Population PME Project Proposal

Robert Zielinski

July 12, 2023

1 Motivation

The PME algorithm is a valuable approach for estimating a low-dimensional parameterization of a manifold in high-dimensional space. The LPME approach extends this to estimate how the embedding map found using PME changes over time. However, both of these methods are limited in the sense that the PME algorithm can only be used for dimension reduction for a single object, and the LPME algorithm can be used for that single object over multiple time points.

There are many cases where it would be useful to allow an estimate from one object to inform an estimate of a similar object, or to compare between estimates of two similar objects. For example, we have focused on the use of the LPME algorithm to estimate the surface of the hippocampus or other subcortical structures from MRI images. The hippocampi of one person are very similar in form to the hippocampi of another person, so it could be useful to allow the parameterization of the hippocampus from one individual to inform the estimation of an embedding map for another person or a larger group of people. This could improve the accuracy of our estimates and allow for consistent parameterizations (meaning that similar parameter values would correspond to similar locations on the hippocampus) across people, which is difficult to achieve given the current approaches.

We may also be interested in using the estimates from the PME and LPME algorithms to compare between similar objects for multiple people. As an example, we can again consider the hippocampus: because people with Alzheimer's disease (AD) tend to experience hippocampal atrophy, we would expect that the hippocampi of someone with AD would have different characteristics than the hippocampi of a cognitively healthy individual who is similar in most other respects. This could be visible at one time point, but the comparison would be even more powerful when viewed over time. Currently, the coefficients of the higher-level spline function estimated by the LPME algorithm would be useful for this purpose, if we had a means of ensuring a consistent parameterization across individuals.

Ultimately, it would be valuable to extend this type of analysis to comparison of multiple groups. For instance, we may be interested in understanding the aggregate change in the hippocampi of cognitively healthy individuals versus that of a group of people with AD. One potentially powerful use case could be in the analysis of clinical trials, where we may be interested in comparing changes in structures between two study groups. A modeling approach that allows for principal manifold estimation over large, potentially heterogeneous groups or populations would address these use cases.

2 Proposed Approach

The modeling approach I am considering right now is influenced by the Latent Trajectory Model in Schulam and Saria 2016. Essentially, rather than using a single spline function to represent the embedding function, I consider using several spline functions, each representing a different level of the model in question. For example, if we were interested in using the ADNI data to model the hippocampus at one time point, we may consider representing the embedding as a sum of three spline functions. One spline would represent the population level, another would represent the subgroup level (for instance, whether someone belongs to the cognitively healthy group or the group with AD), with the spline modeling any differences between the population and each subgroup, and the final spline would represent individual-level variability from the subgroup they belong to. The number of splines used, and their structure, would ultimately be flexible depending on the situation being modeled. However, I believe the general structure could be flexible enough to adapt to a number of distinct scenarios. Because the hierarchical structure is over a large number of spline coefficients that each interact with each other, the additive approach will be helpful for keeping the model flexible, and having coefficients clearly assigned to different groups of interest will improve interpretability.

To fit the model, it will be necessary to consider two separate scenarios: one where we consider only a single time point, and one where we attempt to model change over time. Both cases will

largely resemble the PME algorithm. We begin with the single time point case.

- Initialization: Run PME on data from one observation, then use the estimated manifold to find parameterizations for all other observations.
- Data Reduction: Run HDMDE on each observation individually.
- Fitting: Iterate between estimating spline coefficients in alignment with the model structure determined for the given problem and updating the parameterization for each observation
- Tuning: For each spline function, select the most appropriate smoothing value

The estimation mechanism for the spline coefficients is currently unclear. In Schulam and Saria 2016, the EM algorithm is used to estimate the parameters, so this may be a viable option in this case as well.

The situation where we account for longitudinal change becomes more complicated.

3 Toy Data

We have discussed using the MNIST dataset as a possible example dataset, and I think this would be a good case for it. We could consider trying to find an aggregate estimate of a manifold for the figures "0", "1", "3", "5", and "7", where differences in the shape of the character from the group-wide estimate would be handled by the random-effects type of structure described above. The model structure in this case would have two levels, one for the population level and one for the individual level variations. If we wanted to make the situation more complex, we could also divide images of the figures "2", "4", and possibly "9" into the various methods people use to write them. These different writing methods could then represent subgroups, which would be represented by another level in the model structure.

4 Application

A clear use case that would extend closely from our work on the LPME algorithm would be to develop estimates of the hippocampi of those in ADNI's healthy control, mild cognitive impairment, and AD groups. This would provide a way to test whether the method is capable of producing meaningful results given a real-world dataset that would closely reflect a potential use case.

5 Potential Challenges

Likely the most significant challenge to be addressed in this project will be how computation time will intersect with the ease of estimating the model parameters. Right now, I have seen the most extensive discussion of fitting hierarchical models from a Bayesian perspective, and there are software packages that may allow us to fit the model with relatively minor adjustments. However, given the iterative nature of the algorithm and the size of the datasets under consideration, it is unlikely that Bayesian inference will be a viable option due to time constraints. This will instead require a different approach to fitting the model (likely EM-related), which will require more math and more custom code, with benefits in terms of time that are unclear at the moment.

Another challenge that may arise is the prospect of overparameterization. Because smoothing spline coefficients are considered the model output, we will need to use models that have a very high number of parameters. If we are able to use Bayesian computation to fit the model, then priors on those parameters may help to alleviate this concern. However, in the more likely scenario that we will need to use frequentist approaches for estimation, overparameterization may become a concern. It is unclear to me how we would resolve that issue should it arise. In any case, the high number of parameters in the model will likely make interpretation difficult, although it may be helpful to have a different set of spline coefficients assigned to each category of interest.

References

Alexander, Monica and Leontine Alkema (2018). "Global Estimation of Neonatal Mortality Using a Bayesian Hierarchical Splines Regression Model". In: *Demographic Research* 38, pp. 335–372. ISSN: 1435-9871. JSTOR: 26457049. URL: https://www.jstor.org/stable/26457049 (visited on 07/06/2023).

Deng, Jiansong et al. (July 1, 2008). "Polynomial Splines over Hierarchical T-meshes". In: *Graphical Models* 70.4, pp. 76–86. ISSN: 1524-0703. DOI: 10.1016/j.gmod.2008.03.001. URL: https://www.sciencedirect.com/science/article/pii/S1524070308000039 (visited on 07/06/2023).

Forsey, David R. and Richard H. Bartels (Apr. 1995). "Surface Fitting with Hierarchical Splines". In: *ACM Transactions on Graphics* 14.2, pp. 134–161. ISSN: 0730-0301, 1557-7368. DOI: 10.1145/221659.221665. URL: https://dl.acm.org/doi/10.1145/221659.221665 (visited on 07/06/2023).

Gelman, Andrew, John B. Carlin, et al. (2014). *Bayesian Data Analysis*. 3rd. Texts in Statistical Science. Boca Raton: CRC Press/Taylor & Francis Group.

Gelman, Andrew and Jennifer Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge: Cambridge University Press.

Green, P. J. and B.W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Monographs on Statistics and Applied Probability 58. Springer.

He, Ying et al. (2006). "Manifold T-Spline". In: *Geometric Modeling and Processing - GMP 2006*. Ed. by Myung-Soo Kim and Kenji Shimada. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 409–422. ISBN: 978-3-540-36865-6. DOI: 10.1007/11802914_29.

Hendricks, Wallace and Roger Koenker (Mar. 1, 1992). "Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity". In: *Journal of the American Statistical Association* 87.417, pp. 58–68. ISSN: 0162-1459. DOI: 10.1080/01621459.1992.10475175. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475175 (visited on 07/06/2023).

Meng, Kun and Ani Eloyan (2021). "Principal Manifold Estimation via Model Complexity Selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83.2, pp. 369–394. ISSN: 1467-9868. DOI: 10.1111/rssb.12416. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12416 (visited on 11/29/2022).

Schulam, Peter and Suchi Saria (2016). "Integrative Analysis Using Coupled Latent Variable Models for Individualizing Prognoses". In: *Journal of Machine Learning Research* 17.232, pp. 1–35. ISSN: 1533-7928. URL: http://jmlr.org/papers/v17/15-436.html (visited on 07/06/2023).

Xie, Z. and G.E. Farin (Jan. 2004). "Image Registration Using Hierarchical B-splines". In: *IEEE Transactions on Visualization and Computer Graphics* 10.1, pp. 85–94. ISSN: 1941-0506. DOI: 10.1109/TVCG.2004.1260760.