Bayesian Matrix Factor Models for Demographic Analysis Across Age and Time

Gregor Zens*

International Institute for Applied Systems Analysis (IIASA)
Wittgenstein Centre for Demography and Global Human Capital (WIC)

[LINK TO FULL PAPER]

October 29, 2025

1. Introduction

Statistical demographic research increasingly relies on high-dimensional mortality, fertility or migration data that encompass multiple small subpopulations observed over both age groups and time. Such fine-grained demographic data can, in principle, be used to produce forecasts and track demographic trends over time, with numerous real-world applications.

However, analyzing such data poses substantial methodological challenges due to the often pronounced heterogeneity across subpopulations, significant stochastic variation in smaller populations, and high-dimensional parameter spaces. As a motivating example, consider Figure 1, showing data on emigration counts in Austrian districts. The subpopulations (stratified by sex and districts) are heterogeneous, exhibit outliers, and are often rather small, making it difficult to distinguish signal from stochastic noise.

In such contexts, making conclusive statements, providing insights into demographic processes, and producing forecasts typically requires model-based solutions. To be useful in this context, models must strike a delicate balance: They must be simultaneously robust enough to withstand stochastic noise and flexible enough to capture 'true' demographic heterogeneity. Additionally, modern demographic modeling frameworks need to remain computationally feasible across a potentially large number of populations, with more and more data sets covering up to hundreds or thousands of subpopulations.

*International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, 2361 Laxenburg, Austria. Mail: zens@iiasa.ac.at.

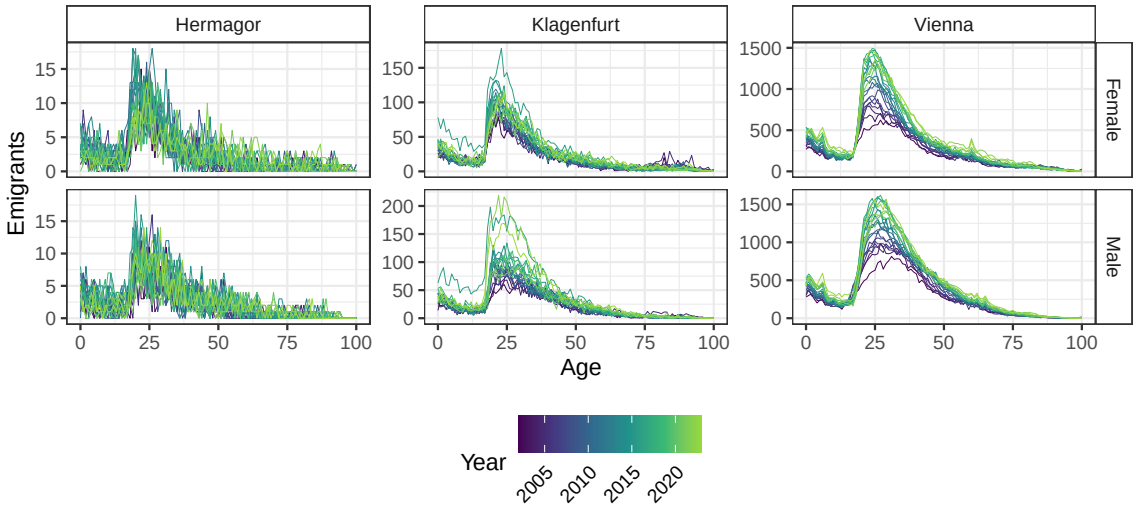

Fig. 1: Emigration counts for eight example subpopulations. Populations are stratified by sex (rows) and four selected Austrian districts (columns). The full dataset covers 190 subpopulations. Data sourced from the Austrian national statistical office.

2. Demographic Factor Models

A promising avenue to address these modeling issues is the use of low-rank representations, which have a long-standing tradition in statistical demography (e.g., Lee and Carter, 1992 and extensions). By exploiting the regularities in demographic age profiles and time trends, principal components analysis, singular value decompositions, factor models, and similar frameworks can provide parsimonious representations that avoid overfitting and reduce the computational burden.

However, classical demographic factorization approaches are often designed to reduce dimensionality *either* exploiting shared variation in the age dimension (as, e.g., in Alexander et al., 2017 or Zens, 2025) *or* shared variation in the time dimension (as, e.g., in multi-population Lee-Carter frameworks such as Li and Lee, 2005). While both strategies have their benefits and can handle moderately sized datasets effectively, they may struggle if the number of subpopulations grows large, or if fine age groups and long time series are considered simultaneously. In particular, when modeling data across hundreds or thousands of subpopulations — each with 100 age groups spanning multiple decades of observations — either factorization perspective alone can result in an extremely large number of parameters, raising computational and overfitting concerns. In this paper, we introduce a Bayesian matrix factorization that allows to adress these issues effectively and efficiently.

3. Main Contributions

Instead of relying on either compressing data in the age *or* time dimension, this paper proposes a matrix factor model that simultaneously summarizes shared age and time variation through two sets of low-dimensional factors. We treat the observed demographic counts (or rates) for each subpopulation i as a matrix of size $(T \times A)$, where T is the number of time points and A the number of age groups. A bilinear factor structure for counts and rates is then introduced, decomposing each $(T \times A)$ matrix into the product of a low-dimensional matrix of common time factors, a subpopulation-specific loading matrix, and a low-dimensional matrix of common age factors. This three-way factorization ensures that the number of population-specific parameters remains manageable even if the overall dataset is very large. Informative priors ensure simultaneous smoothing in the age and time dimensions. An additional latent Gaussian error term is introduced to account for sampling noise and overdispersion in the observed counts. The model also serves as a fully probabilistic forecasting tool for demographic processes, allowing practitioners to easily quantify uncertainty in future demographic trends.

We introduce a straight-forward computational algorithm for fully probabilistic inference and illustrate the model using Austrian district-level emigration data. We show that the model accurately recovers demographic patterns and leads to interpretable inference. In a pseudo-out-of-sample exercise, the proposed approach substantially outperforms classical demographic factorizations in predictive terms, even though it compresses the number of parameters to a fraction of their parameter spaces.

4. Illustration Using High-Dimensional Austrian Emigration Data

To illustrate the approach, we apply the proposed model to data on age-specific emigration counts from the 95 political districts of Austria, covering a 22-year window from 2002 to 2023. These data are stratified by single years of age (0 to 100) and sex, yielding 190 distinct subpopulations in total. On a single core of an standard laptop with AMD Ryzen 5500U (2.1GHz) CPU, and without any advanced computational optimizations, a single MCMC run takes less than one hour. We show that our matrix factor model accurately recovers well-known age patterns resembling, among others, a Rogers-Castro profile as well as highly interpretable time-varying factors, capturing for instance the influx of a large number of refugees from the middle east in 2015/16 and from Ukraine in 2022.

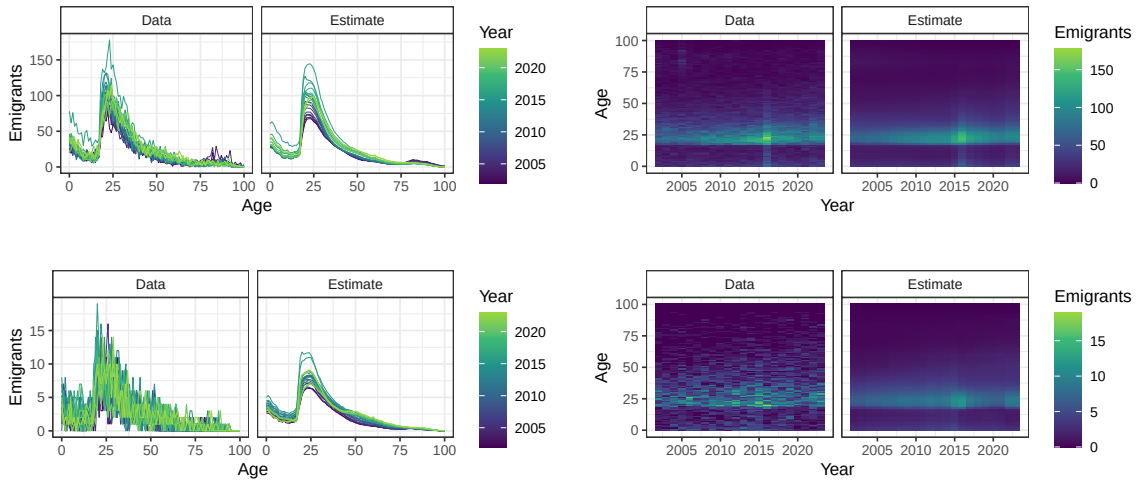

Fig. 2: Observed and fitted emigration counts for two subpopulations. The top row displays data for females emigrating from Klagenfurt, representing a district with a large population and relatively low stochastic noise levels. The bottom row shows data for males emigrating from Hermagor, a sparsely populated rural area with significant stochastic variation in the raw data. The plot pairs in the left and right columns display the same data and estimates from different perspectives.

Fig. 2 presents observed emigrant counts along with fitted values for both a small and a large subpopulation. This demonstrates how the model effectively smoothes out stochastic noise in smaller subpopulations by borrowing information on demographic patterns from other subpopulations, while closely following the patterns observed in more informative subpopulations.

References

Alexander, M., Zagheni, E., and Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography*, 54(6):2025–2041.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American statistical association*, 87(419):659–671.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42:575–594.

Zens, G. (2025). Flexible Bayesian modelling of age-specific counts in many demographic subpopulations. *Journal of the Royal Statistical Society Series A: Statistics in Society*.