Hierarchical Semi-Markov Smooth Models of Latent Neural States

Joshua Krause, Jacolien van Rij, and Jelmer P. Borst
Bernoulli Institute for Mathematics, Computer Science, and Artificial Intelligence

Hidden (semi-) Markov Models (HsMMs) are increasingly being used to segment neuro-physiological signals into sequences of latent cognitive processes. The idea: different processes will leave distinct traces in trial-level recordings of (multivariate) neuro-physiological signals. Markov models, equipped with an *emission model* of these traces and a *latent process model* describing the progression through the different latent processes involved in a task, can then be used to infer the most likely process for any time-point and trial. However, the currently used HsMMs remain limited in two important ways. First, they cannot account for subject-level heterogeneity in the latent and emission process. Instead, a single group-level model is assumed to explain the entire data. Second, they cannot account for the potentially non-linear effects of experimental covariates on the latent and emission process. To address these problems, we present a modeling framework in which the HsMM parameters of the emission and latent process are replaced with mixed additive models, including smooth functions of experimental covariates and random effects. We derive all necessary quantities for empirical Bayes and fully Bayesian inference for all parameters and provide a Python implementation of all estimation algorithms. To demonstrate the advantages offered by this framework, we apply such a multi-level model to an existing lexical decision dataset. We show that, even in such a simple task, not all subjects rely on the same processes equally and that at least two semi-Markov states, previously believed to reflect distinct processes, might actually relate to the same cognitive process.

Keywords: HsMMs, General Smooth Models, Multi-level Models, Bayesian Inference, Model Selection, Brain States, Multivariate Neuro-physiological Signals

1 Introduction

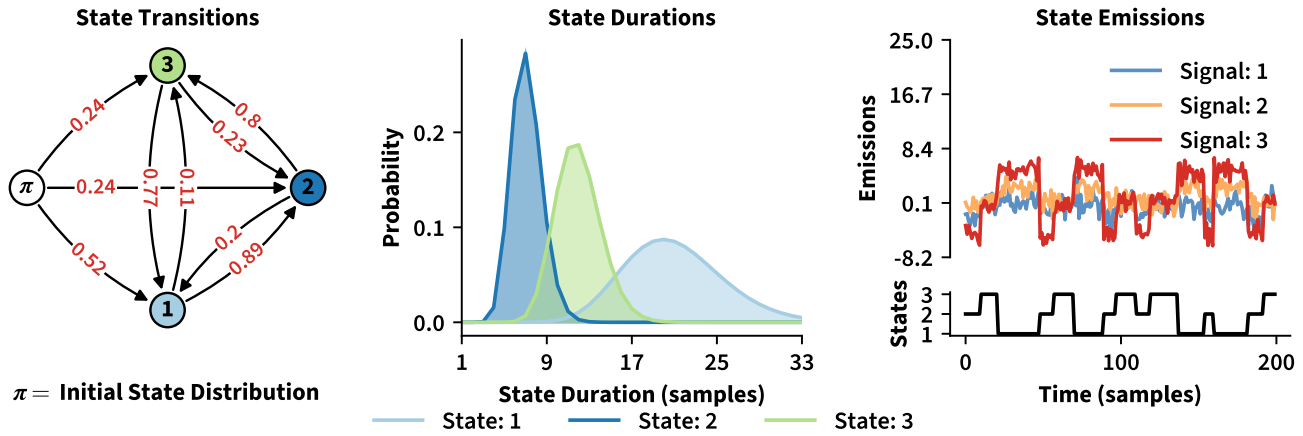
Scientists have long wanted to recover the cognitive processes involved in task completion, with early attempts tracing back to Sternberg (1969) and Donders (1868). One challenge is that even simple tasks can conceivably be solved in many different ways. Successfully completing a lexical decision task for example, requiring participants to decide whether a presented character string corresponds to a word or not, implies completion of *at least* two distinct cognitive processes: initial string encoding and response initialization. Before participants can initialize the response, however, they will first have to weigh the available evidence such as the string's orthographic plausibility and general similarity to other well-known words. However, exactly which information ultimately drives the decision, and whether this information is evaluated in a single overall decision process or in a sequence of decision-related processes, remains up for debate (e.g., Berberyan et al., 2021; Krause et al., 2024; Mahé et al., 2015; Ratcliff et al., 2004; Wagenmakers et al., 2008). This problem extends beyond lexical decision making to most cognitive tasks, with theories often disagreeing on the nature of the cognitive processes involved: their number, content, and on whether they have to be completed sequentially or can overlap at least partially (e.g., McClelland, 1979).

In this paper we present a novel modeling framework, based on hierarchical Hidden semi-Markov Models, that can be used to decode cognitive processes from time-varying neuro-physiological signals and thus help to find answers to these and related questions. The idea to decode processes from experimental data is far from new. Early attempts, which largely relied on manipulations of the experimental design and involved models of behavioral data such as reaction time (RT) measurements, date back to Sternberg (1969) and Donders (1868). Because behavioral data only provides information about the combined duration of processes required to achieve task performance, researchers recently started to rely on neuro-physiological signals to get more direct access to the processing time-course (see Anderson et

Joshua Krause Jacolien van Rij Jelmer P. Borst
Correspondence concerning this article should be addressed to Joshua Krause, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands. E-mail: j.krause@rug.nl

Figure 1

Hidden Semi-Markov Model Overview



Note. Figure 1 visualizes the properties of the *latent* and *emission* processes characterizing a Hidden Semi Markov Model – here assumed to have three latent states. The left-most panel shows the transition graph implied by the *state transition probabilities*. Note, that self-transitions are considered impossible in many HsMMs and thus not visualized here. To account for prolonged stays in a given state, HSMMs introduce explicit *state sojourn distributions* (e.g., Yu, 2010). State-specific Gamma distributions, as visualized in the central panel, are a popular choice to model these distributions. The panel on the right shows how the latent process relates to the emission process. The panel on the bottom-right shows a possible realization of the Markov chain parameterized by the aforementioned state transition and duration probabilities. While in a given state, observable emissions are generated according to the state's emission distribution (top-right). For this example, each state's emission distribution was chosen as a multivariate Gaussian with distinct mean vectors and covariance matrices – in practice, differences between the emissions of different states are often more subtle.

al., 2016, for more detailed discussions).

The basic premise of these approaches – including the one presented here – is that different processes either rely on different neural substrates altogether or utilize them differently (for example to perform different computations), and thus leave distinct traces in neuro-physiological signals (e.g., Anderson et al., 2016; Borst & Anderson, 2015; Coquelet et al., 2022; Pieramico et al., 2025; Quinn et al., 2018; Vidaurre et al., 2016). Note, that *process* here does not necessarily refer to a single (type of) cognitive computation but also applies to *mixtures* of temporally overlapping computations associated with a distinct neuro-physiological trace (e.g., Baker et al., 2014; Pieramico et al., 2025; Vidaurre et al., 2025). Machine learning models can then be used to detect these traces and to infer the process most likely to have generated them. Recent attempts to decode processes have for example involved models of the BOLD response in different cortical regions (e.g., Anderson & Fincham, 2014; Anderson et al., 2016; Dang et al., 2017; Vidaurre et al., 2016, 2025), or multi-electrode EEG or MEG recordings (e.g., Anderson et al., 2016; Borst & Anderson, 2015; Coquelet et al., 2022; Krause et al., 2024; Lehmann et al., 1998; Vidaurre et al., 2016, 2018; Weindel et al., 2024).

Successfully decoding processes requires algorithms or methods that can reliably detect evidence of these processes in continuous time-varying signals. Conceptually, Hidden (semi) Markov Models (H(s)MMs) are ideally suited for this task. These models are characterized by two processes: a *latent process*, represented by a discrete Markov chain, and an *emission process*, producing observable output dependent on the current latent state of the Markov chain (e.g., Rabiner, 1990; Yu, 2010; Yu & Kobayashi, 2006, see also Figure 1).

The Markov chain can be considered a model of the potential progression through unobservable cognitive processes: just like a participant progresses through a sequence of processes to solve a certain task, the chain progresses through a sequence of distinct states according to an initial probability distribution and state transition probabilities (see left-most panel in Figure 1). As discussed in more detail in the upcoming methods section, the time spent in each state is typically modeled explicitly in Hidden semi-Markov Models – allowing these models to account for the fact that the duration of cognitive processes often varies considerably depending on the purpose they fulfill (see central panel in Figure 1).

Similarly, the emission process can be considered a formalization of the *linking hypothesis*, detailing how residence in a particular cognitive process manifests in the observable neurological signal (Anderson et al., 2016; Vidaurre et al., 2016, 2018). Conveniently, as shown in the right-most panel of Figure 1, the emission process can readily accommodate

the multi-dimensional nature of most neuro-physiological data.

Another advantage offered by these models is that well-tested algorithms exist to estimate the parameters involved in the emission and latent processes from collected data. Subsequently, these parameter estimates can be used to decode the most likely realization of the Markov chain for the given data (see Yu, 2010; Zucchini et al., 2017, for an overview). Considering the conceptual parsimony and theoretical maturity of Markov models, it is thus perhaps unsurprising that they are popular among researchers wanting to recover processing stages from neuro-physiological data (e.g., Anderson et al., 2016; Borst & Anderson, 2015; Coquelet et al., 2022; Dang et al., 2017; Krause et al., 2024; Pieramico et al., 2025; Vidaurre et al., 2016, 2018, 2025; Weindel et al., 2024).

However, in contrast to the models presented in this paper, the specific (Semi) Markov models currently used to recover processing stages remain limited in two important ways. First, these models typically do not account directly for subject-level heterogeneity in the latent and emission process. Second, they cannot readily account for potentially non-linear effects of experimental covariates on the latent and emission processes.

Specifically, while data is commonly collected from different subjects, all of whom complete multiple experimental trials, it is common to estimate only a single set of parameters applicable to the entire group. While such a group-level model can still be used to decode the separate (order and duration of) processes completed on individual trials, doing so assumes that the time-series from individual trials reflect different realizations of the same Markov chain and emission process. This assumption is difficult to justify, since different subjects can generally be expected to vary systematically in the time it takes them to complete individual cognitive processes.

Similarly, some participants might simply not complete a specific process at all or might gradually stop to do so because they become more familiar with the task or adapt their strategy over time (e.g., Baayen et al., 2017). Even if data were to be collected from a single expert participant, any experimental covariate designed to influence task difficulty would almost certainly influence the properties of the latent process as well – and might do so non-linearly. Indeed, in our recent work we observed strong non-linear effects of word-likeness on the duration of virtually all processes involved in lexical decision making (Krause et al., 2024). There we also discussed the possibility of word-likeness influencing the neural pattern indicating residence in a particular cognitive process, and that it is plausible that different subjects might deviate differently from this pattern as well.

In summary, subject-level heterogeneity and non-linear covariate effects both introduce dependencies into the emission and latent process, which current models cannot readily account for. To address these dependencies we propose to replace constant group-level parameters of HsMMs with additive mixed (or multi-level) models of the parameters (or a known function of them; cf. Michelot, 2025; Wood et al., 2016). The multi-level models of individual parameters can include any combination of smooth functions of experimental covariates and i.i.d random effects to capture subject-level heterogeneity.

We start by formally introducing the specific semi-Markov model used most commonly by researchers wanting to detect cognitive processes in the upcoming section 2.1. We then outline how this model can be expanded to include non-linear and random effects (section 2.2) and derive the quantities necessary to estimate the resulting models, such as the log-likelihood function and corresponding gradient (sections 2.2.1 - 2.2.3). We also provide an implementation of the discussed algorithms in the open-source Python toolbox `mssm` (version $\geq$ 1.2.0), available on GitHub and distributed via the Python Packaging Index[1].

In section 3 we demonstrate the usefulness of the framework presented here: we present a re-analysis of the lexical decision dataset collected by Krause et al. (2024), showing that only a multi-level model accounting for subject-level heterogeneity and non-linear covariate effects could have generated the data. While the improved model generally makes similar predictions about group-level effects of word-likeness and word-type, it also reveals that not all subjects seem to rely on the same processes equally and that at least two semi-Markov states, previously believed to reflect distinct processes, might actually relate to the same cognitive process.

Finally, in section 4 we discuss the implications of the framework presented here for the task of *model selection*: identifying the model most likely to have generated the data from a set of candidates. The space of possible models grows substantially once random effects and smooth functions of covariates are taken into account, which significantly complicates this task. We discuss different options to address this problem and their respective limitations. Finally, we discuss possible extensions and applications of the framework presented here.

2 Methods

Recent attempts to recover processing stages mostly rely on some kind of Hidden (semi-) Markov model. The particular semi-Markov model most commonly used is the "Explicit Duration Hidden Markov Model" (EDHMM), which is a specific type of HsMM (see Levinson, 1986; Rabiner, 1990; Yu, 2010). In the next section we conduct a formal review of this type of model. To simplify notation for this review, we assume that only a single time-series is observed.

[1] See https://github.com/JoKra1/mssm.

In section 2.1.3 we outline how the EDHMM can be generalized to account for multiple time-series (see also Rabiner, 1990).

2.1 Review: Explicit Duration Hidden Markov Models

Like other HMMs (and HsMMs), the EDHMM is characterized by a *latent process*, taking on the form of a discrete Markov chain of latent states $\mathcal{S} = \{1, ..., M\}$, and an *emission process*, producing observable output o_t (e.g., EEG amplitudes or BOLD responses) dependent on the current state S_t of the Markov chain (e.g., Rabiner, 1990; Yu, 2010; Yu & Kobayashi, 2006). The main difference between HsMMs, including the EDHMM, and conventional HMMs, is that HsMMs also include an explicit model of the time spent in each state as part of the latent process model (e.g., Rabiner, 1990; Yu, 2010). In the next two sub-sections we provide a formal review of the properties of the latent and emission process of an EDHMM.

2.1.1 The Latent Process of an EDHMM

The latent process of an EDHMM is a Markov chain with the extension that time spent in each state is modeled via explicit distributions (e.g., Rabiner, 1990; Yu, 2010). We first consider the properties of the Markov chain, which is defined by a probability mass function over the initial states and state transition probabilities (Yu, 2010). It is assumed that at $t = 1$, the Markov chain is in state j with probability $\boldsymbol{\pi}_j \stackrel{\text{def}}{=} p(S_1 = j)$. t by convention denotes the current discrete time-point. In the context of this paper, it thus makes sense to use t to refer to a particular sample of neuro-physiological data. For example, $t = 1$ will typically reflect the sample coinciding with the onset of a trial or stimulus. Similarly, the last observable output o_T is generated by state S_T at time-point $t = T$, which in the context of this paper typically reflects the last sample collected during a trial (for example when a participant gives a response).

To simplify the computation of the likelihood, it is conventionally assumed that a new state *begins* at $t = 1$ – in contrast to continuing from $t = 0$, $t = -1$, etc – and that the final state ends at $t = T$, after omitting the final emission (e.g., Yu, 2010). This assumption is not always easy to justify when latent states are assumed to correspond to cognitive processes. For example, in some experiments participants might come to anticipate the onset of the next stimulus and it would be unreasonable to assume that the onset of the first cognitive process coincides with stimulus onset. We return to these "censoring" issues, and how they can be addressed in section 2.2.1 (see also Yu, 2010, for an overview of different censoring issues).

The chain then repeatedly transitions from one state to another state according to the *state transition probabilities* (see left-most panel of Figure 1). We denote the transition from state i to state j with $\boldsymbol{\Delta}_{ij} \stackrel{\text{def}}{=} p(S_{t+1} = j | S_t = i, S_{t+1} \neq i)$ – all collected in the $M * M$ matrix $\boldsymbol{\Delta}$. For conventional HMMs, self-transitions are the mechanism to account for prolonged stays in a given state. When working with HsMMs, the time spent in each state (i.e., the "state sojourn time "; Yu, 2010) is modeled explicitly instead. As such, self-transitions are considered impossible by convention for the EDHMM, so that $\boldsymbol{\Delta}_{ii} = 0 \; \forall \; i \in S$ (cf. Yu, 2010).

The state sojourn time is treated as a discrete random variable d with state-dependent probability mass function $p_j(d)$ for $j \in \mathcal{S}$ and $d \in \{1, ..., D\}$ (e.g., Levinson, 1986; Yu, 2010; Yu & Kobayashi, 2006). Alternatively, a (normalized) probability density function can be used for the model of the sojourn time instead (e.g., Levinson, 1986; Yu, 2010). That is, $d|S = j \sim \mathcal{F}^d(\nu_j, ..., \gamma_j)$, where $\mathcal{F}^d$ is a state-specific duration or state sojourn time distribution parameterized by $\nu_j, ..., \gamma_j$. This drastically lowers the number of parameters required in the model of the durations expected for each state. While different distributions may require a different number of parameters, a mean ν_j parameter and optional scale parameter γ_j are sufficient when relying on a (normalized) member of the exponential family for $\mathcal{F}^d$ (see Yu, 2010, and the central panel of Figure 1).

2.1.2 The Emission Process of an EDHMM

The output o_t observed at time-point t is assumed to depend on the current state of the chain (e.g., Yu, 2010). Conventionally, the EDHMM assumes univariate emissions, believed to be independent and identically distributed (i.i.d) given knowledge about the current state of the system (this "conditional independence" assumption is common when working with other H(s)MMs as well, see; Levinson, 1986; Rabiner, 1990; Yu, 2010; Yu & Kobayashi, 2006). Hence, $o_t|S_t = j \sim \mathcal{F}^o(\mu_j, ..., \phi_j)$, where $\mathcal{F}^o$ denotes a state-specific *emission distribution* or *emission model*. Common choices for $\mathcal{F}^o$ are again members of the exponential family, such as Gaussian or Gamma distributions (e.g., Yu, 2010), which only require a mean μ_j and optional scale parameter ϕ_j.

When using EDHMMs to segment neuro-physiological signals into cognitive processes, it is typically necessary to account for multivariate output (e.g., a K vector of EEG amplitudes recorded from multiple electrodes). Multiple strategies exist to ensure that the EDHMM generalizes to multivariate output.

A popular option is to assume that the multivariate output vectors $\mathbf{o}_t$ are generated by state-specific multivariate normal distributions with mean vectors $\boldsymbol{\mu}_j$ and covariance matrices $\boldsymbol{\Sigma}_j$ – which also permits for multivariate auto-regressive emission models (e.g., Vidaurre et al., 2016). Such emissions models have been applied to trial-level EEG and MEG data directly (e.g., Vidaurre et al., 2016, 2018). More recently however, researchers have started to first transform the raw data, creating "virtual channels", sometimes also called "de-

layed embeddings", that carry information about temporal and spatial relations between and within raw data channels (i.e., electrodes) (see Borst & Anderson, 2015; Vidaurre et al., 2018, 2025). Subsequently, a Principal Component Analysis (PCA) is applied to the transformed data and a multivariate Gaussian emission model is assumed for the K top components (e.g., Vidaurre et al., 2018, 2025).

Alternatively, it is common to treat the emissions of the K signals as conditionally independent, given knowledge about the current state (e.g., Anderson et al., 2016; Borst & Anderson, 2015; Weindel et al., 2024). This is essentially a generalization of the univariate emission model outlined earlier. Specifically, one assumes that $p(\mathbf{o}_{t1}, ..., \mathbf{o}_{tK}|S_t = j) = \prod_{k=1}^{K} p(\mathbf{o}_{tk}|S_t = j)$. $p(\mathbf{o}_{tk}|S_t = j)$ denotes the density implied by the conditional univariate distribution $\mathbf{o}_{tk}|S_t = j \sim \mathcal{F}^o(\mu_{jk}, ..., \phi_{jk})$, where $\mathbf{o}_{tk}$ is the emission generated by signal k at time-point t (i.e., element k of emission vector $\mathbf{o}_t$, see also; Zucchini et al., 2017). Anderson et al. (2016) rely on this assumption for example, to derive their "multivariate pattern" emission model of EEG data. This model of the emissions assumes that the onset of a new process is assumed to result in the addition of a weighted half-sine of fixed duration to the EEG signal at each electrode, which can otherwise be described as a multivariate oscillatory noise process with zero expectation. To estimate an EDHMM, the authors compute a PCA of the raw EEG data and subsequently treat the K top components as independent (see Appendix A for more details, but also Anderson et al., 2016).

To avoid notational clutter we write $\mathbf{o}_t$ from now on, for both univariate and multivariate emissions and independent of whether a multivariate Gaussian or multiple independent densities are used as the model of multivariate emissions.

2.1.3 Multiple Time-Series, Subject-level Differences, and Covariate Effects

The EDHMM outlined above can account for data collected from multiple time-series and for possible trial-level variability (also between conditions or subjects) in the onset of cognitive processes – if we assume that different time-series reflect independent realizations of the *same* Markov chain and emission process (e.g., Zucchini et al., 2017). This assumption is attractive because only a single set of *group-level* parameters applicable to all collected time-series, subjects, etc. has to be estimated. Additionally, the independence assumption drastically simplifies computation of the model's likelihood (e.g., Yu, 2010; Zucchini et al., 2017).

While it is common to assume independence of different time-series given group-level parameter estimates only, this is not justified when dealing with the problem of recovering processing stages from neuro-physiological data due to the reasons outlined in the introduction: All of the aforementioned emission models assume a distinct neural pattern of some form, indicating the onset of or residence in a particular cognitive process, and different subjects might deviate from this pattern differently. Additionally, experimental covariates (e.g., any continuous or trial-specific manipulation of task difficulty) are likely to influence this pattern as well – and they might do so non-linearly.

Similarly, it is all but guaranteed that there will be systematic differences between subjects in the timing and the order of the underlying cognitive processes, i.e., the latent process. Finally, experimental covariates are just as likely to affect the properties of the latent process, as they are to affect the pattern in the observable data.

Previous work typically addresses the possibility of subject-level differences and non-linear covariate effects with a two-step or dual estimation process (e.g., Krause et al., 2024; Vidaurre et al., 2025). In a first step, group-level parameters of the EDHMM are estimated. These parameters are used to decode the most probable state sequence for every subject. In a second estimation step, subject-specific EDHMM parameters and functions of covariates are estimated given the recovered state sequences, which are kept fixed. While effective, such procedures are limited by the fact that there is no guarantee that similar state sequences would be produced by a model in which the EDHMM parameters are allowed to vary between subjects and as a function of experimental covariates. As such, subject-level and function estimates obtained from group-level state sequences remain biased.

Notably, if the true number of states M (i.e., processes) is assumed to be known in advance, this bias might not be of practical relevance. This will depend on factors such as the degree of subject-level variability and the amount of data available to estimate the group-level models. In contrast, the bias might be far more substantial when M is to be estimated alongside the Markov model parameters, for example via leave-one-out cross-validation (LOOCV; e.g., Krause et al., 2024). In that case, a model including only group-level parameters might already result in a different estimate for M than a model in which parameters are allowed to vary. Ultimately, two-step estimates from this model could thus be vastly different from the subject-level and covariate effects produced by a model with varying parameters.

As such, it would generally be desirable to be able to estimate a model, not unlike a multi-level EDHMM, that permits non-linear covariate effects on model parameters and for these parameters to vary between subjects. Crucially, assuming independence of different time-series given model parameters would be far less problematic for such a model, since the parameters (and hence the latent and emission processes) themselves can vary between subjects and as a function of experimental covariates (e.g., Michelot, 2025). The next section outlines how such a model, which we will refer to as a *multi-level semi-Markov smooth model*, can be formalized.

2.2 Multi-level Semi-Markov Smooth Models

As mentioned briefly at the end of the previous section, we propose to borrow the latent and emission processes of the EDHMM but to model their definining parameters, or a known function of them, as an additive combination of smooth functions of covariates and i.i.d random effects to capture subject-level heterogeneity (cf. Wood et al., 2016). Technically, the class of multi-level semi-Markov smooth models (SMSMs) generalizes beyond the emission and latent processes assumed by an EDHMM to those assumed by any of the other common HsMMs discussed by Yu (2010). Even more generally, the SMSM framework is applicable to other latent variable models, potentially including additional latent variables related to the emission process (e.g., Taghia et al., 2018), requiring only that the gradient of the log-likelihood function can be computed. Here we consider an EDHMM-like SMSM as an example, but much of the theory will apply to SMSM extensions of other models. Additionally, the specific SMSM presented here can be considered a semi-Markov extension of the models discussed by Michelot (2025), who focused largely on conventional HMMs without explicit models of the state sojourn times[2].

To formalize an EDHMM-like SMSM, we first consider the parameters required by emission and state sojourn time distributions. Let μ_j denote any parameter involved in the model of the emission process of state j that is required by emission distributions $\mathcal{F}^o$. For example, μ_j could be a mean of a (multivariate) normal distribution, the scale parameter of a Gamma distribution, or a covariance parameter in the precision matrix of a multivariate normal distribution (i.e., an off-diagonal element in $\mathbf{\Sigma}^{-1}$). In addition, let ν_j denote any parameter required by the duration distribution $\mathcal{F}^d$ of state j. For example, ν_j could be the mean of a Gamma or Poisson distribution, or the scale parameter of an inverse Gaussian distribution.

These and the remaining parameters, the state transition probabilities in $\mathbf{\Delta}$ and initial state probabilities in $\boldsymbol{\pi}$, are considered constant in the conventional EDHMM. In contrast, in a SMSM they are each parameterized by a *linear predictor* η (e.g., Michelot, 2025; Wood et al., 2016). Here we are interested in linear predictors of the form outlined in Equation (1).

$$\eta = f_1(x_1) + f_2(x_{21}, x_{22}) + \ \dots \ + b \tag{1}$$

Here, b is a random coefficient so that $b \sim N(0, \lambda_b^{-1})$ and f_1 and f_2 are smooth functions of covariates x_1 and x_{21}, x_{22} respectively. Functions like f are represented as weighted sums of δ known *basis functions* b so that (e.g., Krause et al., 2025; Wood, 2017; Wood et al., 2016)

$$f(x_1, x_2, \dots) = \sum_l \delta_l(x_1, x_2, \dots)\beta_l.$$

To prevent overfitting, functions like f_1 and f_2 are regularized by placing an (improper) *smoothness* prior on the vector of coefficients $\boldsymbol{\beta}_k$ associated with smooth function f_k, so that $\boldsymbol{\beta}_k \sim N(\mathbf{0}, \mathrm{S}_\lambda^{k-})$ (e.g., Kimeldorf & Wahba, 1970; Wood, 2017; Wood et al., 2016). Here, S_λ^{k-} denotes the pseudo-inverse of the semi-definite precision matrix S_λ^k, which has a known structure up to one or more scaling variables $\lambda \in \boldsymbol{\lambda}_k$ (see Wood, 2017, for a more detailed discussion). These λ parameters are essentially regularization hyper-parameters with their magnitude determining how complex functions like f can be. Thus, they will typically have to be estimated like $\boldsymbol{\beta}_k$ – we discuss this in an upcoming section. Note also, that the prior is improper to allow treating sufficiently simple function estimates (i.e., lower-order polynomials) as fixed (see Wood, 2017).

Representing functions like f_1 and f_2 as sums of weighted basis functions reveals that Equation (1) matches an *additive mixed model* of the form shown in Equation (2) (see also Michelot, 2025).

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \text{ with } \boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}_\lambda^-) \tag{2}$$

Here, $\mathbf{X}$ corresponds to the model matrix of $\boldsymbol{\eta}$ and all coefficients (e.g., $\boldsymbol{\beta}_k, b, \dots$) involved in the model of $\boldsymbol{\eta}$ have been collected in $\boldsymbol{\beta}$. As discussed, $\boldsymbol{\beta}$ is subjected to an improper Gaussian prior, with $\mathbf{S}_\lambda^-$ denoting the pseudo inverse of semi-definite precision matrix $\mathbf{S}_\lambda$ (e.g., Krause et al., 2025; Wood, 2017; Wood et al., 2016). $\mathbf{S}_\lambda = \sum_k \mathbf{S}_\lambda^k$ is obtained by summing over embedded[3] versions $\mathbf{S}_\lambda^k$ of all precision matrices S_λ^k involved in the model of $\boldsymbol{\eta}$, including those associated with standard i.i.d random coefficients (e.g., like b in Equation 1) for which S_λ^k is simply a scaled identity (Wood, 2017).

Based on the definition of linear predictors in Equation (2), we define models of parameters $\boldsymbol{\mu}_j$ and $\boldsymbol{\nu}_j$ as shown in Equation (3).

$$g^{\mu_j}(\boldsymbol{\mu}_{jst}) = \boldsymbol{\eta}_{st}^{\mu_j} \text{ and } g^{\nu_j}(\boldsymbol{\nu}_{jst}) = \boldsymbol{\eta}_{st}^{\nu_j} \tag{3}$$

Here, g are known, optionally state-specific, smooth *link* functions with known inverse so that, for example, $\boldsymbol{\mu}_{jst} = g^{\mu_j}(\boldsymbol{\eta}_{st}^{\mu_j})^{-1}$ (e.g., Faraway, 2016; Wood, 2017). Subscripts indicate the value of a parameter (e.g., $\boldsymbol{\mu}_j$) or predictor (e.g., $\boldsymbol{\eta}^{\mu_j}$) for time-series (i.e., trial) s and at time-point t. Specifically, $\boldsymbol{\eta}_{st}^{\mu_j} = \mathbf{X}_{st}^{\mu_j}\boldsymbol{\beta}^{\mu_j}$, where $\mathbf{X}_{st}^{\mu_j}$ indexes the specific row in the subset of model matrix $\mathbf{X}^{\mu_j}$ associated with parameter μ and state j for time-series s and time-point t. Finally, $\boldsymbol{\beta}^{\mu_j}$ denotes the set of coefficients involved in the model of $\boldsymbol{\eta}^{\mu_j}$.

[2]Notably, Michelot (2025) discussed how to *approximate* explicit duration models with HMMs. In contrast, the algorithms presented here are exact.

[3]So that $\boldsymbol{\beta}^\top \mathbf{S}_\lambda^k \boldsymbol{\beta} = \boldsymbol{\beta}_k^\top \mathrm{S}_\lambda^k \boldsymbol{\beta}_k$, where S_λ^k is any (semi-definite) precision matrix associated with the (improper) prior placed on $\boldsymbol{\beta}_k \subset \boldsymbol{\beta}$.

We now turn to the state transition probabilities $\boldsymbol{\Delta}_{ij}$. As mentioned, these are also expressed in terms of a linear predictor in a SMSM. To achieve this, first note that the off-diagonal probabilities in row i of matrix $\boldsymbol{\Delta}$ define the mass function for a Multinomial distribution of the transitions away from state i to state $j \in \mathcal{S}/i$ (e.g., Michelot, 2025). Thus, we can write

$$\boldsymbol{\Delta}_{i\,j\,s\,t} = \frac{exp(\boldsymbol{\eta}_{s\,t}^{\boldsymbol{\Delta}_{ij}})}{\sum_{j' \in \mathcal{S}/i} exp(\boldsymbol{\eta}_{s\,t}^{\boldsymbol{\Delta}_{ij'}})} \; \forall j \in \mathcal{S}/i. \tag{4}$$

Note, that for every state i only $M-2$ linear predictors $\boldsymbol{\eta}_{s\,t}^{\boldsymbol{\Delta}_{ij}}$ are identifiable, requiring the additional constraint that at least one of those is set to zero. The choice is essentially arbitrary and can also differ for different states i (see also Michelot, 2025). A similar parameterization is chosen for $\boldsymbol{\pi}$. Specifically, the elements in $\boldsymbol{\pi}$ again form the mass function for a Multinomial distribution over the initial state. Hence, we again write:

$$\boldsymbol{\pi}_{j\,s} = \frac{exp(\boldsymbol{\eta}_{s}^{\boldsymbol{\pi}_{j}})}{\sum_{j' \in \mathcal{S}} exp(\boldsymbol{\eta}_{s}^{\boldsymbol{\pi}_{j'}})} \; \forall j \in \mathcal{S}. \tag{5}$$

Again, only $M-1$ linear predictors $\boldsymbol{\eta}_{s}^{\boldsymbol{\pi}_{j}}$ are identifiable, so we enforce the constraint that $\boldsymbol{\eta}_{s}^{\boldsymbol{\pi}_{1}} = 0$. Also, note that the t subscripts have been omitted in this equation, since the initial state distribution cannot change over time.

With models specified for the different parameters as shown above, it is now possible to account for subject-level heterogeneity in the emission and latent processes as well as for non-linear covariate effects. Note, that the models of the parameters outlined above have been kept extremely general – considerable simplifications will often be possible (and even necessary, if the model is to be estimated in a reasonable amount of time). For example, many experiments will only contain covariates varying between time-series but not within (i.e., with time). In that case, the t sub-scripts can be omitted in either (or both) of Equations (3) and (4). Similarly, prior knowledge might be available regarding the structure and elements of $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$. In that case, models could be specified only for the parameters involved in the emission and state duration distributions (e.g., μ_j and ν_j). The next section outlines how the log-likelihood of the specific EDHMM-like SMSM considered here can be computed efficiently.

2.2.1 *The Log-likelihood of the EDHMM-like Semi-Markov Smooth Model*

To facilitate the likelihood definition, let $\boldsymbol{\beta}$ denote the total set of coefficients, required by the models of all parameters. Next, we show how the log-likelihood $\mathcal{L}_s(\boldsymbol{\beta}) = log(p(\mathbf{o}_{s\,[1:T]}|\boldsymbol{\beta}))$ of $\boldsymbol{\beta}$ can be computed given the emission sequence $\mathbf{o}_{s\,[1:T_s]}$ corresponding to an individual time-series s. We use square brackets in sub-scripts to index a sequence of elements for a given dimension (i.e., $[1:T_s]$ indexes time-points $1, ..., T_s$). Additionally, T_s has received a subscript to emphasize that the number of time-points collected can differ between series. Given a way to compute the log-likelihood for individual series, the total log-likelihood is then simply $\mathcal{L} = \sum_s \mathcal{L}_s(\boldsymbol{\beta})$, due to the assumption that different time-series are mutually independent given $\boldsymbol{\beta}$.

As for other Markov models, $\mathcal{L}_s$ is best computed recursively for the model considered here, utilizing the so-called *forward variables* denoted by α (e.g., Rabiner, 1990; Yu, 2010). The forward variables presented here are scaled versions of the ones defined previously by Yu and Kobayashi (2006). The scaling, previously introduced by Lystig and Hughes (2002) in the context of conventional HMMs, drastically facilitates computation of the gradient of the log-likelihood later on. For $t = 1$ the variables are initialized as shown in Equation (6).

$$\begin{aligned} \alpha_{s\,1}(j,d) &\overset{\text{def}}{=} p(\mathbf{o}_{s\,1}, S_{s\,1} = j, \tau_{s\,1} = d|\boldsymbol{\beta}) \\ &= \boldsymbol{\pi}_{j\,s}\, p_j(d)\, b_j(\mathbf{o}_{s\,1}) \end{aligned} \tag{6}$$

Here, τ_t is an auxiliary variable denoting the remaining or residual duration in the current state at time-point t and $\boldsymbol{\pi}_{j\,s}$ is as defined in Equation (5). Additionally, $p_j(d)$ and $b_j(\mathbf{o}_{s\,t})$ are again used as shorthands for $p(\mathbf{o}_{s\,t}|S_{s\,t} = j, \boldsymbol{\beta}_j^o)$ and $p_j(d) \overset{\text{def}}{=} p(d|S = j, \boldsymbol{\beta}_j^d)$ respectively, where $\boldsymbol{\beta}_j^o \subset \boldsymbol{\beta}$ and $\boldsymbol{\beta}_j^d \subset \boldsymbol{\beta}$ denote the sets of coefficients required to parameterize the emission and state sojourn time distributions $\mathcal{F}^o$ and $\mathcal{F}^d$ of state j [4]. Note that conditioning on $\boldsymbol{\beta}_j^o$ and $\boldsymbol{\beta}_j^d$ implicitly defines t and s subscripts for the densities/probabilities $p_j(d)$ and $b_j(\mathbf{o}_{s\,t})$, since the values of parameters required by $\mathcal{F}^o$ and $\mathcal{F}^d$ can change for different series and time-points as shown in Equation (3).

For $T_s \geq t \geq 2$, the forward variables can then be computed recursively as shown in Equation (7).

$$\begin{aligned} \alpha_{s\,t}(j,d) &\overset{\text{def}}{=} p(\mathbf{o}_{s\,t}, S_{s\,t} = j, \tau_{s\,t} = d|\mathbf{o}_{s\,1:t-1}, \boldsymbol{\beta}) \\ &= \Bigg\{ \alpha_{s\,t-1}(j, d+1)\, b_j(\mathbf{o}_{s\,t}) \\ &\quad + \left[\sum_{i \in S/j} \alpha_{s\,t-1}(i,1)\, \boldsymbol{\Delta}_{i\,j\,s\,t-1} \right] b_j(\mathbf{o}_{s\,t})\, p_j(d) \Bigg\} \Lambda_{s\,t-1}^{-1} \end{aligned} \tag{7}$$

Here $\boldsymbol{\Delta}_{i\,j\,s\,t-1}$ is defined as in Equation (4) and

$$\Lambda_{s\,t} \overset{\text{def}}{=} p(\mathbf{o}_{s\,t}|\mathbf{o}_{s\,[1:t-1]}) = \sum_j \sum_d \alpha_{s\,t}(j,d). \tag{8}$$

[4] So $\boldsymbol{\beta}_j^o$ denotes the set of coefficients required by the models of all parameters (e.g., mean, scale, etc.) required to parameterize $\mathcal{F}^o$

Intuitively, the forward variable defined in Equation (7) thus computes the desired probability $p(\mathbf{o}_{st}, S_{st} = j, \tau_{st} = d|\mathbf{o}_{s[1:t-1]}, \boldsymbol{\beta})$ based on the probability of the same state j continuing (i.e., having duration $\tau_{st-1} = d + 1$ at the previous time-point $t - 1$) and the probability of any state $i \in S/j$ ending on the previous time-point (i.e., having remaining duration $\tau_{st-1} = 1$) with the chain then transitioning from state i into j for duration d (Yu & Kobayashi, 2006).

Note, that some forward variables can be set to zero and thus do not have to be computed, in case the possibility of right-censoring can be discounted (e.g., Yu, 2010). Specifically, if there is prior knowledge suggesting that the last state ends at the last time-point $t = T$, we can set $\alpha_{st}(j, d) = 0 \; \forall \; t + d > T_s + 1$. Similarly, if there is prior knowledge suggesting that the last state is j, we can set $\alpha_{sT_s}(i, d) = 0 \; \forall \; d, i \in \mathcal{S}/j$.

Accounting for left-censoring (e.g., Yu, 2010) is more complicated. Specifically, use of $p_j(d)$ in both Equations (6) and (7) implies the assumption that the first state begins at the first time-point $t = 1$ (i.e., that there is no left-censoring). If this is not the case, then the mass associated with duration d at $t = 1$ could be very different than at other time-points, and use of $p_j(d)$ in both Equations will no longer be justified. To this end we propose to generally[5] replace $p_j(d)$ with $p_{\pi j}(d)$ in Equation (6). $p_{\pi j}(d)$, like $p_j(d)$, is parameterized as outlined in Equation (3), but comes with a separate set of coefficients $\boldsymbol{\beta}^d_{\pi j}$. If there is prior evidence suggesting that the possibility of left-censoring can be discounted, we can set $\boldsymbol{\beta}^d_{\pi j} = \boldsymbol{\beta}^d_j$. Otherwise these extra coefficients will have to be estimated (see also Yu, 2010).

Finally, note that the definition of the forward variables given in Equations (6) and (7) also permits emission models for which the density $p(\mathbf{o}_{st}|S_{st} = j)$ cannot easily be computed, but $p(\mathbf{o}_{st}|S_{st} = j, \tau_{st} = d)$ can. As discussed in more detail in Appendix A, this is necessary for example for the emission model of EEG proposed by Anderson et al. (2016).

With the forward variables defined, the log-likelihood of the EDHMM-like latent smooth model for series s can then be computed based on the Λ_{st} variables defined in Equation (8). Specifically, relying on the fact that any likelihood can always be factorized into a product of conditional probabilities (e.g., Lystig & Hughes, 2002),

$$\mathcal{L}_s(\boldsymbol{\beta}) = \sum_t log(\Lambda_t) \text{ and } \mathcal{L}(\boldsymbol{\beta}) = \sum_s \mathcal{L}_s(\boldsymbol{\beta}). \tag{9}$$

Note, that while the computational complexity of the forward variables is considerable, Equations (6) and (7) can be evaluated in parallel for different series s, which can drastically speed up the computation of $\mathcal{L}(\boldsymbol{\beta})$. The forward variables are also useful for a broad range of other tasks. For example, they are required to compute the most likely state sequence, given estimates of $\boldsymbol{\beta}$, for every series via the Viterbi algorithm (e.g., Rabiner, 1990). Similarly, they are required for sampling state sequences given data and estimates of $\boldsymbol{\beta}$ (e.g., Dewar et al., 2012). The next section considers how to obtain such estimates of $\boldsymbol{\beta}$ for a SMSM.

2.2.2 *Estimating Semi-Markov Smooth Models*

The main challenge complicating estimation of SMSMs is that the (inverted) variance/regularization parameters $\boldsymbol{\lambda}$, involved in the priors placed on random effects and smooth functions f in any of the models defined in Equations (3-5), will have to be estimated as well. Here we consider two possible solutions to this problem: an empirical Bayes approach and fully Bayesian inference.

The *empirical approach* is to (approximately) optimize the Bayesian marginal likelihood $p(\mathbf{O}|\boldsymbol{\lambda})$ (e.g., Krause et al., 2025; Wahba, 1985; Wood, 2017; Wood et al., 2016). Here we use $\mathbf{O}$ to denote the concatenated observation sequences from all time-series and recorded signals (i.e., all the data) and $\boldsymbol{\lambda}$ again denotes the set of all variance/regularization parameters required by the model. Let $p(\boldsymbol{\beta}|\boldsymbol{\lambda})$ denote the density implied by the (improper) prior Gaussian prior $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}^-_\lambda)$ placed on the total coefficient vector $\boldsymbol{\beta}$. Note, that $\mathbf{S}^-_\lambda$ now denotes the block-diagonal matrix obtained by stacking the $\mathbf{S}^-_\lambda$ matrices defined for each linear predictor (see Equation 2; and also Krause et al., 2025; Wood et al., 2016, for a more detailed description). Then, the Bayesian marginal likelihood can be computed (approximately) as shown in Equation (10).

$$\begin{aligned} p(\mathbf{O}|\boldsymbol{\lambda}) &= \int p(\mathbf{O}|\boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{\lambda}) d\boldsymbol{\beta} \\ &= \int p(\mathbf{O}, \boldsymbol{\beta}|\boldsymbol{\lambda}) d\boldsymbol{\beta} \\ &\approx p(\mathbf{O}, \hat{\boldsymbol{\beta}}|\boldsymbol{\lambda}) \frac{\sqrt{2\pi}^{N_p}}{|\mathcal{H}|^{0.5}} \end{aligned} \tag{10}$$

Here, $p(\mathbf{O}|\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta})$, N_p is the number of elements in $\boldsymbol{\beta}$, and the approximate result of the third line follows from a Laplace approximation to the conditional posterior $\boldsymbol{\beta}|\mathbf{O}, \boldsymbol{\lambda}$ (see Wood, 2017; Wood et al., 2016, for more detailed discussions). $\hat{\boldsymbol{\beta}}$ is the maximizer of the joint log-likelihood $log(p(\mathbf{O}, \boldsymbol{\beta}|\boldsymbol{\lambda}))$. Similarly, $\mathcal{H}$ is the negative Hessian of the joint log-likelihood, evaluated at said maximizer. Considering that the density of $p(\boldsymbol{\beta}|\boldsymbol{\lambda})$ is a zero-mean multivariate Gaussian, it can readily be shown that $\hat{\boldsymbol{\beta}}$ is obtained by maximizing the *penalized* log-likelihood $\mathcal{L}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T \mathbf{S}_\lambda \boldsymbol{\beta}$ while $\mathcal{H} = \mathbf{H} + \mathbf{S}_\lambda$, where $\mathbf{H} = -\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\Big|_{\hat{\boldsymbol{\beta}}}$ is the negative Hes-

[5] As an exception, this will not be necessary if $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$ are set up in a way that guarantees that the Markov chain never re-enters the first state. Examples are strict left-to-right models for which $\boldsymbol{\pi}_1 = 1$ and $\boldsymbol{\Delta}$ is a supra-diagonal matrix.

sian of the log-likelihood evaluated at $\hat{\boldsymbol{\beta}}$ (e.g., Krause et al., 2025; Wood et al., 2016).

Maximizing $log(p(\mathbf{O}|\boldsymbol{\lambda}))$, or the Laplace approximate version $\mathcal{V}(\boldsymbol{\lambda}) = log(p_L(\mathbf{O}|\boldsymbol{\lambda}))$ given in the third line of Equation (10), thus automatically provides point-estimates $\hat{\boldsymbol{\beta}}$ alongside point-estimates $\hat{\boldsymbol{\lambda}}$ (e.g., Wood, 2017). Additionally, the Laplace approximation provides a normal approximation to the conditional posterior $\boldsymbol{\beta}|\mathbf{O}, \boldsymbol{\lambda} \sim N(\hat{\boldsymbol{\beta}}, \mathcal{H}^{-1})$ (Wood, 2017; Wood et al., 2016). Combined with the asymptotic approximation $\boldsymbol{\rho} \sim N(\hat{\boldsymbol{\rho}}, \mathbf{V}_\rho)$, where $\boldsymbol{\rho} = log(\boldsymbol{\lambda})$ and $\mathbf{V}_\rho = \left(\frac{\partial^2 \mathcal{V}(exp(\boldsymbol{\rho}))}{\partial \boldsymbol{\rho} \partial \boldsymbol{\rho}^\top}\Big|_{\hat{\boldsymbol{\rho}}}\right)^{-}$, this enables approximate inference about $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ (Krause et al., 2025; Wood, 2017; Wood et al., 2016).

Different approaches have been proposed to maximize $\mathcal{V}(\boldsymbol{\lambda})$ numerically. Wood et al. (2016) outline how exact gradients and the hessian of $\mathcal{V}(\boldsymbol{\lambda})$ can be computed. Both, $\mathcal{L}_\lambda(\boldsymbol{\beta})$ and $\mathcal{V}(\boldsymbol{\lambda})$ can then be maximized via Newton's method. However, this approach generally requires up to fourth order derivatives of the log-likelihood $\mathcal{L}$. In contrast, the Extended Fellner Schall (EFS) method described by Wood and Fasiolo (2017) only requires the gradient of the log-likelihood and $\mathbf{H}$ (i.e., up to second-order derivatives) to optimize $\mathcal{L}_\lambda(\boldsymbol{\beta})$ via Newton's method and to approximately maximize $\mathcal{V}(\boldsymbol{\lambda})$.

Michelot (2025) show that automatic differentiation of the computer code of the log-likelihood can be used to obtain the gradient and $\mathbf{H}$, again enabling optimization of $\mathcal{L}_\lambda(\boldsymbol{\beta})$ and computation of $\mathcal{V}(\boldsymbol{\lambda})$. The latter can then again be maximized based on the EFS method or via quasi-Newton methods relying on numerical approximations to the derivatives of $\mathcal{V}(\boldsymbol{\lambda})$ (see Michelot, 2025; Nocedal & Wright, 2006). Any computations involving the negative Hessian $\mathbf{H}$ will however become expensive for complex models involving many random effects. To address this problem, Krause et al. (2025) recently showed that $\mathbf{H}$ in the EFS update can also be replaced with a limited-memory symmetric rank one approximation, computed from accumulated gradients of the log-likelihood and modified to remain semi-definite (see also Nocedal & Wright, 2006). The resulting limited-memory quasi-EFS (L-qEFS) update thus avoids explicitly forming $\mathbf{H}$, and only requires the gradient of $\mathcal{L}$ to approximately optimize $\mathcal{L}_\lambda(\boldsymbol{\beta})$, via quasi-Newton methods, and $\mathcal{V}(\boldsymbol{\lambda})$ via the modified EFS update.

While the gradient of the log-likelihood $\mathcal{L}(\boldsymbol{\beta})$ of a SMSM could be derived via automatic differentiation (e.g., Michelot, 2025), an analytic solution can be computed recursively for the specific EDHMM-like model considered here, just like the forward variables defined in Equations (6) and (7). The analytic solution was previously derived by Lystig and Hughes (2002) for conventional HMMs – here we generalize their approach to latent smooth models. First, Lystig and Hughes (2002) point out that

$$\begin{aligned}\partial \mathcal{L}_s / \partial \boldsymbol{\beta} &= \frac{\partial log(P(\mathbf{o}_{s\,[1:T_s]}|\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = \frac{\partial P(\mathbf{o}_{s\,[1:T_s]}|\boldsymbol{\beta})\partial \boldsymbol{\beta}}{P(\mathbf{o}_{s\,[1:T_s]}|\boldsymbol{\beta})} \\ &= \left(\frac{\partial P(\mathbf{o}_{s\,[1:T_s]}|\boldsymbol{\beta})\partial \boldsymbol{\beta}}{P(\mathbf{o}_{s\,[1:T_s-1]}|\boldsymbol{\beta})}\right)\left(\frac{P(\mathbf{o}_{s\,[1:T_s]}|\boldsymbol{\beta})}{P(\mathbf{o}_{s\,[1:T_s-1]}|\boldsymbol{\beta})}\right)^{-1}.\end{aligned} \tag{11}$$

Based on the definitions given in Equation (8),

$$\frac{P(\mathbf{o}_{s\,[1:T_s]}|\boldsymbol{\beta})}{P(\mathbf{o}_{s\,[1:T_s-1]}|\boldsymbol{\beta})} = \Lambda_{s\,T_s}.$$

Crucially, Lystig and Hughes (2002) then point out that the remaining quantity in Equation (11), $\frac{\partial P(\mathbf{o}_{s\,[1:T_s]}|\boldsymbol{\beta})\partial \boldsymbol{\beta}}{P(\mathbf{o}_{s\,[1:T_s-1]}|\boldsymbol{\beta})}$, can be computed based on the partial derivatives of the forward variables with respect to $\boldsymbol{\beta}$. To achieve this for the specific definitions of the forward variables shown in Equations (6) and (6), we first initialize ψ variables as shown in Equation (12).

$$\begin{aligned}\psi_{s\,1}(j,d) &\stackrel{\text{def}}{=} \partial p(\mathbf{o}_{s\,1}, S_{s\,1} = j, \tau_{s\,1} = d|\boldsymbol{\beta})\partial \beta_r \\ &= \frac{\partial \boldsymbol{\pi}_{j\,s}}{\partial \beta_r}\, p_j(d)\, b_j(\mathbf{o}_{s\,1}) + \boldsymbol{\pi}_{j\,s}\, \frac{\partial p_j(d)}{\partial \beta_r}\, b_j(\mathbf{o}_{s\,1}) \\ &\quad + \boldsymbol{\pi}_{j\,s}\, p_j(d)\, \frac{\partial b_j(\mathbf{o}_{s\,1})}{\partial \beta_r}\end{aligned} \tag{12}$$

Here β_r can be any coefficient in $\boldsymbol{\beta}$. Note, that although it is possible for different linear predictors to share some (or all) coefficients (see also Wood et al., 2016), it will usually not be the case that both emission densities and state duration probabilities depend on the same coefficient β_r. Hence, it will typically not be necessary to evaluate all three sums in Equation (12) for any given β_r. Instead, it will commonly be the case that for any given β_r all but one of the sums evaluate to zero. Note, that similar simplifications will typically be possible when computing the general definition of the recursive update to ψ for $T_s \geq t \geq 2$, shown in Equation (13). Generally,

$$\begin{aligned}\psi_{s\,t}(j,d) &\stackrel{\text{def}}{=} \frac{\partial P(\mathbf{o}_{s\,[1:t]}, S_{s\,t} = j, \tau_{s\,t} = d|\boldsymbol{\beta})\partial \beta_r}{P(\mathbf{o}_{s\,[1:t-1]}|\boldsymbol{\beta})} \\ &= \Bigg\{\psi_{t-1}(j,d+1)\, b_j(\mathbf{o}_{s\,t}) + \alpha_{s\,t-1}(j,d+1)\, \frac{\partial b_j(\mathbf{o}_{s\,t})}{\partial \beta_r} \\ &\quad + \frac{\partial \left(\sum_{i \in S/j} \alpha_{s\,t-1}(i,1)\, \boldsymbol{\Delta}_{i\,j\,s\,t-1}\right)}{\partial \beta_r}\, b_j(\mathbf{o}_{s\,t})\, p_j(d) \\ &\quad + \left(\sum_{i \in S/j} \alpha_{s\,t-1}(i,1)\, \boldsymbol{\Delta}_{i\,j\,s\,t-1}\right) \frac{\partial b_j(\mathbf{o}_{s\,t})}{\partial \beta_r}\, p_j(d) \\ &\quad + \left(\sum_{i \in S/j} \alpha_{s\,t-1}(i,1)\, \boldsymbol{\Delta}_{i\,j\,s\,t-1}\right) b_j(\mathbf{o}_{s\,t})\, \frac{\partial p_j(d)}{\partial \beta_r}\Bigg\}\, \Lambda_{t-1}^{-1},\end{aligned} \tag{13}$$

where

$$\frac{\partial\left(\sum_{i\in S/j}\alpha_{s\,t-1}(i,1)\,\mathbf{\Delta}_{i\,j\,s\,t-1}\right)}{\partial\beta_r} = \sum_{i\in S/j}\left(\psi_{s\,t-1}(i,1)\,\mathbf{\Delta}_{i\,j\,s\,t-1} + \alpha_{s\,t-1}(i,1)\frac{\partial\mathbf{\Delta}_{i\,j\,s\,t-1}}{\partial\beta_r}\right). \tag{14}$$

Based on the definitions in Equations (12) and (13) it is clear that

$$\frac{\partial P(\mathbf{o}_{s\,[1:T_s]}|\boldsymbol{\beta})\partial\boldsymbol{\beta}}{P(\mathbf{o}_{s\,[1:T_s-1]}|\boldsymbol{\beta})} = \sum_{j\in S}\sum_{d}\psi_{s\,T_s}(j,d)$$

and thus,

$$\partial\mathcal{L}_s\partial\boldsymbol{\beta} = \frac{\sum_{j\in S}\sum_{d}\psi_{s\,T_s}(j,d)}{\Lambda_{s\,T_s}}.$$

Finally, relying again on the mutual independence of different time-series given $\boldsymbol{\beta}$,

$$\partial\mathcal{L}\partial\boldsymbol{\beta} = \sum_{s}\partial\mathcal{L}_s\partial\boldsymbol{\beta}. \tag{15}$$

Computing the partial derivatives in Equations (12), (13), and (14) requires partial derivatives of the different densities/probabilities parameterized through different linear predictors. As discussed by Wood et al. (2016), these can typically be evaluated by means of transformation. Appendix B outlines how this can be achieved for the specific derivatives required here. Additionally, note that Equations (12) and (13) can again be computed in parallel for different series s, just like the forward variables. Similarly, it is not actually necessary to store $\psi_{s\,t}(j,d)$ for all time-points t. As becomes evident from inspecting Equations (12) and (13), computing ψ for any given time-point t only requires access to the ψ variables from the previous time-point $t-1$. Thus, only the ψ variables of the previous time-point $t-1$ need to be kept in memory when evaluating the forward recursion.

The information provided so far is sufficient to obtain point-estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}$ for SMSMs via the L-qEFS update by Krause et al. (2025). The empirical approach, to optimize the Laplace approximate criterion $\mathcal{V}(\boldsymbol{\lambda})$, will likely work well for moderately complex models involving a modest number of smooth functions f and random effects. For more complex models, opting for fully Bayesian inference might be preferable instead. For once, the Laplace approximation assumes that, in the vicinity of $\hat{\boldsymbol{\beta}}$, the log-likelihood is approximately quadratic (i.e., that the third and fourth order derivatives are negligible compared to the elements in $\mathbf{H}$; see Wood et al., 2016). Additionally, it is necessary that the maximizer of the penalized log-likelihood, $\hat{\boldsymbol{\beta}}$, can reliably be obtained for any $\boldsymbol{\lambda}$ – otherwise the Laplace approximation will not hold and updates to $\boldsymbol{\lambda}$ might simply fail. Unfortunately, the surface defined by $\mathcal{L}(\boldsymbol{\beta})$ will typically be all but smooth for complex SMSMs and while $\mathcal{L}_\lambda$ will typically be slightly easier to navigate, estimation of $\hat{\boldsymbol{\beta}}$ might still get stuck in a local maximum (see Michelot, 2025, for similar discussions).

Theoretically, it is possible to mitigate this risk by restarting the quasi-Newton optimization routine for $\hat{\boldsymbol{\beta}}$, described in more detail by Krause et al. (2025), multiple times from different starting values. Alternatively, a stochastic global optimization algorithm (e.g., simulated annealing; Kirkpatrick et al., 1983) could be used before (or alternated with) the quasi-Newton routine to optimize $\hat{\boldsymbol{\beta}}$. However, for more complex models the associated computational overhead will often be comparable to fully Bayesian inference as discussed in the upcoming paragraphs.

Fully Bayesian inference requires the ability to sample from the posterior $p(\boldsymbol{\beta},\boldsymbol{\lambda}|\mathbf{O})$ (see Gelman et al., 2013, for a detailed introduction to Bayesian inference). Since $\partial\mathcal{L}\partial\boldsymbol{\beta}$ is available, a No-U-Turn sampler could be used to complete this task efficiently (NUTS; Hoffman & Gelman, 2014). For the NUTS sampler to work, we require the gradient of the joint log-likelihood

$$log(p(\mathbf{O},\boldsymbol{\beta},\boldsymbol{\lambda})) = log(p(\mathbf{O}|\boldsymbol{\beta})) + log(p(\boldsymbol{\beta}|\boldsymbol{\lambda})) + log(p(\boldsymbol{\lambda}))$$

with respect to $[\boldsymbol{\beta},\boldsymbol{\lambda}]^\top$. Here $p(\boldsymbol{\lambda})$ denotes an additional prior distribution placed on $\boldsymbol{\lambda}$ or $\boldsymbol{\rho}$ – common choices are wide Gamma and bounded Uniform priors respectively (e.g., Wood, 2020). The partial derivatives of $log(p(\mathbf{O},\boldsymbol{\beta},\boldsymbol{\lambda}))$ with respect to $\boldsymbol{\beta}$ are simply the elements in the gradient $\partial\mathcal{L}_\lambda(\boldsymbol{\beta})\partial\boldsymbol{\beta}$ of the penalized log-likelihood $\mathcal{L}_\lambda(\boldsymbol{\beta})$, which can be computed given $\partial\mathcal{L}\partial\boldsymbol{\beta}$ (see for example Wood et al., 2016, for the derivations). Similarly, the partial derivatives of $log(p(\mathbf{O},\boldsymbol{\beta},\boldsymbol{\lambda}))$ with respect to $\boldsymbol{\lambda}$ are the elements in the gradient of $log(p(\boldsymbol{\beta}|\boldsymbol{\lambda})) + log(p(\boldsymbol{\lambda}))$ with respect to $\boldsymbol{\lambda}$. The latter is straightforward to compute, considering that the density of $p(\boldsymbol{\beta}|\boldsymbol{\lambda})$ is multivariate normal and the aforementioned choice of standard prior distributions (see Wood, 2017; Wood et al., 2016).

Given these quantities, the NUTS sampler can be used to produce samples $\boldsymbol{\beta}^*$, and $\boldsymbol{\lambda}^*$ from $p(\boldsymbol{\beta},\boldsymbol{\lambda}|\mathbf{O})$. These can be used to derive credible intervals, maximum a posterior (MAP) estimates $\hat{\boldsymbol{\beta}}^*$, and $\hat{\boldsymbol{\lambda}}^*$, and posterior mean estimates for all quantities involved in the model (i.e., for the coefficients but also individual parameters; see Gelman et al., 2013; Michelot, 2025, for more detailed introductions). For very large models, the requirement to compute $\partial\mathcal{L}\partial\boldsymbol{\beta}$ repeatedly in order to generate samples from the posterior might however still become prohibitive.

For very complex models, the alternate sampling strategy proposed by Dewar et al. (2012) will thus be more appropriate. The authors show how to sample from the posterior distribution $p(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{Z},\mathbf{D}|\mathbf{O})$ by alternating between sampling $\mathbf{Z}^*,\mathbf{D}^*$ from $p(\mathbf{Z},\mathbf{D}|\mathbf{O},\boldsymbol{\beta})$ followed by sampling $\boldsymbol{\beta}^*,\boldsymbol{\lambda}^*$ given

$\mathbf{Z}^*, \mathbf{D}^*$ from $p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{O}, \mathbf{Z}, \mathbf{D})$. Here $\mathbf{Z}$ and $\mathbf{D}$ are random variables corresponding to the concatenated state sequences and durations spent in all states across all time-series respectively (e.g., Dewar et al., 2012). This approach thus also automatically produces samples from the posterior of state sequences and state durations, which can be useful for further analysis.

Sampling $\mathbf{Z}^*$ and $\mathbf{D}^*$ for the specific EDHMM-like model considered here only requires computing the forward variables. We refer to Dewar et al. (2012) for a discussion of this step and instead focus on the problem of sampling $\boldsymbol{\beta}^*, \boldsymbol{\lambda}^*$ given $\mathbf{Z}^*, \mathbf{D}^*$. Conveniently, assuming that no coefficients are shared between different linear predictors, this sampling step can be performed separately for subsets of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. For example, let $\boldsymbol{\beta}_j^o \subset \boldsymbol{\beta}$ again denotes the subset of coefficients involved in the emission model of state j (i.e., required by $\mathcal{F}^o$) and let $\boldsymbol{\lambda}_j^o \subset \boldsymbol{\lambda}$ denote the subset of regularization parameters involved in any priors place on a subset of $\boldsymbol{\beta}_j^o$. Samples $\boldsymbol{\beta}_j^{o*}, \boldsymbol{\lambda}_j^{o*}$ can be drawn from $p(\boldsymbol{\beta}_j^o, \boldsymbol{\lambda}_j^o|\mathbf{O}, \mathbf{Z}^*, \mathbf{D}^*)$. Conveniently, if $\mathcal{F}^o$ is a member of the exponential family of distributions, the corresponding joint likelihood $p(\boldsymbol{\beta}_j^o, \boldsymbol{\lambda}_j^o, \mathbf{O}|\mathbf{Z}^*, \mathbf{D}^*)$ coincides with the one of a Generalized Additive Mixed Model (GAMM) of Location, Scale, and Shape (GAMMLSS, if linear predictors are specified for multiple parameters of $\mathcal{F}^o$; Rigby & Stasinopoulos, 2005; Wood, 2017; Wood et al., 2016) of the subset of emissions $\mathbf{O}$ that $\mathbf{Z}^*$ indicates have been generated by state j (i.e., the rows in $\mathbf{O}$ for which $\mathbf{Z}^*$ contains j) – with any prior on $\boldsymbol{\lambda}_j^o$ carrying over.

Setting up a NUTS or Gibbs sampler for such a GAMM(LSS) is possible, so that samples $\boldsymbol{\beta}_j^{o*}, \boldsymbol{\lambda}_j^{o*}$ can be generated very efficiently (see Wood, 2016, 2020, for examples). This strategy also generalizes to subsets of coefficients involved in the duration distribution of a given state – the joint likelihood here is of a GAMMLSS of the durations of state j in $\mathbf{D}^*$ – and those involved in either the state transition or initial state probabilities – the joint likelihood of coefficients involved in the models of $\boldsymbol{\Delta}_{ij} \forall\, j \in \mathcal{S}/i$ for example, is of a Multinomial GAMMLSS of all transitions away from state i in $\mathbf{Z}^*$.

An even more efficient alternative is available if we are content with approximate inference about $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. Specifically, using again the example of $\boldsymbol{\beta}_j^o$ and $\boldsymbol{\lambda}_j^o$, it is also possible to generate samples $\boldsymbol{\beta}_j^{o*}, \boldsymbol{\lambda}_j^{o*}$ from the large sample Gaussian approximations $\boldsymbol{\beta}_j^o|\mathbf{O}, \boldsymbol{\lambda}_j^o, \mathbf{Z}, \mathbf{D} \sim N(\hat{\boldsymbol{\beta}}_j^o, (\mathcal{H}_j^o)^{-1})$ and $\boldsymbol{\rho}_j^o|\mathbf{O}, \mathbf{Z}, \mathbf{D} \sim N(\hat{\boldsymbol{\rho}}_j^o, \mathbf{V}_{\rho\, j}^o)$, discussed earlier in the context of the Laplace approximation (see also Wood et al., 2016). Here, $\hat{\boldsymbol{\beta}}_j^o$ and $\mathcal{H}_j^o$ are the maximizer and negative Hessian of the joint log-likelihood $log(p(\boldsymbol{\beta}_j^o, \mathbf{O}|\boldsymbol{\lambda}_j^o, \mathbf{Z}, \mathbf{D}))$. Similarly, $\hat{\boldsymbol{\rho}}_j^o = log(\hat{\boldsymbol{\lambda}}_j^o)$ with $\hat{\boldsymbol{\lambda}}_j^o$ being the maximizer of the corresponding Laplace approximate marginal log-likelihood $\mathcal{V}(\boldsymbol{\lambda}_j^o)$. Finally, $\mathbf{V}_{\rho\, j}^o$ again corresponds to the pseudo-inverse of the negative Hessian of $\mathcal{V}(\boldsymbol{\lambda}_j^o)$ evaluated with respect to $\boldsymbol{\rho}_j^o$ at $\hat{\boldsymbol{\rho}}_j^o$ (e.g., Wood et al., 2016). Obtaining these quantities given samples $\mathbf{Z}^*, \mathbf{D}^*$ again only requires estimating a GAMM(LSS) of the subset of emissions $\mathbf{O}$, generated by state j according to $\mathbf{Z}^*$. Then, $\boldsymbol{\lambda}_j^{o*}$ can be sampled from $N(\hat{\boldsymbol{\rho}}_j^o, \mathbf{V}_{\rho\, j}^o)$ in a first step and $\boldsymbol{\beta}_j^{o*}$ can be sampled from $N(\hat{\boldsymbol{\beta}}_j^o, (\mathcal{H}_j^o)^{-1})$ where $\hat{\boldsymbol{\beta}}_j^o$ is the maximizer of $log(p(\boldsymbol{\beta}_j^o, \mathbf{O}|\boldsymbol{\lambda}_j^{o*}, \mathbf{Z}^*, \mathbf{D}^*))$ (see Wood, 2017).

This concludes our discussion of the two options, empirical Bayes and fully Bayesian inference, to estimate latent smooth models with a log-likelihood as defined in Equation (9) in the previous section. In section 3 we use the Laplace-approximate approach of the previous paragraph to sample from a multi-level SMSM of the lexical decision data collected previously by Krause et al. (2024). The Supplementary GitHub repository for this paper (see section 3) includes additional examples on how the different estimation strategies presented in this section can be achieved with the mssm Python toolbox (Krause et al., 2025). A critical step of each analysis involving such models, will be to validate whether the model assumptions are reasonable given recorded data. The next section outlines how to achieve this.

2.2.3 Validating Semi-Markov Smooth Models

When working with Generalized Linear Mixed Models (GLMMs) or GAMM(LSS), *model validation* typically involves inspecting the model residuals. Informally, the latter can be considered as the difference between the observed data and the model prediction (e.g., Faraway, 2016; Wood, 2015, 2017). Intuitively, this difference should correspond to random noise if the true data-generating model would be used to generate the prediction. Thus, estimated models producing model residuals that are in clear violation of the expected randomness can be rejected as possible explanations of the data, because they could not possibly have generated it (e.g., Wood, 2015, 2017).

While residuals are commonly used as a tool to validate conventional regression models such as GLMMs or GAMMLSS (see Faraway, 2016; Wood, 2017, for examples), this approach has not yet been adopted widely by researchers using Markov models to recover cognitive processes. This might in part be due to the fact that many different definitions of residuals for HMMs and HsMMs can be found in the literature (e.g., Buckby et al., 2020; Zucchini et al., 2017). Additionally, it is not always clear how these different residual definitions generalize to a multivariate emission process (cf. Kalliovirta, 2008). Thus, we briefly discuss two types of (multivariate) residuals that are applicable to standard EDHMMs and SMSMs (see also Dunn & Smyth, 1996; Kalliovirta, 2008; Langrock et al., 2012; Zucchini et al., 2017). The results section includes practical examples, showing how these residuals can be used to detect problems with these models.

Most studies involving HMMs report some type of pseudo- or quantile-residual (e.g., Buckby et al., 2020; Dunn

& Smyth, 1996; Kalliovirta, 2008; Langrock et al., 2012). The specific pseudo-residual considered here, which has also been referred to as *forward* pseudo-residual (Langrock et al., 2012, e.g.,), is defined formally in Equation (16) (see also Dunn & Smyth, 1996). To remain applicable to multivariate emissions, the residual defined in Equation (16) assumes that the K signals can be treated as mutually independent, given $\boldsymbol{\beta}$ (e.g., Langrock et al., 2012). As discussed, this assumption is typically justifiable for SMSMs: assuming sufficient random effects are included in the emission models, it is often reasonable to assume that each signal has been generated by a univariate distribution (see also Michelot, 2025). If a multivariate normal distribution is assumed for the emissions instead, then the matrix $N*K$ matrix $\mathbf{O}$ first needs to be transformed, based on the estimated covariance matrix for state j, so that the K columns can again be expected to be mutually independent if the model is correct (see Appendix A).

$$\boldsymbol{\epsilon}_{s\,t\,k} \stackrel{\text{def}}{=} \Phi^{-1}(F(\mathbf{o}_{s\,t\,k}|\mathbf{o}_{s\,[1:t-1]}, \boldsymbol{\beta})) \tag{16}$$

Here Φ^{-1} denotes the quantile function of a standard normal and $F(\mathbf{o}_{s\,t\,k}|\mathbf{o}_{s\,[1:t-1]}, \boldsymbol{\beta})$ is the cumulative probability of observation $\mathbf{o}_{s\,t\,k}$ at time-point t for signal k and series s, given the history of all previous observation vectors $\mathbf{o}_{s\,[1:t-1]}$ for series s. For the specific EDHMM-like SMSM considered here, $F(\mathbf{o}_{s\,t\,k}|\mathbf{o}_{s\,[1:t-1]}, \boldsymbol{\beta})$ can be computed using the forward variables, requiring only a minor modification of the definitions provided in Equations (6) and (7) (see Zucchini et al., 2017).

Additionally, if the model is specified correctly $F(\mathbf{o}_{s\,t\,k}|\mathbf{o}_{s\,[1:t-1]}, \boldsymbol{\beta})$ should be distributed uniformly, so that transforming via Φ^{-1} produces a standard normal variable (see for example Dunn & Smyth, 1996). If the model is specified correctly, the content of each of the K vectors $\boldsymbol{\epsilon}_k$ should thus look approximately like i.i.d samples from $N(0, 1)$. Each vector $\boldsymbol{\epsilon}^k$ can thus be inspected like the residual vector of any conventional GLMM and GAMM (see Faraway, 2016; Wood, 2015, 2017, for practical residual plots). A limitation of the residual is that it is typically less sensitive to miss-specified models of the parameters involved in the state sojourn time distributions (see the Supplementary GitHub repository for this paper for a simulation of this tendency).

To address this problem, it will generally be helpful to also consider residuals which utilize an estimate of the latent state sequence $\mathbf{Z}^*$. This includes so-called Viterbi-path residuals, computation of which involves Viterbi decoding to obtain $\mathbf{Z}^*$ (e.g., Buckby et al., 2020). Given $\mathbf{Z}^*$, one can then for example compute deviance or quantile residuals – with the computation depending on the choice for $\mathcal{F}^o$ – of the observations indicated to be in state $j \in \mathcal{S}$ (e.g., Dunn & Smyth, 1996; Faraway, 2016; Wood, 2017). Naturally, Viterbi-path residuals are limited in that they are conditioned on the particular choice for $\mathbf{Z}^*$(e.g., Buckby et al., 2020). However, any residual plots can be generated multiple times for different samples $\mathbf{Z}^*$ obtained either from $p(\mathbf{Z}|\mathbf{O}, \boldsymbol{\beta}, \boldsymbol{\lambda})$ or $p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{Z}|\mathbf{O})$.

Given $\mathbf{Z}^*$ and $\mathbf{D}^*$, again denoting an estimate of the duration of all states across all series, residuals can also be defined specifically for properties of the latent process. For example, let $\mathbf{D}^*_{j\,s\,t}$ denote the estimated duration of state j which started at time-point t for series s. Based on the definition of the latent process model considered here, we can expect that $\mathbf{D}^*_{j\,s\,t} \sim F^d(\boldsymbol{\nu}_{j\,s\,t}, ...)$ – if the model of the state's sojourn time is correctly specified. This again enables computation of deviance or quantile residuals for the decoded durations of state j. Sampling different $\mathbf{Z}^*$ can again help to alleviate the risk of drawing wrong conclusions based on a particular choice for $\mathbf{Z}^*$ (e.g., Buckby et al., 2020). Consider for example strict left-to-right models: for these models, each state is visited exactly once for every time-series so that only a single residual $\boldsymbol{\epsilon}^d_{j\,s}$ needs to be computed per time-series for state j. An average duration residual can then be computed based on samples $\mathbf{D}^*$, which again should approximately be distributed like a standard normal.

In the upcoming Results section we rely on both types of residuals to improve on the model of lexical decision-making proposed originally by Krause et al. (2024).

3 Results

In this section we perform a re-analysis of the dataset[6] collected in the lexical decision experiment presented previously by Krause et al. (2024). Originally, this dataset was analyzed using a standard group-level HsMM approach, here we apply the new framework to be able to account for subject and covariate effects. For a detailed discussion of the experimental setup and covariates used, we refer to the original paper. Here we summarize the aspects most relevant for the re-analysis. The experiment originally featured three different types of stimuli (words, pseudo-words, and random strings) and a continuous measure of each stimulus' word-likeness, based on Google search result counts. While pseudo-words and random strings are, by definition, non-words, this does not imply that they are all equally word-like (or rather unlike): some pseudo-words might look like miss-spellings of common words while others might feel entirely unfamiliar. Google search result counts act as a continuous proxy measure of these differences in word-likeness and, in contrast to conventional word-frequency measures, can easily be obtained for both words and non-words (see Hendrix & Sun, 2020; Krause et al., 2024, for a more detailed discussion of this measure).

[6]The pre-processed data is openly available at https://github.com/JoKra1/HSMM_LD_EEG. Code to replicate the analysis and all figures is included in the Supplementary GitHub repository for this paper, available at https://github.com/JoKra1/SMSM_supplementary.

Pre-processing involved low-pass (40 Hz) and high-pass (0.5 Hz) filtering, visual artifact rejection, an independent component analysis to remove eye-related artifacts, and interpolation of missing channels. Afterwards, data was available from 10,948 unique trials (5,402 word, 2,668 pseudoword, and 2,878 random string trials) collected from 24 subjects. The EEG data were then down-sampled to 100 Hz, base-line corrected by subtracting the mean amplitude in the 100 ms preceding stimulus onset, and epoched so that individual time-series last from stimulus to response onset. Krause et al. (2024) then performed a principal component analysis on the different electrodes and retained the 10 components contributing the most variance (> 90% variance explained), which were normalized per participant. These pre-processing steps are typical for analyses involving the EDHMM model architecture proposed by Anderson et al. (2016), which Krause et al. (2024) used for their analysis as well (see Berberyan et al., 2021; Portoles et al., 2022; van Maanen et al., 2021, for other examples).

Following their approach, the current EDHMM assumes that cognitive processes proceed sequentially from the first to the last process, without the possibility to skip a state. That is, the model is a strict left-to-right EDHMM with a supra-diagonal state transition matrix $\mathbf{\Delta}$ and $\boldsymbol{\pi}_1 = 1$. Thus, the model only requires estimating parameters involved in the state sojourn time distributions and emission models.

The emission model Krause et al. (2024) used was originally proposed by Anderson et al. (2016), and assumes that the EEG process can be characterized sufficiently by *flat* states – periods of variable duration during which the EEG electrodes (or PCA components) have zero mean amplitude and are mutually independent – and *bump* states – microstates of fixed duration during which the EEG is dominated by a stable topography visible across electrodes (or components) (see also Weindel et al., 2024). A more detailed description is provided in Appendix A, alongside the necessary information to compute the density for observed PCA scores during flat and bump states. Notably, state sojourn time distributions only need to be defined for flat states, since bump states are of fixed duration. Krause et al. (2024) used Gamma distributions with state-specific mean parameters but fixed scale[7] parameters of 0.5 (cf. Anderson et al., 2016; Borst & Anderson, 2015).

Given the number M of flat states, estimating the model thus only requires estimating M scale parameters and $K * (M - 1)$ weights[8] μ_{jk}, corresponding to the expected value of component k during bump state j. Krause et al. (2024) used leave-one-subject-out cross-validation to determine the optimal number M, which revealed that a model with six flat-states $M = 6$ generalized best to held-out data from all word types (see also Berberyan et al., 2021; Mahé et al., 2015). Notably, they also found that the duration of the penultimate flat state, assumed to reflect the actual decision process, and multiple topographies differed between word types.

3.1 Replicating the Six Flat-State Group-level Model

In a first step, we re-estimated their six flat-state model. Because Krause et al. (2024) revealed that the topographies preceding flat states four and five differed between word-types, we define the μ_{jk} for bump states three and four as shown in Equation 17.

$$\mu_{jk} = \alpha_{jk} + \beta_{1\,jk}\, i_{RS}(s) + \beta_{2\,jk}\, i_{PW}(s) \text{ if } j \in \{6, 8\} \quad (17)$$

Here α_{jk} is a state- and PCA component-specific intercept. $i_{RS}(s)$ and $i_{PW}(s)$ are indicator variables that evaluate to 1, if the stimulus of series s was a random string and pseudoword respectively, else 0. $\beta_{1\,jk}$ and $\beta_{2\,jk}$ thus correspond to state-, signal- and word-type-specific intercept adjustments, allowing for the topographies of bump states 3 and 4 (*overall* states $j = 6$ and $j = 8$) to vary between word-types. For the remaining bump states, Krause et al. (2024) used the same topography for all word-types. The μ_{jk} for bump states one, two and five are thus defined only in terms of a state-, and signal-specific intercept α_{jk} as shown in Equation 18.

$$\mu_{jk} = \alpha_{jk} \text{ if } j \in \{2, 4, 10\} \quad (18)$$

The models for the mean parameters ν_j of the state sojourn time distributions are similar, but specified for $log(\nu_j)$ rather than ν_j directly. Krause et al. (2024) showed that the expected duration of flat states four and five varied between word-types. The models of the (log of the) mean parameters ν_j for these flat states thus also included state- and word-type-specific intercept adjustments, while the models for the remaining states again included only a state-specific intercept. Note, that the μ_{jk} (and ν_j) lack time-series s and time-point t subscripts, because the original model by Krause et al. (2024) only includes group-level effects. Later in this section we present a more appropriate hierarchical model alternative.

To estimate the resulting model, the quasi-Newton approach utilized by the L-qEFS update (e.g., Krause et al., 2025) was used to maximize the log-likelihood as defined in the Methods section. As expected, the resulting model produces estimates virtually indistinguishable from those obtained by Krause et al. (2024), which are shown in Figure 2.

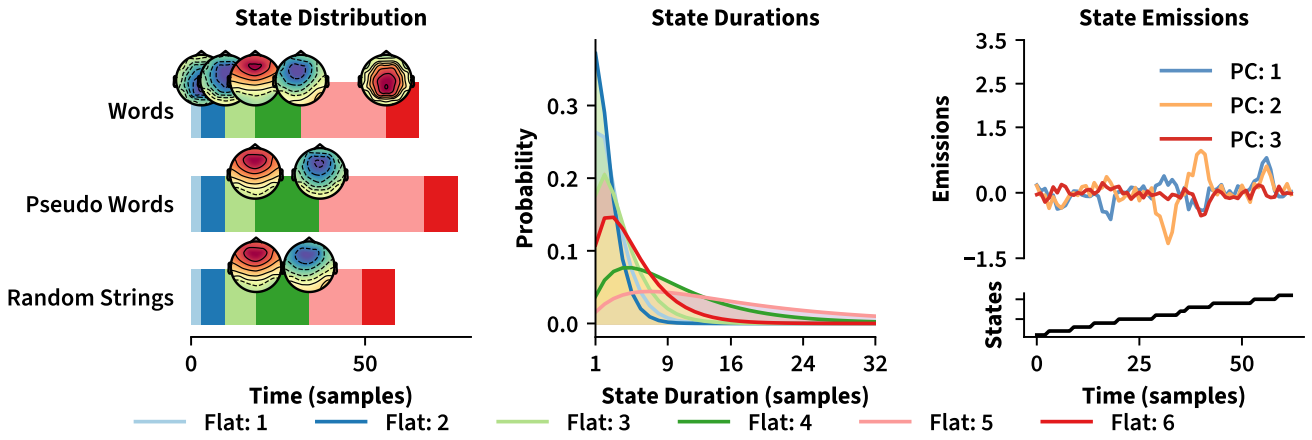
Notably, four different topographies can be distinguished for words in the left-most panel of Figure 2: two topographies featuring central-posterior negativity, a third topogra-

[7]Corresponding to an alpha parameter of $\alpha = 2$ if the Gamma distribution were to be defined in terms of an alpha and scale parameter.

[8]There are always $M - 1$ bump states and M flat states in the EDHMM defined by Anderson et al. (2016), since a topography is assumed to precede the onset of each process but the very first.

Figure 2

Overview Six Flat-State Model



Note. Figure 2 shows an overview of the re-estimation of the six flat-state model of lexical decisions proposed originally by Krause et al. (2024). The left-most panel shows the average duration of each flat state separately for every word-type. These were estimated from the Viterbi-decoded state sequences. The plot also contains the estimated bump topographies, preceding the onset of each flat state. The central panel shows the average (over word-types) probability mass function for the duration of each flat state – note how later flat states are generally associated with wider distributions, in particular the penultimate flat state. The right-most panel visualizes the emission and latent process assumed by the model: during a flat state the signal is noisy with an expected amplitude of zero. Entering bump states results in a weighted half-sine being added to the noisy process. The peaks of the weighted half-sines from all electrodes form the topographies shown in the left-most panel.

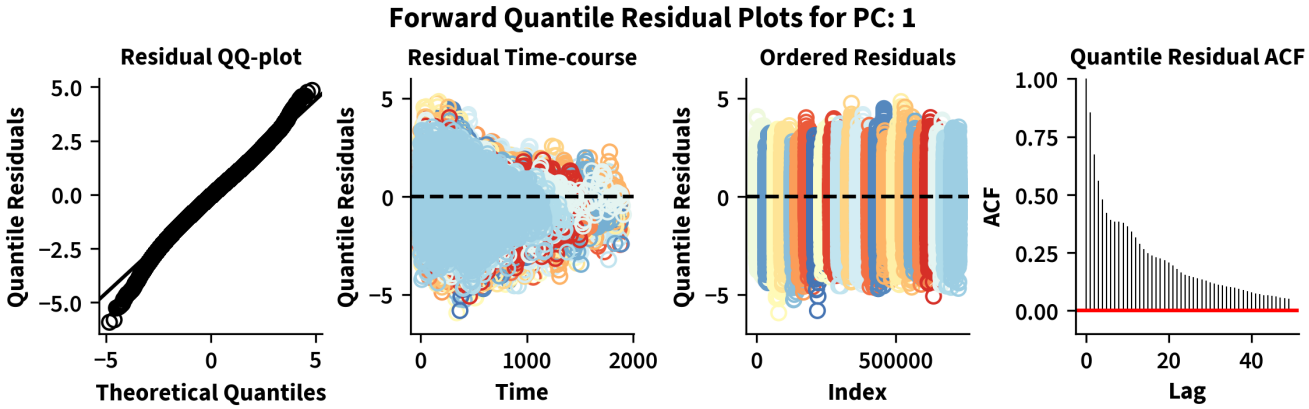
phy resembling the mid/frontal old-new (N170) ERP component, a fourth topography featuring central-anterior negativity, and a final, fifth, topography featuring parietal-posterior positivity (see the top bar of the bar-plot). Previous work suggests similar patterns to be related to the onset of visual processing, orthographic familiarity assessment, decision-making, and response initialization respectively. The duration of the fifth flat generally dominated the overall reaction times for all word types (see central panel of Figure 2 for average duration mass functions). However, the duration of this flat state also varied strongly between word types (see the left-most panel of Figure 2), a pattern that previous models suggest is typical for (lexical) decision-making (e.g., Ratcliff et al., 2004; Wagenmakers et al., 2008).

In a second step we computed forward pseudo-residuals for this six stage model, as defined in the methods section. As discussed in section 2.2.3, pseudo-residuals are typically computed for each of the K multivariate signals (PCA components in this case). Because it is difficult to show plots for all 10 PCA components, we only show the plots for the first component, with the understanding that the plots for the other components show similar patterns unless otherwise noted. Additionally, the Supplementary GitHub repository for this paper contains code to create the plots for all components. Figure 3 shows quantile-quantile (QQ-plot), time-course, and ordered residual plots as well as an estimate of the auto-correlation function of the residuals for different time-lags. While the residuals appear to be approximately normally distributed marginally (see the QQ-plot; other PCA components show stronger deviations), the variance of the residuals appears to change over time (second plot from the left) and to differ between subjects (third plot from the left). Similarly, the residuals continue to show strong temporal correlations, even for distant time-lags (right-most plot). As such, there appear to be remaining dependencies in the data that the model cannot account for.

A possible explanation for the remaining temporal correlation of the residuals is the comparatively simple emission model used here (see the right-most panel of Figure 2 for a visualization of the model's prediction for the emissions for the first three PCA components): instead of assuming that emission deviations from the expected value (zero during flat states, topography-dependent during bump state) are i.i.d realizations of a standard normal, it might be more appropriate to assume that these deviations are temporally correlated. The strong temporal correlations visible in Figure 3 persist over multiple lags, which suggests that a multi-lag auto-correlation model of the deviations might be required. However, because such a model would be very expensive to compute, we considered an auto-correlation model of lag 1 (i.e., an ar-1 model) instead – the details of this model are provided in Appendix A. Based on a visual inspection of the

Figure 3

Residuals of Six Flat-State Model



Note. Figure 3 shows different plots of the forward pseudo-residuals computed for the six flat-state model of lexical decisions shown in Figure 2. The left-most panel plots the observed residual quantiles against the theoretical (standard normal) quantiles to assess deviations from normality. The second plot from the left plots the residuals against time, with residuals from different subjects being colored differently. Note, that the variance in the residuals appears to decrease over time. The third plot from the left shows the residuals in the order of the data, again colored by subject. Note, that there appears to be a difference in the variance of the residuals between subjects. The final plot shows the estimated auto-correlation function of the residuals for different time-lags. Evidently, residuals of this model remain highly correlated.

auto-correlation function in Figure 3, we chose $\rho = 0.7$ for the value of the auto-correlation weight.

Figure 4 shows the updated residuals for the six flat-state model including an ar-1 model of the emission deviations. The auto-correlation function has improved substantially, although weak temporal correlations between shorter time-lags remain visible (see right-most panel of Figure 4). Residuals also appear to deviate slightly more from marginal normality, and their variance continues to decrease with time and to differ between different subjects. This might in part be a result of an overly simplistic model of the state sojourn time distributions: In a follow-up analysis Krause et al. (2024) showed the expected duration of all flat states to vary as a word-type-specific function of word-likeness. Additionally, the expected duration of flat states varied smoothly over trials – and did so differently for different subjects. The current model of the state sojourn times, a Gamma distribution with state-specific mean parameter, cannot account for either of these patterns.

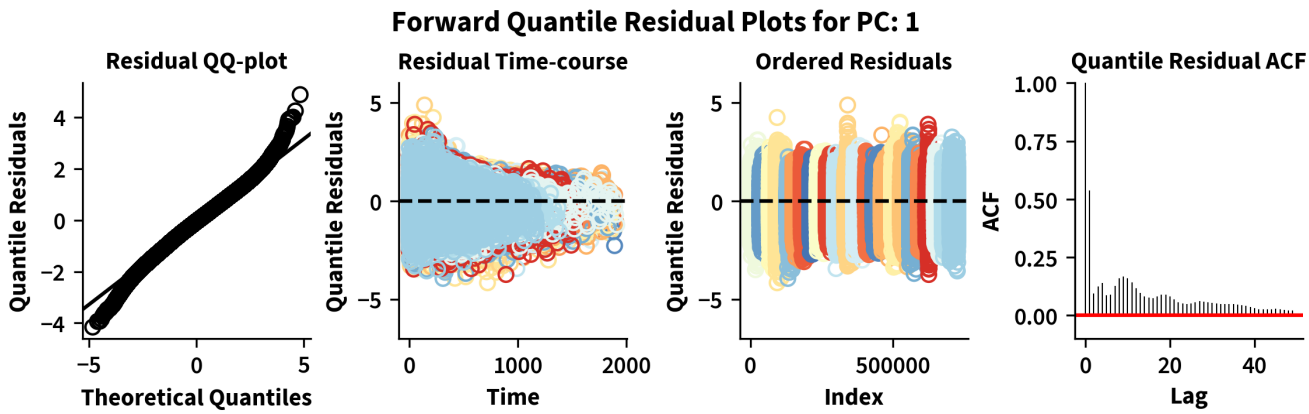
To test for the correctness of the model of the state sojourn times, we computed average residuals of the duration of each flat state for every time-series, as discussed in the final part of section 2.2.3. The average duration residuals were computed based on 100 samples obtained from the posterior of state sequences given observed data and maximum likelihood estimates of all model parameters. Different visualizations of these residuals are depicted in Figure 5 for flat states 3-6 (flat states 1 & 2 showed similar patterns as flat state 3 and were thus omitted from the Figure to improve legibility).

The first row of Figure 5 visualize estimates of the state- and word-type specific effects of word-likeness on the expected duration residual. Similarly, the second row shows subject-specific changes in the expected residual over the time-course of the experiment. These were obtained by fitting additive mixed models (e.g., Wood, 2017) to the state-specific average duration residuals. If the model of the state sojourn times were correct, the expected value of the duration residuals should be zero and constant, i.e., insensitive to changes in word-likeness or time spent in the experiment. Instead, the clear deviations from this pattern, visible in the top rows of Figure 5, suggest that the expected duration of the different flat-states is not constant, but rather an additive combination of state-, subject-, and word-type-specific smooth non-linear functions of word-likeness and the experimental time-course. The fact that the state-specific duration residuals deviate from normality and show strong temporal correlations across multiple trials further suggests that the current model of the sojourn time distributions is too simple (see last two rows in Figure 5).

As mentioned, the fact that the current model of the state sojourn times does not account for any subject-specific differences, could partially explain why the variance of the pseudo-residuals, shown in Figure 4, appears to vary between subjects. It is more likely however, that the emission process model also needs to account for subject-specific differences. The current model assumes that the state topographies preceding flat states are the same for all subjects (see Equations 17 and 18). Even if all subjects complete the same sequence

Figure 4

Residuals of Six Flat-State Model Including an ar-1 Model of the Emissions



Note. Like Figure 3, Figure 4 shows different plots of the forward pseudo-residuals, but computed for the the six flat-state model including an ar-1 model of the emission deviations. Notably, while the residuals of this model appear to deviate slightly more from marginal normality (see left-most plot) temporal correlations between residuals are substantially weaker (see right-most plot). The variance of the residuals still appears to decrease over time and to differ between subjects however (see central plots).

of processes however, the topographies marking the onset of a new process could very well be different for different subjects. In fact, it might not even be reasonable to assume that all subjects complete the same sequence of processes (on all trials): even for a simple task as lexical decision-making, different subjects might employ slightly different strategies.

3.2 A Hierarchical Six Flat-State Model

To account for these possibilities, a multi-level or hierarchical latent semi-Markov smooth model is required with additive mixed models for the μ and ν parameters. For the μ parameters we keep the group-level structure the same as before and only add subject-specific random effects. Therefore we again need to define different effect structures for bump states three and four on the one hand and bump states one, two, and six on the other. The updated models are shown in Equations 19 and 20.

$$\mu_{jks} = \alpha_{jk} + b_{p(s)\,jk} \text{ if } j \in \{2, 4, 10\} \text{ and,} \tag{19}$$

$$\begin{aligned}\mu_{jks} =& \alpha_{jk} + \beta_{1\,jk}\, i_{RS}(s) + \\ & \beta_{2\,jk}\, i_{PW}(s) + b_{p(s)\,jk} \text{ if } j \in \{6, 8\}.\end{aligned} \tag{20}$$

Here, $b_{p(s)\,jk}$ denotes a state- and PCA component-specific random intercept (adjustment) for subject $p(s)$ from which series s was recorded. The inclusion of the $b_{p(s)\,jk}$ allows for different topographies to be estimated for each subject. Additionally, we again need to define mixed models for the (log of the) mean of the state sojourn time distribution $\nu_{j\,s}$, which, as shown in Equation (21), can now differ for series s and flat state j.

$$log(\nu_{j\,s}) = f_{w(s)\,j}(wl(s)) + f_{p(s)\,j}(s) \tag{21}$$

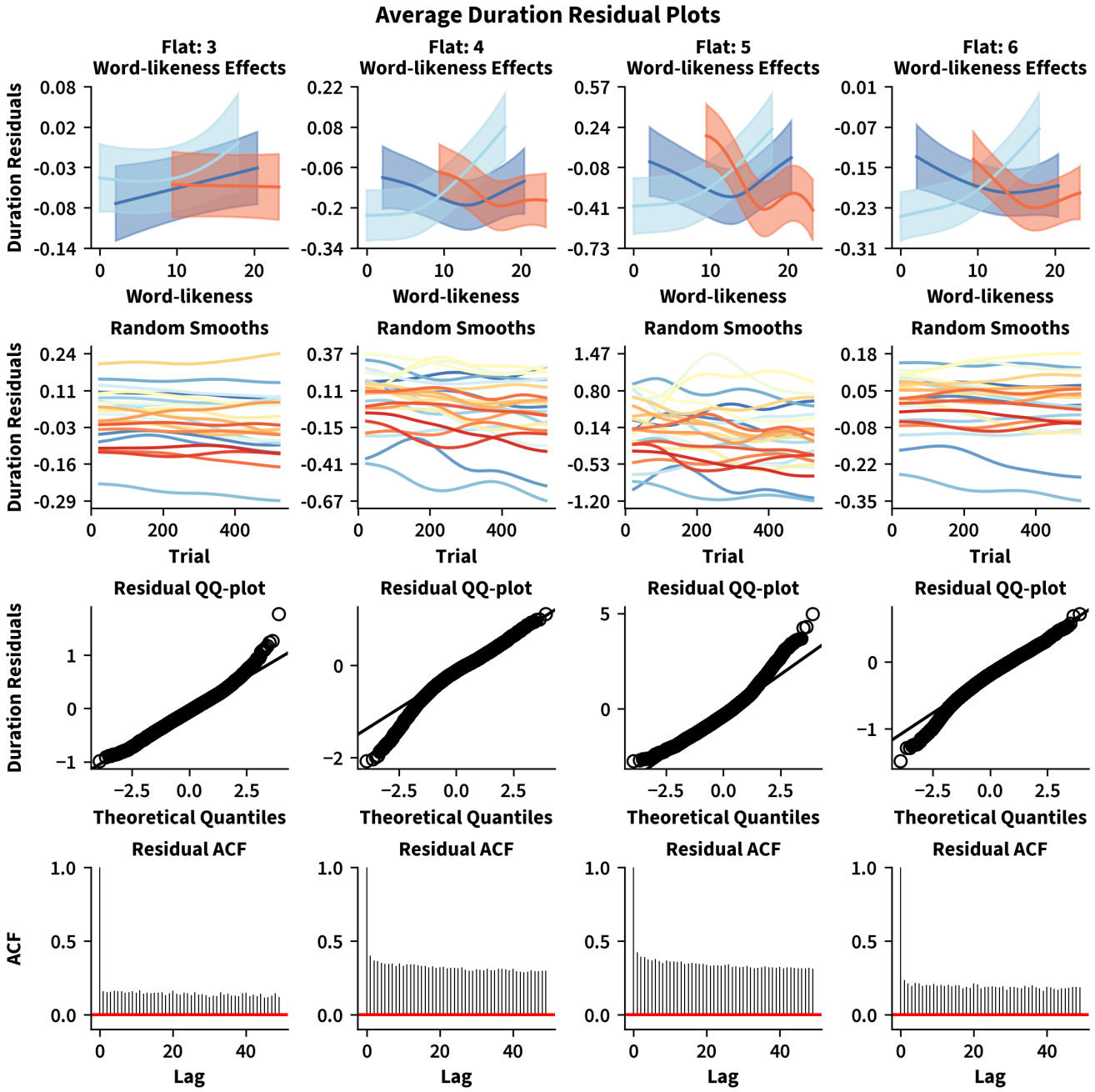
Here, $wl(s)$ defines the word-likeness of the stimulus present for series s. Correspondingly, $f_{w(s)}(wl(s))$ denotes a state- and word-type ($w(s)$ returns the word-type of the stimulus present for series s) specific smooth function (including an offset) of the word-likeness $wl(s)$ of the stimulus present for series s. Finally, $f_{p(s)\,j}(s)$ is a state- and subject-specific *random smooth function* (i.e., subjected to a proper full-rank prior) of the experimental time-course (i.e., s; see Wood, 2017).

The resulting latent smooth model is quite complex, requiring estimation of 2910 coefficients $\boldsymbol{\beta}$ and 80 inverted variance/regularization parameters $\boldsymbol{\lambda}$. Therefore, we relied on the Laplace-approximate Bayesian estimation approach discussed in the methods section, alternating between sampling state and state sojourn time sequences $\mathbf{Z}^*, \mathbf{D}^*$ and using normal approximations to sample $\boldsymbol{\lambda}|\mathbf{O}, \mathbf{Z}^*, \mathbf{D}^*$ and $\boldsymbol{\beta}|\mathbf{O}, \boldsymbol{\lambda}, \mathbf{Z}^*, \mathbf{D}^*$ (see also Dewar et al., 2012).

Two chains were sampled for 1250 iterations, with the first half being subsequently discarded and only the second half of each chain being used for analysis. Trace plots suggested that the marginal posteriors for some random intercepts (mostly those estimated for subjects 1 and 18) and λ values might be multi-modal, with both chains converging to different but stationary modes. To test for the possibility of additional modes, both chains were re-initialized based on averages taken over different sampling windows of the original chains and sam-

Figure 5

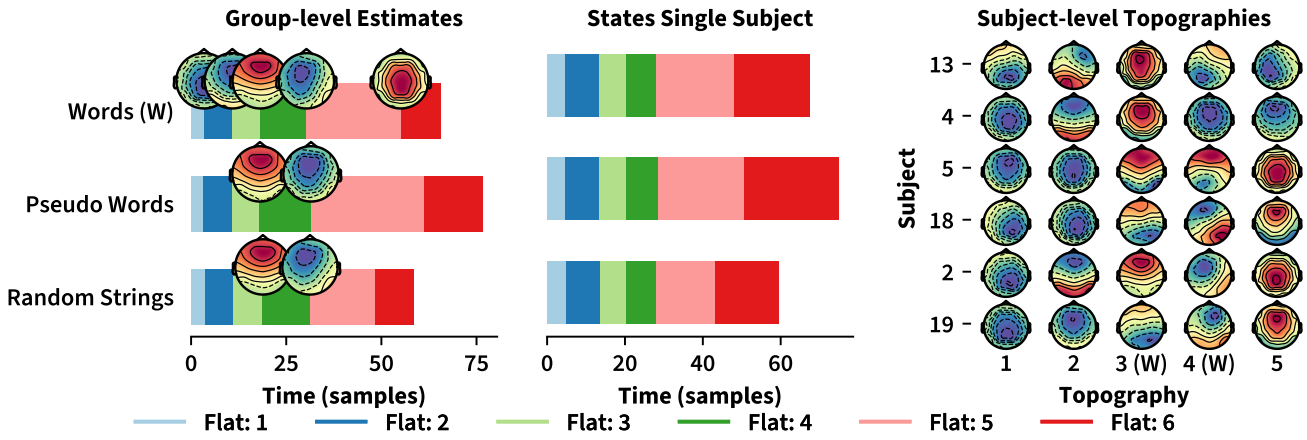
Average Flat-State Duration Residuals



Note. Figure 5 shows plots of the average duration residuals for all flat states, computed for the the six flat-state model including an ar-1 model of the emission deviations. The first row shows, separately for each word-type, the estimated effect of word-likeness on the expected residual. As observed previously by Krause et al. (2024), later flat states show more complex effects of word-likeness (note the y-axis differences). The second row shows estimates of the subject-specific change in the expected residual over trials. The deviations also appear to be greater for later flat states – in particular the penultimate flat state. The last two rows depict QQ-plots and estimates of the auto-correlation function of the residuals: the average duration residuals appear to deviate slightly from normality and generally show strong temporal correlations across different lags (i.e., trials). This suggests that expected flat durations vary between subjects and over the time-course of the experiment.

Figure 6

Six Flat-State Multi-level Semi-Markov Smooth Model



Note. Figure 6 shows an overview of the six flat-state multi-level semi-Markov smooth model of lexical decisions. The left-most plot matches the left-most plot in Figure 2 and shows the estimated duration of each flat state, as well as the estimated topographies preceding these states. Importantly, the topography and state duration estimates were obtained using only the posterior mean estimates for the *group-level coefficients* (i.e., α_{jk}, $\beta_{1\,jk}$, $\beta_{2\,jk}$, and the coefficients involved in the $f_{w(s)\,j}(wl(s))$) in the models of μ and ν defined in Equations (19-21). In contrast, the central plot shows the state duration estimate obtained for a particular subject, based on the posterior mean estimates for *all* coefficients (i.e., without zeroing random coefficients). Similarly, the right-most plot shows a matrix of the posterior mean estimates of the subject-specific bump topographies for words ("W") and a subset of select subjects (see Figure C1 in Appendix C for a version including all subjects). Note, that while some subjects show bump topographies similar to the group-level estimates in the left-most panel, other subjects show bump topographies that differ systematically from the group-level estimates – which is particularly evident for the bump topography preceding the fifth flat (i.e., topography 4, here only shown for words).

pling was continued for 50 additional iterations: for coefficients showing multi-modality, both chains virtually immediately jumped back to one of the previous modes. For the remaining coefficients, visual inspection of the trace plots suggested that the original chains converged approximately to the same mode. However, $\hat{R}$ remained (slightly) above 1.1 for some of these coefficients, suggesting that sampling would have benefited from additional iterations to really establish approximate convergence (see Gelman et al., 2013, for an introduction to convergence statistics and a more detailed discussion).

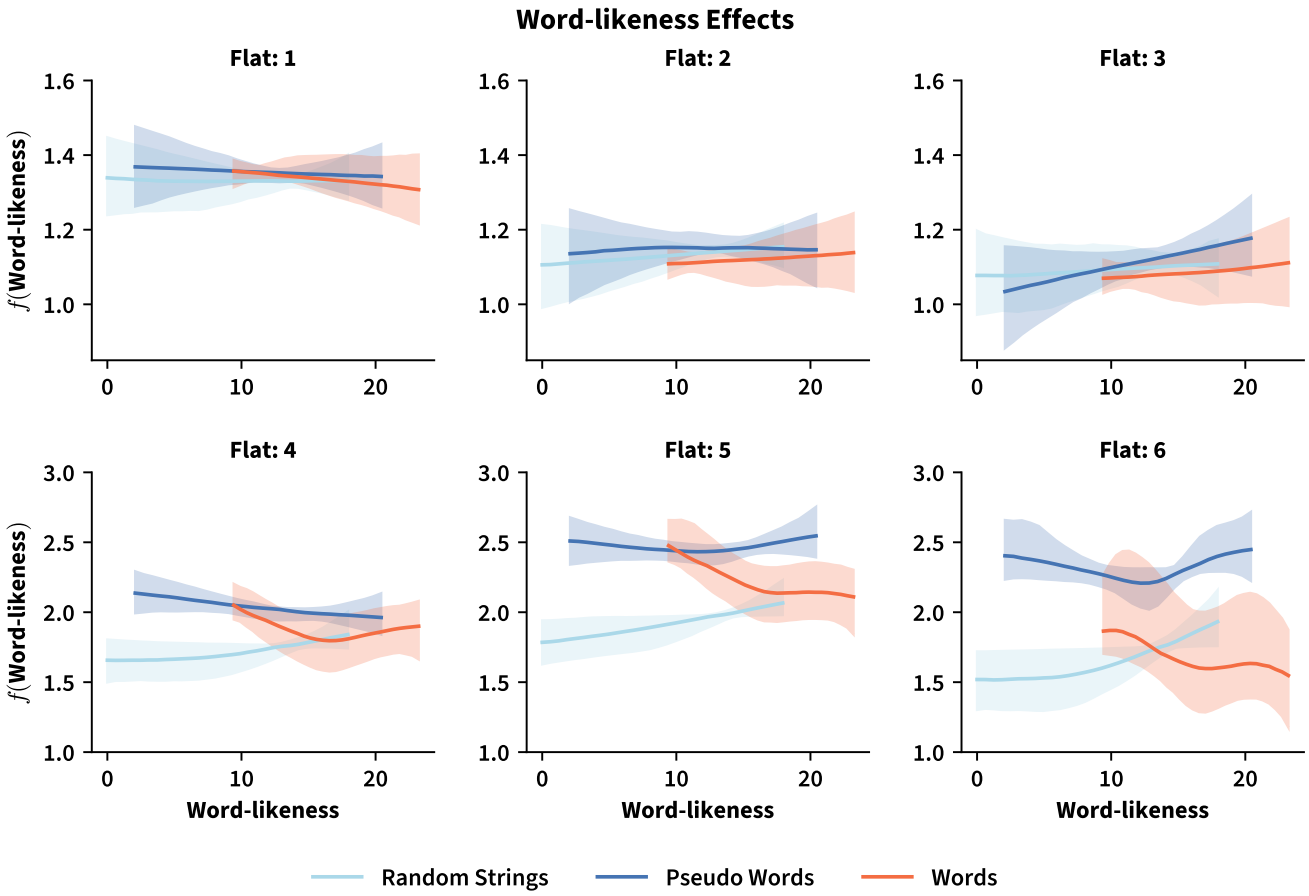
Figure 6 shows an overview for this hierarchical six flat-state latent smooth model. The left-most panel shows posterior means of the group-level bump topographies and flat state duration estimates. While the group-level estimates of the bump topographies are comparable to those obtained with the simpler model (visualized in Figure 2), the estimated durations of the flat states show greater differences compared to the simpler model. Similarly, for some participants the subject-level estimates of flat state durations (see central panel in Figure 6) and topographies alike (see right-most panel in Figure 6 and Figure C1 in Appendix C1) deviate substantially from the group-level estimates obtained from the simpler model.

The differences caution against over-interpreting decoding estimates from a model including only group-level effects: in such a model the decoding is forced to determine state estimates based on topographies/emission model parameters that might not actually be optimal for a particular subject. As a result, decoding estimates from such a model might exaggerate differences between states, suggesting that distinct states might reflect distinct processes.

In contrast, the subject-level variations in the estimated topographies suggest that flat-states four and five might actually reflect a mixture of partially overlapping processes or even the evolution of a single (decision) process over time, rather than distinct processes: for a sub-group of subjects (e.g., subjects 5 and, to a lesser degree, 13 in Figure 6 as well as subjects 3, 8, 15, 21, 23, and 24 in Figure C1 included in Appendix C), the subject-level estimates (also based on posterior means) of the fourth topography actually show strong positivity, rather than negativity. These estimates thus remain qualitatively more similar to the group-level estimate of the third topography, again resembling the familiarity-related N170 ERP component (e.g., Krause et al., 2024). A possible explanation: topographies three and four are snapshots of an evolving decision process, with some some subjects (e.g., 5 and 13 in Figure 6) relying more than others on evidence ac-

Figure 7

Word-likeness Estimates from Multi-level Semi-Markov Smooth Model



Note. Figure 7 shows an overview of the posterior mean estimates of the smooth functions $f_{w(s)\,j}(wl(s))$ involved in the models of the mean parameter ν_j of the state sojourn times. Each panel corresponds to a different flat state, and visualizes the estimated smooth for all three word-types. Shaded areas are 95% posterior credible intervals for the shape of the word-likeness smooth.

cumulated early during this process. This could suggest that a sub-group of participants might have employed a different strategy than the rest of the group, to reach the final decision.

Further evidence for the aforementioned hypothesis stems from the posterior mean estimates of the $f_{w(s)\,j}(wl(s))$, the state- and word-type-specific smooth functions of word-likeness included in the models of the expected duration of state sojourn times. The corresponding estimates are visualized in Figure 7, and differ slightly from the estimates obtained in a follow-up analysis by Krause et al. (2024). Specifically, the posterior mean estimates computed here mainly show differences between the first three and the last three flat states. As observed by Krause et al. (2024), the effect of word-likeness is generally more attenuated in these first three flat states. However, in contrast to their previous findings there appear to be no differences between word-types (see Appendix D for credible intervals of the difference between word-types) in these first three flat states. Additionally, the estimated effects of word-likeness (and word-type) in the last flat states are generally more similar across states – particularly for flat states five and six. This also explains why both flat states five and six now show similar main effects of word-type on the average duration (see pink and red bars in left-most panel in Figure 6). While Krause et al. (2024) also observed similar effects of word-likeness in flat states five and six, they generally observed more pronounced differences between flat states – particularly between flat states four and five.

A possible explanation for these differences is the fact that the results for the multi-level model are based on posterior mean estimates – the MAP estimates might produce estimates more comparable to those reported by Krause et al.

(2024). However, the fact that word-likeness effects are more similar in the last three flat states also supports the aforementioned hypothesis that some or all of these flats might actually reflect a continuously evolving process, rather than distinct cognitive processes.

More work will be necessary to establish this, since even the complex multi-level model presented here neglects an important source of subject-level differences: Specifically, for some subjects (e.g., 4 and 13 in Figure 6 as well as 1, 7, 12, and 14 in Figure C1 included in Appendix C), the final topography, preceding the final flat state, continues to show strong negativity and thus remains more similar to the word-type-specific group-level estimates of the fourth topographies preceding the fifth flat. This observation aligns with the findings by Krause et al. (2024), who showed that an additional sub-state might be visible between flat states five and six. The authors suggested that this sub-state might only be necessary for pseudo-words, reflecting a more complicated decision process for these types of non-words. The analysis performed here suggests, that this extra sub-state might also be present only for a subset of subjects. This could imply that this sub-group of subjects dedicated additional processing time to the decision process or that these subjects relied on additional information, not considered by other subjects.

Importantly, the multi-level model presented here did not account for this sub-state. As a result, parts of the duration of the fifth flat state might have been (incorrectly) attributed to the duration of flat states four and/or six. Specifically, we would expect that, when including an extra topography for selected subjects for which the final topography currently shows strong negativity, their estimates of the final topography would again become more similar to the current group-level estimate showing parietal positivity (see left-most panel in Figure 6). In consequence, the estimated effects of word-likeness on the duration of the different flat states might changes as well and potentially become more similar again to the effects observed by Krause et al. (2024).

Additionally, while the average residuals of flat state durations have improved considerably for this multi-level model (see Appendix E), the forward pseudo-residuals improved only slightly (see Appendix E), with some subjects continuing to show greater variance in the residuals. Exploring an extra sub-state (i.e., topography plus flat state) for the aforementioned individuals might be sufficient to address this, but it might also be necessary to consider a more complex random-effect structure for the models of the μ in general.

To summarize, the results presented here generally support previous findings, which suggested that five prototypical EEG topographies precede the response in lexical decision tasks (e.g., Krause et al., 2024). Rather than reflecting distinct processes, the topographies preceding flat states four and five might reflect the continuous evolution of a single decision process over time. Additionally, even for a simple lexical decision task there appear to be systematic strategic differences between subjects: some appear to rely more strongly, or even exclusively, on evidence available early during the decision process, while others seem to dedicate extra processing time to reach an accurate decision. Extensions of the model presented here remain necessary, but are outside the scope of the current paper, as the main goal was to illustrate the new framework.

While the modeling framework presented here is sufficiently general to accommodate even complex future additions, such as subject-specific extra states, practical application is complicated by the considerable computing resources required to estimate these models. This and other open challenges will be discussed in more detail in the upcoming section.

4 Discussion

In this paper we discussed hierarchical semi-Markov smooth models (SMSMs) of latent neural states, and how to estimate them. SMSMs improve on existing methods to discover cognitive processes from neuro-physiological recordings because i.i.d random effects and smooth functions of covariates can be included in the models of individual parameters. This allows SMSMs to account for subject-level heterogeneity and potentially non-linear covariate effects in the latent and emission process.

To demonstrate the usefulness of SMSMs, we re-analyzed the lexical decision data presented previously by Krause et al. (2024). The primary objective of this analysis was to give an impression of the kind of research questions that can be investigated with these models; "Do all subjects complete the same processes in the same order?" "Do the expected duration of processes or state transition probabilities change smoothly over the time-course of an experiment or as a function of experimental covariates?" – because they can incorporate random and smooth effects, SMSMs can help answer these and related questions.

Additionally, we wanted to demonstrate that existing HsMM models, even very specific ones like the left-to-right EDHMM presented by Anderson et al. (2016) and used by Krause et al. (2024), can quickly be transformed into SMSMs. Conceptually, the SMSM framework presented here is not a replacement for a specific method or type of model used. Instead, the framework presented here can be applied to extend any kind of HsMM (or HMM; see Michelot, 2025) for which the log-likelihood and gradient can be computed. As such, the framework is agnostic to the specific choice of emission density and whether or not initial state and state transition probabilities are to be estimated or not. Moreover, SMSMs can be applied to univariate and multivariate emission processes alike: cross-temporal dependencies between multivariate signals can either be accounted for by choosing an emission model explicitly accounting for

such dependencies (e.g., Vidaurre et al., 2016, 2018) or by including sufficient random effects in the emission models so that the signals can reasonably be treated as realizations from univariate densities (see Michelot, 2025, for similar arguments).

Because the focus was on introducing the SMSM framework, the analysis presented here involved some shortcuts: the six flat-states solution determined by Krause et al. (2024) was used as a starting point and the number of states was subsequently kept fixed. Despite these limitations, which we discuss in more detail in a later part of this section, the analysis performed here provided important theoretical insights, further highlighting the usefulness of the SMSM framework.

First of all, the analysis revealed substantial between-subject differences for some of the *EEG topographies* – essentially the emission process parameters of the specific EDHMM used here (see right-most panel of Figure 6). Previously, Anderson et al. (2016) argued that these topographies reflect the onset of a distinct "processing stage". However, many of the subject-level estimates of the word-type specific topographies preceding the fifth flat state deviated considerably from the group-level estimates featuring strong central negativity (see left-most panel of Figure 6). For some subjects, the estimates could better be described as a mixture of the topographies preceding the fourth and fifth flat state (see subjects 3, 5, 8, 13, 15, 21, 23, and 24 in the right-most panel of Figure 6 and Figure C1 in Appendix C). For other subjects, the estimated topographies preceding the fifth flat state showed no sign of central negativity at all. As discussed in the Results section, the topographies preceding flats four and five could thus also reflect a time-continuous emission process of a single latent process rather than two distinct processing stages.

The emission process model by Anderson et al. (2016) assumes that the EEG returns to an average amplitude of zero after briefly resembling one of the state-specific topographies, and thus does not account for such a continuous emission process. Instead it might be more appropriate to merge flat states four and five and to model the emissions of this merged state with a multivariate Gaussian with a mean vector varying smoothly as a function of the duration of the state. Theoretically, this only requires minor modifications of the likelihood function defined in the Methods section. Yet another alternative, that does not require any modification of the log-likelihood defined here, would be to rely on a multi-variate auto-correlation model for the emissions of this merged state (e.g., Vidaurre et al., 2016).

Additionally, the analysis revealed that including an extra topography and flat state for some subjects might considerably improve the fit for these subjects (see subjects 1, 4, 7, 12, 13, and 14 Figure 6 and Figure C1 in Appendix C). Specifically, the analysis revealed that, for some subjects, the subject-level estimate of the final topography was less similar to the group-level estimate of the final topography, featuring strong parietal positivity. Instead, the subject-level estimates were more similar to the group-level estimate of the earlier topography, preceding the fifth flat and featuring strong central negativity. This ability, to accommodate subject-specific deviations from the group-level state trajectory, is a major advantage SMSMs have over pure group-level models.

As discussed briefly in the Results section, we hypothesize that the corresponding subjects take more time to reach the actual decision or consider additional information not accessed by most subjects. Hence, we would expect that the final topography of these subjects becomes more similar to the final group-level topography again, when estimating an extra topography and flat state for these subjects. Intriguingly, Krause et al. (2024) suggested that such an extra state might actually be completed by all subjects, but only for pseudo-words. With the SMSM framework presented here it is possible to investigate which of these competing hypotheses, or perhaps a combination, is more likely.

More generally, the subject-level deviations observed here also caution against over-interpreting estimates obtained from models including only group-level effects: state estimates decoded from group-level estimates alone might, for example, exaggerate differences in the duration or emissions of individual states which can generate the impression that these states reflect distinct processes. This reinforces the point raised briefly in the Introduction, that Semi-Markov states can, at best, be considered a reflection of a combination of temporally (partially) overlapping processes (e.g., Baker et al., 2014; Pieramico et al., 2025; Vidaurre et al., 2025).

An important factor that contributed to the decision not to investigate a model with a subject-specific extra state or a different emission process model for a combination of the fourth and fifth flat state, is the considerable computational complexity associated with estimating SMSMs. Obtaining 1250 samples from the final model reported here took more than a week and a dedicated server with 40 CPUs. This drastically complicates the task of selecting between two or more candidate models – even when an estimate of or prior knowledge about the number of states and appropriate emission distribution/model is available so that the candidate models differ only in their effect structure. While previous work typically considers only a single emission distribution/model shared by all states, often motivated by theoretical considerations, it is common to estimate the effect structure and number of states alongside model parameters (e.g., Borst & Anderson, 2015; Vidaurre et al., 2025; Weindel et al., 2024). We return to the problem of determining the number of states in the final part of this discussion and for now focus instead on the problem of determining an appropriate (hierarchical) effect structure for the models of the different parameters given (an estimate of) the number of states (and emission distributions/models).

4.1 Determining an Appropriate Effect Structure

This "simpler" problem is already far from trivial, because the space of possible SMSMs is extremely vast – even for simple experiments involving only a few manipulations and covariates. Taking into account the aforementioned high computational complexity of the log-likelihood and gradient functions, a brute-force cross-validation approach over a grid of different effect structures for parameters involved in the emission and latent processes and any remaining hyper-parameters (e.g.,the ρ value of an ar-1 process) will often be impossible. For that reason, we instead chose the value for the ρ parameter of the ar-1 model of emission deviations and the effect-structure of the state sojourn time models based on inspections of residual visualizations and previous work.

To be clear, modifying a model based on patterns visible in the residuals is a valid strategy – not every extension has to be informed by a significance test or a selection procedure. Indeed, the results from the analysis performed here again highlight the importance and usefulness of residual visualizations and model validation in general: the forward pseudo-residuals (see Langrock et al., 2012) of the six flat-state model showed strong temporal dependencies. Adjusting the specific emission model proposed by Anderson et al. (2016) to incorporate an ar-1 model of the PCA deviations (see Appendix A) largely accounted for these temporal dependencies – providing immediate justification for the added complexity of the resulting model.

In line with this, we would recommend to generally rely heavily on model validation for the task of selecting an appropriate (hierarchical) effect structure given (an estimate of) the number of states. Ideally, this search would already start with a reasonably complex effect structure, involving sufficient random effects in the models of all parameters (c.f. Bates et al., 2018), determined based on theoretical considerations. Subsequently, the different types of residuals discussed in the methods section could then be used to check if the data contains substantial remaining dependencies that the model cannot account for. Additionally, a formal selection among a small number of candidate models can be performed without having to complete an expensive cross-validation procedure.

Instead, the selection could be based on any of the available Bayesian selection criteria which can be computed from sampling output (see Gelman et al., 2013, for an overview). Alternatively, a conditional version of the Akaike Information Criterion (AIC) can be computed as $-2\mathcal{L}(\hat{\boldsymbol{\beta}}) + tr(\mathbf{VH})$, where $\hat{\boldsymbol{\beta}}$ is the maximizer of the penalized log-likelihood given $\hat{\boldsymbol{\lambda}}$ obtained as discussed in section 2.2.2, $\mathbf{H}$ is the negative Hessian of the log-likelihood (or a quasi-Newton approximation; see Krause et al., 2025) evaluated at $\hat{\boldsymbol{\beta}}$, and $tr(\mathbf{VH})$ denotes the trace of $\mathbf{VH}$ (Wood et al., 2016). $\mathbf{V}$ can either be based on the covariance matrix of the Laplace approximation to the posterior $\boldsymbol{\beta}|\mathbf{O}, \boldsymbol{\lambda}$ or estimated from samples of the posterior $\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{O}$ in which case the AIC will automatically be corrected for uncertainty in the estimate $\hat{\boldsymbol{\lambda}}$ as well (see Wood, 2016; Wood et al., 2016).

4.2 Determining the Number of States

A data-driven analysis designed to determine the optimal number of Semi-Markov states alongside an appropriate effect structure faces the same problems outlined before. However, while computational complexity was already a concern for the task of determining an appropriate effect structure, a brute-force cross-validation approach over a grid of different effect structures for parameters involved in the emission and latent processes, number of states, hyper-parameters, and (optionally) different emission and state sojourn time distributions/models becomes intractable.

A potential alternative might be to adopt a sequential selection strategy – not unlike what was done for the analysis presented here: first, cross-validation could be used to determine the optimal number of states, given a fixed and relatively simple group-level effect structure and choices for emission and state sojourn time distributions. The second step, determining an appropriate hierarchical effect structure as well as values for any hyper-parameters could then be addressed as discussed in the previous section, based on the number of states estimated in the first step. For example, Krause et al. (2024) used leave-one-out cross-validation (LOOCV) and only accepted additional states if this reduced the cross-validation error for a majority of subjects. In consequence, it could be expected that the recovered EEG topographies should make for good group-level estimates in a hierarchical model as well. Indeed, the hierarchical model presented here produced very similar group-level estimates (compare the left-most panels of Figures 2 and 6), despite the fact that the hierarchical model also included an ar-1 model of the emission deviations.

Naturally, there are many ways in which this first step might fail to produce a useful estimate for the number of states. For example, if, in truth, word-likeness had a (non-linear) group-level effect on the EEG topographies, LOOCV, based on a model treating the topographies as constant, would be unlikely to result in a useful estimate of the number of states. As such, success of such a sequential selection strategy cannot be guaranteed[9]. Fortunately, the framework presented in this paper, implemented in the mssm Python toolbox (version ≥ 1.2.0; Krause et al., 2025), allows to thoroughly investigate the usefulness of this and any other strat-

[9]Considering the vast space of possible SMSMs for any given data-set, this will be the case for any automated selection strategy anyway. Careful consideration of existing theories and available experimental evidence to narrow down the potential number of states and effect structure will thus generally be advisable and typically increase the chances of a subsequent selection procedure to generate useful estimates.

egy to estimate the number of states. Additionally, careful model validation, based on the different residual types considered here, will at least allow to determine whether the final model obtained by any such strategy remains highly unlikely to have generated the collected data. Similarly, inspections of these residuals can also guide the search for an appropriate effect structure.

4.3 Conclusion

In summary, the SMSM framework offers a valuable addition to the toolkit of researchers wanting to discover cognitive processes from neuro-physiological recordings. SMSMs can account for subject-level heterogeneity in the latent and emission process and can include smooth functions of experimental covariates in the (mixed) models of individual parameters. Because the computational complexity of estimating (or sampling from) SMSMs is considerable, more research is necessary to validate and determine effective data-driven model selection strategies. To facilitate such research, all the algorithms considered in this paper have been implemented in the mssm Python toolbox.

References

Anderson, J. R., & Fincham, J. M. (2014). Discovering the Sequential Structure of Thought. *Cognitive Science*, *38*(2), 322–352. https://doi.org/10.1111/cogs.12068

Anderson, J. R., Zhang, Q., Borst, J. P., & Walsh, M. M. (2016). The discovery of processing stages: Extension of Sternberg's method. *Psychological Review*, *123*(5), 481–509. https://doi.org/10.1037/rev0000030

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Baker, A. P., Brookes, M. J., Rezek, I. A., Smith, S. M., Behrens, T., Probert Smith, P. J., & Woolrich, M. (2014). Fast transient networks in spontaneous human brain activity (J. C. Culham, Ed.). *eLife*, *3*, e01867. https://doi.org/10.7554/eLife.01867

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious Mixed Models. https://doi.org/10.48550/arXiv.1506.04967

Berberyan, H. S., van Maanen, L., van Rijn, H., & Borst, J. (2021). EEG-based Identification of Evidence Accumulation Stages in Decision-Making. *Journal of Cognitive Neuroscience*, *33*(3), 510–527. https://doi.org/10.1162/jocn_a_01663

Borst, J. P., & Anderson, J. R. (2015). The discovery of processing stages: Analyzing EEG data with hidden semi-Markov models. *NeuroImage*, *108*, 60–73. https://doi.org/10.1016/j.neuroimage.2014.12.029

Buckby, J., Wang, T., Zhuang, J., & Obara, K. (2020). Model Checking for Hidden Markov Models. *Journal of Computational and Graphical Statistics*, *29*(4), 859–874. https://doi.org/10.1080/10618600.2020.1743295

Coquelet, N., De Tiège, X., Roshchupkina, L., Peigneux, P., Goldman, S., Woolrich, M., & Wens, V. (2022). Microstates and power envelope hidden Markov modeling probe bursting brain activity at different timescales. *NeuroImage*, *247*, 118850. https://doi.org/10.1016/j.neuroimage.2021.118850

Dang, S., Chaudhury, S., Lall, B., & Roy, P. K. (2017). Learning effective connectivity from fMRI using autoregressive hidden Markov model with missing data. *Journal of Neuroscience Methods*, *278*, 87–100. https://doi.org/10.1016/j.jneumeth.2016.12.019

Dewar, M., Wiggins, C., & Wood, F. (2012). Inference in Hidden Markov Models with Explicit State Duration Distributions. *IEEE Signal Processing Letters*, *19*(4), 235–238. https://doi.org/10.1109/LSP.2012.2184795

Donders, F. C. (1868). Die Schnelligkeit psychischer Processe: Erster Artikel. *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, 657–681.

Dunn, P. K., & Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244. https://doi.org/10.2307/1390802

Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781315382722

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b16018

Hendrix, P., & Sun, C. (2020). A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*. https://doi.org/10.1037/xlm0000819

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(47), 1593–1623.

Kalliovirta, L. (2008). Quantile Residuals for Multivariate Models.

Kimeldorf, G. S., & Wahba, G. (1970). A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines. *The Annals of Mathematical Statistics*, *41*(2), 495–502. https://doi.org/10.1214/aoms/1177697089

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Krause, J., Borst, J. P., & Rij, J. v. (2025). The Mixed-Sparse-Smooth-Model Toolbox (MSSM): Efficient Estimation and Selection of Large Multi-Level Statistical Models. https://doi.org/10.48550/arXiv.2506.13132

Krause, J., van Rij, J., & Borst, J. P. (2024). Word Type and Frequency Effects on Lexical Decisions Are Process-dependent and Start Early. *Journal of Cognitive Neuroscience*, *36*(10), 2227–2250. https://doi.org/10.1162/jocn_a_02214

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., & Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: Hidden Markov models and extensions. *Ecology*, *93*(11), 2336–2342. https://doi.org/10.1890/11-2241.1

Lehmann, D., Strik, W. K., Henggeler, B., Koenig, T., & Koukkou, M. (1998). Brain electric microstates and momentary conscious mind states as building blocks of spontaneous thinking: I. Visual imagery and abstract thoughts. *International Journal of Psychophysiology*, *29*(1), 1–11. https://doi.org/10.1016/S0167-8760(97)00098-6

Levinson, S. E. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech & Language*, *1*(1), 29–45. https://doi.org/10.1016/S0885-2308(86)80009-2

Lystig, T. C., & Hughes, J. P. (2002). Exact Computation of the Observed Information Matrix for Hidden Markov Models. *Journal of Computational and Graphical Statistics*, *11*(3), 678–689.

Mahé, G., Zesiger, P., & Laganaro, M. (2015). Beyond the initial 140 ms, lexical decision and reading aloud are different tasks: An ERP study with topographic analysis. *NeuroImage*, *122*, 65–72. https://doi.org/10.1016/j.neuroimage.2015.07.080

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*, 287–330. https://doi.org/10.1037/0033-295X.86.4.287

Michelot, T. (2025). hmmTMB: Hidden Markov Models with Flexible Covariate Effects in R. *Journal of Statistical Software*, *114*, 1–45. https://doi.org/10.18637/jss.v114.i05

Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization*. Springer New York. https://doi.org/10.1007/978-0-387-40065-5

Pieramico, G., Makkinayeri, S., Guidotti, R., Basti, A., Voso, D., Lucarelli, D., D'Andrea, A., L'Abbate, T., Romani, G. L., Pizzella, V., & Marzetti, L. (2025). Robustness of brain state identification in synthetic phase-coupled neurodynamics using Hidden Markov Models. *Frontiers in Systems Neuroscience*, *19*. https://doi.org/10.3389/fnsys.2025.1548437

Portoles, O., Blesa, M., Van Vugt, M., Cao, M., & Borst, J. P. (2022). Thalamic bursts modulate cortical synchrony locally to switch between states of global functional connectivity in a cognitive task (D. Bush, Ed.). *PLOS Computational Biology*, *18*(3), e1009407. https://doi.org/10.1371/journal.pcbi.1009407

Quinn, A. J., Vidaurre, D., Abeysuriya, R., Becker, R., Nobre, A. C., & Woolrich, M. W. (2018). Task-Evoked Dynamic Network Analysis Through Hidden Markov Modeling. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00603

Rabiner, L. R. (1990). A tutorial on hidden Markov models and selected applications in speech recognition. In *Readings in speech recognition* (pp. 267–296). Morgan Kaufmann Publishers Inc.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological review*, *111*(1), 159–182. https://doi.org/10.1037/0033-295X.111.1.159

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *54*(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315. https://doi.org/10.1016/0001-6918(69)90055-9

Taghia, J., Cai, W., Ryali, S., Kochalka, J., Nicholas, J., Chen, T., & Menon, V. (2018). Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nature Communications*, *9*(1), 2505. https://doi.org/10.1038/s41467-018-04723-6

van Maanen, L., Portoles, O., & Borst, J. P. (2021). The Discovery and Interpretation of Evidence Accumulation Stages. *Computational Brain & Behavior*, *4*(4), 395–415. https://doi.org/10.1007/s42113-021-00105-2

Vidaurre, D., Hunt, L. T., Quinn, A. J., Hunt, B. A. E., Brookes, M. J., Nobre, A. C., & Woolrich, M. W. (2018). Spontaneous cortical activity transiently organises into frequency specific phase-coupling networks. *Nature Communications*, *9*(1), 2987. https://doi.org/10.1038/s41467-018-05316-z

Vidaurre, D., Masaracchia, L., Larsen, N. Y., Ruijters, L. R., Alonso, S., Ahrends, C., & Woolrich, M. W. (2025). The Gaussian-linear hidden Markov model: A Python package. *Imaging Neuroscience*, *3*, imag_a_00460. https://doi.org/10.1162/imag_a_00460

Vidaurre, D., Quinn, A. J., Baker, A. P., Dupret, D., Tejero-Cantero, A., & Woolrich, M. W. (2016). Spectrally resolved fast transient brain states in electrophysiological data. *NeuroImage*, *126*, 81–95. https://doi.org/10.1016/j.neuroimage.2015.11.047

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159. https://doi.org/10.1016/j.jml.2007.04.006

Wahba, G. (1985). A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. *The Annals of Statistics*, *13*(4), 1378–1402. https://doi.org/10.1214/aos/1176349743

Weindel, G., van Maanen, L., & Borst, J. P. (2024). Trial-by-trial detection of cognitive events in neural time-series. *Imaging Neuroscience*, *2*, 1–28. https://doi.org/10.1162/imag_a_00400

Wood, S. N. (2015). *Core Statistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781107741973

Wood, S. N. (2016). Just Another Gibbs Additive Modeler: Interfacing **JAGS** and **mgcv**. *Journal of Statistical Software*, *75*(7). https://doi.org/10.18637/jss.v075.i07

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781315370279

Wood, S. N. (2020). Inference and computation with generalized additive models and their extensions. *TEST*, *29*(2), 307–339. https://doi.org/10.1007/s11749-020-00711-5

Wood, S. N., & Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*, *73*(4), 1071–1081. https://doi.org/10.1111/biom.12666

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, *174*(2), 215–243. https://doi.org/10.1016/j.artint.2009.11.011

Yu, S.-Z., & Kobayashi, H. (2006). Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Transactions on Signal Processing*, *54*(5), 1947–1951. https://doi.org/10.1109/TSP.2006.872540

Zucchini, W., MacDonald, I. L., & Langrock, R. (2017, December). *Hidden Markov Models for Time Series: An Introduction Using R, Second Edition* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b20790

Appendix A
Working with Emission Densities Conditional on Residual State Duration

As discussed in the main paper, Anderson et al. (2016) set out to define a more appropriate model of EEG emissions, based on the observation that several theories predict that the EEG signal can be described as a multivariate oscillatory noise process with zero expectation most of the time. Only in a brief period following a shift in processing, assumed to coincide with the onset of a new cognitive process, is the signal believed to deviate systematically from this process.

Specifically, the onset of a new process is assumed to result in the addition of a weighted half-sine of fixed duration to the EEG signal at each electrode. Anderson et al. (2016) thus suggest to distinguish between the emissions generated while in a "flat state" and the emissions generated during these "bumps". In an EDHMM or latent smooth model, the former can be represented as a regular (semi-Markov) stage, while the latter can be treated as fixed duration micro-states. Let $\mathcal{S}_F \subset \mathcal{S}$ denote the set of flat states. Considering that the expected value of the signal at each electrode is assumed to be zero and that signals from different electrodes are assumed to be conditionally independent, it is reasonable to assume that

$$p(\mathbf{o}_{st}|S_{st} = j, j \in \mathcal{S}_F) \propto exp\left(-\left\{\sum_{k=1}^{K} \mathbf{o}_{stk}^2\right\}\right), \tag{A1}$$

i.e., that the density of observing $\mathbf{o}_{st}$ during a flat state is proportional to the negative sum of squares (taken over the K electrodes) of the EEG amplitudes (Anderson et al., 2016).

Matters are slightly more complicated for bump "states": in contrast to flat states, for which the expected value is simply zero for all electrodes and at all times, the expected value now depends on the residual duration $\tau_t = d'$ of bump state j and can be different for different electrodes. Specifically, considering that bump states are assumed to have fixed duration $\mathcal{D}'$, the expected value at electrode k is equal to $s(d')\boldsymbol{\mu}_{jk}$ when the latent process is in state $j \notin \mathcal{S}_F$ with residual duration $\tau_t = d'$ and $d' \in \{1, ...D'\}$. $s(d')$ returns the amplitude of a half-sine of duration $\mathcal{D}'$ time-locked to $d' = 1$ at time-point $D' - d' + 1$ and $\boldsymbol{\mu}_{jk}$ corresponds to the aforementioned weight of the half-sine added to the signal of electrode k when the latent process enters state j (Anderson et al., 2016). Note, that the value returned by $s(d')$ depends on the sampling frequency at which data was collected and the assumed frequency of the sine: Anderson et al. (2016) assumed a 10 Hz frequency for the sine, resulting in a 50 ms duration of bump states. Assuming that data is collected at a sampling frequency of 100 Hz, $\mathcal{D}' = 5$ and $s(5) = 0.309$, $s(4) = 0.809$, $s(3) = 1.00$, $s(2) = 0.809$, $s(1) = 0.309$ (Anderson et al., 2016).

Based on these considerations, the density of observing $\mathbf{o}_t$ conditional on the residual duration of state $j \notin \mathcal{S}_F$ can then be defined as $P(\mathbf{o}_{st}|S_{st} = j, \tau_{st} = d', j \notin \mathcal{S}_F) \propto$

$$exp\left(-\left\{\sum_{k=1}^{K} \left(\mathbf{o}_{stk} - s(d')\boldsymbol{\mu}_{jk}\right)^2\right\}\right), \tag{A2}$$

for all $d' \in \{1, ...D'\}$. Fortunately, access to this density is sufficient to compute the forward variables defined in Equations (6) and (7). Note, that this will still require specification of the missing normalization constant to convert the sum-of-square terms on the right-hand side of Equations (A1) and (A2) into actual densities. In practice, this can be achieved by assuming a known distribution for the squared deviation of the amplitude at a given electrode from the expected values (i.e., $\mathbf{o}_{tk}^2$ and $\left(\mathbf{o}_{tk} - s(d')\boldsymbol{\mu}_{jk}\right)^2$). Plausible choices are Exponential, Chi-square, Gamma, or Normal distributions (see Anderson et al., 2016, for a more detailed discussion of plausible distributions). Finally, note that $\boldsymbol{\mu}_{jk}$ can readily be parameterized by an additive mixed model as shown in Equation (3). Notationally, all that remains necessary to use the Emission model by Anderson et al. (2016) as part of a latent smooth is to replace $\boldsymbol{\mu}_{jk}$ with $\boldsymbol{\mu}_{stjk}$ in Equations (A1) and (A2).

ar-1 Model of the Deviations

Assuming a Gaussian model also enables simple auto-correlation (ar) models of the deviations (e.g., Wood, 2017), let $\mathbf{Z}_s^*$ denote the true sequence of bump and flat states visited on trial s. Assuming that the deviations of the amplitude at a given electrode k from the expected values (i.e., $\mathbf{o}_{tk} - 0$ and $\mathbf{o}_{tk} - s(d')\boldsymbol{\mu}_{jk}$) are normally distributed with zero mean and variance σ^2, the conditional density of all T_s amplitude recording $\mathbf{o}_{sk}$ collected on trial s at electrode k is multivariate normal, since $\mathbf{o}_{sk}|\mathbf{Z}_s^* \sim N(\boldsymbol{\mu}_{sk}, \mathbf{I}\sigma^2)$. $\boldsymbol{\mu}_{sk}$ is the vector of expected values containing zeros for every time-point $\mathbf{Z}_s^*$ indicates to be in a flat state and the correct "bump" weights for electrode k and state j for every time-point $\mathbf{Z}_s^*$ indicates to be in bump state j. As discussed in the main paper, here $\boldsymbol{\mu}_{sk}$ is specified in terms of an additive (mixed) model. Thus, we write $\boldsymbol{\mu}_{sk} = \mathbf{X}_s^{\mu_k}\boldsymbol{\beta}^{\mu_k}$. $\mathbf{X}_s^{\mu_k}$

is obtained by stacking rows of zeros (for observations corresponding to flat states) and rows from the state-specific model matrices $\mathbf{X}_s^{\mu_{jk}}$ (for observations corresponding to bump state j). Similarly, $\boldsymbol{\beta}^{\mu_k}$ is obtained by concatenating all of the state-specific coefficient vectors $\boldsymbol{\beta}^{\mu_{jk}}$ for all bump states. Since flat states are assumed to have an expectation of zero, no coefficient vectors are associated with them.

To include an ar model of the deviations we now substitute the covariance matrix $\mathbf{V}$ of an ar process (see Wood, 2017, for details on how to form $\mathbf{V}$) for $\mathbf{I}$ and instead assume that $\mathbf{o}_{sk}|\mathbf{Z}_s^* \sim N(\mathbf{X}_s^{\mu_k}\boldsymbol{\beta}^{\mu_k}, \mathbf{V}\sigma^2)$. Applying standard results about linear transformations of normal random variables, we get $\mathbf{L}^{-1}\mathbf{o}_{sk}|\mathbf{Z}_s^* \sim N(\mathbf{L}^{-1}\mathbf{X}_s^{\mu_k}\boldsymbol{\beta}^{\mu_k}, \mathbf{I}\sigma^2)$, where $\mathbf{L}^{-1}$ is the inverse of the Cholesky factor $\mathbf{L}$ of $\mathbf{V} = \mathbf{L}\mathbf{L}^\top$ (e.g., Wood, 2017). If $\mathbf{Z}_s^*$ were known, estimating $\boldsymbol{\beta}^{\mu_k}$ for an ar model could thus proceed in the same way as before – requiring only that the transformations of $\mathbf{o}_{sk}$ and $\mathbf{X}_s^{\mu_k}$ are computed once, before parameter fitting. Typically, $\mathbf{Z}_s^*$ will however not be known so that the transformation cannot be computed in advance. Importantly, for lower-order ar models (e.g., of lag 1, resulting in an ar-1 model) computing $\mathbf{L}^{-1}\mathbf{o}_{sk}$ and $\mathbf{L}^{-1}\mathbf{X}_s^{\mu_k}$ corresponds to a simple an efficient re-weighting of rows in $\mathbf{o}_{sk}$ and $\mathbf{X}_s^{\mu_k}$ (Wood, 2017). As such, for lower-order ar models it remains possible to compute this re-weighting dynamically, while evaluating the forward variables, without too much computational overhead. Attention has to be paid to computing the expected value during the onset of a bump state: assuming an ar-1 model and a strict left-to-right HsMM with supra-diagonal transition matrix, as suggested by Anderson et al. (2016), the previous expectation, for the last observation emitted during a flat state, will always be zero. Additionally, when computing the gradient of the log-likelihood, it will be necessary to carefully consider cross-state dependencies of the emission densities: again assuming an ar-1 model, the density of the first observation of the current (flat) state will depend on the coefficients of the preceding (bump) state.

Appendix B
Partial Derivatives of Densities and Probabilities Required for Gradient Evaluation

In this appendix we show how the partial derivatives of the different densities/probabilities, all parameterized through different linear predictors, involved in Equations (12), (13), and (14) of the main paper can be obtained via transformation. This relies on the general strategy described by Wood et al. (2016). Let β_r again denote a coefficient included in the total set of coefficients $\boldsymbol{\beta}$ required by the models of all parameters. Then, assuming that β_r is involved in the model of parameter μ_j,

$$\frac{\partial b_j(\mathbf{o}_{st})}{\partial \beta_r} = \frac{\partial b_j(\mathbf{o}_{st})}{\partial \boldsymbol{\mu}_{jst}} \left(\frac{\partial g^{\mu_j}(\boldsymbol{\mu}_{jst})}{\partial \boldsymbol{\mu}_{jst}} \right)^{-1} \frac{\partial \boldsymbol{\eta}_{st}^{\mu_j}}{\partial \beta_r}, \tag{B1}$$

where $\frac{\partial \boldsymbol{\eta}_{st}^{\mu_j}}{\partial \beta_r} = \mathbf{X}_{str}^{\mu_j}$ with $\mathbf{X}_{str}^{\mu_j}$ indexing column r in the row of model matrix $\mathbf{X}^{\mu_j}$ associated with time-series s and time-point t. Fortunately, the derivatives of the densities with respect to their parameters (i.e., $\frac{\partial b_j(\mathbf{o}_{st})}{\partial \boldsymbol{\mu}_{jst}}$) are available for many distributions – including the multinomial distribution, which enables computation of the derivatives involving $\boldsymbol{\pi}$ and $\boldsymbol{\Delta}$ (e.g., Wood et al., 2016).

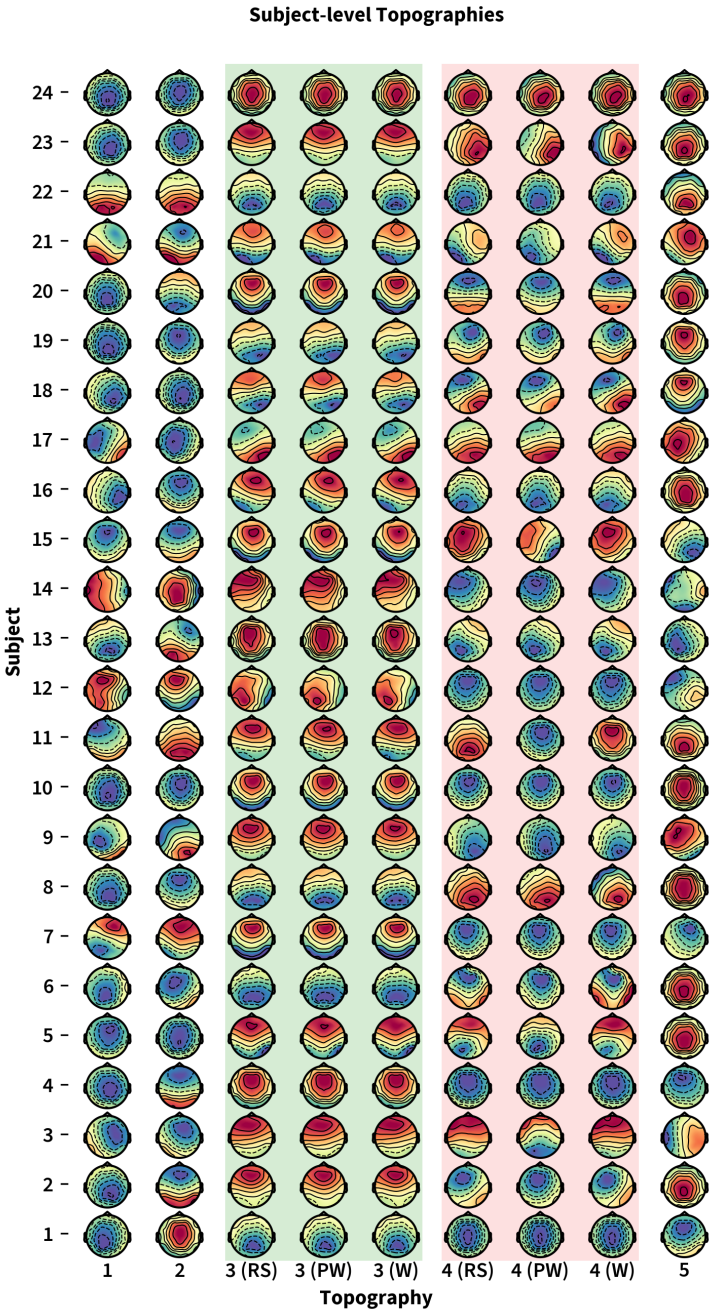
Computation of the derivatives of the state sojourn time densities $\frac{\partial p_j(d)}{\partial \beta_r}$ with respect to β_r might require additional work. If we assume a continuous distribution (i.e., $\mathcal{F}^d$) for state durations d, then the derivative can be computed exactly as shown in Equation (B1). However, it is common to normalize the densities to ensure a proper probability mass function over the discrete interval $\{1, D\}$, in which case $p_j(d) = \frac{p_j^*(d)}{\sum_{d' \in D} p_j^*(d')}$, with $p_j^*(d)$ corresponding to the density of a continuous distribution (i.e., $\mathcal{F}^d$; see also Yu, 2010). In that case, the required derivative has to be computed as shown in Equation (B2) – note that $\frac{\partial p_j^*(d)}{\partial \beta_r}$ can again be computed as outlined above (i.e., Equation B1).

$$\frac{\partial p_j(d)}{\partial \beta_r} = \frac{\left(\frac{\partial p_j^*(d)}{\partial \beta_r} \sum_{d' \in D} p_j^*(d') \right) - \left(p_j^*(d) \sum_{d' \in D} \frac{\partial p_j^*(d')}{\partial \beta_r} \right)}{\left(\sum_{d' \in D} p_j^*(d') \right)^2} \tag{B2}$$

Appendix C
Subject-level Topography Estimates

Figure C1

Subject-level Topography Estimates for all Word-types and Subjects

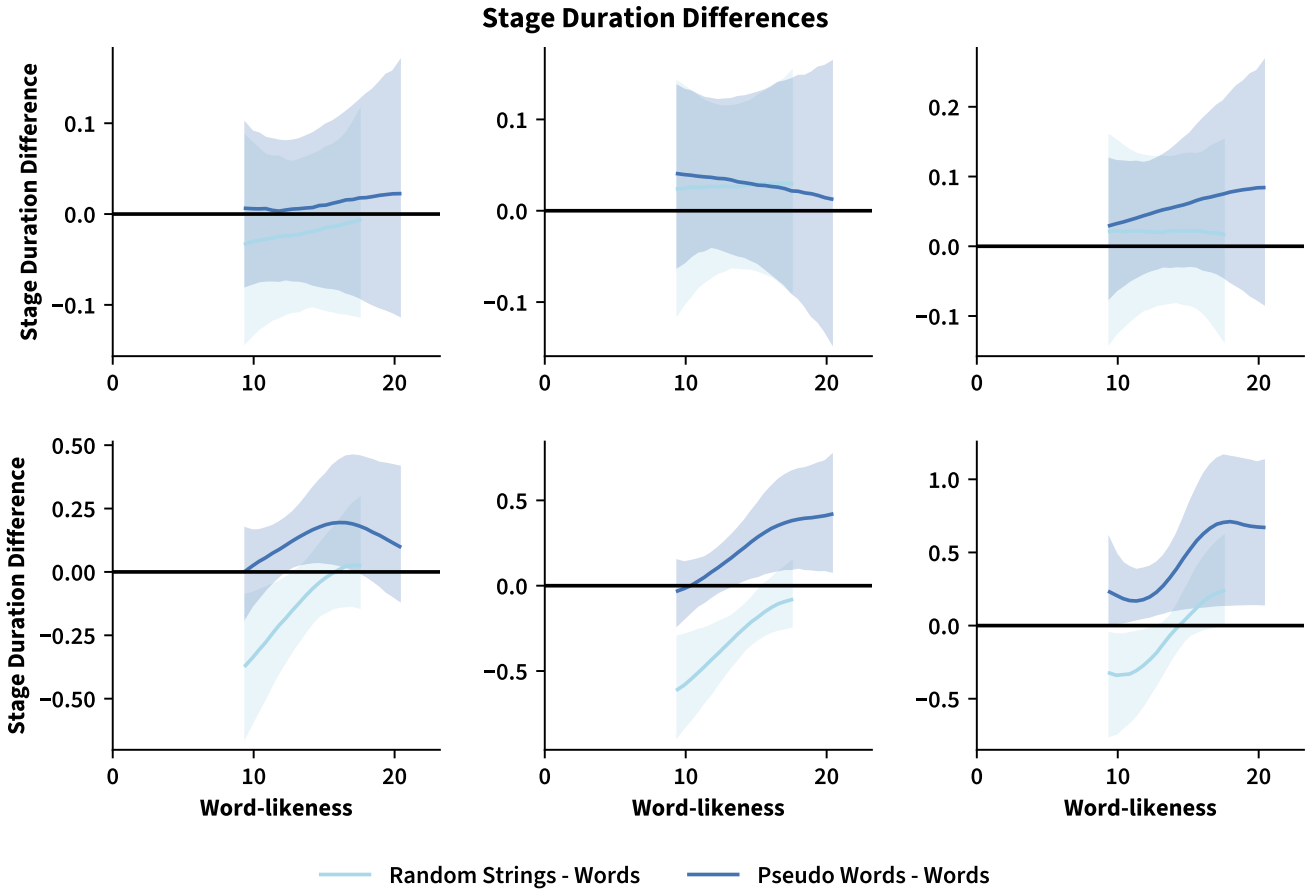


Note. Figure C1 shows the subject-level estimates of the EEG topographies for all subjects and word-types. For the third and fourth topography (highlighted by color bars), separate group-level estimates were obtained per word-type (Words ("W"), Pseudo-words ("PW"), and Random Strings ("RS")), which also results in word-type specific subject-level estimates for these topographies (although the same subject-specific difference is added to the group-level estimates).

Appendix D
Credible Intervals of the Difference in the Effect of Word-likeness

Figure D1

Credible Intervals of the Difference in the Effect of Word-likeness between non-words and words

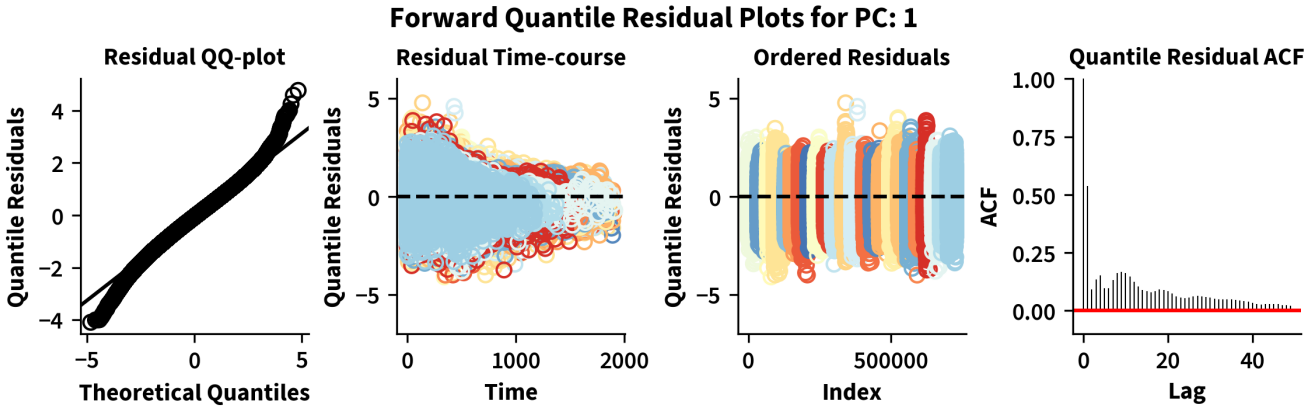


Note. Figure 7 shows an overview of the posterior mean estimate of the difference between the smooth functions $f_{w(s)\,j}(wl(s))$, involved in the models of the mean parameter ν_j of the state sojourn times, between both types of non-words (random strings and pseudo-words) and words. Each panel corresponds to a different flat state, and visualizes the estimated difference smooth for both comparisons. Shaded areas are 95% posterior credible intervals for the difference. Evidently, later states show stronger differences in the effect of word-likeness between non-words and words. Importantly, the magnitude of the difference varies as a word-type specific function of word-likeness, word-type, and flat state – with the strength of the effect (measured by whether the credible interval includes zero or not) varying for different combinations of these three variables.

Appendix E
Residuals of Six Flat-State Multi-level Semi-Markov Model

Figure E1

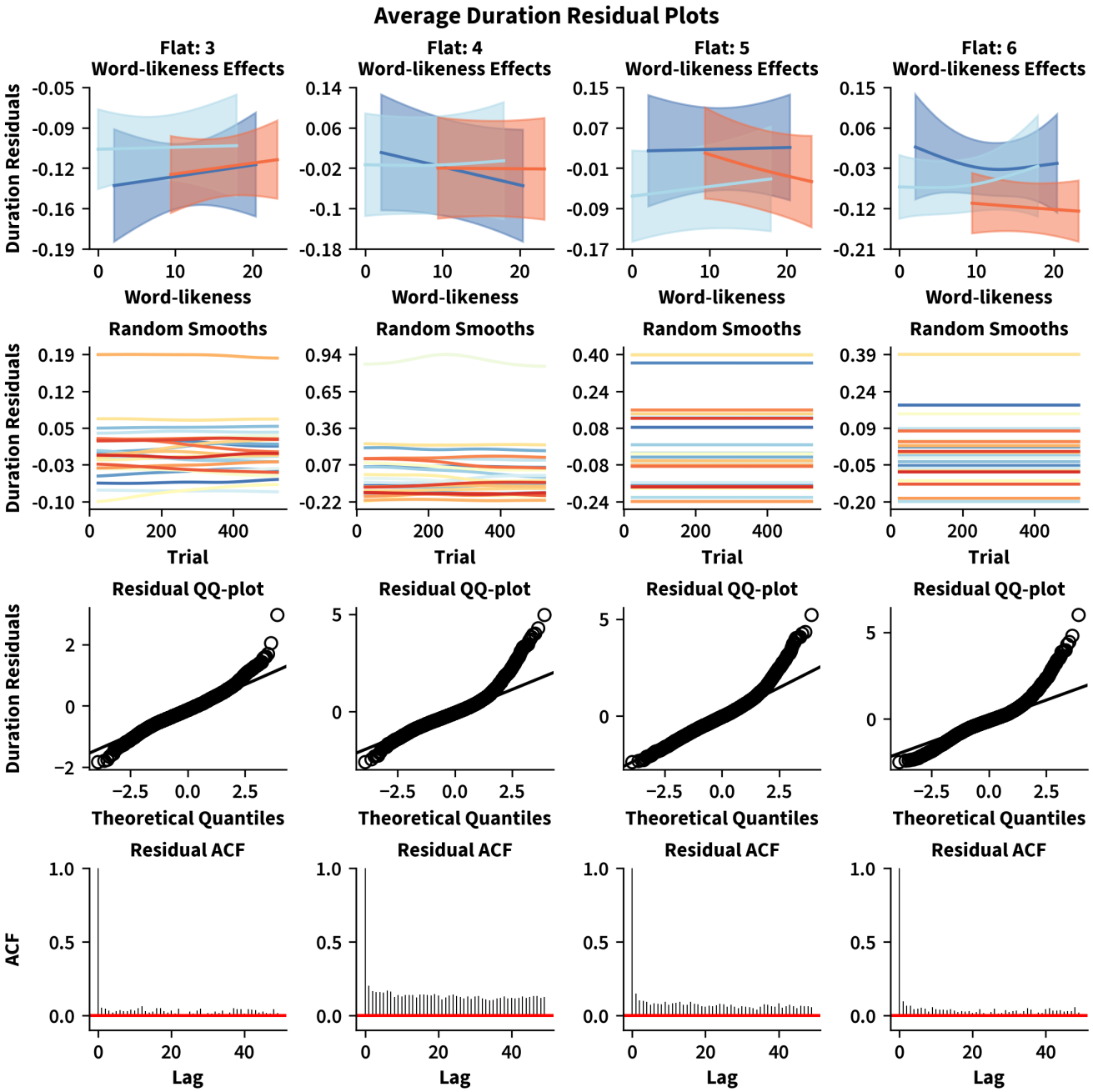
Residuals of Six Flat-State Multi-level Semi-Markov Smooth Model



Note. Like Figures 3 and 4 in the main paper, Figure E1 shows different plots of the forward pseudo-residuals, but computed for the final multi-level semi-Markov Smooth model presented in the main paper. Notably, the forward pseudo-residuals of the multi-level model have only improved marginally. The main difference between the multi-level model and the group-level model (see Figure 4) is that the variance of the multi-level residuals differs less between subjects.

Figure E2

Average Flat-State Duration Residuals of Six Flat-State Multi-level Semi-Markov Smooth Model



Note. Like Figure 5 in the main manuscript, Figure E2 shows plots of the average duration residuals for all flat states, but computed for the final multi-level semi-Markov Smooth model presented in the main paper. Note, that the average duration residuals now provide only week evidence of group-level effects of word-likeness or subject-specific changes in the expected residual over trials (note the y-axis differences when comparing with Figure 5). This is expected, considering that the multi-level model explicitly accounted for these effects. QQ-plots still show deviations from normality, but the residuals are far less correlated across different lags.