STATIONARITY AND INFERENCE IN MULTISTATE PROMOTER MODELS OF STOCHASTIC GENE EXPRESSION VIA STICK-BREAKING MEASURES[*]

WILLIAM LIPPITT[†], SUNDER SETHURAMAN[‡], AND XUEYING TANG[‡]

Abstract. In a general stochastic multistate promoter model of dynamic mRNA/protein interactions, we identify the stationary joint distribution of the promoter state, mRNA, and protein levels through an explicit 'stick-breaking' construction perhaps of interest in itself. This derivation is a constructive advance over previous work where the stationary distribution is solved only in restricted cases. Moreover, the stick-breaking construction allows to sample directly from the stationary distribution, permitting inference procedures and model selection. In this context, we discuss numerical Bayesian experiments to illustrate the results.

Key words. multistate, promoter, mRNA, protein, Bayesian, inference, model validation, stick-breaking, Dirichlet, Markovian, stationary distribution, constructive

MSC codes. 92Bxx, 37N25, 62P10, 62E15

1. Introduction.

Relatively recent models of mRNA creation and degradation in cells incorporate the notion of a stochastic 'promoter' which influences birth rates and serves as a surrogate to the complex underlying structure of chemical reactions. In such models, the stationary distribution of mRNA levels is of interest, given in particular that now readings from cells can be taken.

The multistate promoter process is a more involved model than the simple birth-death process with constant rates in which the evolution is somewhat regular and the stationary distribution is Poisson. In particular, observations in types of cells indicate that the production of mRNA in the multistate process can be 'bursty' and the levels of mRNA in stationarity can have heavy, non-Poissonian tails [1], [18] and references therein. In this respect, the multistate mRNA process can reproduce both phenomena, and is now receiving much attention as a possible complex yet tractable model [5], [6], [17], [19], [30], and references therein.

The general multistate promoter process is a pair evolution (E, M) where the state of the promoter $E \in \mathfrak{X}$ belongs to discrete finite or countably infinite set $\mathfrak{X}$ and the level of mRNA $M \in \{0, 1, 2, \ldots\}$ is a nonnegative integer. The dynamics of the pair is that the promoter $E = i$ switches to a different state $E = j$ with rate $G_{i,j} \geq 0$ for $i \neq j$, not dependent on M. On the other hand, when $E = i$, the birth rate of M to $M + 1$ is $\beta_E = \beta_i \geq 0$ and the death rate of M to $M - 1$ is δM, proportional to M with $\delta > 0$, degradation modeled not dependent on the state E. The parameters $\beta = \{\beta_i\}_{i \in \mathfrak{X}}$ and 'generator' $G = \{G_{i,j}\}_{i,j \in \mathfrak{X}}$, where $G_{i,i} = -\sum_{j \neq i} G_{i,j}$, completely specify the process; see Figure 1 for a chemical reaction representation.

In [3], [15], [20], [25], the stationary distribution for mRNA levels M is identified for multistate processes when $\mathfrak{X} = \{1, 2\}$ as a scaled Beta-Poisson mixture. More generally, in [17], when $\mathfrak{X} = \{1, 2, \ldots, n\}$ and the generator G is such that $G_{i,j} = \alpha_j$ is independent of $i \neq j$, it is shown that the stationary distribution is a scaled Dirichlet-

[*] Submitted to the editors DATE.

Funding: This work was partially supported by the Army Research Office grant ARO-W911NF-18-1-0311.

[†] Department of Biostatistics & Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO (william.lippitt@cuanschutz.edu).

[‡] Department of Mathematics, University of Arizona, Tucson, AZ (sethuram@math.arizona.edu, xytang@math.arizona.edu).

$$\begin{cases} S_i \xrightarrow{G_{ij}} S_j & \text{for } i, j \in \mathfrak{X};\ i \neq j \\ S_i \xrightarrow{\beta_i} S_i + M & \text{for } i \in \mathfrak{X} \\ M \xrightarrow{\delta} \emptyset & \\ M \xrightarrow{\alpha} M + P & \\ P \xrightarrow{\gamma} \emptyset & \end{cases}$$

FIG. 1. *Above is a representation in chemical equations for the multi-state promoter process with protein. Promoter states are represented by chemical species $S_i : i \in \mathfrak{X}$ with transitions between states such that molecule numbers always satisfy $[S_i] \in \{0, 1\}$ for all i and $\sum_i [S_i] = 1$; see [17]. Representing mRNA by a species M and protein by a species P, the promoter process $(E(t), M(t), P(t))$ is determined by the above elementary equations with $E(t) = i$ when $[S_i] = 1$, and $(M(t), P(t)) = ([M], [P])$. We note that the multi-state mRNA process $(E(t), M(t))$, without protein, is well-defined.*

Poisson mixture. In the 'refractory' case when $\beta_i > 0$ for exactly one $i \in \mathfrak{X} = \{1, 2, \ldots, n\}$ and G is now allowed to be general where $G_{i,j}$ may depend on i, [17] derives a scaled Beta product-Poisson mixture. In [29], a general hypergeometric formula is given for general G generators. Although a certain generating function of the stationary distribution in the general model is known to satisfy a PDE in terms of parameters [17], a constructive solution for the stationary distribution is not known for the general multistate promoter process.

1.1. Prospectus.

In this context, the first aim of this article is to consider a multistate promoter process with general class creation rates β and promoter switching rates G in the context of 'stick-breaking'. We identify an explicit form of the stationary distribution of (E, M) in terms of a 'Markovian stick-breaking' mixture distribution, reminiscent of the stick-breaking form of the Dirichlet process used much in nonparametric Bayesian statistics. We also state formulas for certain moments aided by the 'stick-breaking' formulation. That 'stick-breaking' would be involved in such a characterization in this mRNA dynamics context was unexpected. When G is such that $G_{i,j}$ does not depend on i for $j \neq i$, there is a sense of independence in the promotor switches which in retrospect suggests that the Dirichlet process may be involved. It turns out the appropriate generalization, which takes into account the time spent on each promoter state before switching, when $G_{i,j}$'s depend on i, is the 'Markovian Stick-Breaking Measure' discussed in the next section.

A second aim is to conduct Bayesian statistical inference based on synthetic stationarily distributed data, with a future view toward inference with respect to laboratory biological mRNA data. This has been considered in the literature for certain multistate models [18] and references therein; see also [22]. In this respect, we exploit the stick-breaking form of the stationary distribution to perform the inference which seems to allow for good computation and error bounds. Bayesian inference is specifically considered as it allows straightforward uncertainty quantification.

Still a third aim of this work is to extend results from the multistate mRNA model to a general multistate mRNA model with *protein* interactions, namely a process $(E, M, P) \in \mathfrak{X} \times \{0, 1, 2, \ldots\}^2$ where (E, M) rates are the same and the rates of P to $P + 1$ is αM and P to $P - 1$ is γP where $\alpha, \gamma > 0$; see Figure 1 for a representation in terms of chemical reactions. We mention, the identification of the stationary distribution in such a model, was posed as an open problem in [17]. Recently, in [4], protein interactions have been considered in the 'refractory' case where $\beta_i > 0$ for exactly one state $E = i$ in terms of Pólya urn models. We mention also previous work on on-off promoter models [28]. We identify here in the general setting the stationary distribution of (E, M, P) in terms of the 'stick-breaking' apparatus, in particular interestingly 'clumped' versions, and discuss computation of moments.

We now amplify and detail more the considerations in these three aims after a brief discussion of related papers in the literature.

1.2. Related literature. The identification of the stationary distribution of (E, M) connects interestingly to disparate models of inhomogeneous Markov chains of their own interest.

First, by a Poisson representation introduced in [10] for chemical reaction models, one can associate a piecewise deterministic Markov process X where X satisfies an ODE depending on the process E. Here, X represents a mass action kinetics process with respect to the levels of the promoter 'chemicals'. Then, in [17], it is shown at stationarity that $M|(E, X) \sim \text{Poisson}(\beta \cdot X)$. The question of identification becomes now one of the distribution of (E, X); see Theorem 3.6.

Next, consider a discrete time inhomogeneous Markov chain $\{M_n\}_{n\geq 0}$ on $\mathfrak{X}$ where the transition kernel from time n to $n + 1$ is $(I + G/n)$. In a sense, $\{M_n\}_{n\geq 0}$ is a discrete time version of the process E. In [8] (see also [9]), the limit of the empirical distribution of this chain is identified as a Markov stick-breaking measure ν_G; see Theorem 3.4.

Finally, [2], in the study of 'freezing MC's', generalizing those in [9], considered the piecewise deterministic Markov process (E, X) (although its connection with mRNA dynamics was perhaps not known). They showed at stationarity, X has the law of the empirical distribution limit of the chain $\{M_n\}_{n\geq 0}$, among other results; see Theorem 3.5.

1.3. Aim 1. These results from the literature form the basis of 'Aim 1', and have natural combination, leading to identification of the desired stationary distribution of (E, M) as a Poisson mixture (Theorem 3.7). We also give some moment computations in Section 3.1. Then, the notion of the 'identifiability' with respect to the mRNA level M is discussed in Section 3.2.

We comment that our approach through stick-breaking distribution relations differs from the generating function/differential equations one in [17]. Using generating functions, marginal distributions of X could be identified as Beta-products in [17], from which the refractory case solved. The stick-breaking construction allows sampling from the full distribution of X in the general setting. In a nutshell, the stick-breaking approach captures the time-averaged frequencies of the promoter states; see [8] for more discussion of the construction. Nevertheless, both approaches take advantage of a Poisson representation of the multistate promoter dynamics [10], as well as the introduction of a piecewise deterministic Markov process.

The introduction of stick-breaking constructions to the study of the multi-state promoter process however supports the development of statistical tools and the extension of the model to include protein in the remaining Aims.

1.4. Aim 2. In Section 5, we discuss parameter estimation and model selection under a Bayesian framework. We assume that observed data come from the stationary distribution of the associated mRNA model with unknown parameters (β, G). In Section 5.1, we describe a Gibbs sampler to draw samples from the posterior distribution. Empirical posterior means and empirical credible intervals are respectively used as the point estimators and interval estimators of (β, G). Given the estimated parameters from several candidate models with different number of states or different sparsity structures in (β, G), we discuss in Section 5.2 how to use Bayesian Information Criterion (BIC) to select the model underlying the observed data.

The key step in both inference tasks is to evaluate the likelihood function (the probability of observing the data) which are not in a closed-form. We utilize truncations of the stick-breaking form of the stationary distribution to approximate the likelihood with Monte Carlo simulations. The discussed procedures in Section 5.1 and Section 5.2 are applied to synthetic datasets with various choices of (β, G) with $|\mathfrak{X}| = 2$ or 3 for empirical performance evaluation. In our experiments, the model parameters can be estimated accurately and the underlying models can be selected correctly with high probability when the sample size is large.

1.5. Aim 3. The protein model (E, M, P) mentioned earlier can be analyzed by writing the interactions in terms of $\tilde{E} = (E, M)$ and $\tilde{M} = P$, where now $\tilde{E}$ is in the role of being a 'promoter' with respect to protein levels $\tilde{M}$. Since the promoter state space $\mathfrak{X} \times \{0, 1, 2, \ldots\}$ is not finite, direct application of results in [8], [2] will not be possible as the transition operator $(I - \tilde{G}/n)$ will not be stochastic, that is $\tilde{G}$ with respect to $\tilde{E}$ transitions will not be bounded. The idea however in Section 4 is to approximate the infinite promoter space $\tilde{E}$ by finite spaces $\tilde{E}^c = \mathfrak{X} \times \{0, 1, \ldots, c\}$. Then, we take limits as $c \uparrow \infty$ of 'clumped' stick-breaking representations of the associated stationary distributions introduced in [8], perhaps of interest in itself (Theorem 4.4).

1.6. Outline. The plan of the paper is to introduce notation and definitions of stick-breaking measures and their 'clumped' forms in Section 2. In Section 3, we discuss the relationship between certain time-inhomogeneous Markov chains, stick-breaking measures, piecewise deterministic Markov processes and multistate mRNA promoter models and formulate Theorem 3.7; in Section 3.1, some moments are computed, and in Section 3.2, identifiability of parameters is discussed. In Section 4, we discuss models which incorporate protein interactions and state Theorem 4.4 which is then shown in Section 6. In Section 5, we discuss how to utilize the stick-breaking constructions to estimate model parameters based on data from the stationary distribution (Section 5.1) and how to perform model selection (Section 5.2). Then, in Section 7, we conclude.

2. Stick-breaking measure representations and other definitions.

We first introduce notation on spaces and matrices used throughout the article in Section 2.1, before defining the notion of a 'stick-breaking' measure and related ingredients in Section 2.2. In Section 2.3, we discuss the notion of a 'clumped' representation of the stick-breaking measure which will be useful in the later discussion of protein interactions.

2.1. Notation on spaces and conventions. We will concentrate on finite spaces $\mathfrak{X}$, enumerated as $\mathfrak{X} = \{1, 2, \ldots, |\mathfrak{X}|\}$. Denote the space of probability measures on $\mathfrak{X}$ by

$$\Delta_{\mathfrak{X}} = \left\{ (p_i)_{i \in \mathfrak{X}} \in [0,1]^{\mathfrak{X}} : \sum_{i \in \mathfrak{X}} p_i = 1 \right\}.$$

Define also that a *generator matrix* G on $\mathfrak{X}$ is the square matrix or operator $G = (G_{i,j})_{i,j \in \mathfrak{X}}$ such that $G_{i,j} \geq 0$ when $i \neq j$ and $\sum_j G_{i,j} = 0$ for each $i \in \mathfrak{X}$. We say that G is an *irreducible* generator matrix when for each pair $(i, j) \in \mathfrak{X}^2$ there is a power $k = k_{i,j}$ such that $(G^k)_{i,j} > 0$. We say G has a *stationary* distribution $\mu \in \Delta_{\mathfrak{X}}$

when μ is a left eigenvector with eigenvalue 0, that is $\sum_i \mu_i G_{i,j} = 0$ for all $j \in \mathfrak{X}$. When G is irreducible, it has a unique stationary distribution μ.

We remark that the generator matrix can always be (non-uniquely) decomposed as $\theta(Q - I)$ where $\theta > 0$ and Q is a stochastic matrix or operator. When G is irreducible, then Q is irreducible and additionally G and Q have the same stationary probability vector(s) μ (independent of the choice of θ).

We now enumerate several conventions used throughout the article.

- If $v \in \mathbb{R}^{\mathfrak{X}}$, then $D(v)$ denotes a square diagonal matrix or operator over $\mathfrak{X}$ whose ith entry is v_i for each $i \in \mathfrak{X}$. If $A \subset \mathfrak{X}$, then $D(A) = D(v)$ where $v = \sum_{i \in A} e_i$ where $\{e_i\}_{i \in \mathfrak{X}}$ is the standard basis of $\mathbb{R}^{\mathfrak{X}}$.
- $\mathbb{N} = \{1, 2, 3, ...\}$ and $\mathbb{N}_0 = \{0, 1, 2, ...\}$
- We define empty sums $\sum_{\emptyset} = 0$, empty scalar products $\prod_{\emptyset} = 1$, and empty matrix products as the identity $\prod_{\emptyset} = I$.
- Products: For a collection of matrices $\{M_j\}_{j=1}^k$, we denote the standard forward order product as $\prod_{j=1}^k M_j = M_1 \cdot M_2 \cdots M_k$ and the non-standard reverse order product as $\prod_{j=1}^{k;(R)} M_j = M_k \cdot M_{k-1} \cdots M_1$.
- Adjoints: Given a probability vector μ over $\mathfrak{X}$, we define the adjoint A^* of a square matrix or operator A on $\mathfrak{X}$ with respect to μ by $A^* = D(\mu)^{-1} A^T D(\mu)$. For a generator G with $G = \theta(Q - I)$ having unique stationary distribution, we always understand G^* and Q^* to be adjoints taken with respect to the associated stationary distribution.

2.2. Stick-breaking measures. Before stating a generalization of the Dirichlet process with respect to $\theta > 0$ and a probability vector μ on $\mathfrak{X}$, which will form the backbone of our work, we first define basic notions. Recall that the classic Dirichlet process is a distribution on the space of probability measures on a measurable space with the property that a sample measure ν is such that the joint distribution of $\big(\nu(A_1), \ldots, \nu(A_k)\big)$ is that of a Dirichlet distribution with parameters $\big(\theta\mu(A_1), \ldots, \theta\mu(A_k)\big)$ for finite partitions $\{A_i\}_{i=1}^k$ of the measurable space. On finite spaces $\mathfrak{X}$, as considered here, the Dirichlet process with parameters $\big(\theta, \mu\big)$ has *Dirichlet*(θ, μ) distribution. Here, a sample $\nu \in \Delta_{\mathfrak{X}}$ is a probability measure on $\mathfrak{X}$.

Such a process ν admits a 'stick-breaking' representation involving two ingredients: a GEM residual allocation model as well as an independent sequence of i.i.d. random variables $\{T_i\}_{i \geq 1}$ on $\mathfrak{X}$ with common distribution μ. See [14, 24] for more on stick-breaking measures. The GEM model is defined as follows.

Definition 2.1 (GEM residual allocation model). *Let $(W_j)_{j \geq 1}$ be an iid sequence of* Beta$(1, \theta)$ *variables, and define*

$$P_j = W_j \prod_{i=1}^{j-1} (1 - W_i).$$

Then, $\mathbf{P} = (P_1, P_2, \ldots)$ *is said to have* GEM(θ) *distribution.*

Define now the (random) 'stick-breaking' measure on $\mathfrak{X}$,

$$\nu = \sum_{j \geq 1} P_j \delta_{T_j}.$$

It is well-known that the law of ν is that of the Dirichlet process on $\mathfrak{X}$ with parameters (θ, μ).

We now consider a generalization where $\{T_i\}_{i\geq 1}$ is a stationary Markov chain on $\mathfrak{X}$ with stationary distribution μ. Such a generalization was first considered in [8] in the context of empirical distribution limits of 'simulated annealing' time-inhomogeneous Markov chains.

Definition 2.2 (MSBM(G), MSBMI(G)). *Let G be an irreducible generator matrix over $\mathfrak{X}$, with a unique stationary distribution μ. Let $G = \theta(Q - I)$ be a decomposition of G. Let also $\mathbf{P} \sim \mathrm{GEM}(\theta)$ and let $\mathbf{T}$ be a stationary homogeneous Markov chain independent of $\mathbf{P}$ and having kernel Q with stationary distribution μ. Then, the random measure*

$$\nu_G = \sum_{j\geq 1} P_j \delta_{T_j} \tag{2.1}$$

taking values in $\Delta_{\mathfrak{X}}$ is said to have distribution MSBM(G). *Here, MSBM stands for Markovian stick-breaking measure. The pair (T_1, ν_G) is said to have* MSBMI(G) *distribution (MSBM and Initial), and is denoted (T_1, ν_G) ∼MSBMI(G). Note that here T_1 is distributed according to μ.*

The construction of the 'stick-breaking' object with $MSBM(G)$ distribution given in the above definition is many to one as the choice of decomposition $G = \theta(Q - I)$ is not unique, though the distribution $MSBM(G)$ itself is independent of this choice. Valid choices of decomposition, that is those such that Q is stochastic, are indexed by the selection of θ, namely those θ such that $\theta \geq \theta(G)$ where $\theta(G) = \sup_{i\in\mathfrak{X}} |G_{i,i}|$. Specifically, for each $\theta \geq \theta(G)$, let $\mathbf{P}^\theta$ have $GEM(\theta)$ distribution and let $\mathbf{T}^\theta$ independent of $\mathbf{P}^\theta$ be a stationary Markov chain with transition kernel $Q^\theta = I + G/\theta$. Define

$$\nu^\theta(\,\cdot\,) = \sum_{j=1}^{\infty} P_j^\theta \delta_{T_j^\theta}(\,\cdot\,).$$

Then, each pair (T_1^θ, ν^θ) is a stick-breaking representation of $MSBMI(G)$; $\nu^\theta \overset{d}{=} \nu^{\theta(G)}$ for all $\theta \geq \theta(G)$. Here, $\mathbf{T}^\theta$ is a Markov chain which may repeat, that is it may be that $\mathcal{P}\left(T_j^\theta = T_{j+1}^\theta\right) > 0$. We consider a clumped, that is non-repeating, stick-breaking representation in the next section, which will be relevant to models involving protein. The series in the stick-breaking construction of Definition 2.2 has the fastest rate of convergence when $\theta = \sup_{i\in\mathfrak{X}} |G_{i,i}|$ is smallest.

We remark exactly in the situation when G permits a decomposition $G = \theta(Q-I)$ such that Q is constant stochastic with rows μ, then $MSBM(G) = Dirichlet(\theta, \mu)$; see Theorem 2.16 [8]. In this case, the Markov chain $\{T_i\}_{i\geq 1}$ is i.i.d. since Q is constant stochastic, and the MSBM measure is the Dirichlet process. When G cannot be decomposed in this way, intuitively non-independence of $\{T_i\}_{i\geq 1}$ introduces non-Dirichlet process evaluations. See [8] for more discussion.

Moreover, we note that the stick-breaking construction allows to bound the error in truncating the series. This will be useful for later statistical inference. Indeed, for $m \geq 0$, $\sum_{j\geq m+1} P_j \delta_{T_j}(\mathfrak{X}) \leq \sum_{j\geq m+1} P_j = \prod_{j=1}^m (1 - W_j)$. Since $-\log(1 - W_j) \overset{d}{=} Exp(\theta)$, we have that $-\log \prod_{j=1}^m (1-W_j) \overset{d}{=} Y_m := Gamma(m, \theta)$, where m is the shape and θ is the rate parameter. Then, the chance the error is greater than $0 < \lambda < 1$ is

$$P(\exp(-Y_m) \geq \lambda) = P(A_\lambda \geq m) \tag{2.2}$$

where $A_\lambda \overset{d}{=} Poisson(-\theta \log(\lambda))$.

2.3. Clumped stick-breaking constructions. We now recall a 'clumped' stick-breaking construction using the Markov chain $\mathbf{Z}$ whose law corresponds to a non-repeating version of the Markov Chain $\mathbf{T}$ of the *MSBMI*(G) distribution (cf. [8] for more discussion). This construction will later aid in the study of models involving protein. Let $\mathbf{Z}$ be a homogeneous Markov chain with initial distribution μ and transition kernel

$$K_{i,j} = \frac{G_{i,j}}{-G_{i,i}} \mathbb{1}(i \neq j)$$

Next, let $\mathbf{W}$ be a random sequence such that $\mathbf{W}|\mathbf{Z}$ is an independent sequence $\{Beta(1, -G_{Z_j,Z_j})\}_{j\geq 1}$ of variables. Define $\mathbf{R}$ from $\mathbf{W}$ as a residual allocation model

$$R_j = W_j \prod_{i=1}^{j-1}(1 - W_i).$$

Form the associated stick-breaking measure

$$\nu(\ \cdot\ ) = \sum_{j=1}^{\infty} R_j \delta_{Z_j}(\ \cdot\ ).$$

Then, $\nu \overset{d}{=} \nu^\theta$, and moreover we have the following 'clumped' statement.

Proposition 2.3 (cf. Theorems 2.8, 2.13 [8]). *Let G be an irreducible generator matrix over $\mathfrak{X}$ with unique stationary distribution μ. Define stochastic kernel*

$$K_{i,j} = \frac{G_{i,j}}{-G_{i,i}} \mathbb{1}(i \neq j).$$

Let $\mathbf{Z}$ be a homogeneous Markov chain with transition kernel K and initial distribution μ. Let $\mathbf{W}$ be a random sequence of [0,1]-valued random variables such that given $\mathbf{Z}$, $\mathbf{W}$ is an independent sequence with $W_j \sim$Beta$(1, -G_{Z_j,Z_j})$. Form the residual allocation model $\mathbf{R} = \{W_j \prod_{i=1}^{j-1}(1 - W_i)\}_{j\geq 1}$. Then,

$$\left(Z_1,\ \sum_{j=1}^{\infty} R_j \delta_{Z_j}(\ \cdot\ ) \right) \ \sim\ \text{MSBMI}(G)$$

3. Time-inhomogeneous MCs, PDMPs, and multistate mRNA promoter processes. We consider now seemingly unrelated processes, which however in combination bear upon the multistate mRNA promoter process. In the main section, we deduce results on the associated stationary distribution and in Section 3.1 on its moments. We also discuss identifiability of parameters with respect to stationary mRNA levels in Section 3.2.

The first process is a time-inhomogeneous Markov chain, considered in [8], [9] with respect to certain 'simulated annealing' models.

Definition 3.1 (Inhomogeneous Chain $\mathbf{M} = (M_n)_{n\geq 1}$ (cf. [8]). *Let G be an irreducible generator matrix on $\mathfrak{X}$. We associate to G the discrete time Markov chain $\mathbf{M} = (M_n)_{n\geq 1}$ with state space $\mathfrak{X}$ having transition kernels*

$$K_n = I + \frac{G}{n} \mathbb{1}(n > N)$$

for sufficiently large N that K_n is stochastic. We denote the empirical measure of $\mathbf{M}$ *up to time n by*

$$\nu^n = \frac{1}{n}\sum_{j=1}^{n} \delta_{M_j}.$$

In words, the Markov chain $\mathbf{M}$ stays on the state it is at with larger probability as n grows, and switches states with probability of order $O(1/n)$. In this way, the states in $\mathfrak{X}$ can be considered 'valleys' from which it becomes more difficult to leave as time increases. Nevertheless, there will be an infinite number of switches in the chain.

We also remark in passing that Definition 3.1 extends to countable discrete spaces (see [8]), although we do not use this extension in this article, concentrating on finite $\mathfrak{X}$.

The second process is a type of piecewise deterministic Markov process (PDMP)– informally, a pair $(E(t), X(t))$ such that $E(t)$ is a Markov jump process on $\mathfrak{X}$ and, if $\{t_n\}_{n\geq 1}$ are the jump times of $E(t)$, then $X(t)$ evolves deterministically on each interval $[t_n, t_{n+1})$ in a manner determined by $E(t_n)$. Such a process is determined by the jump rates of $E(t)$, the transition measure of $(E(t), X(t))$, and the flows governing the deterministic behavior of $X(t)$ between jumps. See [7] for a more general and precise definition; note that we have switched the standard PDMP order $(X(t), E(t))$ to $(E(t), X(t))$ to be consistent with [17].

Definition 3.2 (Exponential Zig-zag Process (cf. [2])). *An exponential zig-zag process is a PDMP $(E(t), X(t))$ taking values in $\mathfrak{X}\times\Delta_{\mathfrak{X}}$ with infinitesimal generator*

$$\begin{aligned}\mathcal{L}_Z f(i,x) &= (e_i - x)\cdot\nabla_x f(i,x) + \sum_{j\neq i} G_{i,j}[f(j,x) - f(i,x)]\\ &= (e_i - x)\cdot\nabla_x f(i,x) + \sum_{j} G_{i,j} f(j,x)\end{aligned}$$

where G is an irreducible generator matrix on a finite space $\mathfrak{X}$. Such a process has a unique stationary distribution (cf. Section 3 [2]).

In words, the E process switches according to rates G. However, depending on the current state $E = i$, the X_j values decrease at rate proportional to X_j for $j\neq i$ and X_i increases at rate $1 - X_i$.

We now state carefully the multistate mRNA promoter process.

Definition 3.3 (Multistate promoter process (cf. [17])). *Let G be an irreducible generator matrix on a finite space $\mathfrak{X}$. Consider the jump Markov process $(E(t), M(t))$ taking values in $\mathfrak{X}\times\mathbb{N}_0$ with transition rates*

$$(i,m)\to(j,n)\quad \textit{at rate}\quad \begin{cases} G_{i,j} & n=m\\ \beta_i & i=j,\ n=m+1\\ \delta m & i=j,\ n=m-1\\ 0 & \textit{otherwise}\end{cases}$$

for $i,\ j\in\mathfrak{X}$ and $m,\ n\in\mathbb{N}_0$.

We also associate to the multistate promoter process a process $X(t)$ taking values in $\Delta_{\mathfrak{X}}$ which is a solution to

$$\frac{d}{dt}X_i(t) = \delta\big[\mathbb{1}(E(t)=i) - X_i(t)\big].$$

It is known that the joint process $(E(t), M(t), X(t))$ *has a unique stationary distribution (cf. Corollary 3.5 [17]). In particular, we denote the stationary distribution of* $(E(t), M(t))$ *by* $\pi_1(i, m|G, \beta, \delta)$.

The multistate promoter process models mRNA production by a gene promoter which can be in one of a finite collection $\mathfrak{X}$ of states. The Markov jump process $E(t)$ with rates G tracks the state of the promoter over time. The rate of mRNA production while the promoter is in state i is given by $\beta_i \geq 0$. Then, the production of mRNA is a birth-death process, with mRNA produced at rate β_i when $E(t) = i$, while each individual mRNA degrades independently at rate $\delta > 0$. The process $X(t)$, although auxiliary, is helpful in computations and, as alluded to in the introduction, represents a mass action kinetics process on the simplex $\Delta_{\mathfrak{X}}$. By introduction of $X(t)$, one may introduce a system of differential equations, which is more readily analyzed than the combinatorial model without it (cf. Section 3 in [2] for more discussion).

We now state three results on these processes and deduce the stick-breaking representation of the multistate mRNA promoter process in Theorem 3.7.

The first result is that the empirical measure of the time-inhomogeneous MC converges weakly to the MSBM stick-breaking measure. A different characterization for types of G may also be found in [9].

Theorem 3.4 (cf. Theorem 2.13 [8]). *Let* G *be an irreducible generator matrix, with stationary distribution* μ*, over* $\mathfrak{X}$*. Let* $\mathbf{M}$ *be the inhomogeneous chain associated to* G*, and* $(\nu^n)_{n\geq 1}$ *be the empirical measures of* $\mathbf{M}$*. Then*

$$(M_n, \nu^n) \xrightarrow{d} \mathrm{MSBMI}(G^*).$$

The second result is that the stationary distribution of the PDMP is the limit empirical measure for the time-inhomogeneous MC, under Assumptions 2.1, 2.6 in [2] which are satisfied since the transition kernels are in form $I + G/n$ for large n. In [2], one may also find a non stick-breaking characterization of the limit, as well as other interesting results.

Theorem 3.5 (cf. Theorem 2.8 [2]). *Let* G *be an irreducible generator matrix over* $\mathfrak{X}$*. Let* $\mathbf{M}$ *be the inhomogeneous chain associated to* G*, and* $(\nu^n)_{n\geq 1}$ *be the empirical measures of* $\mathbf{M}$*. Let also* $(E(t), X(t))$ *be an exponential zig-zag process parametrized by* G*. Then, the associated stationary distribution* (E, X) *is the limit distribution of the time-inhomogeneous Markov chain:*

$$(E, X) \overset{d}{=} \lim_{n\to\infty} (M_n, \nu^n).$$

We comment that the above limit [2] is identified as the one found in [9] characterized by moments. Later, in [8], this limit was seen in terms of stick-breaking as stated in Theorem 3.4.

The third result finds that the stationary distribution of the multistate promoter process is a certain Poisson mixture. In [17], generating functions of the stationary distribution are also given.

Theorem 3.6 (cf. Proposition 4.1 [17]). *Let* G *be an irreducible generator matrix over* $\mathfrak{X}$*. Let* $\beta \in (\mathbb{R}^+)^{\mathfrak{X}}$ *and* $\delta,\ \lambda > 0$*. Let* $(E(t), M(t))$ *be a multistate promoter process parametrized by* G *with associated process* $X(t)$*. Suppose* $M(0)|E(0), X(0) \sim \mathrm{Poisson}(\lambda(0))$ *where* $\lambda(t)$ *satisfies*

$$\lambda(t) = \delta^{-1}\beta \cdot X(t).$$

Then,

$$M(t)\Big|\big(E(\tau)\big)_{\tau\geq 0} \ \sim\ \text{Poisson}(\lambda(t)) \qquad \textit{where} \qquad \partial_t\lambda(t) = \beta_{E(t)} - \delta\lambda(t).$$

Further, with respect to an observation (E, M, X) from the stationary distribution of $(E(t), M(t), X(t))$, we have

$$M\Big|E, X \ \sim\ \text{Poisson}(\delta^{-1}\beta\cdot X).$$

We remark, in passing, that the stationary distribution of $(E(t), M(t))$ does not depend on the initial distribution of $M(0)$. Indeed, the representation of $M(t)$ as a Poisson mixture is retained after a finite, random burnoff period corresponding with degradation of all mRNA initially present; see [22].

We now straightforwardly combine the three previous results to find a stick-breaking representation of the stationary distribution $\pi_1(i, m|G, \beta, \delta)$, that is of the limit (E, M).

First, by scaling time by δ, we see that the stationary limit of the multistate process components (E, X) is the limit of the PDMP in the work of [2] with generator G/δ. In turn, the work of [8] shows that this limit is *MSBMI*(G^*/δ). Hence, we obtain the main statement of this section, namely the following theorem.

Theorem 3.7. *Let G be an irreducible generator matrix over $\mathfrak{X}$. Let $\beta\in(\mathbb{R}^+)^{\mathfrak{X}}$ and $\delta > 0$. Let $(E(t), M(t))$ be a multistate promoter process parametrized by G with associated process $X(t)$. Then,*

$$\big(E(t), M(t), X(t)\big) \xrightarrow{d} (E, M, X)$$

where

$$(E, X)\sim \text{MSBMI}(G^*/\delta) \qquad \textit{and} \qquad M\Big|E, X \ \sim\ \text{Poisson}(\delta^{-1}\beta\cdot X).$$

Hence, the stationary distribution $\pi_1(i, m|G, \beta, \delta)$ of (E, M) is the law of the mixture Poisson$(\delta^{-1}\beta\cdot X)$.

3.1. Moments with respect to the multistate mRNA promoter process. Using the stick-breaking apparatus we may identify moments of interest. We first state a result found in [23].

Proposition 3.8 (cf. Theorem 4 [23]). *Let $(T,\nu)\sim$ MSBMI(G) for an irreducible generator matrix G with stationary distribution over $\mathfrak{X}$. Let $n\in\mathbb{N}$, $\vec{k}\in\mathbb{N}_0^n$, and $(A_j)_{j=1}^n$ be disjoint collection of subsets of $\mathfrak{X}$. Define $\mathbb{S}(\vec{k})$ to be the collection of all distinct k-lists consisting of k_1 many 1's, k_2 many 2's, and so on to k_n many n's, where $k = \sum_{j=1}^n k_j$. Then,*

$$\left(\mathbf{E}\left[\prod_{j=1}^n \nu(A_j)^{k_j}\middle| T = x\right]\right)_{x\in\mathfrak{X}} = \left(\#\mathbb{S}(\vec{k})\right)^{-1}\sum_{\sigma\in\mathbb{S}(\vec{k})}\left[\prod_{j=1}^{k;(R)}(I - G/j)^{-1}D(A_{\sigma_j})\right]\vec{1}.$$

We may rewrite the above expression in more convenient form.

Corollary 3.9. *Consider the context of the previous proposition. If $(\tilde{T},\tilde{\nu})\sim$ MSBMI(G^*), then*

$$\mathbf{E}\left[\prod_{j=1}^n \tilde{\nu}(A_j)^{k_j}\middle|\tilde{T} = x\right] = \mu_x^{-1}\cdot\left(\#\mathbb{S}(\vec{k})\right)^{-1}\sum_{\sigma\in\mathbb{S}(\vec{k})}\mu^T\left[\prod_{j=1}^k D(A_{\sigma_j})(I - G/j)^{-1}\right]e_x.$$

Proof. By convention,

$$\begin{aligned}&\Big[\prod_{j=1}^{k;(R)}(I-G/j)^{-1}D(A_{\sigma_j})\Big]\vec{1}\\&\quad=D^{-1}(\mu)(I-G^T/k)^{-1}D(\mu)D(A_{\sigma_k})\cdots D^{-1}(\mu)(I-G^T)^{-1}D(\mu)D(A_{\sigma_1})\vec{1}.\end{aligned}$$

Since $D(\mu)$ and $D(A_{\sigma_\cdot})$ commute, the product equals

$$D^{-1}(\mu)(I-G/k)^{-1}D(A_{\sigma_k})(I-G/(k-1))^{-1}\cdots(I-G)^{-1}D(A_{\sigma_1})\mu.$$

The xth entry can be found by taking the transpose and multiplying by e_x. Noting that $D^{-1}(\mu)e_x=\mu_x^{-1}e_x$ finishes the calculation. □

We observe that we can recover a formal, if not particularly useable, expression of the stationary probabilities, and also moments of the mRNA levels M in stationarity; see also [17] for a derivation using a PDE for the generating function.

COROLLARY 3.10. *Let G be an irreducible generator matrix over $\mathfrak{X}$ having unique stationary distribution μ. Let $\beta\in(\mathbb{R}^+)^{\mathfrak{X}}$ and $\delta>0$. Then, the stationary distribution π_1 of the multistate promoter process (E,M) parametrized by G, β, and δ is given as*

$$\begin{aligned}\pi_1(i,m|G,\beta,\delta)&=\pi_1(i,m|G/\delta,\beta/\delta,1)\\&=\mu^T\left[\frac{1}{m!}\prod_{k=1}^{m}D(\beta/\delta)(I-G/(\delta k))^{-1}\right]\\&\qquad\times\left[\sum_{n\geq 0}\frac{(-1)^n}{n!}\prod_{k=m+1}^{m+n}D(\beta/\delta)(I-G/(\delta k))^{-1}\right]e_i.\end{aligned}$$

Further, by the factorial moment property of the Poisson distribution, for each $k\in\mathbb{N}_0$,

$$\begin{aligned}&\mathbf{E}\left[M(M-1)(M-2)\cdots(M-k+1)\right]\\&\quad=\mathbf{E}\left[(\beta\cdot X/\delta)^k\right]=\mu^T\left[\prod_{j=1}^{k}D(\beta/\delta)(I-G/(\delta j))^{-1}\right]\vec{1}.\end{aligned}\tag{3.1}$$

where $X\sim$ MSBM(G^*/δ).

Proof. Note that the Poisson mixture relation $(T,\text{Poisson}(\delta^{-1}\beta\cdot\nu))\sim\pi_1$ for $(T,\nu)\sim$ *MSBMI*(G^*/δ) is stated in Theorem 3.7. The stationary probability formulas now follow from the moment calculation (3.1). To verify these observe $\beta\cdot\nu/\delta=\sum_{i\in\mathfrak{X}}\delta^{-1}\beta_i\nu(i)$ and

$$\mathbb{E}\big[(\beta\cdot X/\delta)^k\big]=\delta^{-k}\sum_{j_1,\ldots,j_k}\beta_{j_1}\cdots\beta_{j_k}\mathbb{E}\big[\nu(j_1)\cdots\nu(j_k)\big]$$

where $1\leq j_1,\ldots,j_k\leq|\mathfrak{X}|$. One can now check, via Corollary 3.9, that the desired formula is obtained. □

3.2. On identifiability of mRNA levels. Consider the multistate mRNA promoter process $(E(t), M(t), X(t))$ parametrized by G, β and δ in Definition 3.3. To begin the discussion, we will scale out the parameter δ and take it as $\delta = 1$. Let (E, X) represent the stationary distribution of the process $(E(t), X(t))$. In [17], it is shown that (E, X) is identifiable by G, that is two different generators cannot give the same stationary distribution.

Indeed, we sketch the argument for the convenience of the reader: The associated Laplace transform of (E, X) is $\phi(s) = (\phi_1 \ldots, \phi_{|\mathfrak{X}|})$, where $\phi_i(s) = E[1(E = i)e^{s\cdot X}]$ satisfies

$$\sum_i s_i \partial_{s_i} \phi(x) = \big(D(s) + G^T\big)\phi(s).$$

Then, for fixed $s = \beta$, the Laplace transform of $(E, \beta \cdot X)$ is $\Phi(w) = \phi(w\beta)$ where

$$w\partial_w \Phi = \big(wD(\beta) + G^T\big)\Phi. \tag{3.2}$$

In Corollary 4.3 of [17], $\Phi(w)$ is developed in power series, $\Phi(w) = \sum_{k\geq 0} c_k(\beta)w^k$, where in particular the characterization $c_1(\beta) = (I - G^T)^{-1}D(\beta)\mu$ is made, where μ is the distribution of E. Note that $\mu = \Phi(0) = \Phi'(0) = \mu'$. Hence, if there are two generators G and H for which $(E, X)_G = (E, X)_H$ in law, then $c_1(s; G) = c_1(s; H)$. Since s is arbitrary and μ has full support as G is irreducible, we conclude $G = H$.

However, one may ask about identifiability of the mRNA level M itself. In Theorem 3.7, we see that the stationary mRNA level M is determined by the distribution of $\beta \cdot X$. Since mRNA level readings are available from lab experiments, for inference purposes, it makes sense to study the identifiability of the distribution of $\beta \cdot X$ in terms of (β, G). Since we are not given the distribution of E and β is not arbitrary, the previous identifiability argument for (E, X) is not sufficient. Moreover, in the refractory case, when only one component $\beta_i > 0$, the rest vanishing, [17] shows that certain eigenvalues of G are identifiable, although G itself cannot be determined. However, see the examples below. Of course, when β is a vector with common entries, $\beta \cdot X = \beta_1$, certainly X cannot be identified. Nevertheless, since the support of $\beta \cdot X$ is $[\min_i \beta_i, \max_i \beta_i]$, both $\min_i \beta_i$ and $\max_i \beta_i$ are identifiable.

To further the discussion, the Laplace transform of $\beta \cdot X$ is $\vec{1} \cdot \Phi(w)$ which satisfies

$$w\vec{1} \cdot \partial_w \Phi = w\vec{1}D(\beta)\Phi + \vec{1}G^T\Phi = w\beta \cdot \Phi$$

since $\vec{1}G^T = \vec{0}$ given that G is a generator matrix. Hence, if the distribution of $\beta \cdot X$ and $\beta' \cdot X'$ with respect to generators G and H respectively match, then

$$\vec{1} \cdot \Phi(w) = \vec{1} \cdot \Phi'(w) \ \text{ and } \ \beta \cdot \Phi(w) = \beta' \cdot \Phi'(w). \tag{3.3}$$

Differentiating the last item $\beta \cdot \Phi$ and multiplying by w, we get that

$$w\beta \cdot \partial_w \Phi(w) = w\beta \cdot D(\beta)\Phi(w) + \beta \cdot G^T\Phi(w) \tag{3.4}$$

and a similar equation with respect to Φ'. One can develop subsequent equations by differentiating in w. One also has expressions for the moments (cf. Corollary 3.10 or [17]).

Despite the nonlinearity of these relations which seem difficult to negotiate, we believe that identifiability of (β, G) holds with respect to irreducible generators G, when the entries of β are strictly ordered, say $\beta = (\beta_1, \ldots, \beta_{|\mathfrak{X}|})$ where $\beta_1 > \beta_2 > \cdots > \beta_{|\mathfrak{X}|} \geq 0$, but we leave this theoretical question to a future investigation. Numerically, in this respect, however, we observe that the work in Section 5 gives positive evidence of this claim.

However, in a 'Dirichlet setting', where $G = \theta(Q - I)$ and $H = \theta'(Q' - I)$ where Q and Q' are stochastic matrices with constant rows μ and μ' respectively, and the scale factors $\theta = \theta'$ agree, the algebra simplifies considerably. In this setting there is no memory in the E promoter transition rates, and these rates sum to the same value.

LEMMA 3.11. *In the Dirichlet setting, when the scale factors $\theta = \theta'$ agree, the parameters are identifiable, that is $(\beta, Q) = (\beta', Q')$.*

The proof of this lemma is left to Appendix A.

We finish the section with two 'refractory' examples, the first showing non-trivial absence of identifiability and the second, studied in the numerical study Section 5.2, where one can actually show identifiability:

Example 3.12. Consider two distinctly parametrized three-state models with generators G and $\check{G}$,

$$G = \begin{pmatrix} -5 & 2 & 3 \\ 2 & -5 & 3 \\ 2 & 2 & -4 \end{pmatrix} \qquad \check{G} = \begin{pmatrix} -5 & 3 & 2 \\ 2 & -3 & 1 \\ 2 & 4 & -6 \end{pmatrix}.$$

and vector $\beta = (1, 0, 0)$. Then, if M is distributed as in Theorem 3.7 according to parameters G and β, and $\check{M}$ is distributed as in Theorem 3.7 according to parameters $\check{G}$ and β, the $M \overset{d}{=} \check{M}$. According to [17], the identifiable parameters of G are the (nonzero) eigenvalues of $-G$ and the eigenvalues of $-G_{(1)}$, where $G_{(1)}$ is a matrix obtained by removing the first row and the first column of G. [Herbach considers $H = G^T$, which gives the same formulas.] Both G and $\check{G}$ share these eigenvalues, with $-G$ and $-\check{G}$ having eigenvalues 0, 7, and 7, and $-G_{(1)}$ and $-\check{G}_{(1)}$ having eigenvalues 2 and 7. As such, the generic three-state model for mRNA levels cannot be identifiable from mRNA levels alone.

Example 3.13. Consider a three-state refractory model where $\beta_1 > 0$ and $\beta_2 = \beta_3 = 0$ and G has zero entries, $G_{13} = G_{31} = 0$. In this case, we claim the model is identifiable. Form

$$G = \begin{pmatrix} -a & a & 0 \\ b & -b-c & c \\ 0 & d & -d \end{pmatrix},$$

where a, b, c, and d are positive real numbers. Again, according to [17], the identifiable parameters of G are the (nonzero) eigenvalues of $-G$ and the eigenvalues of $-G_{(1)}$, where $G_{(1)}$ is a matrix obtained by removing the first row and the first column of G.

Let λ_1 and λ_2 be the two nonzero eigenvalues of $-G$. They are the zeros of the equation (in λ) $\lambda^2 - (a+b+c+d)\lambda + ad + bd + ac = 0$. Let λ_3 and λ_4 be the eigenvalues of $-G_{(1)}$. They are the zeros of the equation (in λ) $\lambda^2 - (b + c + d)\lambda + bd = 0$. Then a, b, c, and d can be expressed by $\lambda_i, i = 1, 2, 3, 4$ as

$$a = \lambda_1 + \lambda_2 - (\lambda_3 + \lambda_4), \quad b = \lambda_3 + \lambda_4 - (\lambda_1\lambda_2 - \lambda_3\lambda_4)/a,$$
$$c = (\lambda_1\lambda_2 - \lambda_3\lambda_4)/a - \lambda_3\lambda_4/b, \quad d = \lambda_3\lambda_4/b,$$

meaning a, b, c, and d are identifiable.

4. Protein production in the multistate mRNA promoter process. We discuss now an extension of the multistate promoter model which incorporates protein production. Specifically, when the multistate promoter process $(E(t), M(t))$ is in state (i, m), we formulate that individual proteins are produced at rate of αm and the protein level p degrades at rate γp. Such an ansatz corresponds to the idea that each individual mRNA independently produces protein at rate $\alpha > 0$ and each protein degrades at rate $\gamma > 0$.

A version of this model was indicated in [17], and stationary distributions in the 'refractory' case, when only one β_i is positive, were found in [4]. In this context, our goal will be to derive the stationary distribution in the general $(\beta, \delta, \alpha, \gamma, G)$ model through the stick-breaking apparatus.

The strategy will be to consider 'bounded joint multistate mRNA-protein processes' which restrict mRNA levels below a capacity level c. Such bounded joint processes have the same abstract finite-state promoter structure as the multistate mRNA promoter process, with stationary distributions given in terms of Markovian stick-breaking measures. The idea is to take a limit now as the capacity level $c \uparrow \infty$ to recover the stationary distribution in the general 'unbounded' model, mentioned in the introduction. Importantly, clumped representations of the MSBM's (Section 2.3) for the bounded joint process will be of use in this regard.

We now define carefully the bounded joint process.

Definition 4.1 (Bounded joint multistate mRNA-protein process). *Let G be an irreducible generator over $\mathfrak{X}$. A bounded joint process is the Markov jump process $(E(t), M(t), P(t))$ on $\mathfrak{X} \times \{0, 1, 2, ..., c\} \times \mathbb{N}_0$ with rates*

$$(i,m,p) \to (j,n,q) \quad \textit{at rate} \quad \begin{cases} G_{i,j} & ; \quad i \neq j, \quad n = m, \qquad\quad q = p \\ \beta_i & ; \quad i = j, \quad n = m+1 \leq c, \quad q = p \\ \delta m & ; \quad i = j, \quad n = m-1, \qquad q = p \\ \alpha m & ; \quad i = j, \quad n = m, \qquad\quad q = p+1 \\ \gamma p & ; \quad i = j, \quad n = m, \qquad\quad q = p-1 \\ 0 & ; \qquad\qquad o.w. \end{cases}$$

We associate to this process the bounded generator matrix over finite state space $\tilde{\mathfrak{X}} = \mathfrak{X} \times \{0, 1, 2, ..., c\}$

$$\begin{aligned} \tilde{G}^c_{(i,m),(j,n)} &= \mathbb{1}(i \neq j, m = n) G_{i,j} + \mathbb{1}(i = j, n = m+1 \leq c)\beta_i \\ &\quad + \mathbb{1}(i = j, n = m-1)\delta m + \mathbb{1}(i = j, m = n)(G_{i,i} - \beta_i \mathbb{1}(m < c) - \delta m) \end{aligned} \tag{4.1}$$

and denote its stationary distribution as $\pi^c_2(i, m, p | G, \beta, \delta, \alpha, \gamma, c)$.

We note also, omitting the protein, the bounded mRNA process $(E(t), M(t))$ on the finite state space $\mathfrak{X} \times \{0, 1, 2, \ldots, c\}$ is well-defined. See Figure 2 for a representation in terms of chemical reactions.

One may understand the bounded joint process as follows. Let $(E(t), M(t), P(t))$ be a bounded joint process with generator G, production rates β and α, death rates δ and γ, and cap c. Denote the same process as

$$(\tilde{E}(t), \tilde{M}(t)) \text{ with } \tilde{E}(t) = (E(t), M(t)) \text{ and } \tilde{M}(t) = P(t). \tag{4.2}$$

Then, we observe $(\tilde{E}(t), \tilde{M}(t))$ is a multistate promoter process taking values in $\tilde{\mathfrak{X}} \times \mathbb{N}_0$ parameterized by generator $\tilde{G}^c$, production rates $\tilde{\beta}_{i,m} = \alpha m$, and death rate

$$\begin{cases} S_i \xrightarrow{G_{ij}} S_j & \text{for } i,j \in \mathfrak{X};\ i \neq j \\ S_i \xrightarrow{\beta_i 1([M]+1 \leq c)} S_i + M & \text{for } i \in \mathfrak{X} \\ M \xrightarrow{\delta} \emptyset & \\ M \xrightarrow{\alpha} M + P & \\ P \xrightarrow{\gamma} \emptyset & \end{cases}$$

FIG. 2. *Above is a representation in chemical equations for a bounded multi-state promoter process with protein. Promoter states are represented by chemical species $S_i : i \in \mathfrak{X}$ with transitions between states such that molecule numbers always satisfy $[S_i] \in \{0,1\}$ for all i and $\sum_i [S_i] = 1$; see [17]. Representing mRNA by a species M and protein by a species P, the promoter process $(E(t), M(t), P(t))$ is determined by the above elementary equations with $E(t) = i$ when $[S_i] = 1$, and $(M(t), P(t)) = ([M], [P])$. Note that always $[M] \leq c$, and that the process $(E(t), M(t))$, without protein, is also well-defined.*

$\tilde{\delta} = \gamma$, where $\tilde{\mathfrak{X}} = \mathfrak{X} \times \{0, 1, \ldots, c\}$. In particular, as $\mathfrak{X}$ and so $\tilde{\mathfrak{X}}$ are finite spaces, we have for all $(i, m, p) \in \mathfrak{X} \times \{0, 1, \ldots, c\} \times \mathbb{N}_0$, that

$$\pi_2^c(i, m, p|G, \beta, \delta, \alpha, \gamma, c) = \pi_1((i, m), p|\tilde{G}^c, \tilde{\beta}, \tilde{\delta}), \tag{4.3}$$

for which there is a stick-breaking relation via Theorem 3.7.

We now state carefully the unbounded joint process.

DEFINITION 4.2 ((Unbounded) joint multistate mRNA-protein process). *Let G be an irreducible generator over $\mathfrak{X}$. The (unbounded) joint process is the Markov jump process $(E(t), M(t), P(t))$ on $\mathfrak{X} \times \mathbb{N}_0^2$ with rates*

$$(i, m, p) \to (j, n, q) \quad \textit{at rate} \quad \begin{cases} G_{i,j} & ; \quad i \neq j, \quad n = m, \quad q = p \\ \beta_i & ; \quad i = j, \quad n = m+1, \quad q = p \\ \delta m & ; \quad i = j, \quad n = m-1, \quad q = p \\ \alpha m & ; \quad i = j, \quad n = m, \quad q = p+1 \\ \gamma p & ; \quad i = j, \quad n = m, \quad q = p-1 \\ 0 & ; \quad o.w. \end{cases}$$

We associate to this process the unbounded generator matrix

$$\begin{aligned} \tilde{G}^{\infty}_{(i,m),(j,n)} &= \mathbb{1}(i \neq j, m = n) G_{i,j} + \mathbb{1}(i = j, n = m+1)\beta_i \\ &\quad + \mathbb{1}(i = j, n = m-1)\delta m + \mathbb{1}(i = j, m = n)(G_{i,i} - \beta_i - \delta m) \end{aligned}$$

and denote its stationary distribution as $\pi_2^{\infty}(i, m, p|G, \beta, \delta, \alpha, \gamma)$.

In the following, we will use the notation $\tilde{G}^{\cdot,T} = (\tilde{G}^{\cdot})^T$ and $\tilde{G}^{\cdot,*} = (\tilde{G}^{\cdot})^*$.

We remark that the existence/uniqueness of π_2^{∞} in the above definition is formulated in the following lemma. The proof of Lemma 4.3 is given in the Appendix B following from an application of Theorem 1.1 [21].

LEMMA 4.3. *There exists a unique stationary distribution π_2^{∞} for the unbounded process in Definition 4.2. Moreover, π_2^{∞} integrates $e^{\epsilon_1 i + \epsilon_2 m + \epsilon_3 p}$, for some constants $\epsilon_1, \epsilon_2, \epsilon_3 > 0$.*

The type of association with a multistate mRNA process made earlier with respect to the bounded joint process cannot be implemented directly with respect to the unbounded joint protein process. Indeed, the generator matrix $\tilde{G}^{\infty}$ associated to the unbounded joint protein process is itself unbounded, and hence cannot be normalized to construct a stochastic kernel as discussed after Definition 2.2. However, we will see that the 'clumped' stick-breaking construction of Proposition 2.3 may still be understood and used in this context.

THEOREM 4.4. *Let $(E(t), M(t), P(t))$ be an unbounded joint process with respect to E-generator G, production rates β and α, and death rates δ and γ. Then, the associated stationary distribution $\pi_2^\infty(i, m, p|G, \beta, \delta, \alpha, \gamma)$ may be sampled as follows:*

Define a stochastic kernel K over $\mathfrak{X} \times \mathbb{N}_0$ by

$$K_{(i,m),(j,n)} = \frac{\tilde{G}^{\infty,*}_{(i,m),(j,n)}}{-\tilde{G}^{\infty,*}_{(i,m),(i,m)}} \mathbb{1}((i,m) \neq (j,n))$$

Let now $\mathbf{Z}$ be a homogeneous Markov chain over $\mathfrak{X} \times \mathbb{N}_0$ with transition kernel K and initial distribution π_1. Conditioned on $\mathbf{Z}$, let $\mathbf{W}$ be a sequence of independent random variables with $W_j \sim \mathrm{Beta}(1, -\tilde{G}^\infty_{Z_j,Z_j}/\gamma)$. Consider the residual allocation model $\mathbf{R} = \{W_j \prod_{i=1}^{j-1}(1 - W_i)\}_{j\geq 1}$, and define a random vector $X \in \Delta_{\mathfrak{X}\times\mathbb{N}_0}$ by

$$X_{(i,m)} = \sum_{j=1}^\infty R_j \delta_{Z_j}((i,m))$$

Then, if

$$P\Big|\mathbf{Z}, \mathbf{R} \sim \mathit{Poisson}\left(\frac{\alpha}{\gamma}\sum_{(i,m)} m X_{(i,m)}\right)$$

and we denote $Z_1 = (E, M) \sim \pi_1(i, m|G, \beta, \delta)$ (as defined in Definition 3.3), we have (E, M, P) is a sample from the stationary distribution of $(E(t), M(t), P(t))$,

$$\pi_2^\infty(\ \cdot\ ,\ \cdot\ ,\ \cdot\ |G, \beta, \delta, \alpha, \gamma),$$

which can be expressed as the limit of the stationary distributions,

$$\pi_2^c(i, m, p|G, \beta, \delta, \alpha, \gamma, c) = \pi_1((i, m), p|\tilde{G}^c, \tilde{\beta}^c, \tilde{\delta}),$$

of bounded joint processes (E^c, M^c, P^c) (cf. (4.3)), that is

$$\pi_2^\infty(i, m, p|G, \beta, \delta, \alpha, \gamma) \overset{d}{=} \lim_{c\to\infty} \pi_1((i, m), p|\tilde{G}^c, \tilde{\beta}^c, \tilde{\delta}).$$

The joint moments of (M, P) can be captured in terms of the limit

$$\mathbf{E}[M^k P^\ell] = \lim_{c\to\infty} \mathbf{E}[(M^c)^k (P^c)^\ell]$$

where $\mathbf{E}[(M^c)^k(P^c)^\ell] = \mathbf{E}\big[(M^c)^k \mathbf{E}[(P^c)^\ell|E^c, M^c]\big]$ has calculation using the relations (4.2) *and Corollary* 3.10.

Remark 4.5. Although the representation of the stationary distribution π_2^∞ of the unbounded mRNA-protein interaction model is given as a limit of 'clumped' representations of Markov stick-breaking measures of the bounded models, we point out that this limit though is *not* in Markov stick-breaking form as given in Definition 2.2. The same strategy of proof could in principle, under assumptions, be used to capture the stationary distribution of an 'infinite' promoter mRNA model where $\mathfrak{X}$ would be countably infinite instead of finite. Given the nature of applications, we did not pursue in this direction however.

5. Bayesian Inference.

Given the possibility to extract mRNA level readings from cells in laboratory, it is natural to explore statistical inference procedures of parameters β and G from data (cf. [1], [18]). In the following, we concentrate on Bayesian estimation of model parameters and selection of an appropriate underlying model based on mRNA readings. The methods are demonstrated by synthetic experiments where the data are samples from a given stationary distribution. The stick-breaking form of the stationary distribution (E, M, X) in the multistate mRNA promoter model with parameters β and G (having scaled $\delta = 1$), will be useful in this regard. In particular, one can directly sample from the stationary distribution to a given level of accuracy by truncating the series.

To compare with literature, in [1] and [18], inference procedures were performed for parameters in a multistate promoter mRNA-protein interaction network where the promoter space $\mathfrak{X} = \{0, 1\}$ has two states. As remarked in the two-state setting, the mRNA level stationary level M is a Poisson-Beta mixture. With certain approximations, extending also to the stationary protein level P, results were found in accord with laboratory data.

From a different point of view, in [22], synthetic data taken at four time points from a multistate promoter mRNA model where $|\mathfrak{X}| = 2, 3$ and some components of the parameters β and G vanish a priori, inference of parameters is carried out.

In the following, we restrict also to $|\mathfrak{X}| = 2, 3$ state multistate promoter mRNA models. Our emphasis will be on understanding the benefit from using an explicit stick-breaking formulation of the stationary measure. Since we also have derived a stick-breaking representation of the stationary distribution in models with protein interactions, the same formalism will apply.

We present, in Section 5.1, a Bayesian approach for estimating the parameters in the multistate promoter model based on data from the stationary distribution. Although the mass function of the stationary distribution is derived in Corollary 3.10, it cannot be used directly for inference due to the slow convergence of the series in the formulation. We overcome the difficulty by approximating the mass function using Monte Carlo simulations according to the stick-breaking representation of the stationary distribution. The estimation performance is examined under various multistate promoter models.

On a different track, the promoter model for a gene is often fixed a priori in the literature. The number of the promoter states and the nonzero parameters in G are often assumed known before data analysis. In this context, we demonstrate in Section 5.2 a data-driven method for selecting the promoter model. This method also relies on the stick-breaking representation of the stationary distribution.

5.1. Parameter estimation..

Given L observations $M_1, \ldots, M_L$ from the following model

$$\begin{aligned} M_l | X_l, \beta, G &\overset{ind.}{\sim} Poisson(\beta \cdot X_l), l = 1, \ldots, L, \\ X_l &\overset{iid}{\sim} MSBM(\mathrm{G}), l = 1, \ldots, L \end{aligned} \tag{5.1}$$

our goal is to estimate the parameters β and G. We first consider the case that β have nonzero and distinct elements and all the entries in G are nonzero. Under this setting, we describe how to estimate β and G in a Bayesian approach [12]. Then we discuss how to modify the procedure when a zero constraint or an equality constraint is desired for some of the parameters.

In a Bayesian framework, parameters are assumed to have a prior distribution, representing experimenter's belief on the parameter values before observing data.

Given data from a certain model, prior beliefs are updated with the information in the data to produce the posterior distribution, that is the conditional distribution of parameters given data.

Without any equality or zero constraints, the parameter space of model (5.1) has n^2 dimensions where $n = |\mathfrak{X}|$. Following the discussion in Section 3.2, we assume $\beta_1 > \beta_2 > \cdots > \beta_n > 0$. This requirement avoids the non-identifiability issue brought by permuting the states in the multistate promoter model, but it, together with the positive constraint on the rate parameters, restricts the parameter space to a subset of the Euclidean space, which brings an extra difficulty to the statistical inference. To get rid of these restrictions, we transform the parameters $(\beta_1, \ldots, \beta_n, G_{1,2}, \ldots, G_{n,n-1})$ into

$$\eta = (\log(\beta_1 - \beta_2), \log(\beta_2 - \beta_3), \ldots, \log \beta_n, \log G_{1,2}, \ldots, \log G_{n,n-1}). \tag{5.2}$$

We first compute the posterior and conduct posterior sampling on η and then apply the inverse of the 'log' transformation (5.2) to obtain the posterior density and the posterior samples of β and G. The parameters are then estimated by the empirical posterior means and the estimation uncertainty are quantified via empirical credible intervals.

We consider independent priors on the elements of η:

$$\pi(\eta) = \prod_{j=1}^{n^2} \pi(\eta_j). \tag{5.3}$$

In all the synthetic experiments presented later in this section, we used independent log-gamma priors for $\eta_j, j = 1, \ldots, n^2$ (that is $\exp(\eta_j) \overset{ind}{\sim} Gamma(a_j, b_j)$) where $a_j = 2.0$, $b_j = 0.01$ for η_j related to β and $a_j = 1.0$, $b_j = 0.1$ for η_j related to G. Other priors such as Gaussian priors can also be used. Given the choice of prior distribution, the posterior distribution of η is

$$\begin{aligned} \pi(\eta \mid M_1, \ldots, M_L) &= \frac{f(M_1, \ldots, M_L|\eta)\pi(\eta)}{\int f(M_1, \ldots, M_L|\eta)\pi(\eta)d\eta} \\ &\propto \prod_{l=1}^{L} E[\exp(-\lambda_l)\lambda_l^{M_l}] \prod_{j=1}^{n^2} \pi(\eta_j), \end{aligned} \tag{5.4}$$

where $f(M_1, \ldots, M_L|\eta)$ is the probability mass function of $M_1, \ldots, M_L$, $\lambda_l = \beta \cdot X_l$, and the expectation is taken with respect to X_l.

Since it is difficult to compute the posterior mean analytically, we use a Gibbs sampler [11], a special Markov Chain Monte Carlo algorithm [26] to draw samples from the posterior distribution. In a Gibbs sampler, parameters are initialized at an arbitrary value $\eta^{(0)} = (\eta_1^{(0)}, \ldots, \eta_{n^2}^{(0)})$. In each iteration, each parameter is sampled from its full conditional distribution given the data and the current value of other parameters. In our case, in iteration g, we should draw $\eta_j^{(g)}$ from

$$\pi(\eta_j|\eta_1^{(g)}, \ldots, \eta_{j-1}^{(g)}, \eta_{j+1}^{(g-1)}, \ldots, \eta_{n^2}^{(g-1)}, M_1, \ldots, M_L). \tag{5.5}$$

However, the distribution (5.5) is difficult to directly sample from. A Metropolis-Hastings (MH) algorithm [16, 26] is adopted to sample η_j from (5.5). More specifically, in iteration g of the Gibbs sampler, given the current value $\eta_j^{(g-1)}$ of η_j, a proposed value η'_j is generated by a Gaussian random walk, that is $\eta'_j \sim N(\eta_j^{(g-1)}, \sigma_j^2)$. Then η'_j is accepted as a new sample with probability

$$\alpha = \min\left\{1, \frac{\pi(\eta'_j \mid \eta_1^{(g)}, \ldots, \eta_{j-1}^{(g)}, \eta_{j+1}^{(g-1)}, \ldots, \eta_{n^2}^{(g-1)}, M_1, \ldots, M_L)}{\pi(\eta_j^{(g-1)} \mid \eta_1^{(g)}, \ldots, \eta_{j-1}^{(g)}, \eta_{j+1}^{(g-1)}, \ldots, \eta_{n^2}^{(g-1)}, M_1, \ldots, M_L)}\right\}. \tag{5.6}$$

If η'_j is not accepted, $\eta_j^{(g-1)}$ is reused as the sample obtained in this iteration. In other words,

$$\eta_j^{(g)} = \begin{cases} \eta'_j & \text{with probability } \alpha; \\ \eta_j^{(g-1)} & \text{with probability } 1-\alpha. \end{cases}$$

The key step of performing the MH step is to evaluate α. As α is determined by the ratio of the full conditional density of η_j at η'_j and $\eta_j^{(g-1)}$ and the full conditional density of η_j is proportional to $\pi(\eta|M_1, \ldots, M_L)$, it is sufficient to evaluate the left hand side of (5.4). Although it can not be computed exactly due to the intractable expectation, we can use Monte Carlo simulations to approximate the expectation. Given the value of η, thus β and G, $E[\exp(-\lambda)\lambda^{M_l}]$ is approximated by

$$\frac{1}{B}\sum_{b=1}^{B} \exp(-\lambda_b)\lambda_b^{M_l},$$

where $\lambda_b = \beta \cdot X^{(b)}$ and $X^{(b)}, b = 1, \ldots, B$ are iid samples from $MSBM(G)$.

Truncations of the stick-breaking constructions in (2.1) are used when drawing samples from $MSBM(G)$. For a given G, the number of terms in the truncated series can be determined explicitly based on the error tolerance. More specifically, to guarantee that the error of truncation is below ε with probability higher than $1-p$, we truncated the series at term $1 + w(G, \varepsilon, p)$ where $w(G, \varepsilon, p)$ is the smallest integer w such that $P(Z \le w) \ge 1 - p$ for a Poisson random variable Z with parameter $-\max_{1\le i\le n} |G_{i,i}| \log(\varepsilon)$ (cf. discussion near (2.2)). In the experiments presented later in this section, we further restrict the maximum number of terms involved in the calculations to avoid extremely long computing time for certain values of G.

Sometimes, it is desirable to obtain estimates of (β, G) with zero constraints or equality constraints on the elements. For example, if one knows from previous investigation that the gene expression of interest follows a two-state refractory promoter model, then the desired estimate should have constraint $\beta_1 > \beta_2 = 0$. If it is known a priori that $X_l, l = 1, \ldots, L$ in (5.1) should follow a Dirichlet distribution in a three-state model, then one would expect an estimate of G with $G_{2,1} = G_{3,1}$, $G_{1,2} = G_{3,2}$, and $G_{1,3} = G_{2,3}$. The estimation procedure described above will not produce estimates satisfying the constraints, but a slight modification to the procedure will suffice as the constraints essentially reduce the dimension of the parameter space. The constrained (β, G) can be transformed to an unconstrained vector η with dimension lower than n^2. In the two-state refractory promoter example, $\eta = (\log\beta_1, \log G_{1,2}, \log G_{2,1})$. In the three-state Dirichlet distribution example, $\eta = (\log(\beta_1 - \beta_2), \log(\beta_2 - \beta_3), \log\beta_3, \log G_{2,1}, \log G_{1,2}, \log G_{1,3})$. Once the unconstrained parameter vector η is identified, a Gibbs sampler similar to the one described above can be applied to obtain posterior samples of η.

We conduct synthetic experiments to study the estimation performance of the above procedure. Four instances of the multistate promoter model described in (5.1) are considered in the experiments:

- a two-state model,
- a three-state model with $MSBM(G)$ being a Dirichlet distribution,
- a three-state model with $MSBM(G)$ having a symmetric structure in G, and
- a three-state model with $MSBM(G)$ having an asymmetric structure in G.

Three choices of sample size, $L = 100, 500, 1000$, are considered to investigate the performance of the sampler. Twenty datasets are generated for each model and each choice of L. The MCMC algorithm is run for five times with randomly generated initial values. In each run, after tuning the step size σ_j^2 in the proposal density, the MCMC algorithm is run for 20,000 iterations. The first 10,000 iterations are discarded as the burn-in period. Every 10 iterations of the remaining 10,000 iterations are kept for inference. The length of the burn-in period is chosen based on a pilot run of the algorithm. The convergence of the MCMCs is further diagnosed according to the Gelman-Rubin statistic [13] computed from the five chains for each dataset. The root mean squared error (RMSE) of the posterior mean estimator for each parameter is recorded for evaluation. It is computed as

$$\text{RMSE} = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(\hat{\theta}_i - \theta)^2},$$

where θ is parameter value used for generating data and $\hat{\theta}_i$ is the estimated value from the ith dataset. Smaller RMSE indicates better estimation performance. We also compute the empirical 95% credible interval of each parameter for each dataset. The (frequentist) coverage of the credible intervals for each parameter across 20 datasets are recorded, where coverage here refers to the proportion of times the interval contained the true value. In the following, we present the results for each model instance.

5.1.1. Two-state model. The parameter setting of the two-state model is adapted from [18]. More specifically, we set $\beta = (1000, 1)^\top$, and $G = \left(\begin{smallmatrix} -10 & 10 \\ .34 & -.34 \end{smallmatrix}\right)$ when generating data.

Table 1

RMSE and the coverage of 95% credible intervals for parameters in the two-state model among 20 datasets.

	L	β_1	β_2	$G_{2,1}$	$G_{1,2}$
RMSE	100	542.73	0.41	0.06	5.84
	500	412.50	0.13	0.03	4.36
	1000	319.84	0.13	0.02	3.36
Coverage	100	0.05	0.90	0.95	0.35
	500	0.45	1.00	0.95	0.70
	1000	0.80	0.90	1.00	0.80

Table 1 presents the RMSE and the coverage of 95% credible interval for parameters in the two-state model. The estimation clearly improves in terms of RMSE as the sample size L increases. The coverage of the credible intervals of β_2 and $G_{2,1}$ stays around 95% for all three choices of sample size while the coverage of β_1 and $G_{1,2}$ is low in the small sample case $L = 100$ and increases significantly as sample size increases. This poor coverage in the small sample case is the consequence of the large bias of the posterior mean estimator (shown in Figure 3) and the insufficient variation in the posterior distribution, as the information about the two parameters in a small dataset is not strong enough to dominate the information in the prior distribution.

It can be difficult to determine the appropriate sample size for estimating parameters in complex models. In our context, $L = 100$ in general poses a difficult statistical inference task. To better see this, Figure 4 displays the prior and the posterior density of the transformed parameters obtained from one dataset. For $\log(\beta_1 - \beta_2)$, the high density region of the prior is to the left of the true value. When $L = 100$, the center of the posterior density and, in fact, most of the posterior probability mass is confined within the high density region of the prior. As a result, the posterior mean of $\log(\beta_1 - \beta_2)$ underestimates the true value and the posterior distribution has insufficient variation. When $L = 1000$, the dataset contains stronger information about $\log(\beta_1 - \beta_2)$ and the posterior is not significantly affected by the biased information in the prior. Such a problem is not obviously reflected in the plot for $\log(G_{1,2})$ as the true value locates in the high density region of the prior. The information of $\log(\beta_2)$ and $\log(G_{2,1})$ in a small dataset is already strong enough to correct the biased information in the prior, producing credible intervals with good coverage and nearly unbiased estimators as shown in Table 1 and Figure 3.

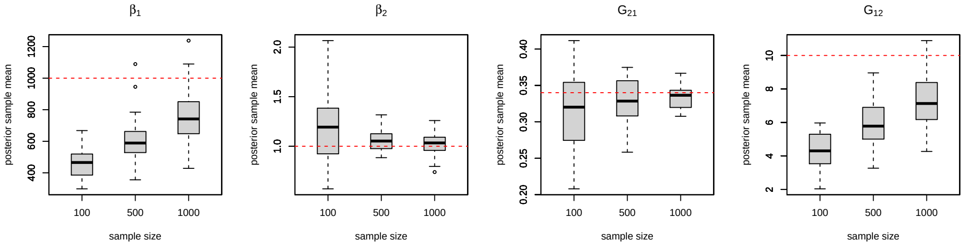

FIG. 3. *Boxplots of the posterior means of the parameters in the two-state model from 20 datasets. The horizontal dashed red lines indicate the true values.*

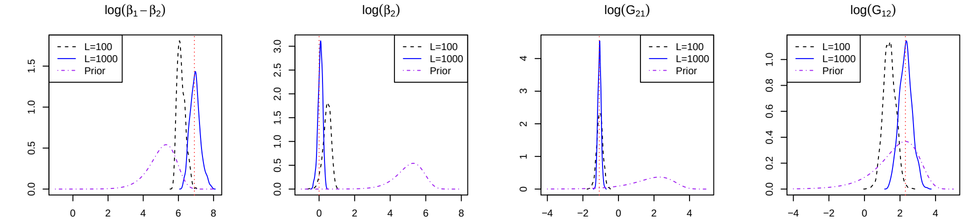

FIG. 4. *Prior and posterior densities of the transformed parameters in the two-state model. The vertical dotted red lines indicate the true values.*

5.1.2. Three-state Dirichlet model. The parameter values for the three-state Dirichlet model are $\beta = (1000, 100, 1)^\top$ and $G = \begin{pmatrix} -11.0 & 1.0 & 10.0 \\ 0.34 & -10.34 & 10.0 \\ 0.34 & 1.0 & -1.34 \end{pmatrix}$. With the equality constraints $G_{1,2} = G_{3,2}$, $G_{1,3} = G_{2,3}$, and $G_{2,1} = G_{3,1}$, the transformed parameter vector is

$$\eta = \big(\log(\beta_1 - \beta_2), \log(\beta_2 - \beta_3), \log(\beta_3), \log(G_{2,1}), \log(G_{1,2}), \log(G_{1,3})\big).$$

The results for this model are presented in Table 2 and Figures 5 and 6. In general, the estimation performance improves as the sample size increases. When $L = 100$, the large bias in the posterior means and the low coverage of the credible intervals are the results of insufficient information in the data and biased information in the prior. We would like to point out that when $L = 100$, the posterior densities of $\log(\beta_1 - \beta_2)$ and $\log(\beta_2 - \beta_3)$ almost match the prior densities, indicating the data provide little information about the parameter.

TABLE 2

RMSE and the coverage of 95% credible intervals for parameters in the three-state Dirichlet model among 20 datasets.

	L	β_1	β_2	β_3	$G_{2,1}$	$G_{3,2}$	$G_{1,3}$
RMSE	100	438.89	118.91	2.13	0.10	0.56	3.49
	500	327.74	112.28	1.17	0.07	0.35	3.38
	1000	269.92	63.50	0.78	0.06	0.23	2.80
Coverage	100	0.65	1.00	0.45	1.00	0.70	1.00
	500	0.90	1.00	0.85	0.95	0.80	1.00
	1000	0.90	1.00	0.90	1.00	0.85	0.95

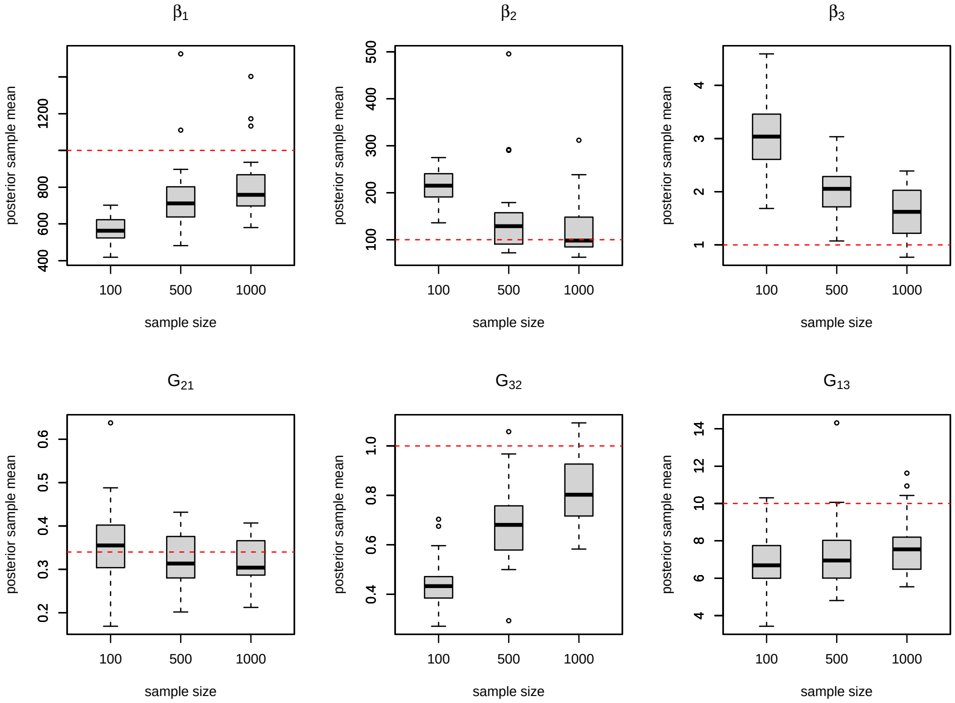

FIG. 5. *Boxplots of the posterior means of the parameters in the three-state Dirichlet model from 20 datasets. The horizontal dashed red lines indicate the true values.*

5.1.3. Symmetric three-state model. The parameters used in the general symmetric three-state model for generating data are $\beta = (300, 150, 20)^\top$ and $G =$

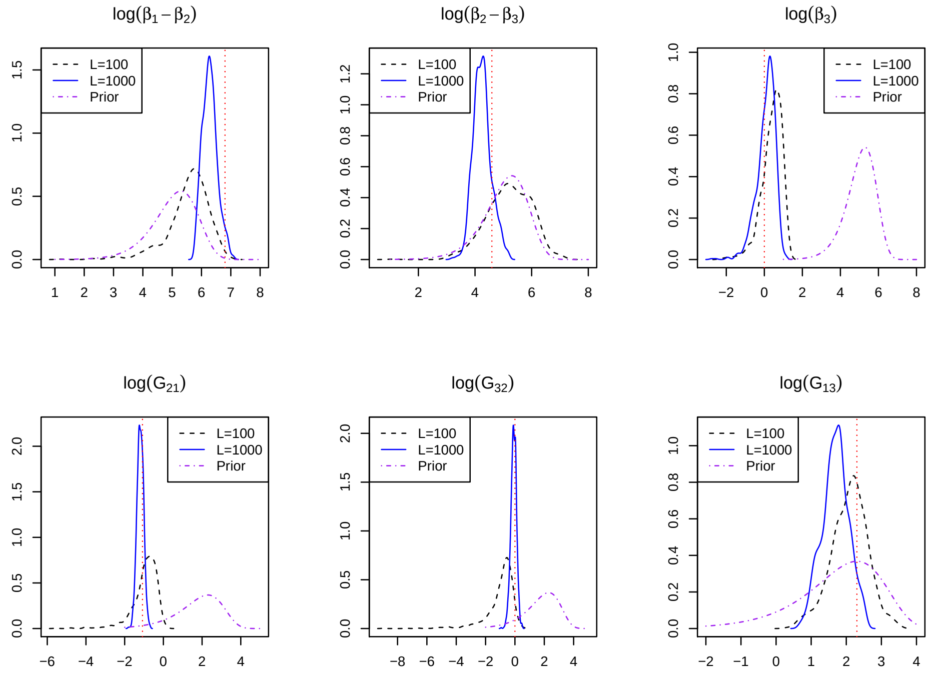

FIG. 6. *Prior and posterior densities of the transformed parameters in the three-state Dirichlet model. The vertical dotted red lines indicate the true values.*

$\begin{pmatrix} -2.0 & 2.0 & 0.0 \\ 0.5 & -1.0 & 0.5 \\ 0.0 & 2.0 & -2.0 \end{pmatrix}$. The model structure of G is known a priori, meaning that $G_{1,3}$ and $G_{3,1}$ are fixed at zero in all iterations of the Gibbs sampler and the transformed parameter vector is

$$\eta = \big(\log(\beta_1 - \beta_2), \log(\beta_2 - \beta_3), \log(\beta_3), \log(G_{1,2}), \log(G_{2,1}), \log(G_{2,3}), \log(G_{3,2})\big).$$

Table 3 and Figures 7 and 8 present the results for this model. In general, similar to the results for the two previous models, the estimation performance for all the parameters improves as the sample size increases. However, the RMSEs of $G_{2,1}$, $G_{2,3}$, and $G_{3,2}$ for $L = 500$ are greater than those for $L = 100$. This happens because one of the 20 datasets produces an estimated value far from the true value. This outlier distorts the values of RMSEs in the case $L = 500$. As the boxplots in Figure 7 show, the estimates from most of the datasets do improve as the sample size increases.

In Figure 8, the posterior densities of $\log(\beta_2 - \beta_3)$, $\log(G_{2,1})$, and $\log(G_{2,3})$ closely resemble the corresponding prior densities when $L = 100$, demonstrating weak information in the data once again.

We observe that the estimated values of the parameters are generally in close vicinity of the true values, especially when the sample size is large. Although we do not theoretically prove the identifiability of the general multistate promoter models, this observation suggests that the model is likely to be identifiable at least for the three-state case considered here.

Table 3

RMSE and the coverage of 95% credible intervals for parameters in the symmetric three-state model among 20 datasets.

	L	β_1	β_2	β_3	$G_{1,2}$	$G_{2,1}$	$G_{2,3}$	$G_{3,2}$
RMSE	100	133.42	51.68	8.28	11.80	7.29	7.80	2.91
	500	132.94	40.69	6.09	9.01	7.88	18.56	4.07
	1000	117.02	10.88	5.11	4.66	0.59	0.16	0.35
Coverage	100	0.70	1.00	1.00	0.20	0.95	0.60	0.70
	500	0.75	1.00	0.95	0.70	0.80	0.90	0.90
	1000	0.95	0.95	1.00	0.95	0.95	0.95	1.00

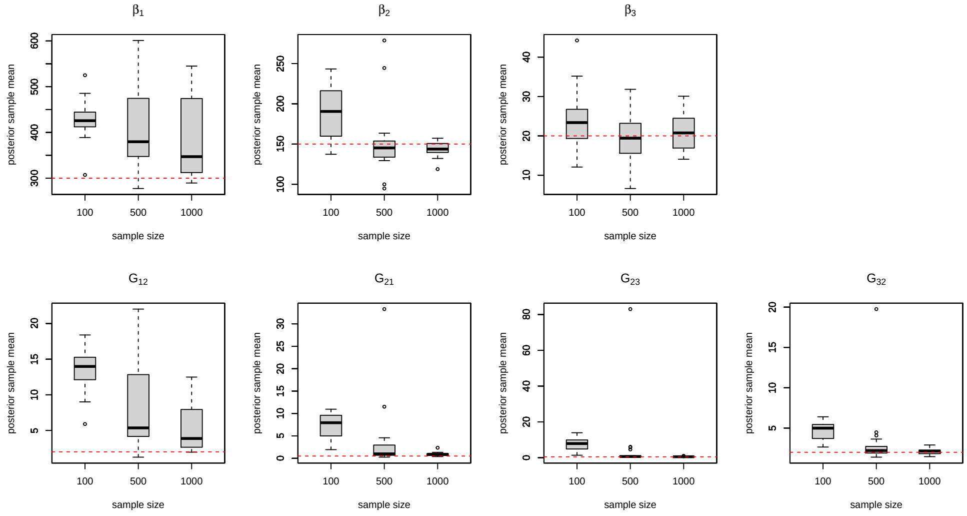

Fig. 7. *Boxplots of the posterior means of the parameters in the symmetric three-state model from 20 datasets. The horizontal dashed red lines indicate the true values.*

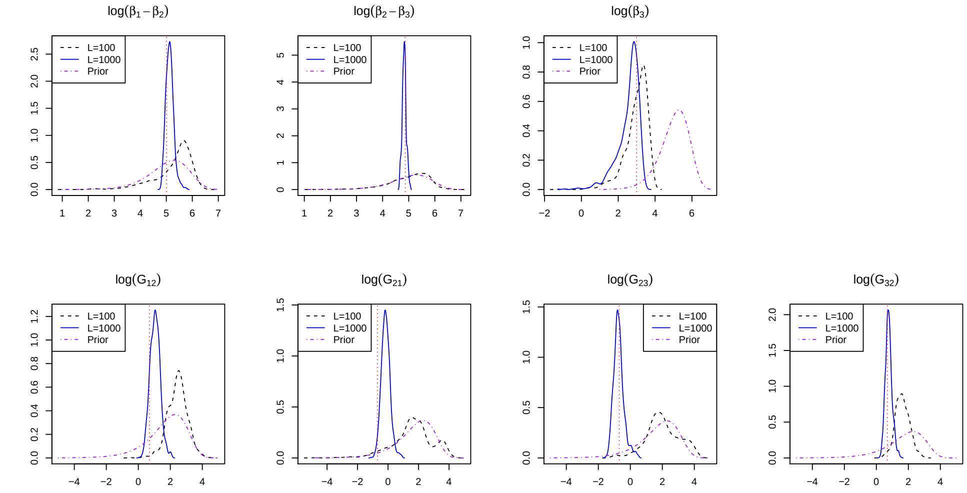

Fig. 8. *Prior and posterior densities of the transformed parameters in the symmetric three-state model. The vertical dotted red lines indicate the true values.*

5.1.4. Asymmetric three-state model. We now consider another general three-state model with $\beta = (300, 150, 20)^\top$ and $G = \begin{pmatrix} -1.0 & 1.0 & 0.0 \\ 0.0 & -0.5 & 0.5 \\ 1.0 & 0.5 & -1.5 \end{pmatrix}$. The G matrix does not have a symmetric pattern as the one in Section 5.1.3. We again assume the structure of G is known when estimating the parameters. The transformed parameter vector is

$$\eta = \big(\log(\beta_1 - \beta_2), \log(\beta_2 - \beta_3), \log(\beta_3), \log(G_{1,2}), \log(G_{2,3}), \log(G_{3,1}), \log(G_{3,2})\big).$$

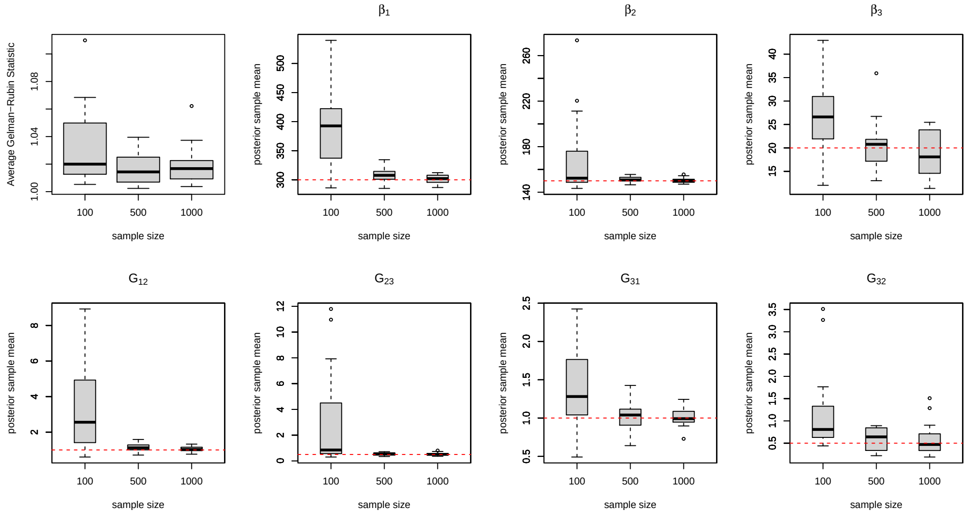

Fig. 9. *Top-left panel: boxplots of the average of Gelman-Rubin statistics of different transformed parameters in the asymmetric three-state model for 20 datasets. Other panels: boxplots of the posterior means of the parameters in the asymmetric three-state model from 20 datasets. The horizontal dashed red lines indicate the true values.*

The Gelman-Rubin statistic averaged over the transformed parameters for each dataset are presented in the first panel of Figure 9 to demonstrate the convergence of the MCMC algorithm. For all three choices of L, the Gelman-Rubin statistic is below 1.2 for all 20 datasets, indicating reasonable convergence of the MCMC algorithm. In addtion, we observe that the algorithm has better convergence (smaller Gelman-Rubin statistic values) when the sample size is larger.

Table 4

RMSE and the coverage of 95% credible intervals for parameters in the asymmetric three-state model among 20 replication.

	L	β_1	β_2	β_3	$G_{1,2}$	$G_{2,1}$	$G_{2,3}$	$G_{3,2}$
RMSE	100	104.24	37.37	10.03	3.29	4.36	0.69	1.05
	500	14.42	2.75	5.13	0.27	0.12	0.18	0.26
	1000	7.21	2.17	5.08	0.15	0.11	0.12	0.34
Coverage	100	0.95	1.00	1.00	0.9	0.80	1.00	1.00
	500	1.00	1.00	1.00	1.00	0.95	0.95	1.00
	1000	1.00	1.00	0.95	1.00	0.85	1.00	0.95

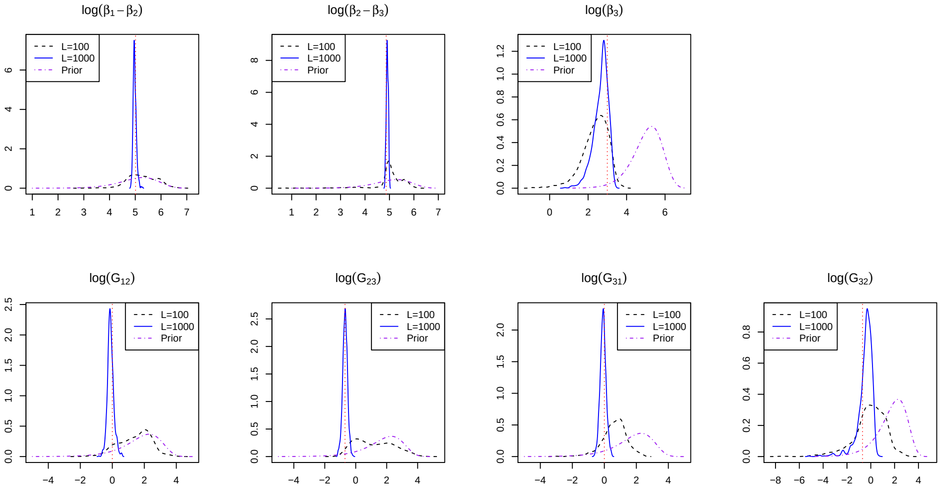

FIG. 10. *Prior and posterior densities of the transformed parameters in the asymmetric three-state model. The vertical dotted red lines indicate the true values.*

The results of parameter estimation are shown in Table 4 and Figures 9 and 10. They once again suggest that the estimation performance improves as the sample size increases and that the model parameters are identifiable.

In addition to the marginal inference presented above, Figures 11 and 12 exhibit pairwise joint posterior densities of the transformed parameters. The densities are obtained by kernel density estimation. The figures suggest that there are varying degree of dependence among the parameters.

5.2. Model selection.. In the previous section, parameters are estimated with the assumption that the structure of the multistate promoter model (the number of states, the position of zero elements, etc.) is known. In this section, we consider selecting an appropriate model structure according to the Bayesian Information Criterion (BIC) [27].

Suppose we want to choose from several models with different structures. Given observed data $M_1, \ldots, M_L$ for each candidate model, the model parameter can be estimated using the procedure described in Section 5.1. Recall that η denotes the unconstrained parameter vector. Let q denote the dimension of η and $\hat{\eta}$ denote the estimated value of η. The BIC for a model with parameter vector η is defined as

$$\text{BIC} = -2 \log f(M_1, \ldots, M_L \mid \hat{\eta}) + \log(L)q,$$

where $f(M_1, \ldots, M_L \mid \hat{\eta})$ is the probability of observing the sample under the model when $\eta = \hat{\eta}$. After computing the BIC for each candidate model, the model with the smallest BIC is chosen as the most appropriate one.

In general, a more complex model (a model with more parameters) produces a higher $f(M_1, \ldots, M_L \mid \hat{\eta})$. However, an overly complex model is undesired in practice as it brings in instability in statistical inference without improving much the explanatory power. If two models give similar $f(M_1, \ldots, M_L \mid \hat{\eta})$, the one with the fewer parameters is favored by BIC due to the term $\log(L)q$. Therefore, BIC can help us select the simplest model that explains the observed data reasonably well.

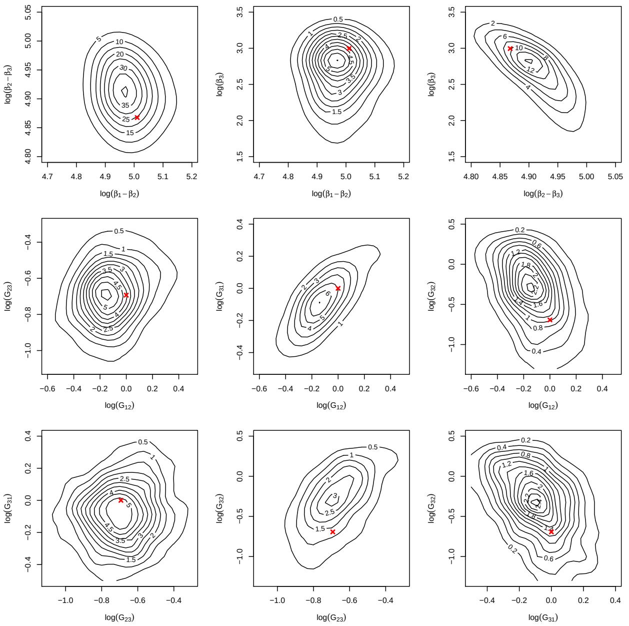

FIG. 11. *Contour plots of the pairwise joint posterior densities from one dataset with $L = 1000$. The two parameters in each plot are from the same parameter group (β or G). The joint densities are obtained by kernel density estimation. The red crosses indicate the true values.*

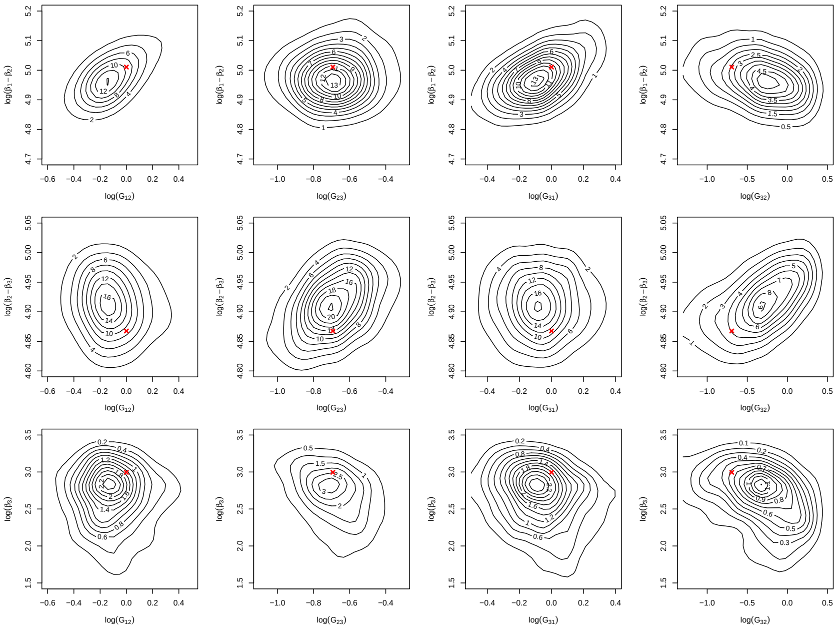

FIG. 12. *Contour plots of the pairwise joint posterior densities from one dataset with $L = 1000$. The two parameters in each plot are from different parameter groups (β or G). The joint densities are obtained by kernel density estimation. The red crosses indicate the true values.*

We demonstrate the model selection approach again using synthetic experiments. We still consider three choices of sample size L, 100, 500, and 1000. For each choice, 20 datasets are generated from the three-state model in Section 5.1.3. There we estimated the nonzero parameters in β and G assuming the number of states and structures of β and G are known. Differently, in this section, we examine whether the correct underlying model (the number of states and the nonzero elements) can be identified from several candidate models using the BIC framework. The candidate models are the following:

- the true model, that is the three-state model with $G_{13} = G_{31} = 0$,
- a two-state model with all parameters being nonzero,
- a three-state refractory promoter model with $\beta_2 = \beta_3 = G_{13} = G_{31} = 0$, and
- a general three-state model with all parameters being nonzero.

Table 5
Distribution of the selected model among 20 datasets.

	True	Refractory	Two-state	General three-state
$L = 100$	0	0	20	0
$L = 500$	14	0	6	0
$L = 1000$	20	0	0	0

For each dataset, we fit all four candidate models and compute BIC for each model fit. The model with the lowest BIC is selected. Table 5 gives the distributions of the selected model among the 20 datasets. It shows that the model selection accuracy improves as the sample size increases. We note that the true model is selected for all 20 datasets when $L = 1000$. As we have seen in Section 5.1.3, the datasets with $L = 100$ in general provide little information about most of the parameters, producing log-likelihood functions that do not vary much for different parameter values. As a result, the log-likelihood component in BIC are comparable for the four candidate models. For 15 out of 20 datasets, the difference in the log-likelihood functions of the four models is within 2.00. Therefore, the penalty term $\log(L)q$ dominates model selection and the simplest model (two-state model) is selected for all 20 datasets.

6. Clumped constructions and proof of Theorem 4.4. Viewing the unbounded model where $(E(t), M(t))$ serves as a 'promoter', define

$$\tilde{e} = (i, m) \qquad \tilde{m} = p \qquad \tilde{\beta}_{(i,m)} = \alpha m \qquad \tilde{\delta} = \gamma$$

$$\begin{aligned}\tilde{G}^{\infty}_{(i,m),(j,n)} = {} & G_{i,j}\mathbb{1}(i \neq j, m = n) + \beta_i \mathbb{1}(i = j, n = m + 1) + \delta m \mathbb{1}(i = j, n = m - 1) \\ & - \mathbb{1}(i = j, m = n)\left[-G_{i,i} + \beta_i + \delta m\right]\end{aligned}$$

As commented before Theorem 4.4, note that $\tilde{G}^{\infty}$ is not a bounded generator matrix since its diagonal entries grow unbounded with m, that is $\theta(\tilde{G}^{\infty}) = \infty$. As a result, $\tilde{G}^{\infty}$ cannot be normalized by some value of θ such that $I + \tilde{G}^{\infty}/\theta$ is a stochastic kernel, preventing consideration of non-clumped stick-breaking measure construction of the stationary distribution π_2^{∞} of the unbounded model parameterized by $\tilde{G}^{\infty}$ as in the discussion of the 'bounded' mRNA-protein model.

We will however derive a type of clumped stick-breaking form of the stationary distribution π_2^{∞} of the process $(E^{\infty}(t), M^{\infty}(t), P^{\infty}(t))$ through a limit with respect to the 'bounded' mRNA $(E^c(t), M^c(t))$ and mRNA-protein $(E^c(t), M^c(t), P^c(t))$ models (cf. Definition 4.1 and Figure 2). To this end, we view the bounded mRNA-protein model and its stationary distribution π_2^c on the full state space $(\tilde{e}, \tilde{m}) \in (\mathfrak{X} \times \mathbb{N}_0) \times \mathbb{N}_0$, explicitly denoting dependence on the integer-valued capacity parameter c:

$$\tilde{e} = (i, m) \qquad \tilde{m} = p \qquad \tilde{\beta}_{(i,m)} = \alpha m \qquad \tilde{\delta} = \gamma$$

and

$$\tilde{G}^c_{(i,m),(j,n)} = \tilde{G}^\infty_{(i,m),(j,n)} \mathbb{1}(m, n \leq c)$$

with the necessary modification of $\tilde{G}^c_{(i,c),(i,c)}$ to accord with the formula (4.1) and to preserve generator structure. Note that for all $m < c$ that

$$\tilde{G}^{c,*}_{(i,m),(i,m)} = \tilde{G}^{\infty,*}_{(i,m),(i,m)} \tag{6.1}$$

By inspection of the rates, we may couple the bounded and unbounded processes so that $E^c(t) \equiv E^\infty(t)$, and also $M^c(t) \leq M^\infty(t)$ and $P^c(t) \leq P^\infty(t)$, and hence also in the $t \uparrow \infty$ limit. Since the stationary distribution π_2^∞ of the unbounded process integrates $e^{\epsilon_1 i + \epsilon_2 m + \epsilon_3 p}$ for some $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ (Lemma 4.3), the stationary measures $\pi_2^c \sim (E^c, M^c, P^c)$ indexed in c are tight and so relatively compact in the space of probability measures on $\mathfrak{X} \times \mathbb{N}_0^2$. In Appendix C, we provide more details to show this tightness.

We now show weak convergence of the clumped stick-breaking forms of π_2^c to π_2^∞. Given uniform exponential moments, then the joint moments of $(M^c)^k (P^c)^\ell$ would converge to those of $(M^\infty)^k (P^\infty)^\ell$. In this way, the last statement of Theorem 4.4 would hold.

Let $\mu^c = \pi_1^c(i, m | G, \beta, \delta)$ be the unique stationary measure of $(E^c(t), M^c(t))$ having support $\{(i, m) : m \leq c\}$. Let also $\mu^\infty = \pi_1(i, m | G, \beta, \delta)$ be the established stationary distribution π_1 of $(E^\infty(t), M^\infty(t))$, the usual mRNA portion of the (unbounded) multistate promoter process. Note that this stationary distribution is unique as the process $(E^\infty(t), M^\infty(t))$ is irreducible.

We now argue that μ^c converges to μ^∞, which is the stationary distribution of $\tilde{G}^\infty$. The measure μ^c is a projection of π_2^c to the pair (E^c, M^c), and so the sequence μ^c indexed in c is tight from the tightness argument of π_2^c. We have $\tilde{G}^c$ converges pointwise to $\tilde{G}^\infty$ as $c \uparrow \infty$,and that $\tilde{G}^c$ is banded (with respect to lexicographical ordering of states (i, m)). By consideration of the balance equation $\mu^c \tilde{G}^c = 0$, it follows that limit points $\mu_{\lim}$ satisfy the balance equation $\mu_{\lim} \tilde{G}^\infty = 0$ which has unique solution μ^∞. Hence, μ^c converges to μ^∞.

Similarly, we argue that π_2^c converges to π_2^∞. Specifically, let $\check{G}^c$ be the generator associated with the process $(E^c(t), M^c(t), P^c(t))$ for $0 \leq c \leq \infty$. The generators $\check{G}^c$ are banded for an appropriate choice of ordering on states (i, m, p) and converge pointwise to $\check{G}^\infty$. Since the sequence $\{\pi_2^c\}_{c \geq 0}$ is also tight, every limit point $\pi_{2,\lim}$ of the sequence must be a distribution which satisfies $0 = \pi_{2,\lim} \check{G}^\infty$. Since π_2^∞ is the only such distribution, π_2^c converges to π_2^∞.

We now consider the clumped stick-breaking construction with respected to the bounded model. For each value of $c < \infty$, define a Markov chain $\mathbf{Z}^c$ on state space $\mathfrak{X} \times \{0, 1, \ldots, c\}$ with initial measure μ^c and non-repeating transition kernel

$$K^c_{\tilde{e}, \tilde{f}} = \frac{\tilde{G}^{c,*}_{\tilde{e}, \tilde{f}}}{-\tilde{G}^{c,*}_{\tilde{e}, \tilde{e}}} \mathbb{1}(\tilde{e} \neq \tilde{f}) + \mathbb{1}(\tilde{e} = \tilde{f} \textit{ and } \tilde{e}_2 > c).$$

Note that this Markov chain only reaches states with $m \le c$.

Given $\mathbf{Z}^c$, let $\mathbf{W}^c$ be an independent sequence of $W_j^c \sim \mathit{Beta}(1, -\tilde{G}^{c,*}_{Z_j^c, Z_j^c}/\gamma)$ variables, and $\mathbf{R}^c = \{W_j^c \prod_{i=1}^{j-1}(1 - W_i^c)\}_{j\ge 1}$ be constructed from $\mathbf{W}^c$ as a residual allocation model. Define the stick-breaking measure

$$X^c(\,\cdot\,) = \sum_{j=1}^{\infty} R_j^c \delta_{Z_j^c}(\,\cdot\,).$$

Now, since by Theorem 3.7 and the clumped representation afforded by Proposition 2.3, we know that if

$$P^c|(\mathbf{Z}^c, \mathbf{R}^c) \sim \mathit{Poisson}\left(\frac{\alpha}{\gamma} \sum_{(i,m)} m\; X^c(i,m)\right)$$

and $Z_1^c = (E^c, M^c) \sim \mu^c$, then the stationary distribution $\pi_2^c \sim (E^c, M^c, P^c)$ for the bounded joint process can be written in terms of μ^c and the Poisson mixture $P^c|(\mathbf{Z}^c, \mathbf{R}^c)$.

We now show that one can take a limit as $c \to \infty$. Since (a) μ^c converges pointwise to μ^∞ of $\tilde{G}^\infty$ and (b) the uniformly banded matrices $\tilde{G}^c$ converge entrywise to $\tilde{G}^\infty$, we have

$$\begin{aligned}
\lim_{c\to\infty} K^c_{\tilde{e},\tilde{f}} &= \lim_{c\to\infty} \frac{\tilde{G}^{c,*}_{\tilde{e},\tilde{f}}}{-\tilde{G}^{c,*}_{\tilde{e},\tilde{e}}} \mathbb{1}(\tilde{e} \ne \tilde{f}) = \lim_{c\to\infty} \frac{[D(\mu^c)^{-1}\tilde{G}^{c,T} D(\mu^c)]_{\tilde{e},\tilde{f}}}{-\tilde{G}^c_{\tilde{e},\tilde{e}}} \mathbb{1}(\tilde{e} \ne \tilde{f}) \\
&= \frac{[D(\mu^\infty)^{-1}\tilde{G}^{\infty,T} D(\mu^\infty)]_{\tilde{e},\tilde{f}}}{-\tilde{G}^\infty_{\tilde{e},\tilde{e}}} \mathbb{1}(\tilde{e} \ne \tilde{f}) \\
&= \frac{\tilde{G}^{\infty,*}_{\tilde{e},\tilde{f}}}{-\tilde{G}^{\infty,*}_{\tilde{e},\tilde{e}}} \mathbb{1}(\tilde{e} \ne \tilde{f}) = K^\infty_{\tilde{e},\tilde{f}}. \qquad (6.2)
\end{aligned}$$

Let now $\mathbf{Z}^\infty$ be a Markov chain with initial distribution μ^∞ and kernel K^∞. Given $\mathbf{Z}^\infty$, let $\mathbf{W}^\infty$ be an independent sequence of $W_j^\infty \sim \mathit{Beta}(1, -\tilde{G}^{\infty,*}_{Z_j^\infty, Z_j^\infty}/\gamma)$ variables, and $\mathbf{R}^\infty$ be constructed from $\mathbf{W}^\infty$ as a residual allocation model.

Let $z^n = \{z_j\}_{j=1}^n$ be a deterministic sequence of states $z_j = (e_j, m_j)$. Since $\tilde{G}^{c,*}_{\tilde{e},\tilde{e}}$ converges to $\tilde{G}^{\infty,*}_{\tilde{e},\tilde{e}}$ as $c \uparrow \infty$, the conditional distribution of $\{R_j^c\}_{j=1}^n|\{Z_j^c\}_{j=1}^n = z^n$ converges to that of $\{R_j^\infty\}_{j=1}^n|\{Z_j^\infty\}_{j=1}^n = z^n$ as $c \uparrow \infty$.

Also, as μ^c converges to μ^∞ and $K^c_{\tilde{e},\tilde{f}}$ converges to $K^\infty_{\tilde{e},\tilde{f}}$, the distribution of $\{Z_j^c\}_{j=1}^n$ converges to that of $\{Z_j^\infty\}_{j=1}^n$ as $c \uparrow \infty$.

We conclude then, since $1 = \sum_{j\ge1} R_j^\infty = \sum_{j\ge1} R_j^c$ for each c, as $c \uparrow \infty$ that $X^c(\cdot)$ converges weakly to

$$X^\infty(\,\cdot\,) = \sum_{j=1}^{\infty} R_j^\infty \delta_{Z_j^\infty}(\,\cdot\,),$$

and $P^c|(\mathbf{Z}^c, \mathbf{R}^c) \sim \mathit{Poisson}\Big(\alpha \sum_{(i,m)} m\; X^c(i,m)\Big)$ converges weakly to

$$P^\infty|(\mathbf{Z}^\infty, \mathbf{R}^\infty) \sim \mathit{Poisson}\left(\frac{\alpha}{\gamma} \sum_{(i,m)} m\; X^\infty(i,m)\right)$$

and $Z_1^c = (E^c, M^c) \sim \mu^c$ converges to $Z_1^\infty = (E^\infty, M^\infty) \sim \mu^\infty$.

Hence,

$$\lim_{c \to \infty} (E^c, M^c, P^c) \overset{d}{=} (E^\infty, M^\infty, P^\infty) \sim \pi_2^\infty(\,\cdot\,,\,\cdot\,,\,\cdot\,|G, \beta, \delta, \alpha, \gamma),$$

and we conclude the proof of Theorem 4.4.

7. Summary and conclusion.

Through relations between seemingly disparate objects, namely stick-breaking Markovian measures, empirical distribution limits of certain time-inhomogeneous Markov chains, and Poisson mixture representations of stationary distributions in multistate mRNA promoter models, we identify the stationary joint distribution of promoter state and mRNA level via a constructive stick-breaking formula. Moreover, we also consider protein interactions influenced by mRNA levels and find a stick-breaking formulation of the joint promoter, mRNA and protein levels. Interestingly, this formula with respect to un-bounded protein levels involves a 'clumped' representation of the stick-breaking measure. These results constitute what seem to be a significant advance over previous work, which approximate stationary distributions or restrict solvable computations to specialized settings.

Importantly, the stick-breaking construction allows to sample directly from the stationary distribution, permitting inference procedures for parameters as well as model selection. Such a feature improves over sampling from the stationary distributions by running the process for a length of time. Our experiments show that, for various choices of the model settings, the inference procedures based on the stick-breaking construction are able to estimate model parameters accurately and select the underlying model correctly when the sample size is sufficiently large. In addition, the form of the stationary distribution allows to compute mixed moments between mRNA and protein levels, which might bear upon correlation analysis as in [1].

Although in principle the 'stick-breaking' apparatus can be used to identify stationary distributions in linear chains of reactions, a natural problem for further study is to understand the role of 'feedback' in constructing the stationary distribution in more general networks, say those where protein or mRNA levels influence promoter switching rates. It would also be of interest to study the stationary distributions and connections with 'stick-breaking' in more general chemical reaction networks, with possibly non-Poisson representations, not necessarily those with Poisson character as in [10].

The experiments in Section 5.1 show that when the sample size is small, the proposed Bayesian inference for the stick-breaking model could be biased due to biased prior information and insufficient information in the data. One might mitigate the problem by placing non-informative priors (e.g. a flat prior) on the transformed parameters so that prior information will not dominate the inference. However, as shown in the experiments, small datasets may contain extremely weak information about model parameters, making the inference mainly to rely on the prior. In this case, if a non-informative prior is used, the MCMC algorithm may encounter convergence issues. In practice, single-cell mRNA data, which are the main motivation and application area of our model, often have a large sample size (a large number of cells). Since the proposed inference procedure tends to produce accurate results for large datasets ($L = 1000$ in our experiments), we do not expect the proposed method to produce results with large bias in practice for weakly or moderately informative priors (e.g. the one used in the our experiments).

We have also discussed the notion of identifiability of parameters and believe mRNA levels M can identify the promoter switching rates G and intensities β when the $\beta = \{\beta_i\}$ components are known to be distinct. Our numerical results indicate that this is the case. We leave the theoretical justification for future investigation. On top of the identifiability, establishing posterior consistency of the proposed Bayesian inference is another piece of future work. Finally, of course, a next step is to understand inference of parameters and model selection from laboratory cell readings.

Appendix A. Proof of Lemma 3.11. First, noting the form of G, (3.2) reduces to

$$w\partial_w\Phi = wD(\beta)\Phi + \theta(\vec{1}\cdot\Phi)\mu - \theta\Phi$$

and (3.4) becomes

$$w\beta\cdot\partial_w\Phi = w\beta\cdot D(\beta)\Phi + \theta\big[(\vec{1}\cdot\Phi)(\beta\cdot\mu) - \beta\cdot\Phi\big].$$

A similar equation holds with respect to Φ'. Note the relations in (3.3). In particular as $\Phi(0) = \mu$ and $\Phi'(0) = \mu'$, note $\beta\cdot\mu = \beta'\cdot\mu'$. Hence, since $w\beta\cdot\partial_w\Phi = w\beta'\cdot\partial_w\Phi'$, $\beta\cdot\Phi = \beta'\cdot\Phi'$ and $\vec{1}\cdot\Phi = \vec{1}\cdot\Phi'$, with respect to the above display, $\beta\cdot D(\beta)\Phi(w) = \beta'\cdot D(\beta')\Phi'(w)$.

Now, by differentiating $\beta\cdot D(\beta)\Phi$, multiplying by w and substituting the previous expression for $w\partial_w\Phi$, one obtains another equation

$$w\beta\cdot D(\beta)\partial_w\Phi = w\beta\cdot D(\beta)^2\Phi + \theta\beta\cdot D(\beta)\big[\mu(\vec{1}\cdot\Phi) - \Phi\big].$$

Then, considering the like equation with respect to Φ', we arrive at $\beta\cdot D(\beta)^2\Phi = \beta'\cdot D(\beta')^2\Phi'$.

One can iterate this process and obtain, for $k \geq 1$, that $\beta\cdot D(\beta)^k\Phi(w) = \beta'\cdot D(\beta')^k\Phi'(w)$. Hence, evaluating at $w = 0$, for $k \geq 1$, we have

$$\sum_i (\beta_i)^k\mu_i = \sum_i (\beta_i')^k\mu_i'. \tag{A.1}$$

Since we know that β_1 and β_1' are the unique positive maxima in the entries of the vectors β, β', one gets from (A.1) that $\beta_1 = \beta_1'$ (confirming what we knew a priori) and so $\mu_1 = \mu_1'$ by considering the asymptotics as k gets large. We may then subtract $\beta_1^k\mu_1$ from both sides, and repeat the argument since the components of β, β' are assumed to be strictly ordered. Successively, we recover that $\beta_j = \beta_j' > 0$ and $\mu_j = \mu_j'$ for $1 \leq j \leq |\mathfrak{X}| - 1$. Finally, from the relation $\vec{1}\cdot\mu = \vec{1}\cdot\mu' = 1$, we get also $\mu_{|\mathfrak{X}|} = \mu'_{|\mathfrak{X}|}$ and so $\beta_{|\mathfrak{X}|} = \beta'_{|\mathfrak{X}|}$. In particular, vectors $\beta = \beta'$ and the constant stochastic operators $Q = \vec{1}\,\mu^t = \vec{1}\,(\mu')^t = Q'$.

Appendix B. Proof of Lemma 4.3. We will verify part (iii) of Theorem 1.1 [21], namely, for a Ψ-irreducible, aperiodic Markov process, “there exists a closed, small set C and an extended-valued non-negative function V satisfying $V(x_0) < \infty$ for some x_0 (in the state-space), such that Condition (V3) holds”, from which the lemma follows by applying the equivalent part (i) of Theorem 1.1 [21]. Here, to satisfy Condition (V3) in [21], we take f in Condition (V3) equal to V specified later–see (B.1).

By inspection, the unbounded joint multistate mRNA-protein continuous-time process in Definition 4.2 is irreducible and aperiodic (Ψ-irreducible when Ψ is counting measure). Recall that we enumerate $\mathfrak{X} = \{1, 2, \ldots, |\mathfrak{X}|\}$. We now check that $C = \mathfrak{X}\times\{0, 1, \ldots, m_0\}\times\{0, 1, \ldots, p_0\} \subset \mathfrak{X}\times\mathbb{N}_0^2$ is a ‘small’ set as in [21] for $m_0, p_0 < \infty$

to be chosen later: Namely, the time t probability starting from $x \in C$ to reach a set $A \in \mathfrak{X} \times \mathbb{N}_0^2$ is bounded below as

$$p^{(t)}(x, A) \geq p^{(s)}(x, y_0)\nu(A),$$

where y_0 is a fixed point in C and s is chosen large enough so that, for a small $\varepsilon > 0$, $\min_{x\in C} p^{(s)}(x, y_0) \geq \varepsilon > 0$, the time $t > s$, and $\nu(A) := p^{(t-s)}(y_0, A)$.

Let L be the generator of the unbounded process, and define the function $V : \mathfrak{X} \times \mathbb{N}_0^2 \to \mathbb{R}$ by $V(i, m, p) = \exp\{\epsilon_1 i + \epsilon_2 m + \epsilon_3 p\}$. Then, by computation,

$$\begin{aligned} LV(i, m, p) = V(i, m, p)\Big[\sum_{j\in\mathfrak{X}} \big(e^{\epsilon_1(j-i)} - 1\big)G_{i,j} \\ + \big(e^{\epsilon_2} - 1\big)\beta_i + \big(e^{\epsilon_3} - 1\big)\alpha m + \big(e^{-\epsilon_2} - 1\big)\delta m + \big(e^{-\epsilon_3} - 1\big)\gamma p\Big]. \end{aligned}$$

Note that

$$\begin{aligned} \max_{i\in\mathfrak{X}} \Big[\sum_{j\in\mathfrak{X}} \big(e^{\epsilon_1(j-i)} - 1\big)G_{i,j} \\ + \big(e^{\epsilon_2} - 1\big)\beta_i + \big(e^{\epsilon_3} - 1\big)\alpha m + \big(e^{-\epsilon_2} - 1\big)\delta m + \big(e^{-\epsilon_3} - 1\big)\gamma p\Big] < -1 \end{aligned}$$

when both (i) $\alpha(e^{\epsilon_3} - 1) + \delta(e^{-\epsilon_2} - 1) < 0$, which for given $\epsilon_2 > 0$ holds for sufficiently small $\epsilon_3 > 0$, and (ii) $m > m_0$ is sufficiently large depending on $\{\epsilon_k\}_{k=1,2,3}$, G, $\{\beta_k\}_{k\in\mathfrak{X}}$, and α, δ, γ.

Then, condition (V3) in [21] holds with $f = V \geq 1$:

$$LV \leq -V + b\mathbb{1}(C) \tag{B.1}$$

where $b = \max_{i\in\mathfrak{X}, m\leq m_0, p\leq p_0} V(i, m, p) + \max_{i\in\mathfrak{X}, m\leq m_0, p\leq p_0} |LV|(i, m, p) < \infty$ say.

Hence, as desired, the lemma statement follows from part (iii) of Theorem 1.1 in [21].

Appendix C. Tightness of $\{\pi_2^c\}$. To show tightness of $\{\pi_2^c\}$, by definition, we show that $\lim_{R\uparrow\infty} \sup_c \pi_2^c(\mathfrak{X} \times \{R, R + 1, \ldots\}^2) = 0$.

By Markov's inequality, and that $M^c \leq M^\infty$, $P^c \leq P^\infty$ with respect to the stationary distributions π_2^c and π_2^∞, we have

$$\begin{aligned} \pi_2^c(\mathfrak{X} \times \{R, R + 1, \ldots\}^2) &\leq e^{-\epsilon_1 - \epsilon_2 R - \epsilon_3 R} \int e^{\epsilon_1 i + \epsilon_2 m + \epsilon_3 p} d\pi_2^c \\ &\leq e^{-\epsilon_1 - \epsilon_2 R - \epsilon_3 R} \int e^{\epsilon_1 i + \epsilon_2 m + \epsilon_3 p} d\pi_2^\infty \end{aligned}$$

for $\epsilon_1, \epsilon_2, \epsilon_3 > 0$.

Since by Lemma 4.3, π_2^∞ integrates $e^{\epsilon_1 i + \epsilon_2 m + \epsilon_3 p}$ for specified $\epsilon_1, \epsilon_2, \epsilon_3 > 0$, the claim follows.

Acknowledgments. The authors would like to thank two anonymous referees for their constructive comments that led to substantial improvement of the manuscript.

REFERENCES

[1] C. Albayrak, C. A. Jordi, C. Zechner, J. Lin, C. A. Bichsel, M. Khammash, and S. Tay, *Digital quantification of proteins and mrna in single mammalian cells*, Molecular cell, 61 (2016), pp. 914–924.

[2] F. Bouguet and B. Cloez, *Fluctuations of the empirical measure of freezing markov chains*, Electronic Journal of Probability, 23 (2018), pp. 1–31.

[3] Z. Cao and R. Grima, *Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells*, Proceedings of the National Academy of Sciences, 117 (2020), pp. 4682–4692.

[4] K. Choudhary and A. Narang, *Urn models for stochastic gene expression yield intuitive insights into the probability distributions of single-cell mrna and protein counts*, Physical Biology, 17 (2020), p. 066001.

[5] A. Coulon, O. Gandrillon, and G. Beslon, *On the spontaneous stochastic dynamics of a single gene: complexity of the molecular interplay at the promoter*, BMC systems biology, 4 (2010), pp. 1–18.

[6] J. Dattani and M. Barahona, *Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization*, Journal of The Royal Society Interface, 14 (2017), p. 20160833.

[7] M. H. Davis, *Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models*, Journal of the Royal Statistical Society: Series B (Methodological), 46 (1984), pp. 353–376.

[8] Z. Dietz, W. Lippitt, and S. Sethuraman, *Stick-breaking processes, clumping, and markov chain occupation laws*, Sankhya A, (2021), pp. 1–43.

[9] Z. Dietz and S. Sethuraman, *Occupation laws for some time-nonhomogeneous markov chains*, Electronic Journal of Probability, 12 (2007), pp. 661–683.

[10] C. W. Gardiner and S. Chaturvedi, *The poisson representation. i. a new technique for chemical master equations*, Journal of Statistical Physics, 17 (1977), pp. 429–468.

[11] A. E. Gelfand and A. F. Smith, *Sampling-based approaches to calculating marginal densities*, Journal of the American statistical association, 85 (1990), pp. 398–409.

[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, Chapman and Hall/CRC, 1995.

[13] A. Gelman and D. B. Rubin, *Inference from iterative simulation using multiple sequences*, Statistical science, (1992), pp. 457–472.

[14] S. Ghosal and A. Van der Vaart, *Fundamentals of nonparametric Bayesian inference*, vol. 44, Cambridge University Press, 2017.

[15] L. Ham, D. Schnoerr, R. D. Brackston, and M. P. Stumpf, *Exactly solvable models of stochastic gene expression*, The Journal of Chemical Physics, 152 (2020), p. 144106.

[16] W. K. Hastings, *Monte carlo sampling methods using markov chains and their applications*, Biometrika, 57 (1970), p. 97.

[17] U. Herbach, *Stochastic gene expression with a multistate promoter: Breaking down exact distributions*, SIAM Journal on Applied Mathematics, 79 (2019), pp. 1007–1029.

[18] U. Herbach, A. Bonnaffoux, T. Espinasse, and O. Gandrillon, *Inferring gene regulatory networks from single-cell data: a mechanistic approach*, BMC systems biology, 11 (2017), pp. 1–15.

[19] G. d. C. P. Innocentini, M. Forger, A. F. Ramos, O. Radulescu, and J. E. M. Hornos, *Multimodality and flexibility of stochastic gene expression*, Bulletin of mathematical biology, 75 (2013), pp. 2600–2630.

[20] J. K. Kim and J. C. Marioni, *Inferring the kinetics of stochastic gene expression from single-cell rna-sequencing data*, Genome biology, 14 (2013), pp. 1–12.

[21] I. Kontoyiannis and S. P. Meyn, *On the f-norm ergodicity of markov processes in continuous time*, Electronic Communications in Probability, 21 (2016), pp. 1–10.

[22] Y. T. Lin and N. E. Buchler, *Exact and efficient hybrid monte carlo algorithm for accelerated bayesian inference of gene expression models from snapshots of single-cell transcripts*, The Journal of chemical physics, 151 (2019), p. 024106.

[23] W. Lippitt and S. Sethuraman, *On the use of markovian stick-breaking priors*, arXiv preprint arXiv:2108.10849, (2021).

[24] P. Müller, F. A. Quintana, A. Jara, and T. Hanson, *Bayesian nonparametric data analysis*, vol. 1, Springer, 2015.

[25] J. Peccoud and B. Ycart, *Markovian modeling of gene-product synthesis*, Theoretical population biology, 48 (1995), pp. 222–234.

[26] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo statistical methods*, vol. 2, Springer, 1999.

[27] G. Schwarz, *Estimating the dimension of a model*, The annals of statistics, (1978), pp. 461–464.

[28] V. Shahrezaei and P. S. Swain, *Analytical distributions for stochastic gene expression*, Proceedings of the National Academy of Sciences, 105 (2008), pp. 17256–17261.

[29] T. Zhou and T. Liu, *Quantitative analysis of gene expression systems*, Quantitative Biology, 3 (2015), pp. 168–181.

[30] T. Zhou and J. Zhang, *Analytical results for a multistate gene model*, SIAM Journal on Applied Mathematics, 72 (2012), pp. 789–818.