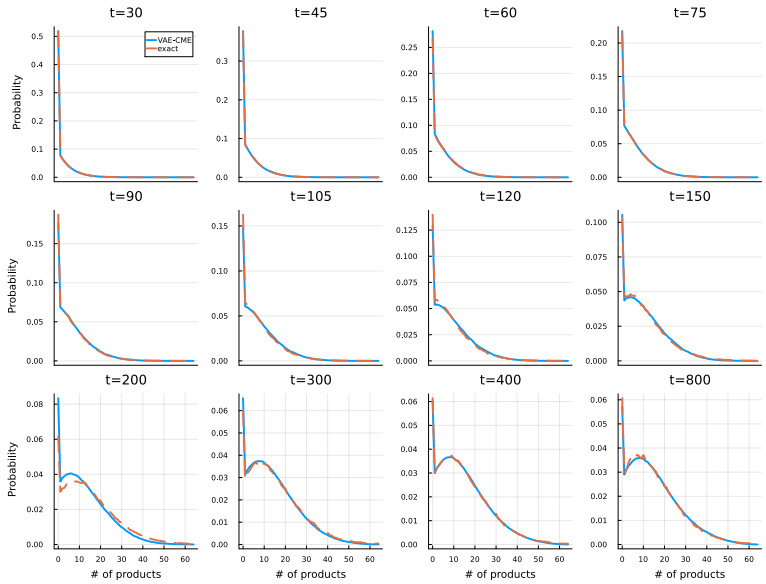

Figure 10: Bursty Model Fitting Variance=4266

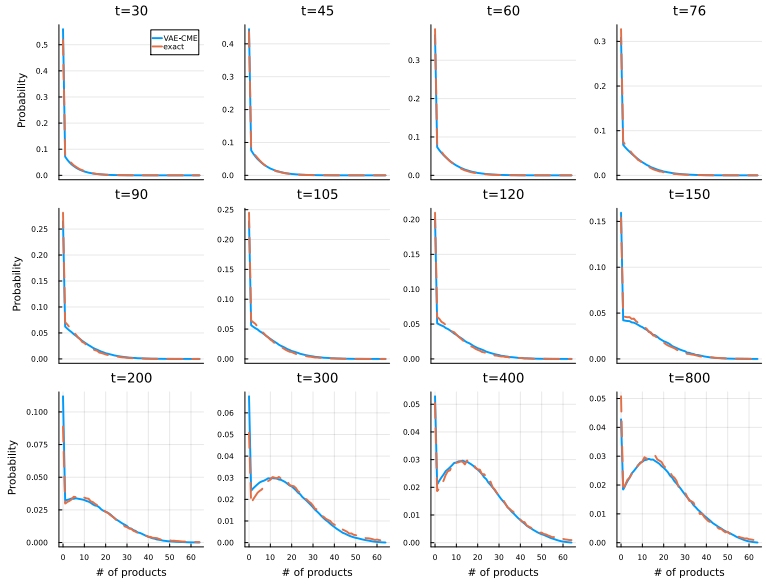

Figure 11: Bursty Model Fitting Variance=11266

The training set is the distribution from 1×10^4 samples using the SSA. Moreover, we assume $\tau = 10$ and truncation $N = 26$.

We only use the simulated trajectories of mature protein Y to train the VAE. In other words, the data of protein X is not required during training. However, an additional regularization term which is the first derivative of the probability distribution of protein X obtained from the chemical master equation (CME), needs to be added to the objective function. Therefore, the final objective function of this experiment is the sum of mean-squared-error of protein Y, KL divergence and regularization term mentioned before. The fitting performance of protein Y after training is shown in Fig. 16 and the predicting performance of protein X is shown in Fig. 17.

4.1 Reducing sample size

Due to the poor predicting performance of MLP [1] on protein X in the Oscillation model, we studied the predicting performance of VAE under small sample conditions. We trained MLP and VAE using probability distributions obtained from simulating SSA for 100, 300, 1000, 3000, and 10000 iterations as training sets, respectively. Since the distribution from 1×10^4 samples using the SSA closely approximate the true probability distribution, we consider it as exact distributions. The measure of the accuracy is the mean squared error between the network(MLP or VAE) and exact distributions.

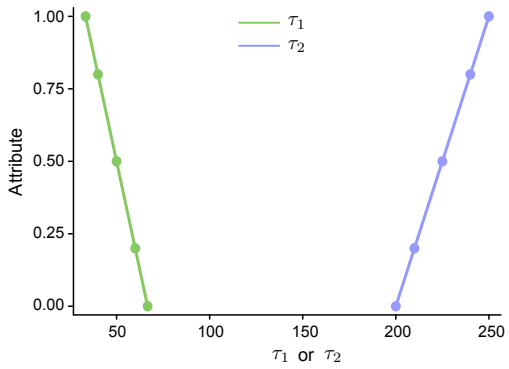

Figure 12: τ_1 or τ_2 Attribute

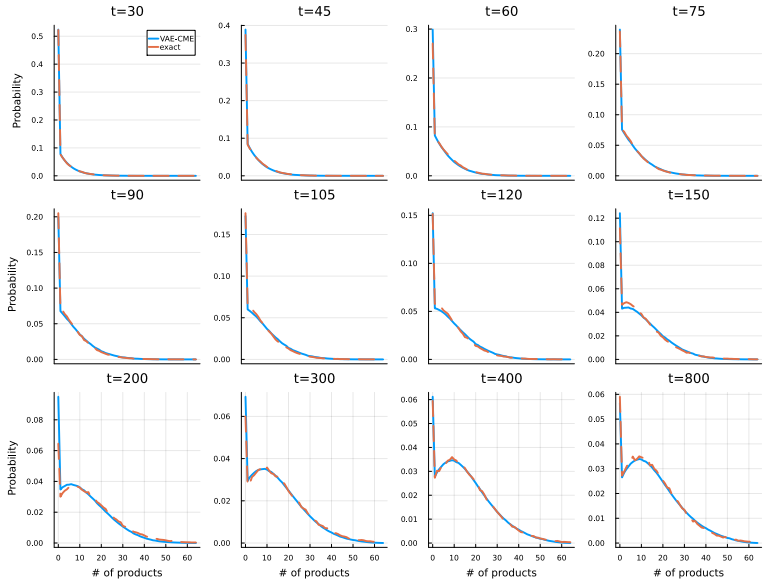

Figure 13: Bursty Model Predicting Variance=5400

We will trained MLP and VAE using three sets of probability distributions obtained from SSA for 100, 300, 1000, 3000 and 10000 iterations respectively. The average mean squared error is calculated and compared(See 18).

Note that the VAE obtained from training with 1×10^2 samples produces a distribution that is as precise as that from 1.6×10^3 samples trained by MLP. In this case the training time of the VAE is also just about $1/3$ of the MLP for every epoch.

5 Exact solution for variable time delay τ

We consider the exact solution that the elongation time τ is a random variable sampled from an arbitrary distribution.

5.1 Birth Death Model

According to Eq. (1), molecules are produced at a rate ρ and are removed from the system after a fxed time delay τ. Let's assume that after the molecules are produced, they disappear after traveling a distance L with a velocity V. Therefore, for each molecule n_i, its $\tau_i = L/V_i$. Note that whenever the molecule encounters a slower moving molecule it must decrease its speed to that of the slower moving molecule.

We denote

$$G(x) = P(\tau_i \leq x) = P(V_i \geq L/x) \tag{6}$$

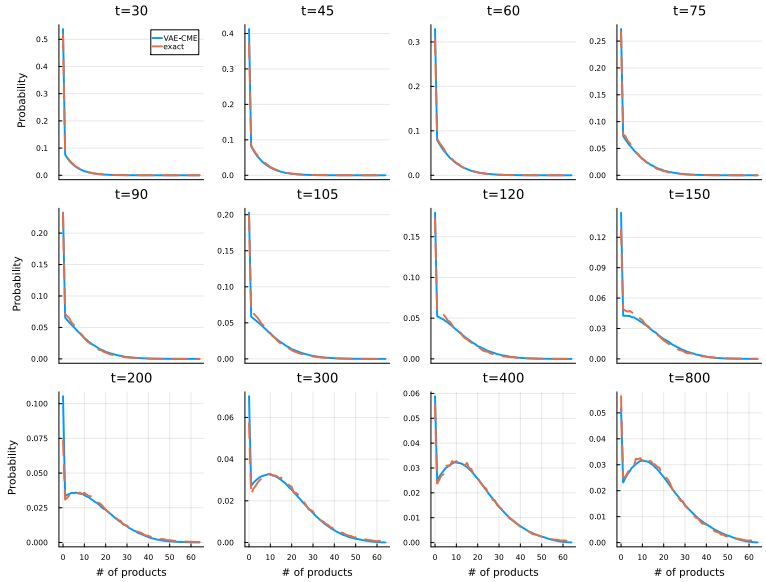

Figure 14: Bursty Model Predicting Variance=7350

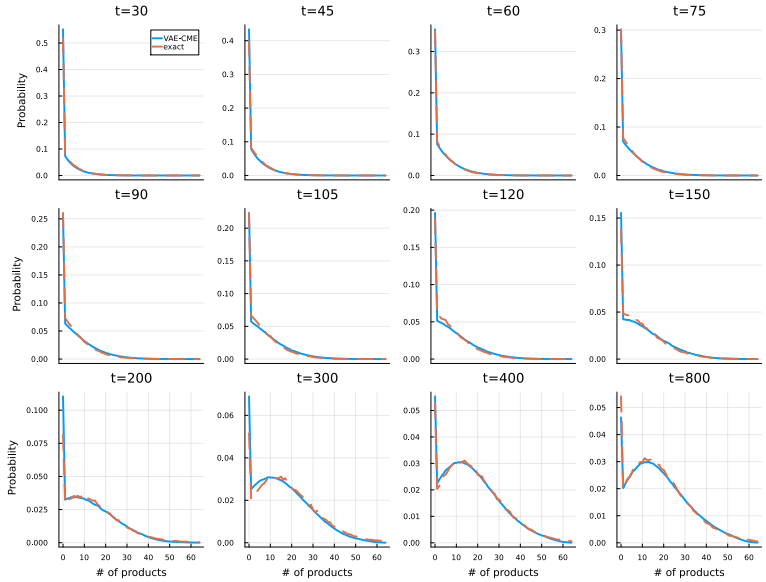

Figure 15: Bursty Model Predicting Variance=9600

according to [2], as for the Birth Death Model, the probability distribution of molecules at time t is given by

$$\begin{aligned} P(N=0) &= e^{-\rho \int_0^t \bar{G}(t-s)ds} = e^{-\rho \int_0^t \bar{G}(u)ds} \\ P(N=n) &= \int_0^t e^{-\rho(t-y)} \frac{(\rho(t-y))^{n-1}}{(n-1)!} \rho \bar{G}(t-y) e^{-\rho \int_0^y \bar{G}(t-s)ds} dy \end{aligned} \tag{7}$$

Note that N is the number of the molecules and $\bar{G}(t-s) = 1 - G(t-s)$.

5.2 Bursty Model

According to Eq. (3), where α stands for the burst frequency and b is the mean burst size. We are interested in the distribution of nascent RNA number $Y(t)$.Intuitively, $Y(t)$ is determined by two factors- the number of "packages" $I(t)$ (where a package stands for an event occurring before time t such that the nascent RNA produced in these events has still not been subject to delayed degradation) and the number of nascent RNAs $X_i \sim \text{Geom}(\frac{1}{1+b})$ in each package i. Therefore, $Y(t)$ can be written in the form

$$Y(t) = \Sigma_{i=1}^{I(t)} X_i \tag{8}$$

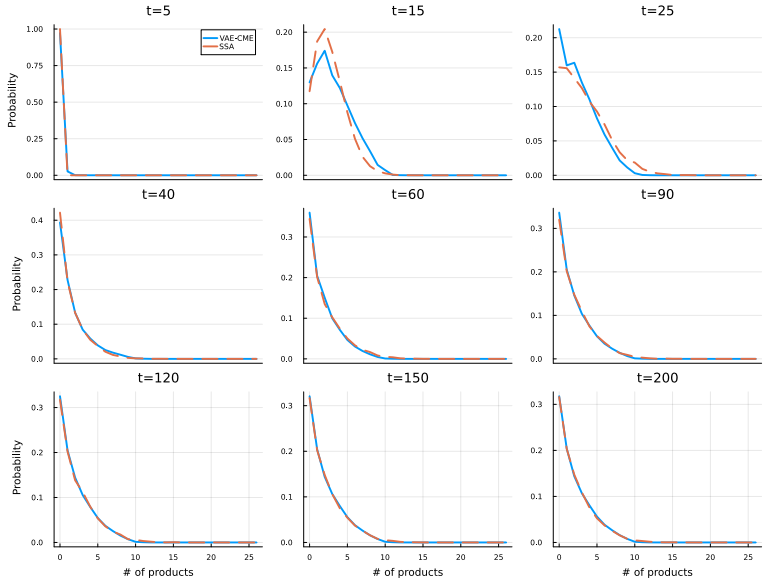

Figure 16: Oscillation Y fitting

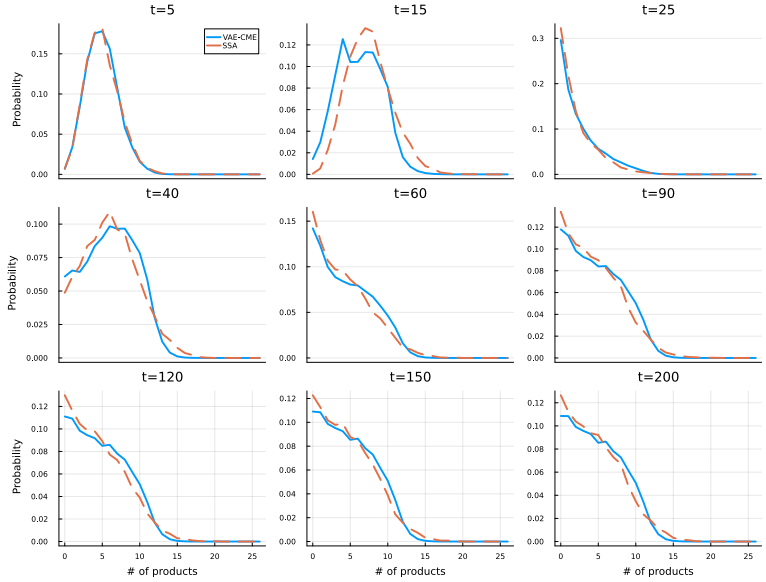

Figure 17: Oscillation X predicting

thereby constituting a compound process. The event number $I(t)$ is determined by the system in Eq. (1). According to Eq. (7), we can obtain the probability distribution of $I(t)$. We first consider $Y(t) > 0$, we can rewrite Eq. (8) as

$$P(Y(t) = n) = \Sigma_{m=1}^{\infty} P(\Sigma_{i=1}^{I(t)} X_i = n | I(t) = m) P(I(t) = m) \tag{9}$$

therefore, we obtain

$$P(Y(t) = n) = \Sigma_{m=1}^{\infty} P(\Sigma_{i=1}^{m} X_i = n) P(I(t) = m) \tag{10}$$

Since $X_i \sim Geo(\frac{1}{1+b})$, we can easily know that $\Sigma_{i=1}^{m} X_i \sim \text{NegativeBinomial}(m, 1/(1+b))$. Here we note that $T \sim \text{NegativeBinomial}(m, 1/(1+b))$, and then we can simplify Eq. (10)

$$P(Y(t) = n) = \Sigma_{m=1}^{\infty} P(T = n) P(I(t) = m), n > 0 \tag{11}$$

Now let's consider $Y(t) = 0$,

$$P(Y(t) = 0) = P(I(t) = 0) + \Sigma_{m=1}^{\infty} P(T = 0) P(I(t) = m), \tag{12}$$

Eq. (11)(12) are equal to

$$\begin{aligned} P(Y(t) = 0) &= P(I(t) = 0) + \Sigma_{m=1}^{\infty} \theta^m P(I(t) = m) = \Sigma_{m=0}^{\infty} \theta^m P(I(t) = m) \\ P(Y(t) = n) &= \Sigma_{m=1}^{\infty} C_{n+m-1}^{n} \theta^m (1-\theta)^n P(I(t) = m), n > 0 \end{aligned} \tag{13}$$

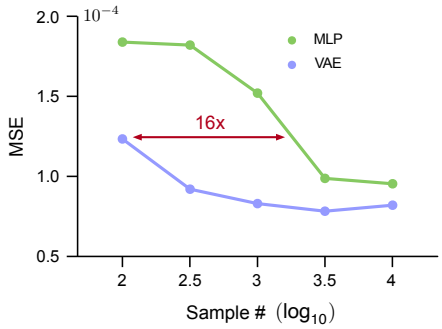

Figure 18: MSE_X VAE and MLP

Note that $\theta = \frac{1}{1+b}$.

References

[1] Qingchao Jiang, Xiaoming Fu, Shifu Yan, Runlai Li, Wenli Du, Zhixing Cao, Feng Qian, and Ramon Grima. Neural network aided approximation and parameter inference of non-markovian models of gene expression. Nature communications, 12(1):2618, 2021.

[2] Sheldon M Ross. Introduction to probability models. Academic press, 2014.