# Supplementary Materials for
# Bayesian Semiparametric Mixed Effects Markov Models with Application to Vocalization Syntax

Abhra Sarkar
Department of Statistics and Data Sciences, University of Texas at Austin,
2317 Speedway D9800, Austin, TX 78712-1823, USA
abhra.sarkar@utexas.edu

Jonathan Chabout
Department of Neurobiology, Duke University, Durham, NC 27710, USA
jchabout.pro@gmail.com

Joshua Jones Macopson
Department of Neurobiology, Duke University, Durham, NC 27710, USA
joshua.jones.macopson@duke.edu

Erich D. Jarvis
Department of Neurobiology, Duke University, Durham, NC 27710, USA
Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
The Rockefeller University, New York, NY 10065, USA
jarvis@neuro.duke.edu

David B. Dunson
Department of Statistical Science, Duke University, Box 90251, Durham NC
27708-0251, USA
dunson@duke.edu

The Supplementary Materials present an efficient Gibbs sampler to draw samples from the posterior, detailed derivations of the auxiliary variable sampler used to update the prior parameters and hyper-parameters are also included. The Supplementary Materials also discuss some theoretical aspects of our model; present additional figures summarizing the results for the real and simulated data sets described in Sections 5 and 6 in the main paper; present MCMC diagnostics for analysis of the real data set; and discuss the contrasting features of generalized linear mixed model based approaches with our proposed approach in additional detail, highlighting the latter's many important advantages over the former.

# S.1 Posterior Inference

## S.1.1 Prior Hyper-parameters and MCMC Initializations

In all our examples, real or synthetic, we set $\alpha_{00} = 1$ and $\lambda_{00}(y_t) = \sum_{s,t} 1\{y_{s,t} = y_t\}/\sum_s T_s$, the overall proportion of syllables among all songs. We set each $\alpha_j$ at the value for which $p_0(H_{0j}) = p_0(k_j = 1) = 1/2$. For $j = 1, \ldots, p$, we initialize $\boldsymbol{\mu}_j$ at $(1/d_j, \ldots, 1/d_j)^{\mathrm{T}}$. We initialize each $z_{j,h}$ at $h$ for $h = 1, \ldots, d_j$. Each level of $x_j$ thus initially forms its own cluster. The associated $\boldsymbol{\lambda}_{h_1,\ldots,h_p}(y_t \mid y_{t-1})$ are initialized at $\sum_{s,t} 1\{y_{s,t} = y_t, y_{s,t-1} = y_{t-1}, x_{s,j} = h_j, j = 1, \ldots, p\}/\sum_{s,t} 1\{y_{s,t-1} = y_{t-1}, x_{s,j} = h_j, j = 1, \ldots, p\}$. Likewise, $\boldsymbol{\lambda}^{(i)}(y_t \mid y_{t-1})$ are initialized at $\sum_{s,t} 1\{y_{s,t} = y_t, y_{s,t-1} = y_{t-1}, i_s = i\}/\sum_{s,t} 1\{y_{s,t-1} = y_{t-1}, i_s = i\}$. For each $y_{t-1}$, $\{\pi_0(y_{t-1}), \pi_1(y_{t-1})\}$ is initialized at $(0.8, 0.2)$. The $v_{s,t}$'s are initialized by sampling from Bernoulli distribution with parameter $\pi_0(y_{s,t-1})$. The parameters $\alpha_0$ and $\alpha^{(0)}$ are both initialized at 1. For the remaining fixed hyper-parameters, we set $a = b = a_0 = b_{\alpha_0} = a^{(0)} = b^{(0)} = 1$. Extensive experiments suggested the results to be highly robust to these choices.

## S.1.2 Posterior Computation

Samples are drawn from the posterior using a Gibbs sampler that exploits the conditional independence relationships depicted in Figure 4 in the main paper. In what follows, $\boldsymbol{\zeta}$ denotes a generic variable that collects all other variables not explicitly mentioned, including the data points. The sampler comprises the following steps.

1. Sample each $z_{j,\ell}$ according to its multinomial full conditional

$$p(z_{j,\ell} = h_j \mid z_{j',\ell} = h_{j'}, j' \neq j, \boldsymbol{\zeta}) \propto \pi_{h_j}^{(j)} \times$$
$$\prod_{y_{t-1}} \prod_{(h_1,\ldots,h_p)} \frac{\beta\{\alpha_0\lambda_0(1 \mid y_{t-1}) + n_{h_1,\ldots,h_p}(1 \mid y_{t-1}), \ldots, \alpha_0\lambda_0(d_0 \mid y_{t-1}) + n_{h_1,\ldots,h_p}(d_0 \mid y_{t-1})\}}{\beta\{\alpha_0\lambda_0(1 \mid y_{t-1}), \ldots, \alpha_0\lambda_0(d_0 \mid y_{t-1})\}},$$

where $n_{h_1,\ldots,h_p}(y_t \mid y_{t-1}) = \sum_{s,t} 1\{y_{s,t} = y_t, y_{s,t-1} = y_{t-1}, v_{s,t} = 0, z_{1,x_{s,1}} = h_1, \ldots, z_{p,x_{s,p}} = h_p\}$.

2. Sample each $\boldsymbol{\mu}_j$ according to its Dirichlet full conditional

$$\{\mu_j(1), \ldots, \mu_j(d_j)\} \mid \boldsymbol{\zeta} \sim \mathrm{Dir}[\alpha_j + n_j(1), \ldots, \alpha_j + n_j(d_j)],$$

where $n_j(h) = \sum_{\ell=1}^{d_j} 1\{z_{j,\ell} = h\}$.

3. Sample each $v_{s,t}$ according to its Bernoulli full conditional

$$p(v_{s,t} = v \mid \boldsymbol{\zeta}) \propto \pi_v(y_{s,t-1}) \times \widetilde{\lambda}_v(y_{s,t} \mid y_{s,t-1}),$$

where $\widetilde{\boldsymbol{\lambda}}_0(\cdot \mid y_{t-1}) = \boldsymbol{\lambda}_{h_1,\dots,h_p}(\cdot \mid y_{t-1})$ with $(z_{1,x_{s,1}},\dots,z_{p,x_{s,p}}) = (h_1,\dots,h_p)$, and $\widetilde{\boldsymbol{\lambda}}_1(\cdot \mid y_{t-1}) = \boldsymbol{\lambda}^{(i)}(\cdot \mid y_{t-1})$.

4. Sample $\boldsymbol{\pi} = \{\pi_0(y_{t-1}), \pi_1(y_{t-1})\}^{\mathrm{T}}$ according to its Beta full conditional

$$\{\pi_0(y_{t-1}), \pi_1(y_{t-1})\} \mid \boldsymbol{\zeta} \sim \mathrm{Beta}\{a_0 + n_0(y_{t-1}), a_1 + n_1(y_{t-1})\},$$

where $n_v(y_{t-1}) = \sum_{s,t} 1\{v_{s,t} = v, y_{s,t-1} = y_{t-1}\}$.

5. Sample each $\boldsymbol{\lambda}^{(i)}(\cdot \mid y_{t-1})$'s according to its Dirichlet full conditional

$$\begin{aligned}&\{\lambda^{(i)}(1 \mid y_{t-1}),\dots,\lambda^{(i)}(d_0 \mid y_{t-1})\} \mid \boldsymbol{\zeta} \sim \\ &\qquad \mathrm{Dir}[\alpha^{(0)}\lambda_0(1 \mid y_{t-1}) + n^{(i)}(1 \mid y_{t-1}),\dots,\alpha^{(0)}\lambda_0(d_0 \mid y_{t-1}) + n^{(i)}(d_0 \mid y_{t-1})],\end{aligned}$$

where $n^{(i)}(y_t \mid y_{t-1}) = \sum_{s,t} 1\{y_{s,t} = y_t, y_{s,t-1} = y_{t-1}, v_{s,t} = 1, i_s = i\}$.

6. Sample each $\boldsymbol{\lambda}_{h_1,\dots,h_p}(\cdot \mid y_{t-1})$ according to its Dirichlet full conditional

$$\begin{aligned}&\{\lambda_{h_1,\dots,h_p}(1 \mid y_{t-1}),\dots,\lambda_{h_1,\dots,h_p}(d_0 \mid y_{t-1})\} \mid \boldsymbol{\zeta} \sim \\ &\qquad \mathrm{Dir}[\alpha_0\lambda_0(1 \mid y_{t-1}) + n_{h_1,\dots,h_p}(1 \mid y_{t-1}),\dots,\alpha_0\lambda_0(d_0 \mid y_{t-1}) + n_{h_1,\dots,h_p}(d_0 \mid y_{t-1})].\end{aligned}$$

7. For $\ell = n_{h_1,\dots,h_p}(y_t \mid y_{t-1})$, sample auxiliary variables $v_{\ell,h_1,\dots,h_p}(y_t \mid y_{t-1})$

$$v_{\ell,h_1,\dots,h_p}(y_t \mid y_{t-1}) \mid \boldsymbol{\zeta} \sim \mathrm{Bernoulli}\left\{\frac{\alpha_0\lambda_0(y_t \mid y_{t-1})}{\ell - 1 + \alpha_0\lambda_0(y_t \mid y_{t-1})}\right\}.$$

Set $v_{h_1,\dots,h_p}(y_t \mid y_{t-1}) = \sum_\ell v_{\ell,h_1,\dots,h_p}(y_t \mid y_{t-1})$. Likewise, for $\ell = 1,\dots,n^{(i)}(y_t \mid y_{t-1})$, sample auxiliary variables $v_\ell^{(i)}(y_t \mid y_{t-1})$ as

$$v_\ell^{(i)}(y_t \mid y_{t-1}) \mid \boldsymbol{\zeta} \sim \mathrm{Bernoulli}\left\{\frac{\alpha^{(0)}\lambda_0(y_t \mid y_{t-1})}{\ell - 1 + \alpha^{(0)}\lambda_0(y_t \mid y_{t-1})}\right\}.$$

Set $v^{(i)}(y_t \mid y_{t-1}) = \sum_\ell v_\ell^{(i)}(y_t \mid y_{t-1})$. Additionally, sample auxiliary variables

$$\begin{aligned}r_{h_1,\dots,h_p}(y_{t-1}) \mid \boldsymbol{\zeta} &\sim \mathrm{Beta}\{\alpha_0 + 1, n_{h_1,\dots,h_p}(y_{t-1})\},\\ s_{h_1,\dots,h_p}(y_{t-1}) \mid \boldsymbol{\zeta} &\sim \mathrm{Bernoulli}\left\{\frac{n_{h_1,\dots,h_p}(y_{t-1})}{n_{h_1,\dots,h_p}(y_{t-1}) + \alpha_0}\right\},\\ r^{(i)}(y_{t-1}) \mid \boldsymbol{\zeta} &\sim \mathrm{Beta}\{\alpha^{(0)} + 1, n^{(i)}(y_{t-1})\},\\ s^{(i)}(y_{t-1}) \mid \boldsymbol{\zeta} &\sim \mathrm{Bernoulli}\left\{\frac{n^{(i)}(y_{t-1})}{n^{(i)}(y_{t-1}) + \alpha^{(0)}}\right\},\end{aligned}$$

where $n_{h_1,\ldots,h_p}(y_{t-1}) = \sum_{y_t} n_{h_1,\ldots,h_p}(y_t \mid y_{t-1})$ and $n^{(i)}(y_{t-1}) = \sum_{y_t} n^{(i)}(y_t \mid y_{t-1})$. Set $v(y_t \mid y_{t-1}) = \sum_{h_1,\ldots,h_p} v_{h_1,\ldots,h_p}(y_t \mid y_{t-1}) + \sum_i v^{(i)}(y_t \mid y_{t-1})$. Also, set $v_0 = \sum_{y_t}\sum_{y_{t-1}}\sum_{h_1,\ldots,h_p} v_{h_1,\ldots,h_p}(y_t \mid y_{t-1})$, $v^{(0)} = \sum_{y_t}\sum_{y_{t-1}}\sum_i v^{(i)}(y_t \mid y_{t-1})$, $\log r_0 = \sum_{y_{t-1}}\sum_{h_1,\ldots,h_p} \log r_{h_1,\ldots,h_p}(y_{t-1})$, $s_0 = \sum_{y_{t-1}}\sum_{h_1,\ldots,h_p} s_{h_1,\ldots,h_p}(y_{t-1})$, $\log r^{(0)} = \sum_{y_{t-1}}\sum_i \log r^{(i)}(y_{t-1})$, and $s^{(0)} = \sum_{y_{t-1}}\sum_i s^{(i)}(y_{t-1})$.

8. Sample $\alpha_0$ and $\alpha^{(0)}$ according to their Gamma full conditionals

$$\begin{aligned}\alpha_0 \mid \boldsymbol{\zeta} &\sim \text{Ga}(a_0 + v_0 - s_0, b_{\alpha_0} - \log r_0),\\ \alpha^{(0)} \mid \boldsymbol{\zeta} &\sim \text{Ga}(a^{(0)} + v^{(0)} - s^{(0)}, b^{(0)} - \log r^{(0)}).\end{aligned}$$

9. Finally, sample $\boldsymbol{\lambda}_0$ according to its Dirichlet full conditional

$$\begin{aligned}&\{\lambda_0(1 \mid y_{t-1}), \ldots, \lambda_0(d_0 \mid y_{t-1})\} \mid \boldsymbol{\zeta} \sim\\ &\qquad \text{Dir}[\alpha_{00}\lambda_{00}(1) + v(1 \mid y_{t-1}), \ldots, \alpha_{00}\lambda_{00}(d_0) + v(d_0 \mid y_{t-1})].\end{aligned}$$

The steps to update the hyper-parameters $\alpha_0$, $\alpha^{(0)}$ and the global transition distributions $\boldsymbol{\lambda}_0$ were adapted from the auxiliary variable sampler of West (1992) and Teh *et al.* (2006). Details are deferred to Section S.1.3 of the Supplementary Materials.

In all our examples, 5,000 MCMC iterations with the initial 2,000 discarded as burn-in and the remaining samples thinned by an interval of 5 produced very stable estimates of the individual and population level parameters of interest. MCMC diagnostic checks were not indicative of any convergence or mixing issues. See Section S.4 in the Supplementary Materials. Our implementation is fully automated, taking in only a single matrix argument - concatenated sequences $y_{s,t}$ with the associated values of the exogenous predictors $x_{s,j}$ and the subject labels repeated $T_s$ times for each sequence $s$ and included as additional columns. For the Foxp2 data set, this required feeding a $148778 \times 4$ dimensional data matrix to the codes and 5,000 MCMC iterations required approximately two hours to run on an ordinary laptop. These codes are available as part of the Supplementary Materials.

### S.1.3 Sampling Prior Parameters and Hyper-parameters

In this section, we detail the sampling of $\boldsymbol{\lambda}_0, \alpha_0, \alpha^{(0)}$ from the posterior using auxiliary variables. We utilize culinary analogies that parallel the Chinese restaurant franchise (CRF) discussed in Teh *et al.* (2006).

In what follows, the notation $\boldsymbol{\lambda} \sim \text{GEM}(\alpha)$ signifies that $\boldsymbol{\lambda}$ is a random probability distribution admitting a stick-breaking construction (Sethuraman, 1994) as

$$\lambda(k) = \pi(k)\prod_{\ell=1}^{k-1}\{1-\pi(\ell)\}, \quad \pi(\ell) \sim \text{Beta}(1,\alpha), \quad k,\ell = 1,\ldots,\infty.$$

Following convention, the law $x \sim G = \sum_{k=1}^{\infty} \lambda(k)\delta_k$ may sometimes be denoted simply as $x \sim \boldsymbol{\lambda}$. Also, when the support $\{1,\dots,d\}$ is easily understood, the law $\mathbf{x} = \{x_1,\dots,x_d\}^{\mathrm{T}} \sim \mathrm{Dir}[\alpha\lambda(1),\dots,\alpha\lambda(d)]$ may be denoted simply as $\mathbf{x} \sim \mathrm{Dir}(\alpha\boldsymbol{\lambda})$.

#### S.1.3.1 A Modified CRF

Let there be $J$ groups, each with $N_j$ observations $\{y_{j,\ell}\}_{\ell=1}^{N_j}$ with $y_{j,\ell} \in \{1,\dots,d_0\}$. Consider a generative model for $y_{j,\ell}$ as

$$\begin{aligned} \boldsymbol{\lambda}_{0j} \mid \alpha_{00}, \boldsymbol{\lambda}_{00} &\sim \mathrm{Dir}(\alpha_{00}\boldsymbol{\lambda}_{00}), \\ \boldsymbol{\lambda}_{j} \mid \alpha_{0}, \boldsymbol{\lambda}_{0j} &\sim \mathrm{Dir}(\alpha_{0}\boldsymbol{\lambda}_{0j}), \\ y_{j,\ell} \mid \boldsymbol{\lambda}_j &\sim \mathrm{Mult}[\lambda_j(1),\dots,\lambda_j(d_0)]. \end{aligned}$$

The model may be reformulated as

$$\begin{aligned} &\boldsymbol{\lambda}_{0j} \mid \alpha_{00}, \boldsymbol{\lambda}_{00} \sim \mathrm{Dir}(\alpha_{00}\boldsymbol{\lambda}_{00}), \\ &G_j = \textstyle\sum_{k=1}^{d_0} \lambda_j(k)\delta_k, \qquad \boldsymbol{\lambda}_{j} \mid \alpha_{0}, \boldsymbol{\lambda}_{0j} \sim \mathrm{Dir}(\alpha_{0}\boldsymbol{\lambda}_{0j}), \\ &y_{j,\ell} \mid G_j \sim G_j. \end{aligned}$$

Another representation is given by

$$\begin{aligned} &G_{0j} = \textstyle\sum_{k=1}^{d_0} \lambda_{0j}(k)\delta_k, \qquad \boldsymbol{\lambda}_{0j} \mid \alpha_{00}, \boldsymbol{\lambda}_{00} \sim \mathrm{Dir}(\alpha_{00}\boldsymbol{\lambda}_{00}), \\ &G_j = \textstyle\sum_{\tau=1}^{\infty} \widetilde{\lambda}_j(\tau)\delta_{\psi_{j,\tau}}, \qquad \widetilde{\boldsymbol{\lambda}}_j \sim \mathrm{GEM}(\alpha_0), \qquad \psi_{j,\tau} \sim G_{0j}, \\ &y_{j,\ell} \mid G_j \sim G_j. \end{aligned}$$

A modified CRF arising from this generative model is as follows. Corresponding to the $J$ groups, we imagine $J$ restaurants, each with infinitely many tables but finitely many dishes $\mathcal{Y} = \{1,\dots,d_0\}$ on their globally shared menu. A customer belonging to the $j^{th}$ group enters restaurant $j$, sits at a table, and is served a dish. While the restaurant assignments are predetermined by group memberships, the table assignment for the $\ell^{th}$ customer in restaurant $j$ is chosen as $\tau_{j,\ell} \sim \widetilde{\boldsymbol{\lambda}}_j$, and each table $\tau$ is assigned a dish via $\psi_{j,\tau} \sim \boldsymbol{\lambda}_{0j}$. Customers sitting at the same table thus eat the same dish. Multiple tables may, however, be served the same dish, allowing two customers enjoying the same dish to be seated at different tables.

Let $n_{j,\tau}$ denote the number of customers in restaurant $j$ at table $\tau$, $n_j(\psi)$ denote the number of customers in restaurant $j$ eating the dish $\psi$, and $n_j$ denote the total number of customers in restaurant $j$. Also, let $n_{j,\tau}(\psi)$ denote the number of customers in restaurant $j$ at table $\tau$ eating dish $\psi$. Clearly, $n_{j,\tau}(\psi) > 0$ only when dish $\psi$ is served at an occupied table $\tau$. Finally, let $v_j(\psi)$ be the number of tables in restaurant $j$ serving dish $\psi$, and $v_j$ be the total number of occupied tables in restaurant $j$.

Given a posterior sample $(\boldsymbol{\tau}, \boldsymbol{\psi})$ of the table and the dish assignments from the

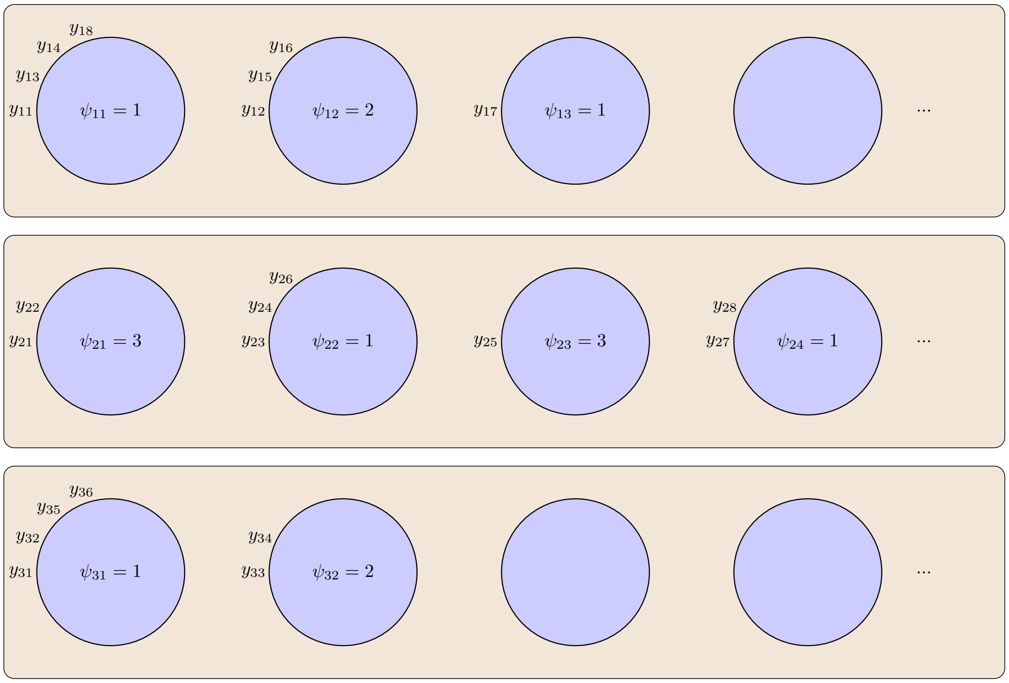


Figure S.1: The modified Chinese restaurant franchise.

modified CRF, we can obtain a draw from the posterior of $\boldsymbol{\lambda}_{0j}$ by noting that a-priori $(\boldsymbol{\lambda}_{0j} \mid \alpha_{00}, \boldsymbol{\lambda}_{00}) \sim \mathrm{Dir}(\alpha_{00}\boldsymbol{\lambda}_{00})$ and that $\psi_{j,\tau}$ for each $\tau$ is a draw from $\boldsymbol{\lambda}_{0j}$. The number of different $\psi_{j,\tau}$'s that are associated with a specific dish $\psi$ is precisely the number of tables in the restaurant $j$ that served the dish $\psi$, that is, $v_j(\psi)$. See Figure S.1. Using Dirichlet-Multinomial conjugacy, we then have

$$(\boldsymbol{\lambda}_{0j} \mid \boldsymbol{\tau}, \boldsymbol{\psi}, \alpha_{00}, \boldsymbol{\lambda}_{00}, \boldsymbol{\zeta}) \sim \mathrm{Dir}[\alpha_{00}\lambda_{00}(1) + v_j(1), \ldots, \alpha_{00}\lambda_{00}(d_0) + v_j(d_0)].$$

Given a data point $y_{j,\ell}$ and the corresponding table assignment $\tau_{j,\ell}$, the dish $\psi_{j,\tau_{j,\ell}}$ assigned to that table is known to be $y_{j,\ell}$ and hence need not be sampled.

The table assignments $\boldsymbol{\tau}$ are, however, latent. To sample $\boldsymbol{\tau}$ from the posterior, we first marginalize out their prior $\widetilde{\boldsymbol{\lambda}}_j \sim \mathrm{GEM}(\alpha_0)$ to obtain

$$(\tau_{j,\ell} \mid \alpha_0, \boldsymbol{\tau}_j^{-\ell}) \sim \textstyle\sum_{\tau \in \mathcal{S}_{j,\tau}^{-\ell}} \frac{n_{j,\tau}^{-\ell}}{n_j - 1 + \alpha_0} \delta_\tau + \frac{\alpha_0}{n_j - 1 + \alpha_0} \delta_{\tau_{new}},$$

where $n_{j,\tau}^{-\ell}$ denotes the number of customers sitting at table $\tau$ in restaurant $j$ excluding the $\ell^{th}$ customer, $\mathcal{S}_{j,\tau}^{-\ell}$ denotes the set of unique values in $\boldsymbol{\tau}_j^{-\ell} = \{\tau_{j,m} : m = 1, \ldots, n_j, m \neq \ell\}$ and $\tau_{new}$ is a generic for any new value of $\tau$ not in $\mathcal{S}_{j,\tau}^{-\ell}$. The distribution of the table assignment $\tau_{j,\ell}$ given $\boldsymbol{\tau}_j^{-\ell}$ and the dish assignments $\boldsymbol{\psi}$ may then be obtained as

$$p(\tau_{j,\ell}=\tau \mid \psi_{j,\tau}=\psi,\alpha_0,\boldsymbol{\psi}_j^{-\ell},\boldsymbol{\tau}_j^{-\ell},\boldsymbol{\lambda}_{0j}) \propto n_{j,\tau}^{-\ell}\delta_{\tau} \quad \text{if } \tau\in\mathcal{S}_{j,\tau}^{-\ell},$$
$$p(\tau_{j,\ell}=\tau_{new} \mid \psi_{j,\tau_{new}}=\psi,\alpha_0,\boldsymbol{\psi}_j^{-\ell},\boldsymbol{\tau}_j^{-\ell},\boldsymbol{\lambda}_{0j}) \propto \alpha_0\lambda_{0j}(\psi) \quad \text{if } \tau_{new}\notin\mathcal{S}_{j,\tau}^{-\ell},$$

where $\boldsymbol{\psi}_j^{-\ell}=\{\psi_{j,\tau_{j,m}}: m=1,\dots,n_j, m\neq\ell\}$. Since these assignments are restricted only to tables serving the dish $\psi$, the distribution reduces to

$$(\tau_{j,\ell}\mid \psi_{j,\tau_{j,\ell}}=\psi,\alpha_0,\boldsymbol{\psi}_j^{-\ell},\boldsymbol{\tau}_j^{-\ell},\boldsymbol{\lambda}_{0j}) \sim \textstyle\sum_{\tau\in\mathcal{S}_{j,\tau}^{-\ell}(\psi)} \frac{n_{j,\tau}^{-\ell}(\psi)}{n_j(\psi)-1+\alpha_0\lambda_{0j}(\psi)}\delta_{\tau} + \frac{\alpha_0\lambda_{0j}(\psi)}{n_j(\psi)-1+\alpha_0\lambda_{0j}(\psi)}\delta_{\tau_{new}},$$

where $\mathcal{S}_{j,\tau}^{-\ell}(\psi)$ denotes the set of unique values in $\boldsymbol{\tau}_j^{-\ell}(\psi)=\{\tau_{j,m}: m=1,\dots,n_j, m\neq\ell, \psi_{j,\tau_{j,m}}=\psi\}$, $n_{j,\tau}^{-\ell}(\psi)$ denotes the number of customers sitting at table $\tau$ in restaurant $j$ and enjoying the dish $\psi$ excluding the $\ell^{th}$ customer, and $\tau_{new}$ is a generic for any new value of $\tau$ not in $\mathcal{S}_{j,\tau}^{-\ell}(\psi)$. This distribution can be identified with a marginalized conditional distribution of assignments of $n_j(\psi)$ observations to different components in a GEM$\{\alpha\lambda_{0j}(\psi)\}$. The full conditional for $\boldsymbol{\lambda}_{0j}$ given $(\boldsymbol{\psi},\boldsymbol{\tau})$ depends on the table assignments only via $v_j(\psi)$ which can be obtained from the table assignments $\boldsymbol{\tau}_j$.

Alternatively, for each of the $n_j(\psi)$ customers in restaurant $j$ enjoying the dish $\psi$, let $v_{j,\ell}(\psi)=0$ if the $\ell^{th}$ customer sits at an already occupied table, and $v_{j,\ell}(\psi)=1$ if the $\ell^{th}$ customer goes to a new table. Then, $v_j(\psi)=\sum_{\ell=1}^{n_j(\psi)} v_{j,\ell}(\psi)$. Using properties of a GEM$\{\alpha\lambda_{0j}(\psi)\}$ distribution, we then have

$$\{v_{j,\ell}(\psi)\mid \mathbf{v}_j^{\ell-1}(\psi),\alpha_0,\boldsymbol{\lambda}_{0j}\} \sim \tfrac{\ell-1}{\ell-1+\alpha_0\lambda_{0j}(\psi)}\delta_0 + \tfrac{\alpha_0\lambda_{0j}(\psi)}{\ell-1+\alpha_0\lambda_{0j}(\psi)}\delta_1,$$

where $\mathbf{v}_j^{\ell-1}(\psi)=\{v_{j,m}(\psi): m=1,\dots,\ell-1\}$. We can then sample the $v_{j,\ell}(\psi)$'s from the posterior by sequentially sampling them as

$$[\{v_{j,\ell}(\psi)\}_{\ell=1}^{n_j(\psi)}\mid \alpha_0,\boldsymbol{\lambda}_{0j}] \sim \textstyle\prod_{\ell=1}^{n_j(\psi)} \text{Bernoulli}\left\{\frac{\alpha_0\lambda_{0j}(\psi)}{\ell-1+\alpha_0\lambda_{0j}(\psi)}\right\}.$$

Next, we derive the full conditional for the hyper-parameter $\alpha_0$, assuming a Ga$(a,b)$ prior and adapting to West (1992). Following Antoniak (1974), integrating out $\boldsymbol{\lambda}_{0j}$, we have $p(v_j\mid\alpha_0,n_j)=\alpha_0^{v_j}s^{\star}(n_j,v_j)\Gamma(\alpha_0)/\Gamma(\alpha_0+n_j)$, where $s^{\star}(n,v)$ are Stirling numbers of the first kind. Letting $\mathbf{n}=\{n_j\}_{j=1}^J$, $\mathbf{v}=\{v_j\}_{j=1}^J$ with $v=\sum_{j=1}^J v_j$, and the restaurants being conditionally independent, we then have

$$p(\alpha_0\mid\mathbf{v},\mathbf{n},\boldsymbol{\zeta}) \propto p_0(\alpha_0\mid a,b)\ p(\mathbf{v}\mid\alpha_0,\mathbf{n}) \propto \exp(-\alpha_0 b)(\alpha_0)^{a-1}\ \textstyle\prod_{j=1}^J\left\{(\alpha_0)^{v_j}\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)}\right\}$$
$$\propto \exp(-\alpha_0 b)(\alpha_0)^{a+v-1}\textstyle\prod_{j=1}^J\left\{\frac{(\alpha_0+n_j)\ \text{Beta}(\alpha_0+1,n_j)}{\alpha_0\ \Gamma(n_j)}\right\}$$
$$\propto \exp(-\alpha_0 b)(\alpha_0)^{a+v-1}\textstyle\prod_{j=1}^J\left\{\left(1+\frac{n_j}{\alpha_0}\right)\int r_j^{\alpha_0}(1-r_j)^{n_j-1}dr_j\right\}$$

$$\propto \exp(-\alpha_0 b)(\alpha_0)^{a+v-1} \textstyle\prod_{j=1}^{J} \left\{ \sum_{s_j=0}^{1} \left(\frac{n_j}{\alpha_0}\right)^{s_j} \int r_j^{\alpha_0}(1-r_j)^{n_j-1} dr_j \right\}.$$

Treating $\mathbf{r} = \{r_j\}_{j=1}^{J}$, $\mathbf{s} = \{s_j\}_{j=1}^{J}$ as auxiliary variables, we have

$$p(\alpha_0, \mathbf{r}, \mathbf{s} \mid \boldsymbol{\zeta}) \propto \exp(-\alpha_0 b)(\alpha_0)^{a+v-1} \textstyle\prod_j \left\{ \left(\frac{n_j}{\alpha_0}\right)^{s_j} r_j^{\alpha_0}(1-r_j)^{n_j-1} \right\}.$$

The full conditionals for $\alpha_0$, $r_j$ and $s_j$ are then obtained in closed forms as

$$(\alpha_0 \mid \boldsymbol{\zeta}) \sim \mathrm{Ga}(a+v-s, b-\log r), \quad (r_j \mid \boldsymbol{\zeta}) \sim \mathrm{Beta}(\alpha_0+1, n_j), \quad (s_j \mid \boldsymbol{\zeta}) \sim \mathrm{Bernoulli}\left(\tfrac{n_j}{n_j+\alpha_0}\right),$$

where $\log r = \sum_{j=1}^{J} \log r_j$, and $s = \sum_{j=1}^{J} s_j$.

#### S.1.3.2 Mixed Effects CRF

In our mixed effects Markov chain setting, when $v_{s,t} = 1, i_s = i, y_{s,t-1} = y_{t-1}$, the customer $(s,t)$ enters the $j \equiv (i, y_{t-1})^{th}$ restaurant, whereas when $v_{s,t} = 0, (z_{1,x_{s,1}}, \ldots, z_{p,x_{s,p}}) = (h_1, \ldots, h_p), y_{s,t-1} = y_{t-1}$, the customer enters the $j \equiv (h_1, \ldots, h_p, y_{t-1})^{th}$ restaurant.

We first focus on the case $v_{s,t} = 1$, leading the customer $(s,t)$ to the $(i, y_{t-1})^{th}$ restaurant. The total number of customers entering the $(i, y_{t-1})^{th}$ restaurant is $n^{(i)}(y_{t-1}) = \sum_{s,t} 1\{y_{s,t-1} = y_{t-1}, v_{s,t} = 1, i_s = i\}$. Among them, the number of customers eating the dish $y_t$ is $n^{(i)}(y_t \mid y_{t-1}) = \sum_{s,t} 1\{y_{s,t} = y_t, y_{s,t-1} = y_{t-1}, v_{s,t} = 1, i_s = i\}$. We define, for each $\ell = 1, \ldots, n^{(i)}(y_t \mid y_{t-1})$, $v_\ell^{(i)}(y_t \mid y_{t-1}) = 0$ if the $\ell^{th}$ customer sits at an already occupied table and $v_\ell^{(i)}(y_t \mid y_{t-1}) = 1$ if the $\ell^{th}$ customer goes to a new table. We can then sample $\{v_\ell^{(i)}(y_t \mid y_{t-1})\}_{\ell=1}^{n^{(i)}(y_t \mid y_{t-1})}$ from the posterior by sampling them sequentially from

$$\{v_\ell^{(i)}(y_t \mid y_{t-1})\}_{\ell=1}^{n^{(i)}(y_t|y_{t-1})} \mid \boldsymbol{\zeta} \sim \textstyle\prod_{\ell=1}^{n^{(i)}(y_t|y_{t-1})} \mathrm{Bernoulli}\left\{\frac{\alpha^{(0)}\lambda_0(y_t|y_{t-1})}{\ell-1+\alpha^{(0)}\lambda_0(y_t|y_{t-1})}\right\}.$$

Then, $v^{(i)}(y_t \mid y_{t-1}) = \sum_{\ell=1}^{n^{(i)}(y_t|y_{t-1})} v_\ell^{(i)}(y_t \mid y_{t-1})$ gives the number of occupied tables serving the dish $y_t$ in the $(i, y_{t-1})^{th}$ restaurant.

The case $v_{s,t} = 0$, leading customer $(s,t)$ to the $j \equiv (h_1, \ldots, h_p, y_{t-1})^{th}$ restaurant can be similarly handled. For instance, if, for each customer $\ell = 1, \ldots, n_{h_1,\ldots,h_p}(y_t \mid y_{t-1})$ eating dish $y_t$ in restaurant $(h_1, \ldots, h_p, y_{t-1})$, we define $v_{\ell,h_1,\ldots,h_p}(y_t \mid y_{t-1}) = 0$ if the customer sits at an already occupied table and $v_{\ell,h_1,\ldots,h_p}(y_t \mid y_{t-1}) = 1$ if the customer goes to a new table. Then, we can sample $\{v_{\ell,h_1,\ldots,h_p}(y_t \mid y_{t-1})\}_{\ell=1}^{n_{h_1,\ldots,h_p}(y_t|y_{t-1})}$

from the posterior by sampling them sequentially from

$$\{v_{\ell,h_1,\dots,h_p}(y_t \mid y_{t-1})\}_{\ell=1}^{n_{h_1,\dots,h_p}(y_t|y_{t-1})} \mid \boldsymbol{\zeta} \sim \textstyle\prod_{\ell=1}^{n_{h_1,\dots,h_p}(y_t|y_{t-1})} \text{Bernoulli}\left\{\frac{\alpha_0\lambda_0(y_t|y_{t-1})}{\ell-1+\alpha_0\lambda_0(y_t|y_{t-1})}\right\}.$$

Then, $v_{h_1,\dots,h_p}(y_t \mid y_{t-1}) = \sum_{\ell=1}^{n_{h_1,\dots,h_p}(y_t|y_{t-1})} v_{\ell,h_1,\dots,h_p}(y_t \mid y_{t-1})$ gives the number of occupied tables serving the dish $y_t$ in the $(h_1,\dots,h_p,y_{t-1})^{th}$ restaurant.

The table assignments in restaurants $(i, y_{t-1})$ and $(h_1,\dots,h_p,y_{t-1})$ with a common subscript $y_{t-1}$ follow $\boldsymbol{\lambda}_0(\cdot \mid y_{t-1})$. Letting $v(y_t \mid y_{t-1}) = \sum_{h_1,\dots,h_p} v_{h_1,\dots,h_p}(y_t \mid y_{t-1}) + \sum_i v^{(i)}(y_t \mid y_{t-1})$ denote the total number of tables serving dish $y_t$ across all such restaurants, we can update $\boldsymbol{\lambda}_0(\cdot \mid y_{t-1})$ using Dirichlet-Multinomial conjugacy as

$$\boldsymbol{\lambda}_0(\cdot \mid y_{t-1}) \mid \boldsymbol{\zeta} \sim \text{Dir}\left\{\alpha_{00}\lambda_{00}(1) + v(1 \mid y_{t-1}),\dots,\alpha_{00}\lambda_{00}(d_0) + v(d_0 \mid y_{t-1})\right\}.$$

To sample the hyper-parameter $\alpha^{(0)}$, we mimic the developments in the modified CRF and introduce auxiliary variables $r^{(i)}(y_{t-1})$ and $s^{(i)}(y_{t-1})$ for each $(i, y_{t-1})$. Let $\mathbf{n}^{(0)} = \{n^{(i)}(y_{t-1}) : i \in \mathcal{I}_0, y_{t-1} \in \mathcal{Y}\}$; $\mathbf{v}^{(0)}, \mathbf{r}^{(0)}, \mathbf{s}^{(0)}$ are similarly defined, $\mathcal{I}_0$ is the set of individuals associated with the sequences. It can then be easily derived that

$$\begin{aligned}
\alpha^{(0)} \mid \boldsymbol{\zeta} &\sim \text{Ga}(a^{(0)} + v^{(0)} - s^{(0)}, b^{(0)} - \log r^{(0)}),\\
r^{(i)}(y_{t-1}) \mid \boldsymbol{\zeta} &\sim \text{Beta}\{\alpha^{(0)} + 1, n^{(i)}(y_{t-1})\},\\
s^{(i)}(y_{t-1}) \mid \boldsymbol{\zeta} &\sim \text{Bernoulli}\left\{\frac{n^{(i)}(y_{t-1})}{n^{(i)}(y_{t-1}) + \alpha^{(0)}}\right\},
\end{aligned}$$

where $v^{(0)} = \sum_{y_t}\sum_{y_{t-1}}\sum_i v^{(i)}(y_t \mid y_{t-1})$, $\log r^{(0)} = \sum_{y_{t-1}}\sum_i \log r^{(i)}(y_{t-1})$, and $s^{(0)} = \sum_{y_{t-1}}\sum_i s^{(i)}(y_{t-1})$.

Likewise, the hyper-parameter $\alpha_0$ and the associated auxiliary variables $r_{h_1,\dots,h_p}(y_{t-1})$ and $s_{h_1,\dots,h_p}(y_{t-1})$ can be sampled from the posterior as

$$\begin{aligned}
\alpha_0 \mid \boldsymbol{\zeta} &\sim \text{Ga}(a_0 + v_0 - s_0, b_{\alpha_0} - \log r_0),\\
r_{h_1,\dots,h_p}(y_{t-1}) \mid \boldsymbol{\zeta} &\sim \text{Beta}\{\alpha_0 + 1, n_{h_1,\dots,h_p}(y_{t-1})\},\\
s_{h_1,\dots,h_p}(y_{t-1}) \mid \boldsymbol{\zeta} &\sim \text{Bernoulli}\left\{\frac{n_{h_1,\dots,h_p}(y_{t-1})}{n_{h_1,\dots,h_p}(y_{t-1}) + \alpha_0}\right\},
\end{aligned}$$

where $v_0 = \sum_{y_t}\sum_{y_{t-1}}\sum_{h_1,\dots,h_p} v_{h_1,\dots,h_p}(y_t \mid y_{t-1})$, $\log r_0 = \sum_{y_{t-1}}\sum_{h_1,\dots,h_p} \log r_{h_1,\dots,h_p}(y_{t-1})$, and $s_0 = \sum_{y_{t-1}}\sum_{h_1,\dots,h_p} s_{h_1,\dots,h_p}(y_{t-1})$.

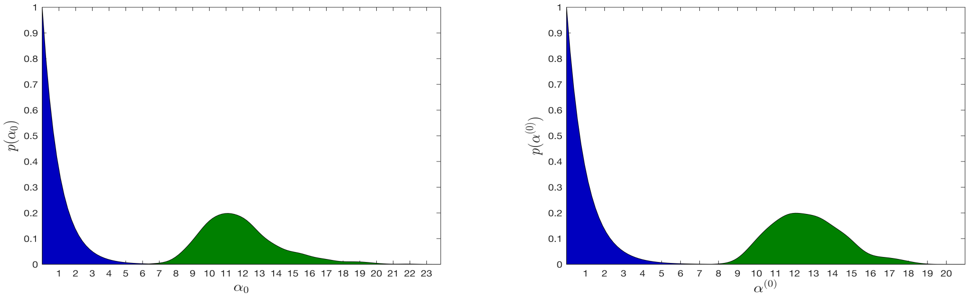


Figure S.2: Estimated posterior densities (green) of the hyper-parameters $\alpha_0$ and $\alpha^{(0)}$ superimposed over their Gamma(1,1) priors (blue).

## S.2 Theoretical Properties

In this section, we discuss some theoretical aspects of our proposed model. We follow the notations and definitions of the main paper.

For Markov sequences with exogenous predictors (MSEPs), the values of the exogenous predictors remain fixed for the entire lengths of the sequences. The notions of ergodicity, stationarity etc for predictor free ordinary Markov sequences thus extend naturally to MSEPs. Let $\mathcal{P}_0 = \{P^{(i)}_{0,x_1,\dots,x_p}(y_t \mid y_{t-1}) : P^{(i)}_{0,x_1,\dots,x_p}(y_t \mid y_{t-1}) = \pi_{0,0}(y_{t-1})\boldsymbol{\lambda}_{0,x_1,\dots,x_p}(y_t \mid y_{t-1}) + \pi_{0,1}(y_{t-1})\boldsymbol{\lambda}^{(i)}_0(y_t \mid y_{t-1})\} \subset \mathcal{P}$ denote the class of transition probability distributions that admit representations similar to our proposed formulation. It is straightforward to check that any $\mathbf{P}^{(i)}_{0,x_1,\dots,x_p}(\cdot \mid y_{t-1}) \in \mathcal{P}_0$ will be ergodic if at least one of the component transition distributions is also so and the associated mixture probability is strictly positive. In particular, if $\boldsymbol{\lambda}_{0,x_1,\dots,x_p}(\cdot \mid y_{t-1})$ and $\boldsymbol{\lambda}^{(i)}_0(\cdot \mid y_{t-1})$ are both ergodic with stationary distributions $\boldsymbol{\pi}_{0,x_1,\dots,x_p} = \{\pi_{0,x_1,\dots,x_p}(1),\dots,\pi_{0,x_1,\dots,x_p}(d_0)\}^{\mathrm{T}}$ and $\boldsymbol{\pi}^{(i)}_0 = \{\pi^{(i)}_0(1),\dots,\pi^{(i)}_0(d_0)\}^{\mathrm{T}}$, respectively, then the stationary distribution of $\mathbf{P}^{(i)}_{0,x_1,\dots,x_p}(\cdot \mid y_{t-1})$, denoted by $\boldsymbol{\pi}^{(i)}_{0,x_1,\dots,x_p} = \{\pi^{(i)}_{0,x_1,\dots,x_p}(1),\dots,\pi^{(i)}_{0,x_1,\dots,x_p}(d_0)\}^{\mathrm{T}}$, has a representation $\pi^{(i)}_{0,x_1,\dots,x_p}(y_t) = \pi_0(y_t)\pi_{0,x_1,\dots,x_p}(y_t) + \pi_1(y_t)\pi^{(i)}_0(y_t)$. Conversely, if $\pi_0(y_{t-1}) \in (0,1)$, $\mathbf{P}^{(i)}_{0,x_1,\dots,x_p}(\cdot \mid y_{t-1})$ can be ergodic even when neither of the two component distributions are so. This can be seen by constructing an example with binary state space $\{1,2\}$ where one of the component transition distributions only allows self transitions $(1 \to 1, 2 \to 2)$ and the other only transitions to the other state $(1 \to 2, 2 \to 1)$. These results all follow from basic definitions of stationarity and also extend naturally to population level transition distributions $\mathbf{P}_{0,x_1,\dots,x_p}(\cdot \mid y_{t-1})$.

We now discuss model flexibility, prior support and posterior consistency. The proposed mixed effect Markov model assumes additivity of predictor effects and individual effects directly on the probability scale. The model assumes an implicit upper bound $\pi_1(y_{t-1})$ on how far the individual effects $\pi_1(y_{t-1})\boldsymbol{\lambda}^{(i)}(y_t \mid y_{t-1})$ can stretch the effects due to the exogenous predictors $\pi_0(y_{t-1})\boldsymbol{\lambda}_{h_1,\dots,h_p}(y_t \mid y_{t-1})$ in modeling $\mathbf{P}^{(i)}_{h_1,\dots,h_p}(y_t \mid y_{t-1})$. This bound can be easily relaxed by allowing the $\pi_1(y_{t-1})$'s to also be individual-specific, assigning additional priors to the parameters of their distribution. We have not pursued such generalizations in this article in favor of simplicity and parsimony. Being based on the partition model for MSEPs introduced in Section 3.1 in the main paper, the model for the population level mean transition probabilities $\mathbf{P}_{h_1,\ldots,h_p}(\cdot \mid y_{t-1})$, on the other hand, is fully nonparametric, taking into account all order interactions between the exogenous and the local predictors. The class $\mathcal{P}_0$, defined above, thus denotes a fairly large class of individual-specific exogenous predictor dependent transition distributions.

It is easy to check that our assumed priors, referred to collectively as $\Pi$, assign positive probability on any arbitrarily close $L_1$ neighborhood of any $\mathbf{P}_0 = \{P^{(i_s)}_{0,x_{s,1},\ldots,x_{s,p}}(y_t \mid y_{t-1})\}_{s=1}^{s_0} \in \mathcal{P}_0$. More formally, with $d(P^{(i)}_{x_1,\ldots,x_p}, P^{(i)}_{0,x_1,\ldots,x_p}) = \sum_{y_{t-1}=1}^{d_0} \sum_{y_t=1}^{d_0} \left| P^{(i)}_{x_1,\ldots,x_p}(y_t \mid y_{t-1}) - P^{(i)}_{0,x_1,\ldots,x_p}(y_t \mid y_{t-1}) \right|$, we have $\Pi\{\mathcal{B}_\delta(\mathbf{P}_0)\} > 0$ for any $\mathbf{P}_0 \in \mathcal{P}_0$ and any $\delta > 0$, where $\mathcal{B}_\delta(\mathbf{P}_0) = \{P^{(i_s)}_{x_{s,1},\ldots,x_{s,p}} : d(P^{(i_s)}_{x_{s,1},\ldots,x_{s,p}}, P^{(i_s)}_{0,x_{s,1},\ldots,x_{s,p}}) \le \delta, s = 1, \ldots, s_0\}$.

Let $\mathcal{P}_{00} \subset \mathcal{P}_0$ be the class of ergodic transition probability distributions $\mathbf{P}_0$ with associated stationary distributions $\pi^{(i)}_{0,x_1,\ldots,x_p}(y_t)$, where $\pi^{(i)}_{0,x_1,\ldots,x_p}(y_t) > 0$ for all $y_t \in \mathcal{Y}$. Assuming $\{y_{s,t}\}_{s=1,t=1}^{s_0,T_s}$ to be ergodic with the true transition dynamics characterized by some $\mathbf{P}_0 \in \mathcal{P}_{00}$, it then follows, using strong law of large numbers for ergodic Markov sequences (Eichelsbacher and Ganesh, 2002), that the posterior $\Pi[\cdot \mid \{x_{s,j}, y_{s,t}\}_{s=1,t=1,j=1}^{s_0,T_s,p}]$ concentrates almost surely in arbitrarily small neighborhoods of the true data generating parameters $\mathbf{P}_0$ as $\min_s T_s \to \infty$ (Ghosh and Ramamoorthi, 2003). Formally, for any $\delta > 0$ and any $\mathbf{P}_0 \in \mathcal{P}_{00}$, $\Pi[\mathcal{B}_\delta(\mathbf{P}_0) \mid \{x_{s,j}, y_{s,t}\}_{s=1,t=1,j=1}^{s_0,T_s,p}] \to 1$ almost surely $\mathbf{P}_0$ as $\min_s T_s \to \infty$.

Asymptotic regimes for classical mixed effects models typically assume the number of subjects to approach infinity while the number of observations for each subject remains fixed. The criteria considered here, on the contrary, assumes the number of subjects to remain fixed but assumes the length of each sequence to approach infinity. This is a more realistic scenario for animal vocalization experiments, since it is practically impossible to study more than a small to moderate number of mice from each genotype. The recording times of the songs, however, can be easily increased.

## S.3 Additional Figures

This section presents additional figures summarizing results for the Foxp2 data set and the simulation experiments from Sections 5 and 6 in the main paper, respectively.

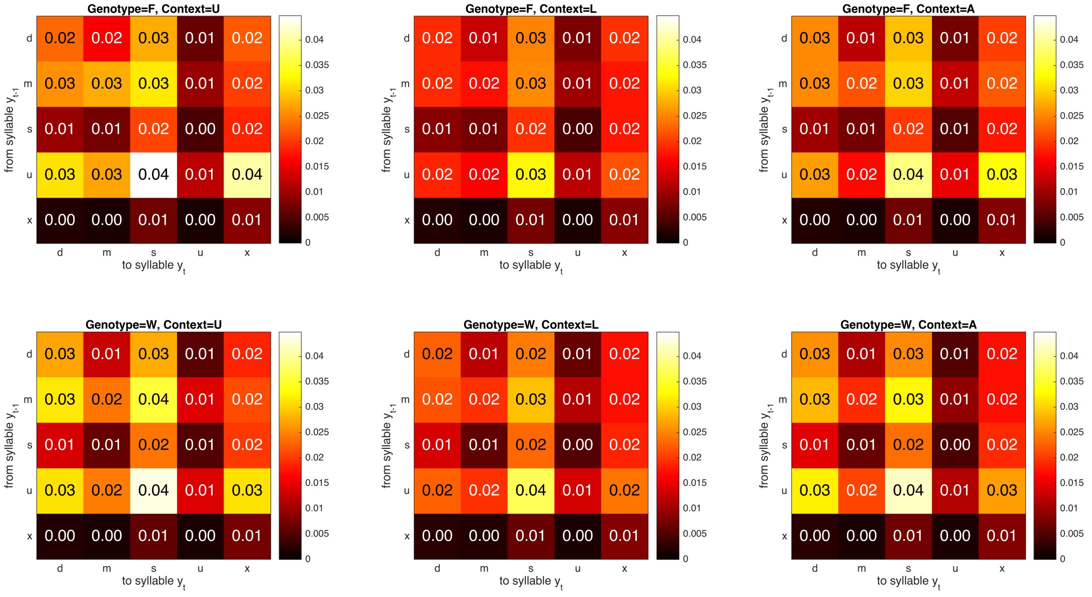


Figure S.3: Results for the Foxp2 data set. Estimated posterior standard deviation of transition probabilities $P_{x_1,x_2}(y_t \mid y_{t-1})$ for syllables $y_t, y_{t-1} \in \{d, m, s, u, x\}$ for different combinations of genotype $x_1 \in \{F, W\}$ and social contexts $x_2 \in \{U, L, A\}$.

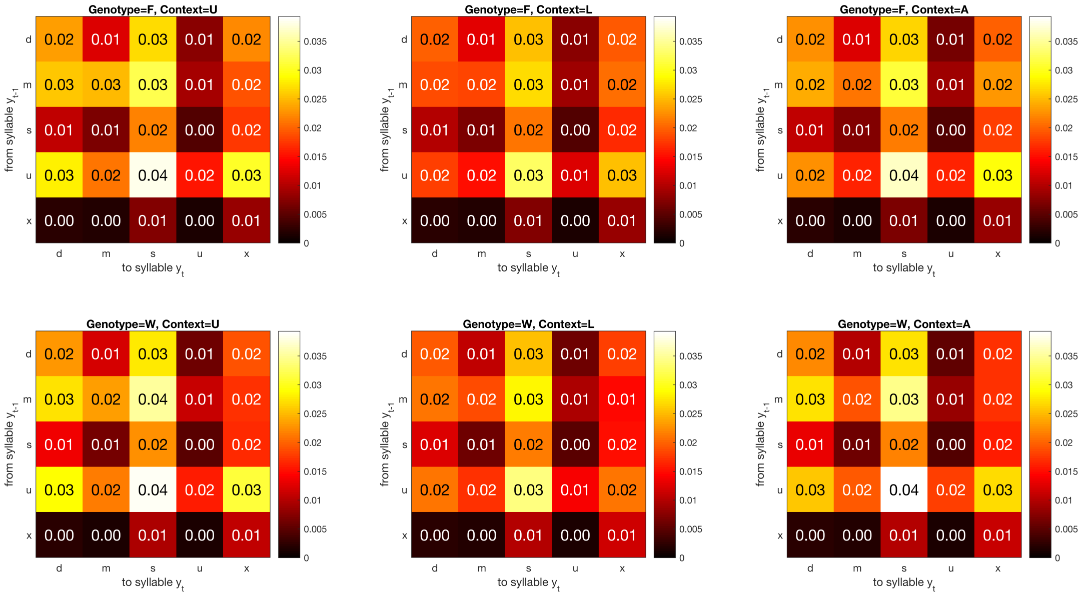
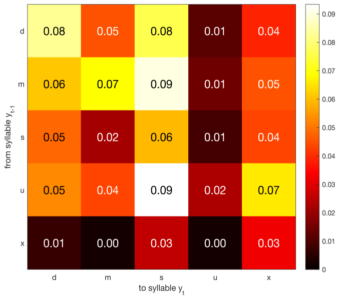
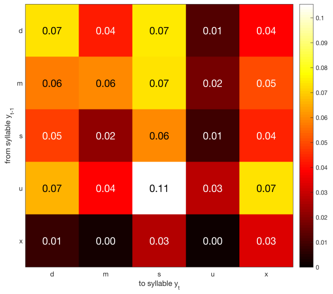


Figure S.4: Results for the simulation scenario D described in Section 6 of the main paper. Estimated posterior standard deviation of transition probabilities $P_{x_1,x_2}(y_t \mid y_{t-1})$ for syllables $y_t, y_{t-1} \in \{d, m, s, u, x\}$ for different combinations of genotype $x_1 \in \{F, W\}$ and social contexts $x_2 \in \{U, L, A\}$.



| | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.08 | 0.05 | 0.08 | 0.01 | 0.04 |
| m | 0.06 | 0.07 | 0.09 | 0.01 | 0.05 |
| s | 0.05 | 0.02 | 0.06 | 0.01 | 0.04 |
| u | 0.05 | 0.04 | 0.09 | 0.02 | 0.07 |
| x | 0.01 | 0.00 | 0.03 | 0.00 | 0.03 |

Figure S.5: Results for the Foxp2 data set. Estimated posterior standard deviation of the random effects parameters $\pi_1(y_{t-1})\lambda^{(i)}(y_t \mid y_{t-1})$ for syllables $y_t, y_{t-1} \in \{d, m, s, u, x\}$.



| | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.07 | 0.04 | 0.07 | 0.01 | 0.04 |
| m | 0.06 | 0.06 | 0.07 | 0.02 | 0.05 |
| s | 0.05 | 0.02 | 0.06 | 0.01 | 0.04 |
| u | 0.07 | 0.04 | 0.11 | 0.03 | 0.07 |
| x | 0.01 | 0.00 | 0.03 | 0.00 | 0.03 |

Figure S.6: Results for the simulation scenario D described in Section 6 of the main paper. Estimated posterior standard deviation of the random effects parameters $\pi_1(y_{t-1})\lambda^{(i)}(y_t \mid y_{t-1})$ for syllables $y_t, y_{t-1} \in \{d, m, s, u, x\}$.

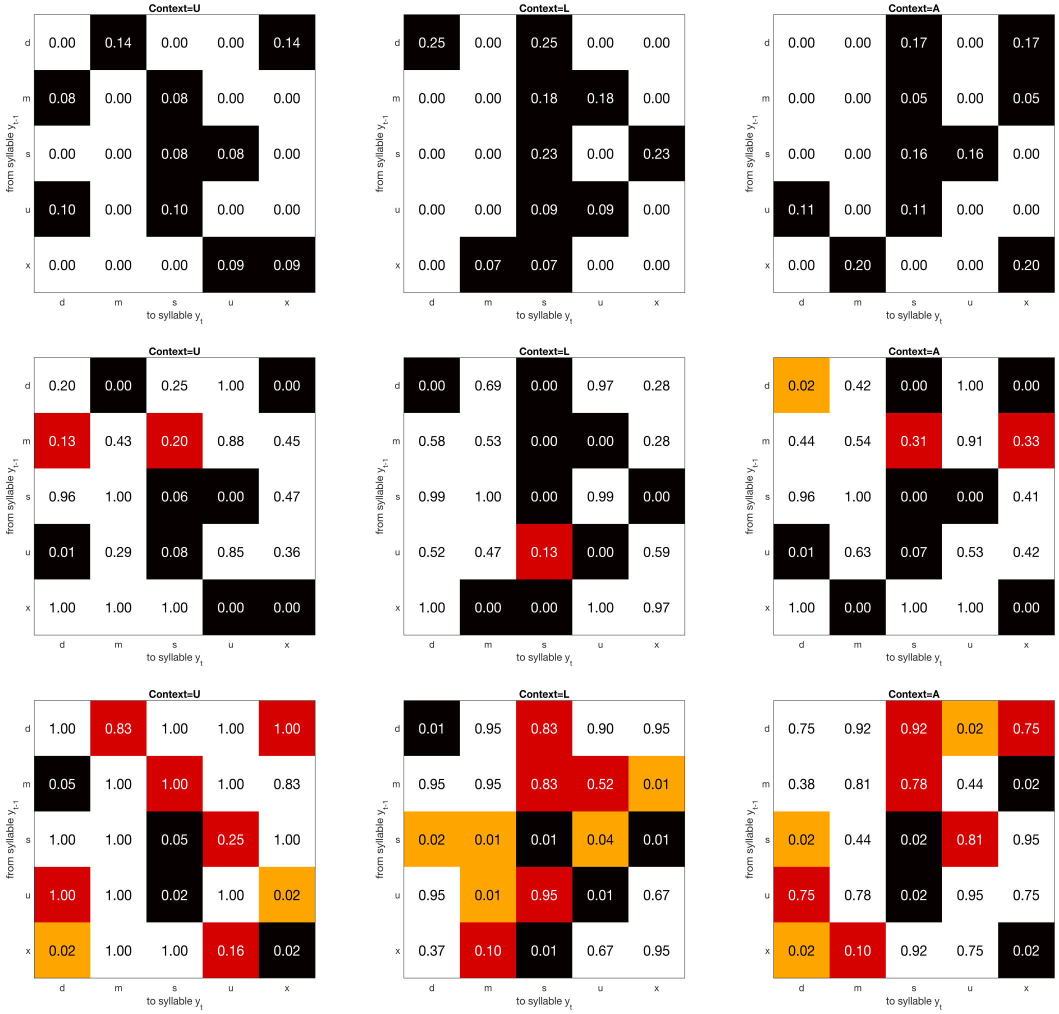


Figure S.7: Results for the simulation scenario F described in Section 6 in the main paper, but for a different random seed. The top row shows the true values of $|\Delta P_{\cdot,x_2}(y_t \mid y_{t-1})| = |P_{1,x_2}(y_t \mid y_{t-1}) - P_{2,x_2}(y_t \mid y_{t-1})|$ for syllables $y_t, y_{t-1} \in \{d, m, s, u, x\}$ and social contexts $x_2 \in \{U, L, A\}$. Positive differences are highlighted in black. The middle row shows the estimated posterior probabilities of $H_{0,y_t|y_{t-1},x_2} : |\Delta P_{\cdot,x_2}(y_t \mid y_{t-1})| \leq 0.02$. The bottom row shows Benjamini-Hochberg adjusted p-values obtained using the method of Chabout *et al.* (2016). Posterior probabilities smaller than 0.1 are considered significant and are highlighted in black and orange. Posterior probabilities greater than 0.1 are presented in white and red. Likewise, p-values smaller than 0.1 are considered significant and are highlighted in black and orange. P-values greater than 0.1 are presented in white and red. White and black cells represent correct decisions, orange cells mark rejections of true $H_{0,\ell}$ (false positives), and red cells mark failures to reject false $H_{0,\ell}$ (false negatives).

## S.4 MCMC Diagnostic Plots for Foxp2 Data Set

This section presents some MCMC diagnostics for samples drawn by the Gibbs sampler described in Section S.1.2 of the Supplementary Materials. The results presented here are for the Foxp2 data set discussed in Section 5 of the main paper. Diagnostics for the simulation experiments were similar and hence not included.

Figure S.8 shows the trace plots of the sampled values of the population level transition probabilities $P_{x_1,x_2}(y_t \mid y_{t-1})$ and is not indicative of any serious convergence or mixing issues. Figure S.9 indicates reasonably fast decays in autocorrelations up lag 20 in thinned samples of $P_{x_1,x_2}(y_t \mid y_{t-1})$.

Assuming an AR(1) model, the Monte Carlo uncertainty in estimating the transition probabilities $P_{x_1,x_2}(y_t \mid y_{t-1})$ using $N_{MC,thinned} = 600$ thinned samples drawn using the Gibbs sampler is given by

$$\widehat{\sigma}_{MC,x_1,x_2}(y_t \mid y_{t-1}) = \frac{\widehat{\sigma}_{x_1,x_2}(y_t \mid y_{t-1})}{\sqrt{N_{MC,thinned}}}\sqrt{\frac{1+\rho_1}{1-\rho_1}}.$$

Here, $\widehat{\sigma}_{x_1,x_2}(y_t \mid y_{t-1})$ is the estimated posterior standard deviation of $P_{x_1,x_2}(y_t \mid y_{t-1})$ shown in Figure S.3 in the main paper, and $\rho_1$ is the lag 1 autocorrelation in thinned samples of $P_{x_1,x_2}(y_t \mid y_{t-1})$. Figure S.10 shows the values of $10^2 \times \widehat{\sigma}_{MC,x_1,x_2}(y_t \mid y_{t-1})$ for different values of $y_t, y_{t-1}, x_1, x_2$. If we want to estimate the posterior means of $P_{x_1,x_2}(y_t \mid y_{t-1})$ with a specified tolerance level $tol$, we need $\widehat{\sigma}_{MC,x_1,x_2}(y_t \mid y_{t-1}) \leq tol$. In this article, the desired accuracy in estimating the posterior means of $P_{x_1,x_2}(y_t \mid y_{t-1})$ is assumed to be $tol = 0.005$. In Figure 5, these estimates are thus presented up to two places after the decimal point. As Figure S.10 shows, the maximum value of $\widehat{\sigma}_{MC,x_1,x_2}(y_t \mid y_{t-1})$ is 0.0025. So the tolerance conditions are all well satisfied.

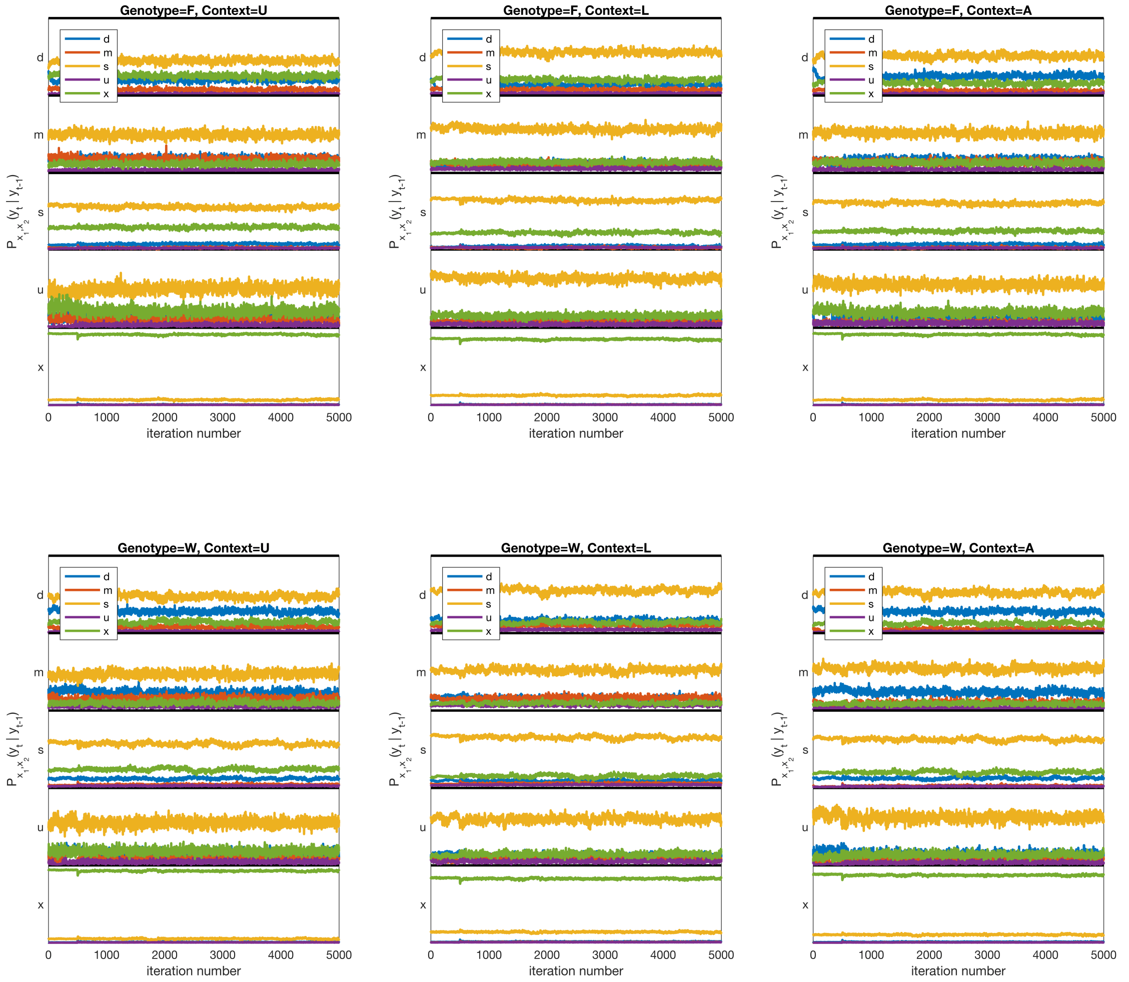
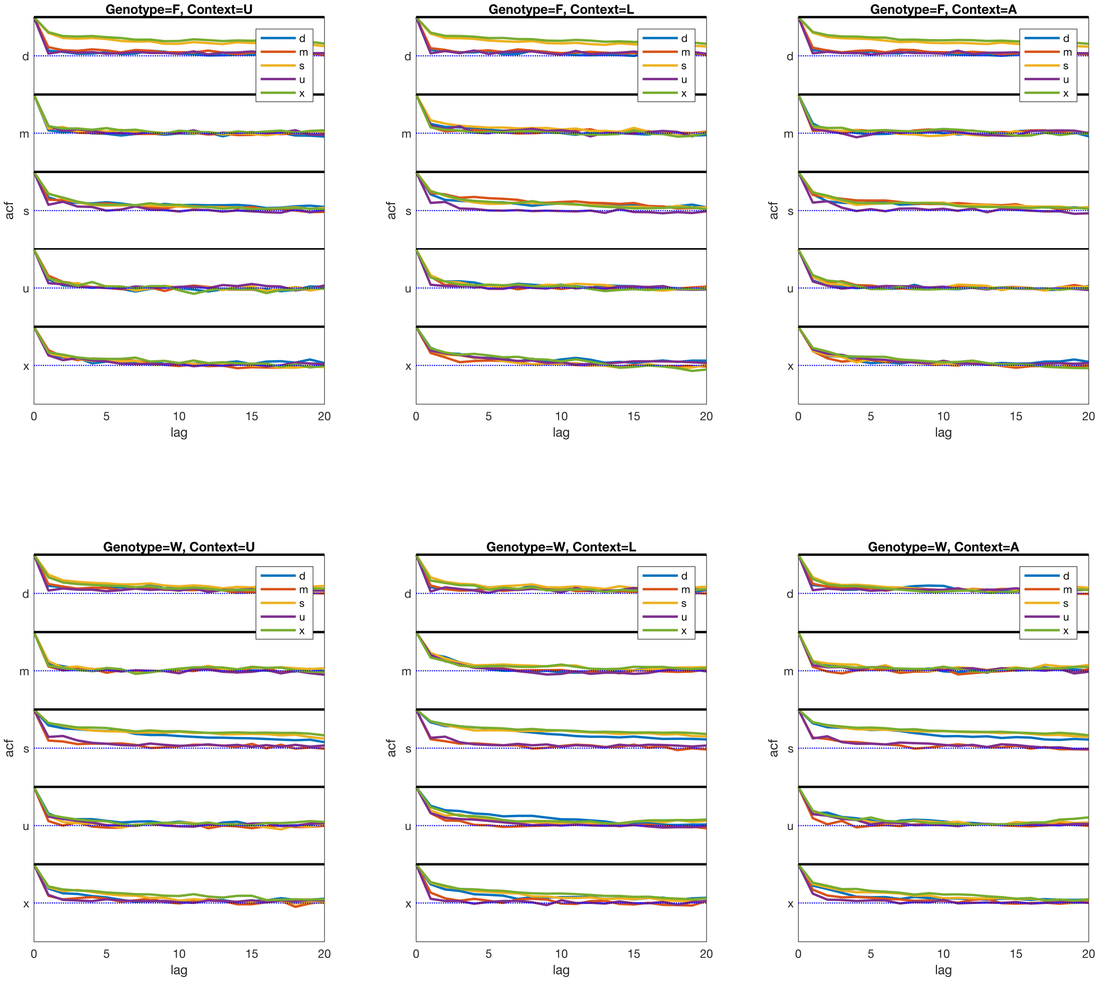


Figure S.8: Results for the Foxp2 data set. Trace plot of sampled values of transition probabilities $P_{x_1,x_2}(y_t \mid y_{t-1})$ for syllables $y_t, y_{t-1} \in \{d, m, s, u, x\}$ for different combinations of genotype $x_1 \in \{F, W\}$ and social contexts $x_2 \in \{U, L, A\}$. In each panel, the preceding syllables $y_{t-1}$ are presented in different horizontal sub-panels. Within each sub-panel, the range of the y-axis is $[0, 1]$ and different syllables $y_t$ are distinguished by different colors.



Genotype=F, Context=U

Genotype=F, Context=L

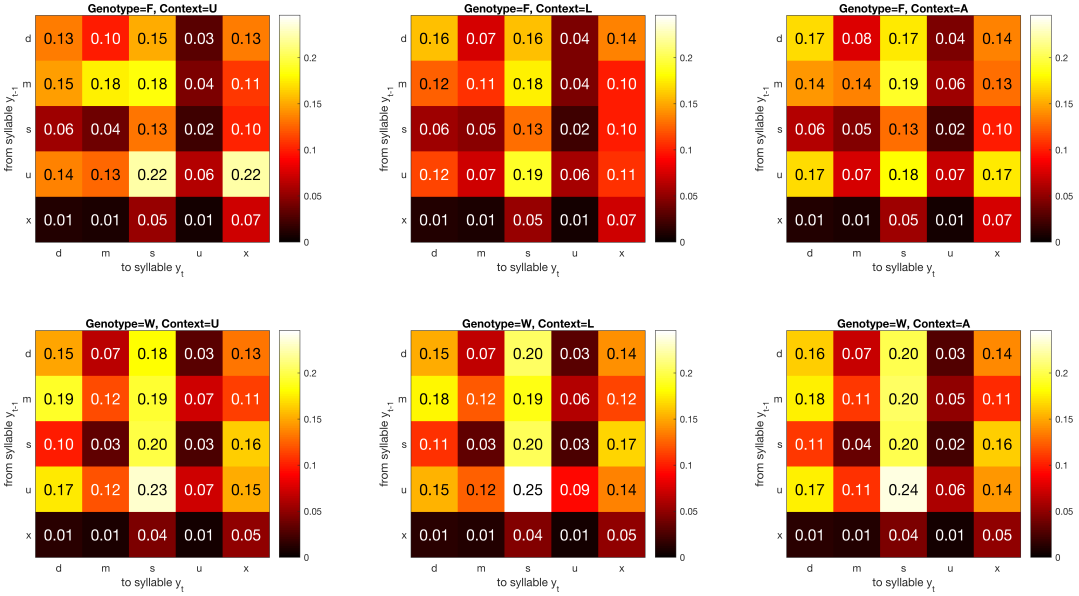
Genotype=F, Context=A

Genotype=W, Context=U

Genotype=W, Context=L

Genotype=W, Context=A

Figure S.9: Results for the Foxp2 data set. Autocorrelation plots of thinned samples of transition probabilities $P_{x_1,x_2}(y_t \mid y_{t-1})$ for syllables $y_t, y_{t-1} \in \{d, m, s, u, x\}$ for different combinations of genotype $x_1 \in \{F, W\}$ and social contexts $x_2 \in \{U, L, A\}$. In each panel, the preceding syllables $y_{t-1}$ are presented in different horizontal sub-panels. Within each sub-panel, the range of the y-axis is $[-1, 1]$ and different syllables $y_t$ are distinguished by different colors.



Genotype=F, Context=U

| | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.13 | 0.10 | 0.15 | 0.03 | 0.13 |
| m | 0.15 | 0.18 | 0.18 | 0.04 | 0.11 |
| s | 0.06 | 0.04 | 0.13 | 0.02 | 0.10 |
| u | 0.14 | 0.13 | 0.22 | 0.06 | 0.22 |
| x | 0.01 | 0.01 | 0.05 | 0.01 | 0.07 |

Genotype=F, Context=L

| | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.16 | 0.07 | 0.16 | 0.04 | 0.14 |
| m | 0.12 | 0.11 | 0.18 | 0.04 | 0.10 |
| s | 0.06 | 0.05 | 0.13 | 0.02 | 0.10 |
| u | 0.12 | 0.07 | 0.19 | 0.06 | 0.11 |
| x | 0.01 | 0.01 | 0.05 | 0.01 | 0.07 |

Genotype=F, Context=A

| | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.17 | 0.08 | 0.17 | 0.04 | 0.14 |
| m | 0.14 | 0.14 | 0.19 | 0.06 | 0.13 |
| s | 0.06 | 0.05 | 0.13 | 0.02 | 0.10 |
| u | 0.17 | 0.07 | 0.18 | 0.07 | 0.17 |
| x | 0.01 | 0.01 | 0.05 | 0.01 | 0.07 |

Genotype=W, Context=U

| | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.15 | 0.07 | 0.18 | 0.03 | 0.13 |
| m | 0.19 | 0.12 | 0.19 | 0.07 | 0.11 |
| s | 0.10 | 0.03 | 0.20 | 0.03 | 0.16 |
| u | 0.17 | 0.12 | 0.23 | 0.07 | 0.15 |
| x | 0.01 | 0.01 | 0.04 | 0.01 | 0.05 |

Genotype=W, Context=L

| | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.15 | 0.07 | 0.20 | 0.03 | 0.14 |
| m | 0.18 | 0.12 | 0.19 | 0.06 | 0.12 |
| s | 0.11 | 0.03 | 0.20 | 0.03 | 0.17 |
| u | 0.15 | 0.12 | 0.25 | 0.09 | 0.14 |
| x | 0.01 | 0.01 | 0.04 | 0.01 | 0.05 |

Genotype=W, Context=A

| | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.16 | 0.07 | 0.20 | 0.03 | 0.14 |
| m | 0.18 | 0.11 | 0.20 | 0.05 | 0.11 |
| s | 0.11 | 0.04 | 0.20 | 0.02 | 0.16 |
| u | 0.17 | 0.11 | 0.24 | 0.06 | 0.14 |
| x | 0.01 | 0.01 | 0.04 | 0.01 | 0.05 |

Figure S.10: Results for the Foxp2 data set. The panels show $10^2 \times \widehat{\sigma}_{MC,x_1,x_2}(y_t \mid y_{t-1})$ for syllables $y_t, y_{t-1} \in \{d, m, s, u, x\}$ for different combinations of genotype $x_1 \in \{F, W\}$ and social contexts $x_2 \in \{U, L, A\}$. Here, $\widehat{\sigma}_{MC,x_1,x_2}(y_t \mid y_{t-1})$ is the estimated Monte Carlo uncertainty in estimating the posterior expectation of the transition probabilities $P_{x_1,x_2}(y_t \mid y_{t-1})$, reported in Figure 5, using thinned MCMC samples.

## S.5 Comparison with GLM Based Approaches

In this section, we revisit GLM based approaches to mixed effects Markov sequences. Adapting to Altman (2007), without any interaction among the local predictor $y_{t-1}$ and the exogenous predictors $x_j, j = 1, \ldots, p$, using the logit link, we are now required to formulate $d_0 - 1$ models, one for each $y_t = 1, \ldots, d_0 - 1$, of the form

$$\log\left\{\frac{P^{(i_s)}_{x_{s,1},\ldots,x_{s,p}}(y_{s,t} = y_t \mid y_{s,t-1})}{P^{(i_s)}_{x_{s,1},\ldots,x_{s,p}}(d_0 \mid y_{s,t-1})}\right\} = \beta_{0,y_t} + \sum_{y_{t-1}=1}^{d_0-1} \beta_{y_t,y_{t-1}} 1\{y_{s,t-1} = y_{t-1}\}$$

$$+ \sum_{j=1}^{p}\sum_{x_j=1}^{d_j-1} \beta_{j,y_t,x_j} 1\{x_{s,j} = x_j\} + \sum_{y_{t-1}=1}^{d_0-1} u^{(i_s)}_{y_t,y_{t-1}} 1\{y_{s,t-1} = y_{t-1}\},$$

where $\mathbf{u}^{(i)} = \{u^{(i)}_{y_t,y_{t-1}}\}_{y_t=1,y_{t-1}=1}^{d_0-1,d_0}$ are random effects to due to the $i^{th}$ individual. Except for the restrictive special case of binary sequences, estimation of the model parameters becomes prohibitively complex, especially in presence of multiple exogenous predictors. Incorporating only second order interactions would require an additional $N_{int} = \sum_{j_1=0}^{p}\sum_{j_2=0,j_1\neq j_2}^{p}(d_{j_1} - 1)(d_{j_2} - 1)$ terms for each of the $d_0 - 1$ models, significantly increasing model complexities. For the Foxp2 application, for instance, this would require $N_{int} = 14$ additional terms in each of the 4 models. We have thus ignored interactions among the exogenous and the local predictors here.

The population average probabilities implied by the model can be obtained by integrating out the random effects as

$$P_{x_1,\ldots,x_p}(y_t \mid y_{t-1}) = \int P^{(i)}_{x_1,\ldots,x_p}(y_t \mid y_{t-1}) f(\mathbf{u}^{(i)}_{y_{t-1}}) d\mathbf{u}^{(i)}_{y_{t-1}}$$

$$= \int \frac{\exp\left(\beta_{0,y_t} + \beta_{y_t,y_{t-1}} + \sum_{j=1}^{p} \beta_{j,y_t,x_j} + u^{(i)}_{y_t,y_{t-1}}\right)}{\sum_{h=1}^{d_0} \exp\left(\beta_{0,h} + \beta_{h,y_{t-1}} + \sum_{j=1}^{p} \beta_{j,h,x_j} + u^{(i)}_{h,y_{t-1}}\right)} f(\mathbf{u}^{(i)}_{y_{t-1}}) d\mathbf{u}^{(i)}_{y_{t-1}},$$

where $\beta_{0,d_0} = 0$, $\beta_{d_0,y_{t-1}} = \beta_{1,d_0,y_{t-1}} = \cdots = \beta_{p,d_0,y_{t-1}} = u^{(i)}_{d_0,y_{t-1}} = 0$ for all $y_{t-1}$, $\mathbf{u}^{(i)}_{y_{t-1}} = (u^{(i)}_{1,y_{t-1}}, \ldots, u^{(i)}_{d_0-1,y_{t-1}})^{\mathrm{T}}$, and $f(\mathbf{u}^{(i)}_{y_{t-1}})$ is the random effects distribution. Typically it is assumed that $f(\mathbf{u}^{(i)}_{y_{t-1}}) = \mathrm{MVN}_{d_0-1}(\mathbf{0}, \boldsymbol{\Sigma}_u)$, where $\mathrm{MVN}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a $q$-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Often such models are further simplified by assuming $\mathbf{u}^{(i)}_{y_{t-1}} = \mathbf{u}^{(i)}$ for all $y_{t-1}$ (Altman, 2007) and the components of $\mathbf{u}_{y_t}$ to be distributed independently with $\boldsymbol{\Sigma}_u = \mathrm{diag}(\sigma^2_{u,1}, \ldots, \sigma^2_{u,d_0-1})$.

Even with such restrictive simplifying assumptions, the population level transition probabilities $P_{x_1,\ldots,x_p}(y_t \mid y_{t-1})$ do not have closed form expressions. Assuming $P_{x_1,\ldots,x_p}(y_t \mid y_{t-1})$ to arise from the same multinomial logit functional form

$$P_{x_1,\ldots,x_p}(y_t \mid y_{t-1}) = \frac{\exp\left(\beta^{\star}_{0,y_t} + \beta^{\star}_{y_t,y_{t-1}} + \sum_{j=1}^{p} \beta^{\star}_{j,y_t,x_j}\right)}{\sum_{h=1}^{d_0} \exp\left(\beta^{\star}_{0,h} + \beta^{\star}_{h,y_{t-1}} + \sum_{j=1}^{p} \beta^{\star}_{j,h,x_j}\right)},$$

an approximation yields $\beta^{\star}_{0,y_t} \approx \beta_{0,y_t}/(1+c^2\sigma^2_{u,y_t})^{1/2}, \beta^{\star}_{y_t,y_{t-1}} \approx \beta_{y_t,y_{t-1}}/(1+c^2\sigma^2_{u,y_t})^{1/2}$ and so on, where $c = (16\sqrt{3})/(15\pi)$ (Zeger *et al.*, 1988). Individual and population level fixed effects parameters are thus different and have to be differently interpreted. Specifically, population level probabilities depend on individual heterogeneity - two populations with different individual heterogeneity will have different population level probabilities even if they have the same individual level fixed effects parameters.

Testing scientific hypotheses related to influences of the predictors using such GLM based models is also complicated. For instance, the global null $H_{0j}$ of no effect of the $j^{th}$ exogenous predictor $x_j$, when translated in terms of the model parameters, becomes a complicated composite hypothesis $H_{0j} : \beta_{j,y_t,x_j} = 0$ for all $y_t = 1,\ldots,d_0-1$ and all $x_j = 1,\ldots,d_j-1$.

Similar exact functional forms for both the marginal and the individual level mixed models can be guaranteed by 'bridge' distributed random effects (Wang and Louis, 2003). Focusing on the simplest binary cases, for any link function $H$ with $h = H'$, the 'bridge' distribution $g$ is given by

$$g(u) = \frac{1}{2\pi}\int \exp\{\iota(k/\varphi - u)\}\frac{\mathcal{F}h(v/\varphi)}{\mathcal{F}h(v)}dv,$$

such that $\int H(u+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})g(u)du = H(k+\varphi\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, where $\mathcal{F}h$ is the characteristic function of $h$ and $\varphi \in (0,1]$ is an attenuating scale factor. For the logistic link $H(x) = \exp(x)\{1+\exp(x)\}^{-1}$, for example, the bridge distribution is given by

$$g(u) = \frac{1}{2\pi}\frac{\sin(\varphi\pi)}{\{\cosh(\varphi u)+\cos(\varphi\pi)\}},$$

a symmetric distribution with variance $\pi^2(\varphi^2-1)/3$. For the probit link $H(x) = \Phi(x)$, the bridge distribution is also Gaussian with rescaling factor $\varphi = (1+\sigma_b^2)^{-1/2}$, where $\sigma_b^2$ is the variance of the bridge distribution Normal$(0,\sigma_b^2)$. Multivariate generalizations are, however, not straightforward. Additionally, although such bridge distributed random effects allow the population and the individual level models to have the same functional forms, other limitations of GLM based approaches, as described above, still remain, making such methods prohibitively complex in settings like ours.

In comparison, our model is highly flexible, parsimoniously accommodating interactions of all orders between the exogenous and the local predictors, while also completely avoiding to have to choose any link function. The random effects in our model for the individual level transition probabilities can be easily integrated out to obtain closed form expressions of the population level transition probabilities. The fixed effects components remain the same in both individual and population level probabilities and hence can be interpreted in the same way. Finally, testing scientific hypotheses related to global influences of the predictors is very straightforward using our approach as they can be translated in terms of a single model parameter.

We implemented the multinomial logit based mixed effects Markov model described above using the MCMCglmm package in R (Hadfield, 2010). Maximum likelihood estimation of the model parameters using other R packages did not produce realistic results. Figure S.11 shows the estimated posterior means of the population level transition probabilities based on $12,000$ samples drawn from the posterior, thinned by an interval of 10 after the initial $2,000$ were discarded as burnin. Comparison with estimates produced by our method, summarized in Figure 5 in the main paper, suggests overall agreement. For reasons detailed above, global significance of the exogenous predictors could not be straightforwardly assessed. We could, however, assess the significance of each $\beta$ parameter from the MCMC output using the minimum of the proportion of samples in which $\beta$ is on one side or the other of zero, referred to as pMCMC in MCMCglmm. For the Foxp2 data set, the four $\beta$ parameters associated with genotype, namely $\beta_{1,1,1}, \beta_{1,2,1}, \beta_{1,3,1}, \beta_{1,4,1}$, had pMCMC values 0.446, 0.436, 0.368 and 0.106, indicating none of them to be marginally significant. To assess local differences in transition probabilities between the two genotypes, we employed the approach developed in Section 4 of the main paper. Figure S.12 summarizes the posterior probabilities of the local null hypotheses $H_{0,y_t|y_{t-1},x_2} : |\Delta P_{\cdot,x_2}(y_t \mid y_{t-1})| \leq 0.02$ estimated from the MCMC output of the GLM based model. Unlike the results produced by our approach, summarized in Figure 6 in the main paper, no local difference was found to be significant at the 0.10 posterior probability level.

To further assess how the multinomial logit based mixed effects Markov model compares with our proposed approach in detecting local differences in transition probabilities between the two genotypes, we compared the results produced by the two methods for data sets simulated under scenario F described in Section 6 in the main paper. The posterior means of the population level transition probabilities estimated by the GLM based approach (not shown here) were quite different from the truth. Figures S.13 and S.14 summarize the estimated posterior probabilities of the local null hypotheses $H_{0,y_t|y_{t-1},x_2} : |\Delta P_{\cdot,x_2}(y_t \mid y_{t-1})| \leq 0.02$ for two different simulated data sets. Compared to the results produced by our approach, summarized in Figure 10 and Figure S.7, there were many more false decisions.

Another possible approach to syntax analysis could be to transform the sample transition proportions and fit a generalized linear mixed effects model to those transformed values. Such an approach would inherit the limitations of summary statistics based approaches, such as the method of Chabout *et al.* (2016) discussed in Section 2 of the main paper, as well as those of the GLM based approaches discussed above. Yet another possibility is to assume an independent Dirichlet model for each mouse under each experimental experimental condition. In experiments with such models, we found the results to be numerically very unstable and highly sensitive to the choice of the hyper-parameters. These issues provided motivation for developing the more structured approach presented in the paper.

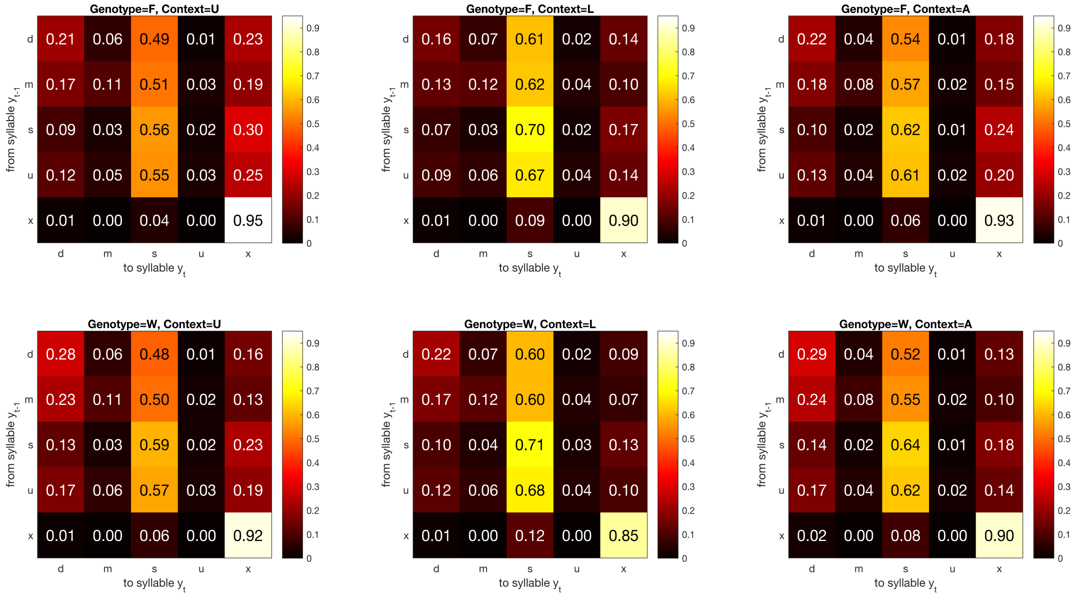


Figure S.11: Results for the Foxp2 data set for the GLM based approach described in Section S.5 in the Supplementary Materials. Estimated approximate posterior mean transition probabilities $P_{x_1,x_2}(y_t \mid y_{t-1})$ for syllables $y_t, y_{t-1} \in \{d, m, s, u, x\}$ for different combinations of genotype $x_1 \in \{F, W\}$ and social contexts $x_2 \in \{U, L, A\}$.

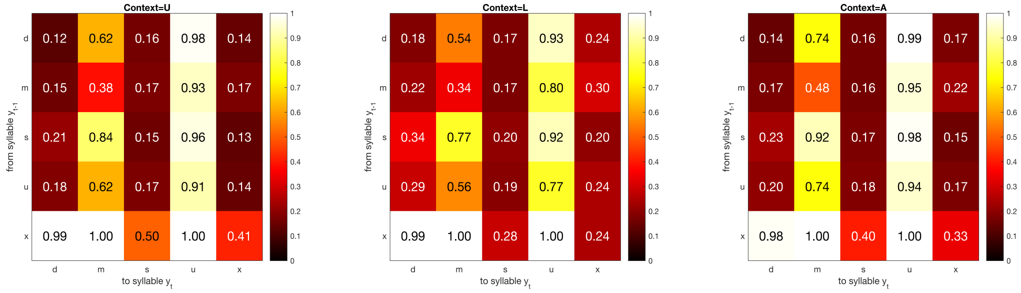
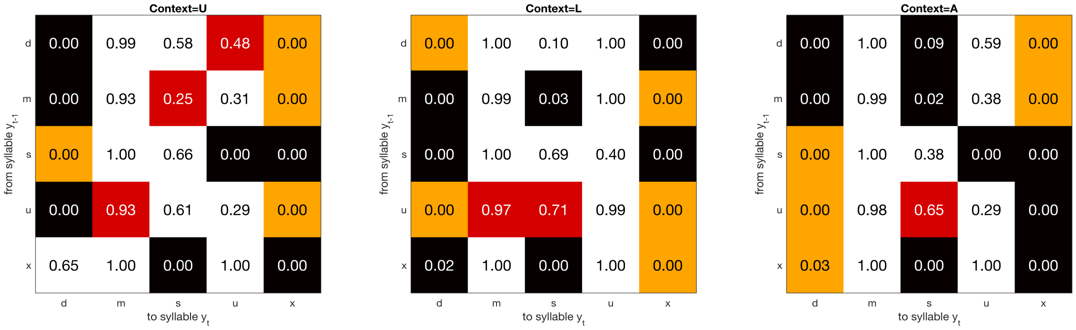


Figure S.12: Results for the Foxp2 data set for the GLM based approach described in Section S.5 in the Supplementary Materials. The estimated posterior probability of $H_{0,y_t|y_{t-1},x_2} : |\Delta P_{\cdot,x_2}(y_t \mid y_{t-1})| = |P_{1,x_2}(y_t \mid y_{t-1}) - P_{2,x_2}(y_t \mid y_{t-1})| \leq 0.02$.



Context=U

| from syllable $y_{t-1}$ \ to syllable $y_t$ | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.00 | 0.99 | 0.58 | 0.48 | 0.00 |
| m | 0.00 | 0.93 | 0.25 | 0.31 | 0.00 |
| s | 0.00 | 1.00 | 0.66 | 0.00 | 0.00 |
| u | 0.00 | 0.93 | 0.61 | 0.29 | 0.00 |
| x | 0.65 | 1.00 | 0.00 | 1.00 | 0.00 |

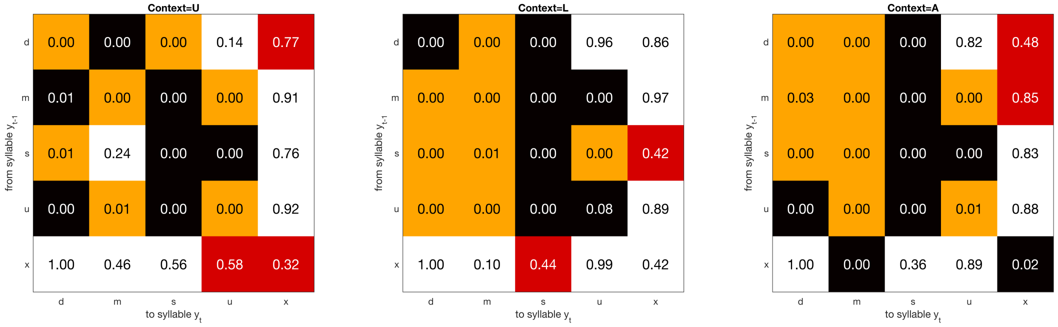
Context=L

| from syllable $y_{t-1}$ \ to syllable $y_t$ | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.00 | 1.00 | 0.10 | 1.00 | 0.00 |
| m | 0.00 | 0.99 | 0.03 | 1.00 | 0.00 |
| s | 0.00 | 1.00 | 0.69 | 0.40 | 0.00 |
| u | 0.00 | 0.97 | 0.71 | 0.99 | 0.00 |
| x | 0.02 | 1.00 | 0.00 | 1.00 | 0.00 |

Context=A

| from syllable $y_{t-1}$ \ to syllable $y_t$ | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.00 | 1.00 | 0.09 | 0.59 | 0.00 |
| m | 0.00 | 0.99 | 0.02 | 0.38 | 0.00 |
| s | 0.00 | 1.00 | 0.38 | 0.00 | 0.00 |
| u | 0.00 | 0.98 | 0.65 | 0.29 | 0.00 |
| x | 0.03 | 1.00 | 0.00 | 1.00 | 0.00 |

Figure S.13: Results for the simulation scenario F described in Section 6 in the main paper. These results were produced by the GLM based approach described in Section S.5 in the Supplementary Materials. The results show the estimated posterior probabilities of $H_{0,y_t|y_{t-1},x_2} : |\Delta P_{\cdot,x_2}(y_t \mid y_{t-1})| \leq 0.02$. Posterior probabilities smaller than 0.1 are considered significant and are highlighted in black and orange. Posterior probabilities greater than 0.1 are presented in white and red. White and black cells represent correct decisions, orange cells mark rejections of true $H_{0,\ell}$ (false positives), and red cells mark failures to reject false $H_{0,\ell}$ (false negatives). Compare with Figure 10 in the main paper.



Context=U

| from syllable $y_{t-1}$ \ to syllable $y_t$ | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.00 | 0.00 | 0.00 | 0.14 | 0.77 |
| m | 0.01 | 0.00 | 0.00 | 0.00 | 0.91 |
| s | 0.01 | 0.24 | 0.00 | 0.00 | 0.76 |
| u | 0.00 | 0.01 | 0.00 | 0.00 | 0.92 |
| x | 1.00 | 0.46 | 0.56 | 0.58 | 0.32 |

Context=L

| from syllable $y_{t-1}$ \ to syllable $y_t$ | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.00 | 0.00 | 0.00 | 0.96 | 0.86 |
| m | 0.00 | 0.00 | 0.00 | 0.00 | 0.97 |
| s | 0.00 | 0.01 | 0.00 | 0.00 | 0.42 |
| u | 0.00 | 0.00 | 0.00 | 0.08 | 0.89 |
| x | 1.00 | 0.10 | 0.44 | 0.99 | 0.42 |

Context=A

| from syllable $y_{t-1}$ \ to syllable $y_t$ | d | m | s | u | x |
|---|---|---|---|---|---|
| d | 0.00 | 0.00 | 0.00 | 0.82 | 0.48 |
| m | 0.03 | 0.00 | 0.00 | 0.00 | 0.85 |
| s | 0.00 | 0.00 | 0.00 | 0.00 | 0.83 |
| u | 0.00 | 0.00 | 0.00 | 0.01 | 0.88 |
| x | 1.00 | 0.00 | 0.36 | 0.89 | 0.02 |

Figure S.14: Results for the simulation scenario F described in Section 6 in the main paper, but for a different random seed. These results were produced by the GLM based approach described in Section S.5 in the Supplementary Materials. The results show the estimated posterior probabilities of $H_{0,y_t|y_{t-1},x_2} : |\Delta P_{\cdot,x_2}(y_t \mid y_{t-1})| \leq 0.02$. Posterior probabilities smaller than 0.1 are considered significant and are highlighted in black and orange. Posterior probabilities greater than 0.1 are presented in white and red. White and black cells represent correct decisions, orange cells mark rejections of true $H_{0,\ell}$ (false positives), and red cells mark failures to reject false $H_{0,\ell}$ (false negatives). Compare with Figure S.7 in the Supplementary Materials.

# References

Altman, R. M. (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102, 201-210.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2, 1152-1174.

Chabout, J., Sarkar, A., Patel, S., Raiden, T., Dunson, D. B., Fisher, S. E., and Jarvis, E. D. (2016). A Foxp2 mutation implicated in human speech deficits alters sequencing of ultrasonic vocalizations in adult male mice. *Frontiers in Behavioral Neuroscience*, 10, 1-18.

Eichelsbacher, P. and Ganesh, A. (2002). Bayesian inference for Markov chains. *Journal of Applied Probability*, 39, 91-99.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics.* Springer Verlag, Berlin.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33, 1-22.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101, 1566-1581.

Wang, Z. and Louis, T. A. (2003). Matching conditional and marginal shapes in binary mixed-effects models using a bridge distribution function. *Biometrika*, 90, 765-775.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. Institute of Statistics and Decision Sciences, Duke University, Durham, USA, Technical report.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.