Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

BAYESIAN OPTIMIZATION FOR CLINICAL PATHWAY DECOMPOSITION FROM AGGREGATE DATA

William Plumb

Dept. of Computing, Imperial College London, UK

ABSTRACT

Data protection rules often impose anonymization requirements on datasets by means of aggregations that hinder the exact simulation of individual subjects. For example, clinical pathways that disclose medical conditions of patients may typically need to be aggregated to preserve anonymity of the subjects. However, aggregation unavoidably results in biasing the simulation process, for example, by introducing spurious pathways that can skew the simulated trajectories. In this paper, we study this problem and develop approximate decomposition methods that mitigate its impact. Our method is shown to produce from the raw aggregates pathways with higher fidelity than sampling a Markov chain model of the aggregate data, even preserving the same length of the original pathways. We observe a relative increase in average cosine similarity of up to 52% with respect to the true pathways compared with aggregate Markov chain sampling.

1 INTRODUCTION

Healthcare data privacy regulations, like GDPR, require anonymization, often through data aggregation. However, this process can create spurious pathways, compromising the accuracy of clinical pathway simulations. To address this, we introduce the Iterated Bi-level Pathway Decomposition (IBPD) method, which uses Bayesian optimization to more accurately reconstruct patient pathways by incorporating pathway lengths and additional statistical information. Our approach significantly improves the fidelity of simulated pathways, achieving up to a 54% increase in cosine similarity over traditional methods.

2 CLINICAL PATHWAY DECOMPOSITION

Clinical pathways represent a patient's journey through healthcare as sequences of medical events (e.g., diagnostic tests, surgeries). Each pathway r is an ordered sequence of nodes $(n_{r,1},\dots,n_{r,S})$, where $n_{r,s} \in N$ represents a medical event, and all pathways share a common exit event e. Pathways can be summarized by a *counting matrix* C_r, which tracks the number of transitions between events, and an initial vector a_r. Aggregated data over multiple pathways is presented in an *aggregate counting matrix* A and an *aggregate initial vector* a, with the standard deviation matrix S reflecting variability across pathways.

Given the aggregate data (a,A), pathway lengths, and S, the goal is to reconstruct the individual counting matrices $C_1,\dots,C_g$ that represent the original patient pathways. The accuracy of this decomposition is measured using *cosine similarity*, which quantifies the similarity between the reconstructed matrices $\widehat{C}_1,\dots,\widehat{C}_g$ and the true matrices $C_1,\dots,C_g$. To find the best matching between reconstructed and true matrices, we solve an optimization problem that maximizes the average cosine similarity across all pairs.

3 DECOMPOSITION METHOD

Our optimization-based method to decompose the aggregate counting matrix A into individual pathways $\widehat{C} = (\widehat{C}_1,\widehat{C}_2,\dots,\widehat{C}_g)$. The constraints to ensure the aggregate matrix is matched, the pathway lengths are preserved, and a single initial event is chosen for each pathway. We also match the sum of initial vectors to

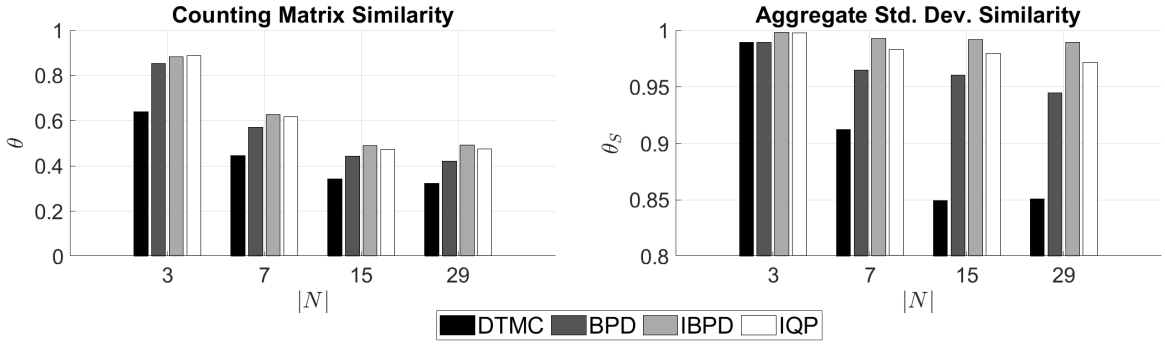

Figure 1: Counting matrix distance and aggregate standard deviation matrix distance results. $|N|$ represents the number of medical events, Θ_S is the cosine similarity of the standard deviation matrices.

the aggregate initial counts. Impose additional constraints that ensure flow balance at each node, meaning all nodes must have equal incoming and outgoing transitions. Enforcing a single transition to the end state.

Using these count constraints, we develop optimization programs for the integer variables $x_r(i,j)$ and $u_r(j)$. The feasible solutions to these Integer Linear Programs (ILPs) will match the aggregate matrix A and pathway lengths l. To improve solution quality, we consider additional consistency with the standard deviation matrix S, requiring a more complex optimizations. Our first method Integer Quadratic Program (IQP), maximizes the cosine similarity $\theta(\widetilde{S},S)$ between the decomposed and true standard deviation matrices through an IQP. To mitigate the difficulty of solving IQPs, we introduce two bi-level optimization strategies.The Bi-level Pathway Decomposition (BPD) uses a two-level optimization. The *inner program* is an ILP with linear constraints and a cost vector c that guides decomposition. The *outer program* optimizes c to align with the standard deviation matrix S. This method ensures feasible solutions are on the convex hull of the feasible set. The Iterated BPD (IBPD) refines BPD by iteratively adjusting the upper bounds of $x_r(i,j)$ to improve feasibility. At each iteration t, the range of variables is restricted to $x_r(i,j) \in \{0,\ldots,\lceil (A(i,j)+t)/g \rceil\}$. This process continues until the best similarity measure θ_S^* is achieved.

4 EXPERIMENTS

We compare our proposed optimization programs (IQP, BPD, IBPD) against a baseline Monte Carlo method using two similarity metrics, Θ and Θ_S for the average cosine similarity to the original counting matrices and the aggregate standard deviation matrices respectively. We use data derived from CRPD GOLD and HES datasets. The original dataset comprises 2712 aggregated pathways related to elective orthopedic surgery, averaged to 6.11 pathways per aggregate. Artificial data is generated with characteristics matching the real dataset, with $g = 6$ pathways and varying medical events ($|N| \in \{3,7,15,29\}$).

Figure 1 summarizes the experimental results. As the number of medical events $|N|$ increases, the cosine similarity generally decreases due to more spurious paths meeting all constraints. Despite this trend, the IQP, BPD, and IBPD methods achieve higher similarity scores for both metrics: counting matrix distance (Θ) and aggregate standard deviation similarity (Θ_S). The IBPD method consistently outperforms others, showing a 16% improvement in Θ_S and a 52% improvement in Θ compared to the DTMC method.

5 CONCLUSION

We address the problem of clinical pathway decomposition from aggregate datasets using three methods: IQP, BPD, and IBPD. The results demonstrate that IBPD offers the best performance in achieving high similarity to original pathways, while BPD provides a balance between accuracy and computational cost. The IQP method, though less effective, allows for more customization. These approaches enhance clinical pathway modeling by enabling data-driven discovery of pathways without manual input, improving simulation accuracy.