APPENDIX TO MONITORING ATHLETE HEALTH THROUGH LATENT STATE MODELLING

FRANK VAN DER MEULEN, LARISA VAN DER GRAAFF, AND EVERT VERHAGEN

Appendix A. OSTRC overuse injuries questionnaire

We consider OSTRC data derived from a longitudinal study on 24 waterpolo players within the Dutch Olympic monitoring program [Verhagen et al., 2021]. The study is carried out through the repeated administration of the OSTRC questionnaire that has been filled in by the athletes on a weekly basis over the 72-week-long follow-up. The data contain athletes' weekly exposure (for different types of training and competition [Injury et al., 2020]) as well as data on athlete's health complaints.

We focus on the dichotomous questionnaire outcomes, which are formulated in a way that responding 1 to any of the questions indicates the presence of a health complaint affecting the corresponding domain. We consider four items (i.e. 4 main questions) – listed in Table 1 – and three time-varying covariates:

- time spent on a sport-specific activity in the last 7 days (in hours),
- time spent on a strength training in the last 7 days (in hours),
- time spent on a competition in the last 7 days (in hours).

The interval between consecutive occasions at which the questionnaire was administered is one week, equal for all participants.

Table 1. The questions selected for the injury risk assessment. The last column denotes the percentage of response 1 (which means the answer to the question was "yes") to each question during the follow-up.

	Question	**%**
Participation:	Have you had any difficulties participating in training due to injury, illness, or other health problems during the past seven days?	22.36
Modification:	Did you have to modify your training due to injury, illness, or other health problems during the past seven days?	12.29
Performance:	Have your injury, illness, or other health problems affected your performance during the past seven days?	14.72
Symptoms:	Have you experienced symptoms/health complaints during the past seven days?	22.89

In Figure 1 we visualise for each athlete the time spent on sport-specific activity and strength training. The responses to each of the 4 questions are visualised in Figure 2. One

Date: November 12, 2025.

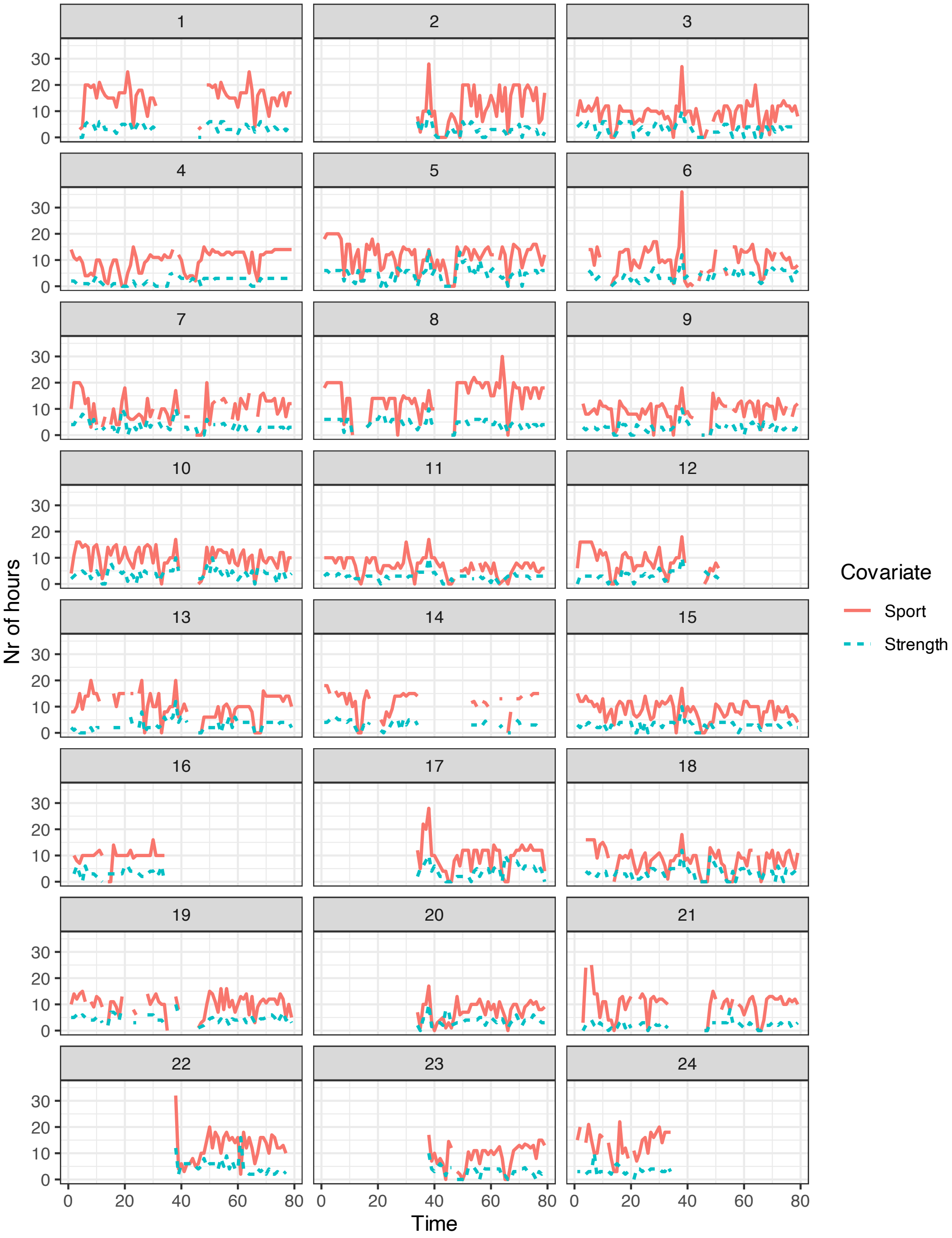

FIGURE 1. Visualisation of the covariates on time spent on sport-specific activity and strength training. Each panel corresponds to an athlete.

can easily see that athletes are monitored over different time periods and that for some athletes there are gaps: periods in which no data were collected.

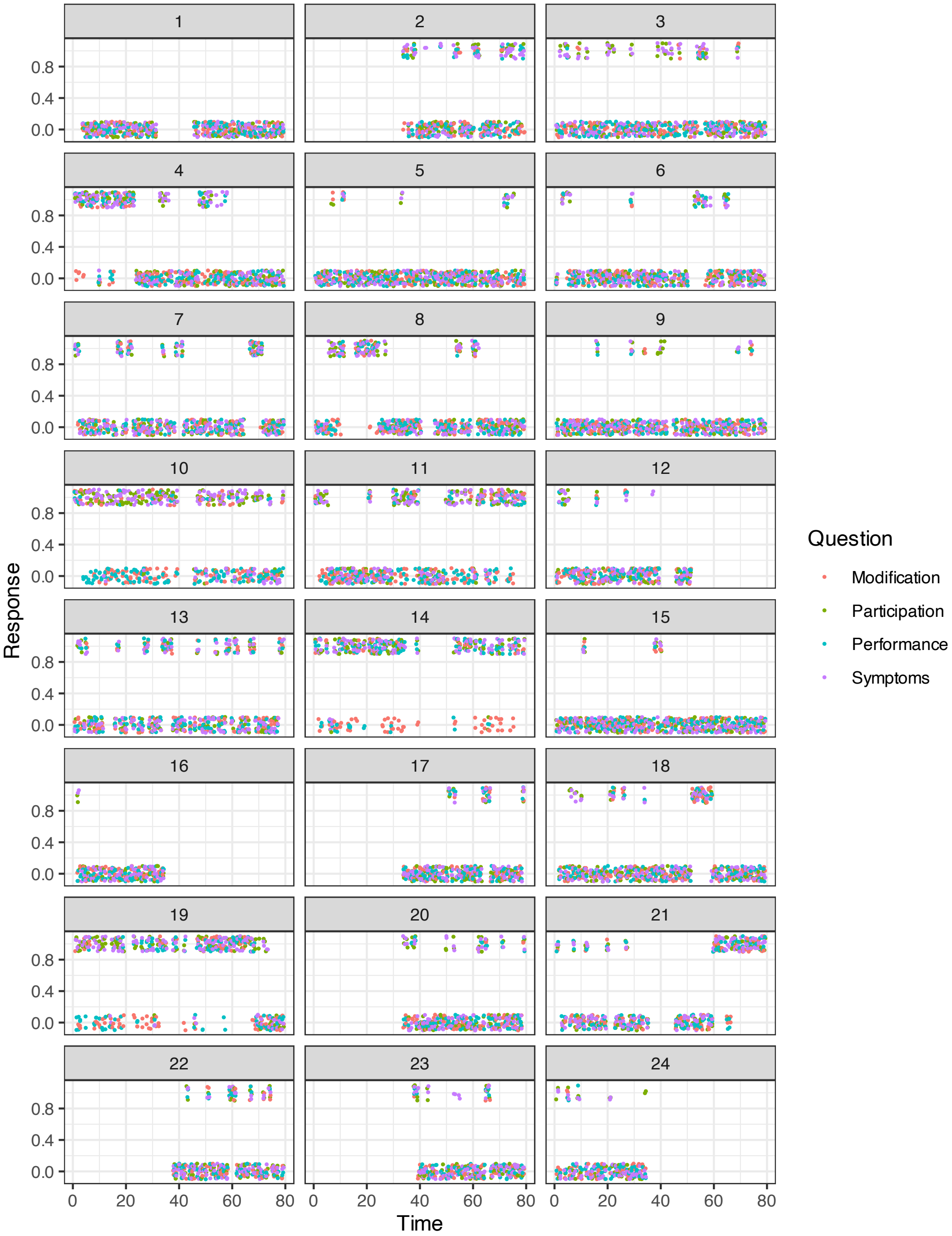

FIGURE 2. Visualisation of the responses to each of the 4 questions. Jitter has been added to the 0/1 outcomes. Each panel corresponds to an athlete.

Appendix B. Statistical methods: latent Markov models

We will start with a non-technical introduction to our approach. Readers that are familiar with hidden Markov models can skip this section. Subsequently, full details for readers with a more extensive background in mathematics and statistics are provided.

B.1. A nontechnical introduction to Latent Markov Models.

For a given athlete, we make a statistical model about the athlete's status, which represents to what extend the athlete is prone to injury. We assume this to be on an ordinal scale, for example "not at risk", "mildly at risk" and "severely at risk". We assume that the states are ordered such that the first state corresponds to least injury prone and the last state to most injury prone. Over time, for example a day or week, an athlete's status may change. We assume this change is probabilistically described by a Markov chain. This entails that the status at time $t+1$ only depends on the present status (time t) and not on the further past. Probabilities to change from one state at time t to another state at time $t+1$ are summarized in a *transition matrix*. For example

$$\Pi_{t+1} = \begin{bmatrix} 0.8 & 0.2 & 0.0 \\ 0.1 & 0.9 & 0.1 \\ 0.01 & 0.04 & 0.95 \end{bmatrix}. \tag{1}$$

Note that all numbers in this matrix are in $[0,1]$ and rows add up to one. Regarding the interpretation of this matrix, the number 0.2, is the probability to become "mildly at risk" at time $t+1$, while being "not at risk" at time t. Similarly, the number 0.95 is the probability to remain in "severely at risk". Each of the elements in the matrix is denoted by $\Pi_{t+1}(d \mid u)$, where $d \in \{1,2,3\}$ and $u \in \{1,2,3\}$. Hence, $\Pi_{t+1}(2 \mid 1) = 0.2$ and $\Pi_{t+1}(3 \mid 3) = 0.95$.

In reality, these probabilities are in fact unknown and part of the model specification. We assume covariates are measured over time. For simplicity, let's assume just one covariate x_t, which denotes training intensity at time t. We will assume the elements in Π_{t+1} to depend on x_t. For this, we need to ensure a specification such that the rows of Π_{t+1} sum to one. Let's see how this works for row 1. We take

$$\log \frac{\Pi_{t+1}(2 \mid 1)}{\Pi_{t+1}(1 \mid 1)} = \gamma_{1,2} x_t \quad \text{and} \quad \log \frac{\Pi_{t+1}(3 \mid 1)}{\Pi_{t+1}(1 \mid 1)} = \gamma_{1,3} x_t.$$

Together with $\Pi_{t+1}(1 \mid 1) + \Pi_{t+1}(2 \mid 1) + \Pi_{t+1}(3 \mid 1) = 1$ this ensures that for any choice of the parameters $\gamma_{1,2}$ and $\gamma_{1,3}$ all probabilities are in $[0,1]$ and sum to 1. For the second and third rows of Π_{t+1}, we can do the same thing using parameters $(\gamma_{2,1}, \gamma_{2,3})$ and $(\gamma_{3,1}, \gamma_{3,2})$ respectively.

Regarding interpretation of these coefficients, if for example $\gamma_{1,3} > 0$, then an increase in training intensity will increase $\Pi_{t+1}(3 \mid 1)$, which is the probability to become "severely at risk" at time $t+1$, given the athlete is "not at risk" at time t. The size of $\gamma_{1,3}$ quantifies the effect, though it is difficult to interpret this size because of the log-scale. This is similar to difficulty in interpretation of estimated coefficients in a logistic regression model.

To recap, once the parameters $\gamma_{i,j}$, $i \neq j$, have been specified, then for any choice of training intensity x_t over time period $1, \ldots T$, we have a probabilistic model for the status on the athlete. By repeated stochastic simulation –simulating from the model using random numbers– we can assess the effect of a particular training schedule on an athlete's status.

Of course, the parameters $\gamma_{i,j}$ are unknown and need to be determined by information provided by data. Let's for simplicity assume at each time instance the athlete answers just one question on a binary scale that provides information about his/her status (the OSTRC results includes multiple questions, and this can be dealt with in a similar way). Let Y_t denote the reply to the one question at time t. Let's consider answer "1" as an affirmative answer. Let U_t denote the status at time t. We assume

$$\mathrm{P}(Y_t = 1 \mid U_t = u) = \lambda(u),$$

where $0 < \lambda(1) < \lambda(2) < \lambda(3) < 1$. This means that if the status of the athlete is "not at risk" ($u = 1$), the athlete will answer in an affirmative manner with probability $\lambda(1)$. Similarly, if he/she is "mildly at risk" ($u = 2$) or "severely at risk" ($u = 3$), an affirmative answer is given with probabilities $\lambda(2)$ and $\lambda(3)$ respectively. The constraint $\lambda(1) < \lambda(2) < \lambda(3)$ therefore implicitly assumed the question is formulated such that an athlete is more likely to answer "yes" if experiencing some degree of risk.

Over time, we observe the results of the posed question. The athlete's status however, is never observed: it is latent. Statistical methods enable to obtain parameters estimates for $\gamma_{i,j}$ $(i \neq j)$ and $(\lambda(1), \lambda(2), \lambda(3))$. Moreover, estimates for the latent paths can be constructed.

The postulated model can be extended to include multiple athletes, multiple covariates and multiple questions in a rather straightforward way, though the bookkeeping using indices can be a bit cumbersome. In the upcoming section i indexes athletes, j indexes questions and t indexes time.

B.2. **General setup.** In the following, as common in statistics, we denote random quantities by capital letters and their realisations (i.e. observations) by lower case letters. Let $y_{ijt} \in \{0, 1\}$ denote the response variable to the jth question in the OSTRC questionnaire administered by the ith subject in week t, with $i = 1, \dots, n$, $j = 1, \dots, J$ and $t \in 1, \dots, T_i$. We use the convention that $y_{ijt} = 1$ means that the question has been answered by "yes". The response vector for subject i at time t is given by the vector of all answers $\mathbf{y}_{it} = (y_{i1t}, \dots, y_{iJt}) \in \{0, 1\}^J$. For subject i, let $\mathbf{x}_{it}$ be the vector of time-varying covariates at time t. In our application, we take these to be the times spent in the last 7 days (in hours) on sport-specific activity and strength training, where we have standardised both covariates. We define $\tilde{\boldsymbol{x}}_{it} = [1, \boldsymbol{x}_{it}]$, which is useful in defining the statistical model that includes an intercept.

Following the latent Markov approach, for each subject i, we assume the existence of a latent process $(U_{\mathrm{i1}}, ..., U_{\mathrm{iT}})$. The latent variable U_{it} represents the injury status of the ith subject in week t. The sequence of latent variables $U_{\mathrm{i1}}, ..., U_{\mathrm{iT}}$ is assumed to follow a (first-order) Markov chain with state space $\{1, \dots, k\}$, where k is the number of latent states. For mathematical convenience, we assume that the responses at time t for subject i, $Y_{i1t}, \dots, Y_{iJt}$, are independent conditional on the latent state U_{it}.

B.2.1. *Approach.* We adopt a fully Bayesian approach. This will facilitate predicting risk to injury of athletes, while taking uncertainty in parameter estimates into account. We show how missing values in either covariates or questionnaire answers can be dealt with. In particular, observing different athletes over different time spans poses no restriction and therefore there is no need to artificially add missing data, as would be needed for example when fitting the model with the R-package `LMest` ([Bartolucci et al., 2017]). The Bayesian approach entails that all unknown parameters in the model are equipped with a prior distribution, which reflects information (or lack of information) about each parameter. Once specified, the Bayesian paradigm postulates that all inferential conclusions are based on the posterior distribution. As this distribution is not available in closed form, we use the probabilistic programming language `Turing` ([Ge et al., 2018]) within the `Julia`-language ([Bezanson et al., 2017]) to sample from the posterior.

B.2.2. *Specification of the latent process.* We assume that the number of latent states equals $k = 3$. For each subject, we assume its initial state can be any of the three latent states with equal probability (these prior probabilities can be adjusted, if information is available). Let $\Pi_{i,t+1}$ denote the transition matrix for individual i for time t to time $t+1$. We assume only transitions to adjacent states are possible and parametrise the transition matrix by vectors $\boldsymbol{\gamma}_{12}$, $\boldsymbol{\gamma}_{21}$, $\boldsymbol{\gamma}_{23}$ and $\boldsymbol{\gamma}_{32}$ such that

$$\Pi_{i,t+1} = \text{softmax}\,.\begin{bmatrix} 0 & \langle \tilde{\mathbf{x}}_{it}, \boldsymbol{\gamma}_{12}\rangle & -\infty \\ \langle \tilde{\mathbf{x}}_{it}, \boldsymbol{\gamma}_{21}\rangle & 0 & \langle \tilde{\mathbf{x}}_{it}, \boldsymbol{\gamma}_{23}\rangle \\ -\infty & \langle \tilde{\mathbf{x}}_{it}, \boldsymbol{\gamma}_{32}\rangle & 0 \end{bmatrix}, \tag{2}$$

where softmax. denotes that the softmax function is to be applied to each row of the matrix (recall softmax: $\mathbb{R}^d \to \mathbb{R}^d$ is defined by $\text{softmax}(x) = (e^{x_1}, \dots, e^{x_d})/\sum_{i=1}^d e^{x_i}$). If any component of $\boldsymbol{x}_{it}$ is missing, we set $\Pi_{i,t+1}$ equal to the identity matrix. This corresponds to assuming the latent state does not change between times t and $t+1$.

B.2.3. *Specification of conditional response probabilities.* We denote the number of questions in the questionnaire by J. In the olympic waterpolo dataset $J = 4$. In case subject i at time t answers "yes" to question j, then $y_{ijt} = 1$, else $y_{ijt} = 0$. We assume that the distribution of the response variables depends only on the latent status by imposing

$$\mathrm{P}(Y_{\mathrm{ijt}} = 1 \mid U_{\mathrm{it}} = u, \mathbf{x}_{\mathrm{it}}) = \lambda_{\mathrm{j}}(u) \tag{3}$$

for each i, j, t and $u \in \{1, \dots, k\}$. Moreover, we require these conditional probabilities to satisfy the constraint

$$0 < \lambda_{\mathrm{j}}(1) < \lambda_{\mathrm{j}}(2) < \dots < \lambda_{\mathrm{j}}(k) < 1 \tag{4}$$

for $j = 1,\dots,J$. This assumption has been used before in [Bartolucci et al., 2009] and ensures identifiability. The constraint (4) implies that the latent states are ordered such that the individuals in the first state are those with the best status (least injury prone) and individuals in the last state are those with the worst injury status. Note that while (3) is the same for all subjects, each subject's injury status is modelled by a separate latent process.

B.2.4. *Prior specification.* For $\boldsymbol{\gamma}_{12}$, $\boldsymbol{\gamma}_{21}$, $\boldsymbol{\gamma}_{23}$ and $\boldsymbol{\gamma}_{32}$ we impose conditionally independent standard multivariate normal priors with covariance matrix σ times the identity matrix. We assign σ the Exponential distribution with mean 1. The underlying ideas is to provide tractable mildly "uninformative" priors.

For each question $j \in \{1, \dots, J\}$ we need to specify a prior on $\boldsymbol{\lambda}_j := (\lambda_j(1), \dots, \lambda_j(3))$ satisfying the ordering constraint in (4). We give a construction for that. Let $Z_j(1)$, $Z_j(2)$, $Z_j(3)$ be independent random variables with the standard Exponential distribution. Set $\psi(x) = 2\,\text{logistic}(3x/4) - 1$, where $\text{logistic}(x) = 1/(1 + e^{-x})$ and note that ψ maps $[0, \infty)$ to $[0, 1)$. Then set

$$\lambda_j(\ell) = \psi\left(\sum_{i=1}^{\ell} Z_j(i)\right), \quad j = 1, \dots, J, \qquad \ell = 1, \dots, k. \tag{5}$$

In Figure 3 we show histograms based on 10_000 samples from the prior. As all $Z_j(i)$ are supported on the positive halfline and ψ is increasing, (4) is satisfied.

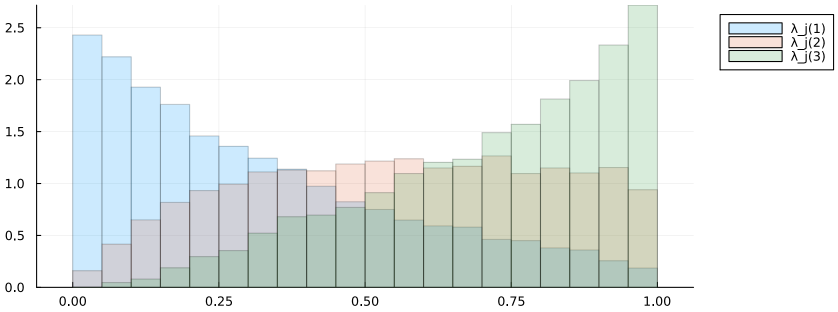

FIGURE 3. Visualisation of the distribution of $\boldsymbol{\lambda}_j$ by Monte-Carlo simulation.

B.2.5. *Path prediction.* Given the estimated model, interest lies in predicting the most likely injury state based on the latest state occupied and the scheduled weekly training exposure. The predicted states over time follow a path that illustrates the progression of injury status over time based on the time-varying training exposure. This indicates the predicted injury risk for an individual athlete at the next time point and serves as a support for sports practitioners and coaches in decision-making during the training process.

Assume that at time T parameters have been estimated based on all subjects in the study. Additionally assume for subject i a training schedule has been determined for the upcoming S weeks, i.e. weeks $T+1,\dots,T+S$. This implies $\mathbf{x}_{i,T+1},\dots,\mathbf{x}_{i,T+S}$ have been specified. Then we can forward simulate scenarios from the latent process $\{U_{it}\}_{t=T}^{T+S}$ to assess injury risk for subject i under the proposed training schedule. This is simply an instance of Monte Carlo simulation where we use a large number of forward simulations which are initialised according to inferred probabilities for U_{iT}.

B.2.6. *Implementation.* Recursive computation of the loglikelihood was implemented in the `Julia`-language in the package `LatentMarkovQuest` (`https://github.com/fmeulen/LatentMarkovQuest`). Subsequently the package `Turing` was used to draw from the posterior using the No-U-Turn-Sampler (NUTS). Cf. [Hoffman et al., 2014]. We used multithreading and in all reported results ran 1000 iterations for each of 6 independent chains (note that by default this implies 2000 iterations are performed, of which the first half are discarded as burnin). Details can be found in Section D.

APPENDIX C. APPLICATION

We ran the NUTS-sampler on the olympic waterpolo dataset. Rhat values (Gelman-Rubin diagnostics, see e.g. Chapter 13 in [Lambert, 2018]) were clearly distinct from 1, indicating lack of convergence of the MCMC chain. Figures 4 and 5 show trace- and density plots for $\boldsymbol{\gamma}_{12}$ and $\boldsymbol{Z}_3 := (Z_3(1), Z_3(2), Z_3(3))$. Corresponding plots for other parameters show similar behaviour. Clearly, depending on how the chain was initialised, the chain samples close to either of two modes. This explains the indicated lack of convergence. It appears that there are two model explanations for the data and it is not directly clear which of those is the more likely one. By default, the chain is initialised by sampling parameter values from the prior distribution. To better understand the sampler's performance, we ran the sampler twice more, with all chains initialised from a parameter vector close to either of the modes.

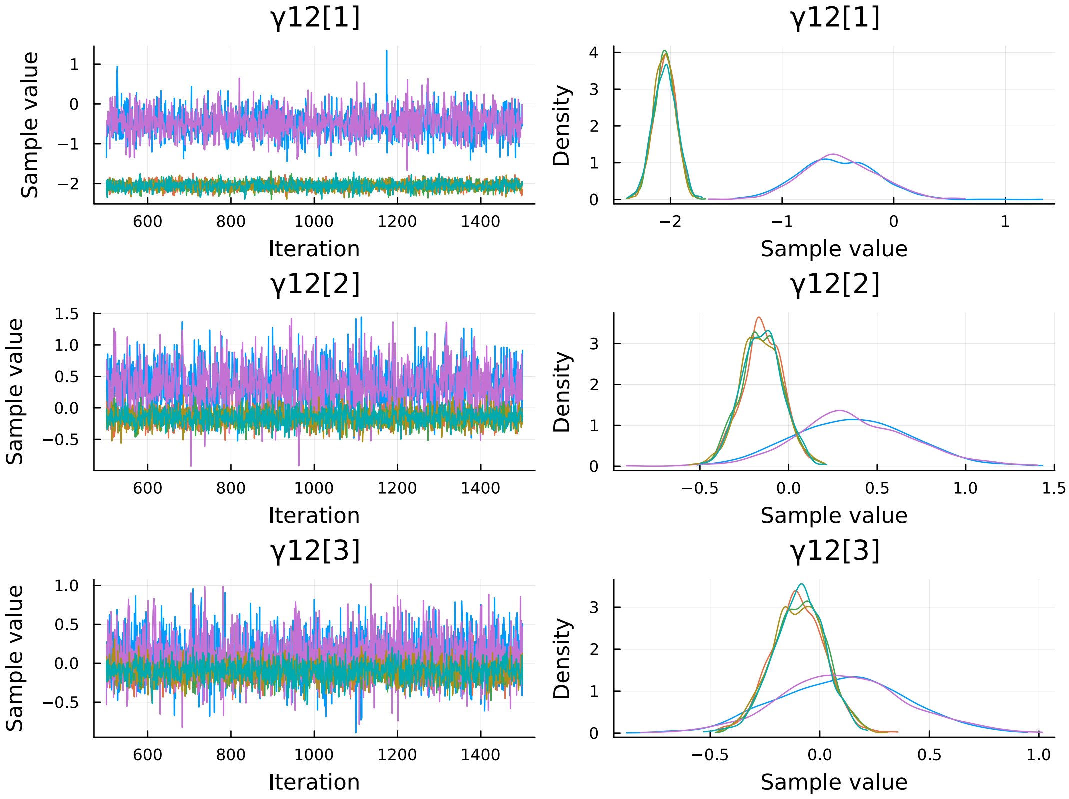

FIGURE 4. Trace- and density plots of iterates for $\boldsymbol{\gamma}_{12}$. Burnin was removed. Initialisation of chains by sampling from the prior.

In Figures 6 and 7 we show trace- and densityplots for $(Z_3(1), Z_3(2), Z_3(3))$ when initialising all chains from either of the modes (which we will call mode 1 and mode 2). The reported Rhat values are close to 1 in both cases, indicating convergence. We conclude that we are faced with a rather challenging stochastic simulation problem where the posterior distribution is (at least) bimodal. It turns out that there is a nice interpretation of either of the modes. For each component of $\boldsymbol{Z}_j$ $(j = 1, \ldots, 4)$ we computed its posterior median values and converted that to $\boldsymbol{\lambda}_j := (\lambda_j(1), \lambda_j(2), \lambda_j(3))$. The resulting values are summarised in Table 2. One can see that mode 2 essentially corresponds to collapsing the first two latent states. That is, whether being in injury state 1 or 2, the athlete will virtually always answer "no" to all questions. For mode 1, it stands out that questions 2 and 3 (on modification and performance respectively) distinguish the 3 latent states.

Ideally, one would implement a sampler that traverses both modes corresponding to their likelihood. Parallel tempering (Brooks et al. [2011], Chapter 11) would be an option for this. Here, as the modes are rather distinct, we have used Laplace approximation as a simpler alternative. From this, it turns out that mode 1 has overwhelming posterior probability (near 1). In Section E we provide mathematical details. It demonstrates that a 3 latent state model is to be preferred over a 2 latent state model. It is interesting to note that mode 2 corresponds to the posterior mode reported by `Turing` when no MCMC-sampling is used.

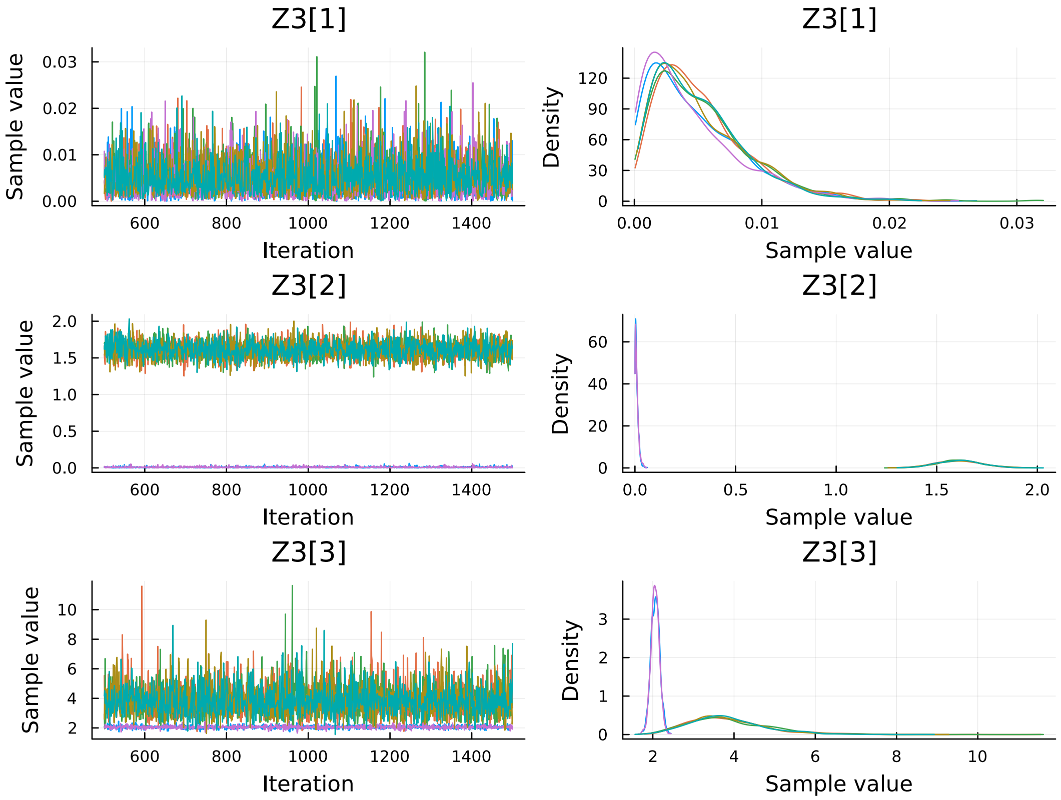

FIGURE 5. Trace- and density plots of iterates for $\boldsymbol{Z}_3$. Burnin was removed. Initialisation of chains by sampling from the prior.

		Mode 1			Mode 2		
		1	2	3	1	2	3
j	1	0.004	0.970	0.991	0.002	0.012	0.982
	2	0.002	0.400	0.964	0.001	0.004	0.543
	3	0.002	**0.541**	0.964	0.001	0.004	0.651
	4	0.011	0.979	0.992	0.006	0.023	0.988

TABLE 2. Median posterior values for mode 1 (left) and mode 2 (right). Each table shows $\lambda_j(k)$, where $j = 1, \ldots, 4$ corresponds to rows and $k = 1, \ldots, 3$ corresponds to columns. For example, under mode 1, the estimated median probability that an athlete in latent class 2 will answer affirmative to question 3 is 0.541.

In the following, we therefore report results obtained by initialising all chains near mode 1. In Figure 9 we visualise for each regression parameter its posterior mean and 2.5%, 25%, 75% and 97.5% posterior percentiles. A similar visualisation is shown in Figure 8 for all parameters appearing in the response probabilities.

C.1. **Interpretation of the results.**

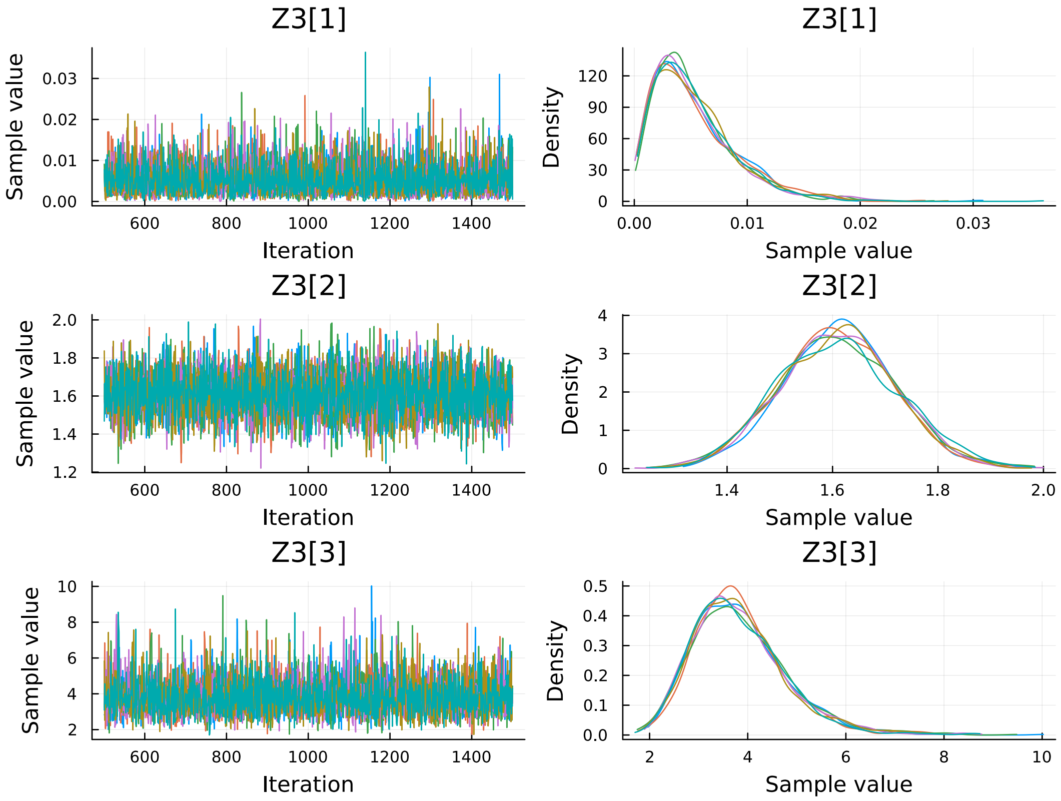

Figure 6. Trace- and density plots of iterates for $\boldsymbol{Z}_3$. Burnin was removed. Initialisation of chains by sampling from mode 1.

(1) As we have standardised the covariates, setting $\boldsymbol{x}_{it}$ to the zero vector yields transition probabilities when each of the covariates is in fact set to its average (over all athletes). Then, the posterior mean estimates for $\boldsymbol{\gamma}_{ud}$ $(u \neq d)$ can be substituted in the transition probability matrix displayed in Equation (2). This gives

$$\begin{bmatrix} 0.89 & 0.11 & 0.00 \\ 0.38 & 0.55 & 0.07 \\ 0.00 & 0.24 & 0.76 \end{bmatrix}. \tag{6}$$

We view these probabilities as a baseline as these are the transition probabilities for an athlete doing average sport-specific activity and strength-training.

(2) Now for each off-diagonal element, we can assess how its value if influenced by the covariate vector. For example, element $(2, 3)$, which is equal to 0.07 depends on the inner-product of $\tilde{\boldsymbol{x}}_{it}$ and $\boldsymbol{\gamma}_{23}$. In Figure 9 all coefficients ending at "[1]", "[2]" and "[3]" relate to the model's intercept, the covariate sport and the covariate strength respectively. Clearly, all credible intervals for the intercept do not contain zero. From all other coefficients, the only coefficients for which the 50% credible interval does not contain zero are $\gamma 12[2]$, $\gamma 12[3]$ and $\gamma 32[3]$. From this, we cautiously conclude (note that there is considerable uncertainty, most likely due to small sample size):

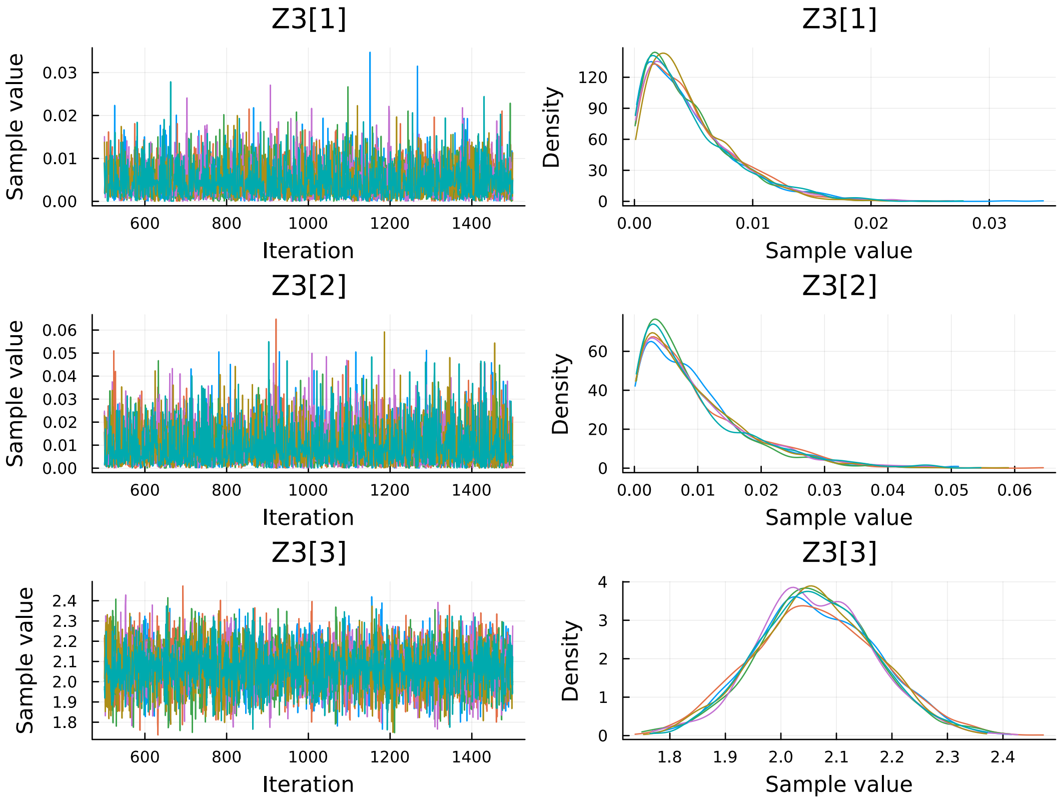

FIGURE 7. Trace- and density plots of iterates for $\boldsymbol{Z}_3$. Burnin was removed. Initialisation of chains by sampling from mode 2.

(a) For the transition from state 1 to state 2, the effects of both sport and strength are negative: this means that increasing these will reduce the probability of transitioning.

(b) For the transition from state 3 to 2, the effect of strength is positive: this means that increasing strength will reduce the probability of transitioning.

C.2. **Future scenario predictions.** We consider athletes starting from either of the initial states combined with either of the following trainingschedules:

- Low intensity where we consider 1 hour spent on sport-specific activity both and 2 hours on strength training on each of the 14 days. On standardised scale, this corresponds to $\boldsymbol{x}_{it} = [-1.80, -0.68]$.
- Average intensity where we set $\boldsymbol{x}_{it} = [0, 0]$ for all 14 days.
- High intensity where we set $\boldsymbol{x}_{it} = [1, 1]$, which means both sport-specific activity and strength training are at the average plus one times the standard deviation in the dataset (on all 14 days). This corresponds to 15.33 hours of sport-specific activity and 5.57 hours of strength training.

We forward simulated the latent states for 14 weeks for all posterior draws of the regression coefficients (in total 6000, as 6 chains ran for 1000 iterations each). In Figure 10 we show the (marginal) distribution over each of the three states at each time.

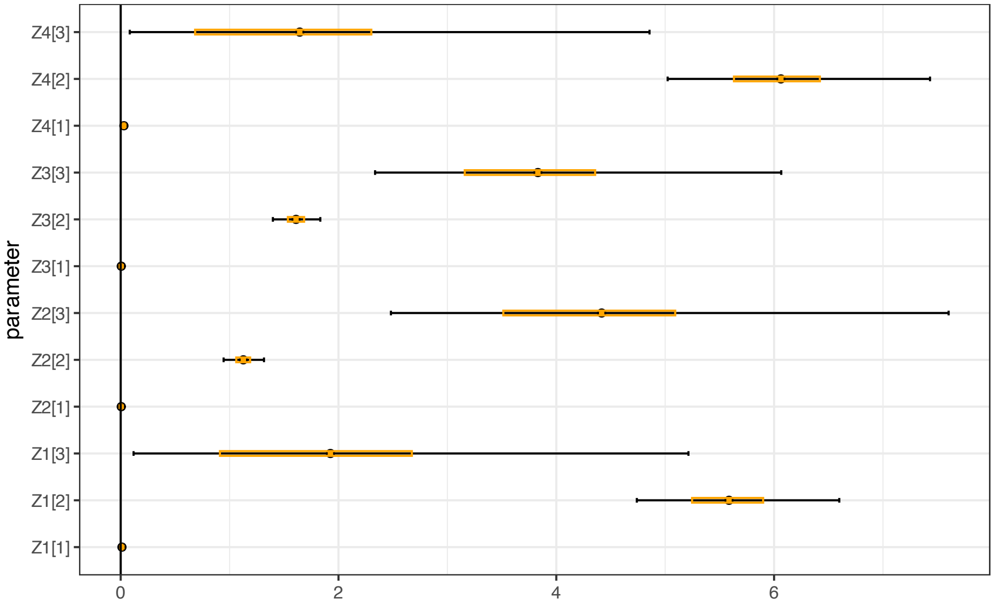

Figure 8. Summary of posterior distribution for all parameters appearing in the response probabilities. The circle depicts the posterior mean. The orange bar connects the 25% and 75% posterior percentiles, while the black bar connect the 2.5% and $97,5\%$ posterior percentiles.

Appendix D. Details on implementation of Bayesian method

D.1. **Implementation using probabilistic programming.** We use Hamiltonian Monte Carlo to sample from this distribution (more specifically, the No-U-Turn-Sampler). This can be conveniently done in the probabilistic programming language `Turing`.

The basic implementation reads as follows:

```
@model function logtarget(𝒪s, p)
    σ ~ Exponential(1.0)

    γ12 ~ filldist(Normal(0.0, σ), p.DIM_COVARIATES)
    γ23 ~ filldist(Normal(0.0, σ), p.DIM_COVARIATES)
    γ21 ~ filldist(Normal(0.0, σ), p.DIM_COVARIATES)
    γ32 ~ filldist(Normal(0.0, σ), p.DIM_COVARIATES)

    Z1 ~ filldist(Exponential(), p.NUM_HIDDENSTATES)
    Z2 ~ filldist(Exponential(), p.NUM_HIDDENSTATES)
    Z3 ~ filldist(Exponential(), p.NUM_HIDDENSTATES)
    Z4 ~ filldist(Exponential(), p.NUM_HIDDENSTATES)

    θ = ComponentArray(γ12 = γ12, γ23 = γ23, γ21 = γ21, γ32 = γ32,
                       Z1=Z1, Z2=Z2, Z3=Z3, Z4=Z4)

    Turing.@addlogprob! loglik(θ, 𝒪s, p)
end
```

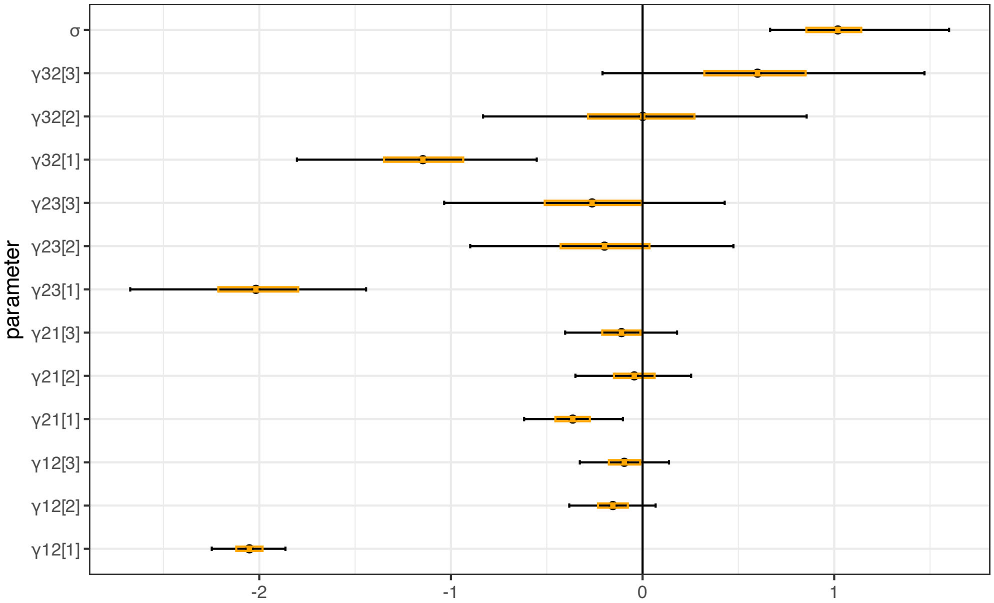

FIGURE 9. Summary of posterior distribution for regression parameters. The circle depicts the posterior mean. The orange bar connects the 25% and 75% posterior percentiles, while the black bar connect the 2.5% and $97,5\%$ posterior percentiles.

```
model = logtarget(𝒪s, p)

chain = sample(model, Turing.NUTS(), MCMCThreads(), 1000, 6)
```

In the first part of the model definition the prior specification is given. The observations are in the data-structure $\mathcal{O}s$ and all that is needed is a function that computes the loglikelihood efficiently. We provide details on this computation in Section D.2. This function, called `loglik`, needs to implemented such that automatic-differentiation libraries can operate on it to compute the gradient of the log posterior density. `p` contains the number of covariates, the number of hidden states and number of questions. Once the model has been specified, MCMC-sampling can be carried out to draw from the posterior.

Remark 1. Chapter 29.4.4 in [Murphy, 2023] considers Bayesian Hidden Markov Models and remarks that a Gibbs sampler that alternates sampling from the smoothing distribution and updating the parameter θ may suffer from bad mixing due to high correlation between the latent path and θ. Here, we follow what he calls "collapsed" inference, where the latent states of each person have been integrated out.

D.2. **Recursive likelihood computation.** It is well known that the likelihood can be computed efficiently in a recursive way. Here we propose to use the backward information filter, which can be viewed as a message passing algorithm. This is well known in the literature, see e.g. [Cappé et al., 2005] and [Van der Meulen, 2022]. Below, we denote the

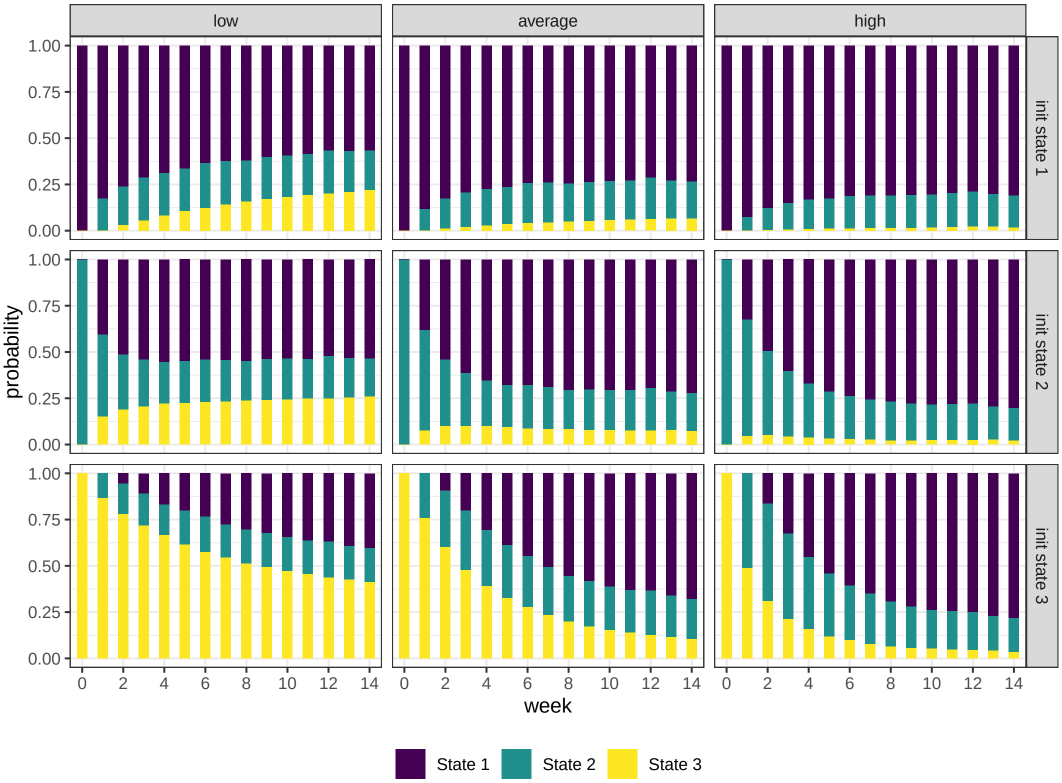

FIGURE 10. Risk trajectories under three different starting configurations and three different training scenarios. Each bar shows the predicted (marginal) probabilities (obtained from the method outlined in Section B.2.5) for being in either of the 3 states during 14 weeks, where week "0" represents the initial state of the athlete.

entrywise product of two vectors by $\odot$: for $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$, $\boldsymbol{a} \odot \boldsymbol{b} = (a_1 b_1, \dots, a_k b_k)$. Let $\mathbf{1} \in \mathbb{R}^3$ denote the vector with all elements equal to 1.

To reduce notational overhead, first assume just one subject, with responses $y_1, \dots, y_T$, where $y_t = (y_{t1}, \dots, y_{t4})$, and latent process $u_1, \dots, u_T$.

Define $u \mapsto h_t(u) = \mathrm{P}(\boldsymbol{Y}_t = \boldsymbol{y}_t, \dots, \boldsymbol{Y}_T = \boldsymbol{y}_T \mid U_t = u)$. As $u \in \{1, \dots, k\}$ this map can be identified with the vector $\boldsymbol{h}_t = (h_t(1), \dots, h_t(k))$. The *backward information filter* consists of the following steps:

- for $t = 1, \dots, T$, let

$$\boldsymbol{g}_{tj} = \begin{cases} \boldsymbol{\lambda}_j & \text{if} \quad y_{tj} = 1 \\ \mathbf{1} - \boldsymbol{\lambda}_j & \text{if} \quad y_{tj} = 0 \end{cases}$$

  and set $\boldsymbol{g}_t = \odot_{j=1}^{J} \boldsymbol{g}_{tj}$;
- set $\boldsymbol{h}_T = \boldsymbol{g}_T$ and

$$\boldsymbol{h}_t = \boldsymbol{g}_t \odot (\Pi_{i,t+1} \boldsymbol{h}_{t+1}), \qquad t = T-1, \dots, 1; \tag{7}$$

- set $h_0 = \Pi_1 \boldsymbol{h}_1$, where Π_1 is the prior on the initial latent state.

The output of this scheme, h_0 is the likelihood. Notationally, we have suppressed any dependence on the parameter vector $\boldsymbol{\theta}$ which under model specification described in the section 5.1. is given by the $\boldsymbol{\theta}$ obtained by concatenating γ_{12}, γ_{21}, γ_{23}, γ_{32} and $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_J$. If we would do so, then indeed $\boldsymbol{\theta} \mapsto h_0(\boldsymbol{\theta})$ is the likelihood function. The extension to multiple subjects/participants is straightforward, we run the backward information filter for each subject and multiply the resulting likelihoods.

If an element in y_{tj} is missing, we can simply set g_{tj} equal to a vector of length k containing only ones. If a covariate vector x_t is missing, we need to specify Π_t separately. Here, we have chose to set Π_t equal to the identity matrix in that case, meaning no latent state-transition at times of a missing covariate.

Remark 2. Direct implementation of the scheme in (7) is numerically unstable. Instead, each time h_t is computed we normalise it, i.e. we divide each element in the vector by the sum of all elements. If we denote this sum by $S_{\boldsymbol{\theta}}(h_t)$, then it follows that

$$\log L(\boldsymbol{\theta}) = \log h_0(\boldsymbol{\theta}) + \sum_{t=1}^{T} \log S_{\boldsymbol{\theta}}(\boldsymbol{h}_t). \tag{8}$$

This avoids numerical underflow problems.

Appendix E. Laplace approximation to approximate the posterior probability of each of the modes

Suppose we have a multimodal distribution with an unnormalized density function $\tilde{p}(\mathbf{x})$, such that the true density is $p(\mathbf{x}) = c^{-1}\tilde{p}(\mathbf{x})$, where c is an unknown normalisation constant. Assume that the locations of K modes, $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_K^*$, are known and that the Hessian matrix of $-\log \tilde{p}(\mathbf{x})$ can be evaluated at each mode. Using Laplace approximation we can estimate the probability mass associated to each mode.

E.1. **Laplace Approximation at a Single Mode.** For a single mode at $\mathbf{x}_k^*$, the Laplace approximation constructs a Gaussian distribution $q_k(\mathbf{x})$ that matches the local curvature of the mode. The steps for mode k are:

(1) Compute the Hessian matrix of the negative log-probability at the mode:

$$\mathbf{H}_k = -\nabla^2 \log \tilde{p}(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_k^*}$$

This matrix $\mathbf{H}_k$ is positive definite if $\mathbf{x}_k^*$ is a local maximum.

(2) The Laplace approximation for the region around mode k is the Gaussian distribution:

$$q_k(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \mathbf{x}_k^*, \mathbf{H}_k^{-1}) = \frac{|\mathbf{H}_k|^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_k^*)^T \mathbf{H}_k (\mathbf{x}-\mathbf{x}_k^*)\right)$$

E.2. **Approximating the full multimodal distribution.** The full distribution is approximated as a mixture of these Gaussians:

$$p(\mathbf{x}) \approx \sum_{k=1}^{K} w_k \, q_k(\mathbf{x})$$

where w_k is the weight of mode k, with $\sum_{k=1}^{K} w_k = 1$. The weight w_k is proportional to the integral of $p(\mathbf{x})$ over the region of mode k. Using the Laplace approximation, this integral is estimated as:

$$\int_{\text{mode } k} p(\mathbf{x})d\mathbf{x} \approx \int q_k(\mathbf{x})d\mathbf{x} \cdot \frac{\tilde{p}(\mathbf{x}_k^*)}{|\mathbf{H}_k|^{1/2}} \cdot \frac{(2\pi)^{d/2}}{c}$$

Thus, approximately, $w_k \propto \tilde{p}(\mathbf{x}_k^*) \cdot |\mathbf{H}_k|^{-1/2}$. The unknown constant c and the factor $(2\pi)^{d/2}$ cancel out when the weights are normalized. Thus,

$$\mathbb{P}(\text{mode } k) = w_k = \frac{\tilde{p}(\mathbf{x}_k^*)\,|\mathbf{H}_k|^{-1/2}}{\sum_{j=1}^{K}\tilde{p}(\mathbf{x}_j^*)\,|\mathbf{H}_j|^{-1/2}}$$

References

F. Bartolucci, M. Lupparelli, and G. E. Montanari. Latent markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *The Annals of Applied Statistics*, pages 611–636, 2009.

F. Bartolucci, S. Pandolfi, and F. Pennoni. Lmest: An r package for latent markov models for longitudinal categorical data. *Journal of Statistical Software*, 81:1–38, 2017.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. URL https://doi.org/10.1137/141000671. Publisher: SIAM.

S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, Boca Raton, FL, 2011. ISBN 978-1-4200-7941-8.

O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

H. Ge, K. Xu, and Z. Ghahramani. Turing: a language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1682–1690, 2018. URL http://proceedings.mlr.press/v84/ge18b.html.

M. D. Hoffman, A. Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

I. O. C. Injury, I. E. C. Group, R. Bahr, B. Clarsen, W. Derman, J. Dvorak, C. A. Emery, C. F. Finch, M. Hägglund, A. Junge, S. Kemp, et al. International olympic committee consensus statement: methods for recording and reporting of epidemiological data on injury and illness in sports 2020 (including the strobe extension for sports injury and illness surveillance (strobe-siis)). *Orthopaedic journal of sports medicine*, 8(2): 2325967120902908, 2020.

B. Lambert. A student's guide to bayesian statistics. 2018.

K. P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL http://probml.github.io/book2.

F. Van der Meulen. Introduction to automatic backward filtering forward guiding. *arXiv preprint arXiv:2203.04155*, 2022.

E. Verhagen, M. Lang, R. Watson, and M. Moen. Injuries and illness in olympic level water polo athletes–a three-season prospective study. *Dtsch Z Sportmed*, 72:195–202, 2021.

Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands
Email address: f.h.van.der.meulen@vu.nl

Amsterdam UMC, The Netherlands
Email address: l.vandergraaff@amsterdamumc.nl

Amsterdam Collaboration on Health & Safety in Sports, Department of Public and Occupational Health, Amsterdam Movement Sciences, Amsterdam UMC, The Netherlands
Email address: Evert.Verhagen@uefa.ch