
Individuality in the hive - Learning to embed lifetime social behavior of honey bees

Benjamin Wild

Department of Computer Science
Freie Universität Berlin
b.w@fu-berlin.de

David M Dormagen

Department of Computer Science
Freie Universität Berlin
david.dormagen@fu-berlin.de

Michael L Smith

Department of Collective Behaviour
Max Planck Institute of Animal Behavior
msmith@ab.mpg.de

Tim Landgraf

Department of Computer Science
Freie Universität Berlin
tim.landgraf@fu-berlin.de

Abstract

Honey bees are a popular model for complex social systems, in which global behavior emerges from the actions and interactions of thousands of individuals. While the average life of a bee is organized as a sequence of tasks roughly determined by age, there is substantial variation at the individual level. Using a unique dataset containing lifetime trajectories of all individuals over multiple generations in two honey bee colonies, we propose a new temporal matrix factorization model that jointly learns the average developmental path and structured variations of individuals in the social network over their entire lives. Our method yields inherently interpretable embeddings that are biologically plausible and consistent over time, which allow one to compare individuals regardless of when, or in which colony, they lived. Our method provides a quantitative framework for understanding behavioral heterogeneity in complex social systems applicable in fields such as behavioral biology, social sciences, neuroscience, and information science.

1 Introduction

Animals living in large groups often coordinate their behaviors, resulting in emergent properties at the group level, from flocking birds to democratic elections. In most animal groups, the role an individual plays in this process is thought to be reflected in the way it interacts with group members. Technological advances have made it possible to track all individuals and their interactions in animal societies, ranging from social insects to primate groups (Mersch et al., 2013; Gernat et al., 2018; Mathis et al., 2018; Graving et al., 2019; Pereira et al., 2019). These datasets have unprecedented scale and complexity, but understanding these data has emerged as a new and challenging problem in itself (Pinter-Wollman et al., 2014; Krause et al., 2015; Brask et al., 2020).

A popular approach to understand high-dimensional data is to learn semantic embeddings (Frome et al., 2013; Asgari and Mofrad, 2015; Camacho-Collados and Pilehvar, 2018; Nelson et al., 2019). Such embeddings can be learned without supervision, are interpretable, and are useful for accomplishing downstream tasks. Individuals in animal societies can be described with semantic embeddings extracted from social interaction networks using matrix factorization methods. For example, in symmetric nonnegative matrix factorization (SymNMF), the dot products of any two animals' factor vectors reconstruct the interaction matrix (Wang et al. (2011); Shi et al. (2015)). If the embeddings allow us to predict relevant behavioral properties, they serve our understanding as *semantic* representations. In temporal settings, i.e. when the interaction matrices change over time, there is no

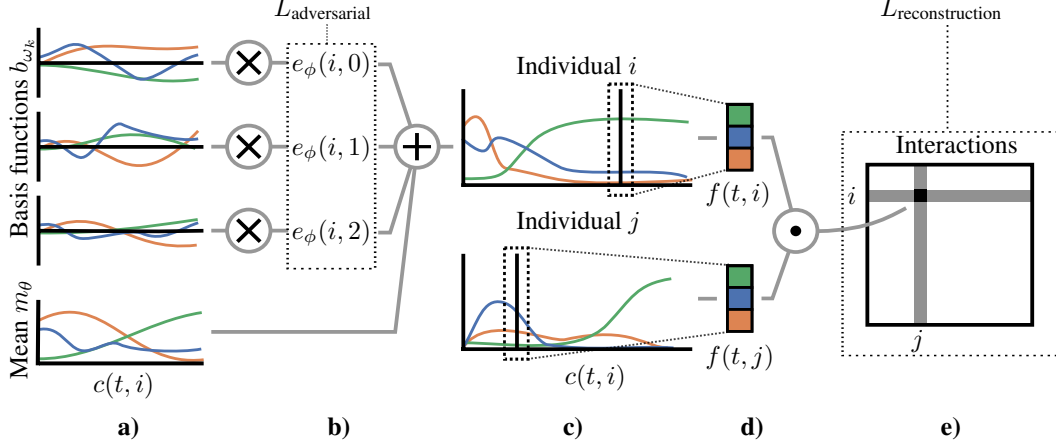


Figure 1: Overview of the method: We learn a parametric function describing the average life trajectory and a set of basis functions of individual variation (a). For each individual, a learned embedding captures to what extent the lifetime can be described by the basis functions (b and c). At each time point, factors can be extracted from the individual lifetime trajectories (d) to reconstruct the interaction affinity between individuals (e). Note that the lifetime trajectories are functions of the individuals’ ages, while interactions can occur at any time point.

straightforward extension of this algorithm. The interaction matrices at different time points can be factorized individually, but there is no guarantee that the embeddings stay semantically consistent over time.

We approach this question using honey bees, a popular model system for studying individual and collective behavior (Elekovich and Roberts, 2005). Bees allocate tasks across thousands of individuals without central control, using an age-based system: young bees care for brood, middle-aged bees perform within-nest labor, and old bees forage outside (Seeley, 1982; Johnson, 2010). We propose to exploit this age-based structure as an inductive bias by jointly learning two meaningful representations of honey bee social behavior: a vector representation of the individuals’ functional position in the social network on a given day, and a parametric representation of their lifetime trajectories through this functional embedding. We show that these representations can be learned in an unsupervised fashion, using only interaction matrices of the individuals over time. We analyze a dataset obtained by tracking thousands of individually marked honey bees in two colonies, at high temporal and spatial resolution over a total of 155 days, covering entire lifespans and multiple generations.

2 Methods

2.1 Temporal NMF algorithm

SymNMF factorizes a matrix $\mathbf{A} \in \mathbb{R}_+^{N \times N}$ such that it can be approximated by the product $\mathbf{F}\mathbf{F}^T$, where $\mathbf{F} \in \mathbb{R}_+^{N \times M}$ and $M \ll N$:

$$\hat{\mathbf{F}} = \underset{\mathbf{F} \geq 0}{\operatorname{argmin}} \left\| \mathbf{A} - \mathbf{F}\mathbf{F}^T \right\|^2 \quad \mathbf{A}_{i,j} \approx \mathbf{f}(i) \cdot \mathbf{f}(j)^T \quad \mathbf{f}(i) = \mathbf{F}_{i,:} \quad \mathbf{f}(i) \in \mathbb{R}_+^M \quad (1)$$

Here we present a novel temporal NMF algorithm (*TNMF*) which extends SymNMF to temporal settings in which $\mathbf{A} \in \mathbb{R}_+^{T \times N \times N}$ changes over time t . We assume that the entities $i \in \{0, 1, \dots, N\}$ follow to some extent a common trajectory depending on an observable property (for example the age of an individual). We represent an entity at a specific point in time t using a factor vector $\mathbf{f}^+(t, i) \in \mathbb{R}_+^M$ such that

$$\hat{\mathbf{A}}_{t,i,j} = \mathbf{f}^+(t, i) \cdot \mathbf{f}^+(t, j)^T \quad \hat{\mathbf{A}} \in \mathbb{R}_+^{T \times N \times N} \quad \mathbf{f}^+(t, i) \in \mathbb{R}_+^M \quad (2)$$

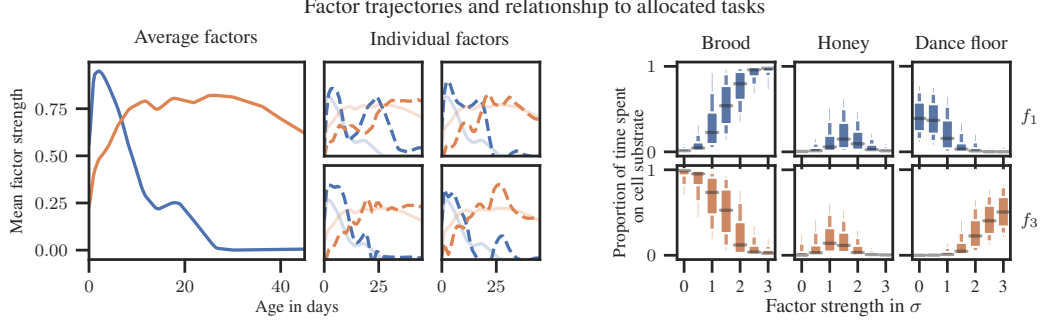


Figure 2: **Left:** Mean lifetime trajectories according to m_θ . The model learns a sparse representation of the functional position of the individuals in the social network. f_1 mostly corresponds to young bees, and f_3 predominantly describes middle aged and older bees. Only factors with a mean magnitude of at least 0.01 are shown. Even though the model uses only these two factors, it is still expressive due to individual variability, as can be seen in randomly sampled individuals’ lifetime trajectories. **Right:** The individual factors f^+ and the proportion of time the individuals spent on different nest substrates. This mapping can be used to interpret the biological meaning of the factors. For example f_1 , which is associated with young bees (see Figure 2), correlates with time spent in the brood area. This is biologically consistent, because young bees tend to care for brood.

In contrast to SymNMF, we do not directly factorize \mathbf{A}_t to find the optimal factors that reconstruct the matrices. Instead, we decompose the problem into learning a common trajectory of factors \mathbf{m}_θ and structured variations from this trajectory $\mathbf{o}_{\phi,\omega}$:

$$\mathbf{f}(t, i) = \mathbf{m}_\theta(\mathbf{c}(t, i)) + \mathbf{o}_{\phi,\omega}(i, \mathbf{c}(t, i)) \quad \mathbf{f}^+(t, i) = \max(0, \mathbf{f}(t, i)) \quad (3)$$

$$\mathbf{c} : \mathbb{N}^{T \times N} \rightarrow \mathbb{N} \quad m_\theta : \mathbb{N} \rightarrow \mathbb{R}_+^M \quad o_{\phi,\omega} : \mathbb{N}^{N \times T} \rightarrow \mathbb{R}^M$$

Note that in the simplest case $c(t, i) = t$, i.e. the trajectory of all entities is aligned with the temporal dimension t of \mathbf{A} . In social networks $c(t, i)$ could map to the age of individual i at time t .

We obtain $\mathbf{o}_{\phi,\omega}$ from a learned set of *individuality basis functions* b_{ω_k} (shared among all entities) that define a coordinate system of possible individual variations and *individuality embeddings* \mathbf{e}_ϕ which capture to what extent each basis function applies to an entity:

$$\mathbf{o}_{\phi,\omega}(i, \mathbf{c}(t, i)) = \sum_{k=0}^{N_k} \mathbf{e}_\phi(i, k) \cdot b_{\omega_k}(\mathbf{c}(t, i)) \quad \mathbf{e}_\phi : \mathbb{N}^{N \times N_k} \rightarrow \mathbb{R}_+^M \quad b_{\omega_k} : \mathbb{N}^T \rightarrow \mathbb{R} \quad (4)$$

where N_k is the number of learned basis functions.

We implement the functions $\mathbf{m}_\theta, b_{\omega_j}$ with small fully connected neural networks with non-linearities and several hidden layers. The parameters of these functions and the entities’ embeddings \mathbf{e}_ϕ are learned jointly using minibatch stochastic gradient descent:

$$\hat{\theta}, \hat{\omega}, \hat{\phi} = \operatorname{argmin}_{\theta, \omega, \phi} \left\| \mathbf{A} - \hat{\mathbf{A}} \right\|^2 + R_{\mathbf{f}^-} \quad R_{\mathbf{f}^-} = N^{-1} T^{-1} \sum_{i=0}^N \sum_{t=0}^T \max(0, -\mathbf{f}(t, i)) \quad (5)$$

where the regularisation term $R_{\mathbf{f}^-}$ is added to encourage the model to learn positive values for all possible inputs because the gradient of the max operation in equation 3 is zero for negative values.

2.2 Regularization

We use several regularization terms to increase the interpretability and semantic consistency of the model. See appendix A.1 for details.

2.3 Data

Two colonies of honey bees were continuously recorded over a total of 155 days. Each individual was manually tagged at emergence, so the date of birth is known for each bee. Locations and identities of all honey bees (N=9286) were extracted from the raw images and used to construct daily aggregated temporal interaction networks based on spatial proximity. We embed all individuals from these two datasets into a common space of *individuality embeddings* e_ϕ and *factor vectors* f^+ .

Table 1: Honey bee datasets

| Dataset | Days | Individuals | Interaction pairs |
|---------|------|-------------|-------------------|
| BN16 | 56 | 2443 | 43 174 748 |
| BN19 | 99 | 6843 | 167 366 381 |

3 Results

We implemented the model using PyTorch (Paszke et al., 2019) and trained it in minibatches of 128 individuals for 100 000 iterations with the Adam optimizer (Kingma and Ba, 2015). See appendix A.3 for the architecture of the learned functions, a precise description of the regularization losses and further hyperparameters. The code is publicly available ¹.

The model learns a sparse representation of the developmental trajectory of a honey bee in the space of social interactions. Only two factors are effectively used (they exceed the threshold value of 0.01). These factors show a clear trend over the life of a bee, indicating that the model captures the temporal aspects of the honey bee division of labor (See Figure 2).

Most individuals can predominantly be described by a single basis function. That means that while each honey bee can collect a unique set of experiences, most can be described with a few *individuality blueprints* which are consistent across cohorts and colonies. In the context of honey bee division of labor, the basis functions are interpretable because the factors correspond to different task groups.

For example, ω_{11} (accounting for $\approx 24.5\%$ of the individuals) describes workers that occupy nursing tasks over their lives, whereas ω_6 (accounting for $\approx 9.9\%$ of the individuals) describes workers that associate with in-hive duties such as honey storage later, but also stronger than average. As the lifetime embeddings e_ϕ only scale the magnitude of the basis functions, they can be interpreted in the same way. Individual lifetime trajectories in the factor space can be computed based on the mean lifetime trajectories (m_θ), individuality basis functions (b_ω) and lifetime embeddings (e_ϕ).

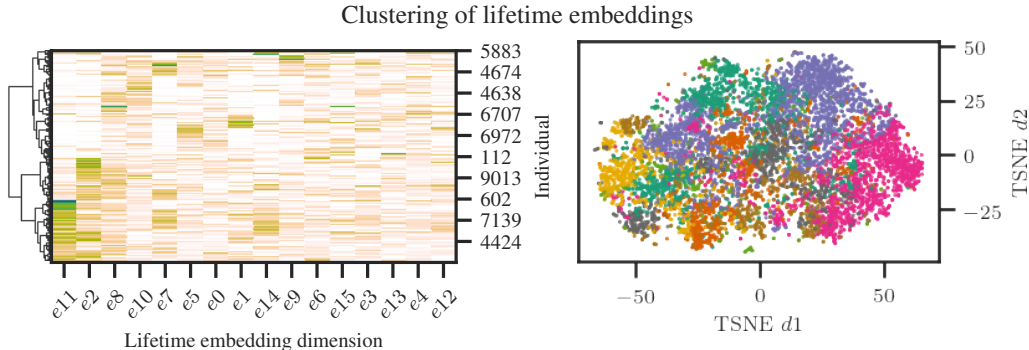


Figure 3: Clustering of lifetime embeddings: Most individuals strongly correspond to a single individuality basis function, making it easy to cluster their lifetime social behavior (i.e. each individual has a high value in a single dimension for their lifetime embedding). Because each cluster is strongly associated with a specific individuality basis function, and because each basis function is interpretable (Figure 2), these blueprints of lifetime development can also be intuitively understood and compared.

¹anonymous.4open.science/r/b2b7e2fc-aa04-4cf8-85c7-646d8dc46400

4 Conclusion

Temporal NMF factorizes temporal matrices with overlapping and even disjointed communities by learning a common embedding of the lifetime development of the individual entities. In the context of honey bees, this embedding is biologically meaningful, consistent over time, and shows that interaction patterns follow a common lifetime trajectory. Differences from the mean are described in a coordinate system of individual variability. The basis functions are interpretable with respect to the division of labor within colonies, and offer a valuable tool to understand and quantify the influence of experimental manipulations on an individual. While we applied our method to honey bees as an exemplary system with many individuals that exhibit an entangled, non-overlapping social structure, our method can be applied to any setting in which some interaction structure follows a general pattern over an observable (such as time) to detect structured deviations at the individual level.

References

- Ehsaneddin Asgari and Mohammad R. K. Mofrad. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE*, 10(11):e0141287, November 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0141287. 00294 Publisher: Public Library of Science.
- Josefine Bohr Brask, Samuel Ellis, and Darren P. Croft. Animal social networks – an introduction for complex systems scientists. *arXiv:2005.09598 [physics, q-bio]*, May 2020. arXiv: 2005.09598.
- Jose Camacho-Collados and Mohammad Taher Pilehvar. From word to sense embeddings: a survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63(1):743–788, September 2018. ISSN 1076-9757. doi: 10.1613/jair.1.11259. 00086.
- Michelle M. Elekonich and Stephen P. Roberts. Honey bees as a model for understanding mechanisms of life history transitions. *Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology*, 141(4):362–371, August 2005. ISSN 1095-6433. doi: 10.1016/j.cbpb.2005.04.014. 00050.
- Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc\textquotesingle Aurelio Ranzato, and Tomas Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc., 2013. 01437.
- Tim Gernat, Vikyath D. Rao, Martin Middendorf, Harry Dankowicz, Nigel Goldenfeld, and Gene E. Robinson. Automated monitoring of behavior reveals bursty interaction patterns and rapid spreading dynamics in honeybee social networks. *Proceedings of the National Academy of Sciences*, 115(7):1433–1438, February 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1713568115. 00000.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.
- Jacob M Graving, Daniel Chae, Hemal Naik, Liang Li, Benjamin Koger, Blair R Costelloe, and Iain D Couzin. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife*, 8: e47994, October 2019. ISSN 2050-084X. doi: 10.7554/eLife.47994. Publisher: eLife Sciences Publications, Ltd.
- Brian R. Johnson. Division of labor in honeybees: form, function, and proximate mechanisms. *Behavioral Ecology and Sociobiology*, 64(3):305–316, January 2010. ISSN 0340-5443. doi: 10.1007/s00265-009-0874-7. 00077.
- Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 43006.
- Jens Krause, Richard James, Daniel W. Franks, and Darren P. Croft. *Animal Social Networks*. Oxford University Press, 2015. ISBN 978-0-19-967904-1. Google-Books-ID: cuAUDAQAQBAJ.
- Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, September 2018. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-018-0209-y. 00001.
- Danielle P. Mersch, Alessandro Crespi, and Laurent Keller. Tracking individuals shows spatial fidelity is a key regulator of ant social organization. *Science*, 340(6136):1090–1093, 2013. tex.ids: mersch_tracking_2013-1 publisher: American Association for the Advancement of Science.

- Walter Nelson, Marinka Zitnik, Bo Wang, Jure Leskovec, Anna Goldenberg, and Roded Sharan. To Embed or Not: Network Embedding as a Paradigm in Computational Biology. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00381. Publisher: Frontiers.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019. 00646.
- Talmo D. Pereira, Diego E. Aldarondo, Lindsay Willmore, Mikhail Kislin, Samuel S.-H. Wang, Mala Murthy, and Joshua W. Shaevitz. Fast animal pose estimation using deep neural networks. *Nature Methods*, 16(1): 117–125, January 2019. ISSN 1548-7105. doi: 10.1038/s41592-018-0234-5. Number: 1 Publisher: Nature Publishing Group.
- Noa Pinter-Wollman, E. A. Hobson, J. E. Smith, A. J. Edelman, D. Shizuka, S. de Silva, J. S. Waters, S. D. Prager, T. Sasaki, G. Wittmyer, J. Fewell, and D. B. McDonald. The dynamics of animal social networks: analytical, conceptual, and theoretical advances. *Behavioral Ecology*, 25(2):242–255, March 2014. ISSN 1045-2249, 1465-7279. doi: 10.1093/beheco/art047.
- Thomas D. Seeley. Adaptive significance of the age polyethism schedule in honeybee colonies. *Behavioral Ecology and Sociobiology*, 11(4):287–293, December 1982. ISSN 1432-0762. doi: 10.1007/BF00299306. 00000.
- Xiaohua Shi, Hongtao Lu, Yangchen He, and Shan He. Community detection in social network with pairwise constrained symmetric non-negative matrix factorization. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 541–546, August 2015. doi: 10.1145/2808797.2809383. 00023.
- Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery*, 22(3):493–521, May 2011. ISSN 1573-756X. doi: 10.1007/s10618-010-0181-y. 00336.

A Appendix

A.1 Regularization terms

We encourage sparsity in both the number of used factors and individuality basis functions by adding L_1 penalties of the mean absolute magnitude of the factors $\mathbf{f}(t, i)$ and basis functions b_{ω_j} to the objective. We encourage individuals' lifetimes to be represented with a sparse embedding using an L_1 penalty of the learned *individuality embeddings* e_ϕ . We also introduce an adversarial loss term to encourage the model to learn embeddings that are semantically consistent over time, i.e. to only represent two entities that were present in the dataset at different times with different embeddings if this is strictly necessary to factorize the matrices \mathbf{A} . We jointly train a discriminative network $d_\psi(e)$ that tries to classify the time of first occurrence of all entities based on their *individuality embeddings* e_ϕ . The cross-entropy loss of this model is added as a regularization term to equation 5 in a training regime similar to generative adversarial networks (Goodfellow et al., 2014). We found that the model performs well without these regularizations, but we are willing to accept a higher methodological complexity for the honey bee data if this allows us to achieve a better interpretability of the results. See appendix A.1 for more details.

$$R_{\text{embeddings}} = \lambda_{\text{embeddings}} N_i^{-1} \sum_{i=0}^{N_i} \sum_{k=0}^{N_k} |e_\phi(i, k)| \quad (6)$$

$$R_f = \lambda_f L_{\text{factors}} N_i^{-1} \sum_{i=0}^{N_i} \sum_{t=0}^{N_t} f^+(t, i) \quad (7)$$

$$R_{\text{basis}} = \lambda_{\text{basis}} N_a^{-1} \sum_{a=0}^{N_a} \sum_{k=0}^{N_k} |b_{\omega_k}(a)| \quad N_a = 60 \quad (8)$$

where N_a can be any number higher than the oldest individual in the dataset at any time.

$$R_{\text{adv}} = \lambda_{\text{adv}} N_i^{-1} \sum_{i=0}^{N_i} \log \left(\frac{\exp(\mathbf{x}_i[c])}{\sum_d^{N_d} \exp(\mathbf{x}_i[c])} \right) \quad (9)$$

where \mathbf{x}_i is a one-hot vector for each individual encoding the day it emerged in the dataset and N_d the number of days in the dataset. E.g., if the data consists of four days and individual 3 emerged on the second day, then $\mathbf{x}_3 = [0, 1, 0, 0]$.

A.2 Network architecture

We use the following neural network architecture for the functions m_θ , b_ω , and d_ψ :

$$\text{Linear}(N_{in}, N_h) \rightarrow \text{LReLU} \rightarrow \underbrace{\text{Linear}(N_h, N_h) \rightarrow \text{LReLU}}_{N_l\text{-times}} \rightarrow \text{Linear}(N_h, N_{out})$$

where *Linear* is an affine transformation $f(x) = Ax + b$ and $\alpha = 0.3$ for the Leaky ReLU activation function. For m_θ and b_ω : $N_{in} = 1$ (the individuals' ages). For m_θ : $N_{out} = N_f$ and for b_ω : $N_{out} = N_f N_k$. For d_ψ : $N_{in} = N_k$ and $N_{out} = N_{\text{labels}}$.

A.3 Hyperparameters

The scaling factors for the regularization losses (see Table 2) were manually selected by increasing each factor until it prevented the model from converging (i.e. the reconstruction loss of the full model $\mathbf{f}^+(t, i)$ did not improve on the age model m_θ). This initial set of hyperparameters was then manually refined such that each regularization loss was still effective (e.g. the factor regularization loss L_f reduced the total number of factors effectively used by the model). Overfitting was not a concern because the model is fitted unsupervised and the goal of the hyperparameter selection was to find a set of parameters that is sparse and interpretable, and not to increase the predictive capabilities of the learned factors.

Table 2: Hyperparameters used in the evaluated models (if not stated otherwise)

| Parameter | Value | Description |
|-------------------------------|---------|--|
| N_l | 3 | Number of hidden layers |
| N_h | 64 | Hidden layer size |
| N_f | 8 | Number of factors |
| N_k | 16 | Number of individuality basis function |
| N_{labels} | 100 | Number of cohorts |
| N_{batch} | 128 | Minibatch size |
| N_{steps} | 100 000 | Number of training iterations |
| λ_f | 0.1 | Factor L_1 regularization |
| λ_{adv} | 0.1 | Factor L_1 regularization |
| λ_{basis} | 0.01 | Basis function L_1 regularization |
| $\lambda_{\text{embeddings}}$ | 0.1 | Embedding L_1 regularization |