Context-Adaptive Statistical Inference: Recent Progress, Open Problems, and Opportunities for Foundation Models

This manuscript (permalink) was automatically generated from LengerichLab/context-review@9fe6f2f on August 16, 2024.

Authors

- **Ben Lengerich**
  0000-0001-8690-9554 · blengerich · ben_lengerich
  Department of Statistics, University of Wisconsin-Madison · Funded by None

- **Caleb N. Ellington**
  0000-0001-7029-8023 · cnellington · probablybots
  Computational Biology Department, Carnegie Mellon University · Funded by None

✉ — Correspondence possible via GitHub Issues

Abstract

Context-adaptive inference offers a promising way to improve statistical methods, particularly by using foundation models to provide context. Recent progress is reviewed, key challenges are identified, and practical opportunities for integration are explored. By combining these approaches, we can unlock new possibilities for making statistical methods more effective and adaptable in real-world scenarios, driving meaningful advances in both research and practice.

Introduction

Personalization aims to solve the problem of *parameter heterogeneity*, where model parameters are *sample-specific*.

$$X_i \sim P(X; \theta_i)$$

From N observations, personalized modeling methods aim to recover N parameter estimates $\hat{\theta}_1, \ldots, \hat{\theta}_N$. Without further assumptions this problem is ill-defined, and the estimators have far too much variance to be useful. We can begin to make this problem tractable by imposing assumptions on the topology of θ, or the relationship between θ and contextual variables.

Population Models

The fundamental assumption of most models is that samples are independent and identically distributed. However, if samples are identically distributed they must also have identical parameters. To account for parameter heterogeneity and create more realistic models we must relax this assumption, but the assumption is so fundamental to many methods that alternatives are rarely explored. Additionally, many traditional models may produce a seemingly acceptable fit to their data, even when the underlying model is heterogeneous. Here, we explore the consequences of applying homogeneous modeling approaches to heterogeneous data, and discuss how subtle but meaningful effects are often lost to the strength of the identically distributed assumption.

Failure modes of population models can be identified by their error distributions.

Mode collapse: If one population is much larger than another, the other population will be underrepresented in the model.

Outliers: Small populations of outliers can have an enormous effect on OLS models in the parameter-averaging regime.

Phantom Populations: If several populations are present but equally represented, the optimal traditional model will represent none of these populations.

Lemma: A traditional OLS linear model will be the average of heterogeneous models.

Context-informed models

Conditional and Cluster Models

While conditional and cluster models are not truly personalized models, the spirit is the same. These models make the assumption that models in a single conditional or cluster group are homogeneous. More commonly this is written as a group of observations being generated by a single model. While

the assumption results in fewer than N models, it allows the use of generic plug-in estimators. Conditional or cluster estimators take the form

$$\hat{\theta}_0, \ldots, \hat{\theta}_C = \arg\max_{\theta_0,\ldots,\theta_C} \sum_{c \in \mathcal{C}} \ell(X_c; \theta_c)$$

where $\ell(X; \theta)$ is the log-likelihood of θ on X and c specifies the covariate group that samples are assigned to, usually by specifying a condition or clustering on covariates thought to affect the distribution of observations. Notably, this method produces fewer than N distinct models for N samples and will fail to recover per-sample parameter variation.

Distance-regularized Models

Distance-regularized models assume that models with similar covariates have similar parameters and encode this assumption as a regularization term.

$$\hat{\theta}_0, \ldots, \hat{\theta}_N = \arg\max_{\theta_0,\ldots,\theta_N} \sum_i \left[\ell(x_i; \theta_i)\right] - \sum_{i,j} \frac{\|\theta_i - \theta_j\|}{D(c_i, c_j)}$$

The second term is a regularizer that penalizes divergence of θ's with similar c.

Parametric Varying-coefficient models

Original paper (based on a smoothing spline function): [1] Markov networks: [2] Linear varying-coefficient models assume that parameters vary linearly with covariates, a much stronger assumption than the classic varying-coefficient model but making a conceptual leap that allows us to define a form for the relationship between the parameters and covariates.

$$\hat{\theta}_0, \ldots, \hat{\theta}_N = \widehat{A} C^T$$
$$\widehat{A} = \arg\max_A \sum_i \ell(x_i; A c_i)$$

Semi-parametric varying-coefficient Models

Original paper: [3] 2-step estimation with RBF kernels: [4]

Classic varying-coefficient models assume that models with similar covariates have similar parameters, or – more formally – that changes in parameters are smooth over the covariate space. This assumption is encoded as a sample weighting, often using a kernel, where the relevance of a sample to a model is equivalent to its kernel similarity over the covariate space.

$$\hat{\theta}_0, \ldots, \hat{\theta}_N = \arg\max_{\theta_0,\ldots,\theta_N} \sum_{i,j} \frac{K(c_i, c_j)}{\sum_k K(c_i, c_k)} \ell(x_j; \theta_i)$$

This estimator is the simplest to recover N unique parameter estimates. However, the assumption here is contradictory to the partition model estimator. When the relationship between covariates and parameters is discontinuous or abrupt, this estimator will fail.

Contextualized Models

Seminal work [5] Contextualized ML generalization and applications: [6], [7], [8], [9], [10], [11], [12], [13]

Contextualized models make the assumption that parameters are some function of context, but make no assumption on the form of that function. In this regime, we seek to estimate the function often using a deep learner (if we have some differentiable proxy for probability):

$$\widehat{f} = \arg\max_{f \in \mathcal{F}} \sum_i \ell(x_i; f(c_i))$$

Latent-structure Models

Partition Models

Markov networks: [14] Partition models also assume that parameters can be partitioned into homogeneous groups over the covariate space, but make no assumption about where these partitions occur. This allows the use of information from different groups in estimating a model for a each covariate. Partition model estimators are most often utilized to infer abrupt model changes over time and take the form

$$\hat{\theta}_0, \ldots, \hat{\theta}_N = \arg\max_{\theta_0,\ldots,\theta_N} \sum_i \ell(x_i; \theta_i) + \sum_{i=2}^{N} \mathrm{TV}(\theta_i, \theta_{i-1})$$

Where the regularizaiton term might take the form

$$\mathrm{TV}(\theta_i, \theta_{i-1}) = |\theta_i - \theta_{i-1}|$$

This still fails to recover a unique parameter estimate for each sample, but gets closer to the spirit of personalized modeling by putting the model likelihood and partition regularizer in competition to find the optimal partitions.

Fine-tuned Models and Transfer Learning

Review: [15] Noted in foundational literature for linear varying coefficient models [3]

Estimate a population model, freeze these parameters, and then include a smaller set of personalized parameters to estimate on a smaller subpopulation.

$$\widehat{\gamma} = \arg\max_{\gamma} = \ell(\gamma; X)$$

$$\widehat{\theta_c} = \arg\max_{\theta_c} = \ell(\theta_c; \widehat{\gamma}, X_c)$$

Context-informed and Latent-structure models

Seminal paper: [16]

Key idea: negative information sharing. Different models should be pushed apart.

$$\hat{\theta}_0, \ldots, \hat{\theta}_N = \arg\max_{\theta_0,\ldots,\theta_N, D} \sum_{i=0}^{N} \prod_{j=0 s.t. D(c_i,c_j)<d}^{N} P(x_j; \theta_i) P(\theta_i; \theta_j)$$

References

1. **Varying-Coefficient Models**
Trevor Hastie, Robert Tibshirani
Journal of the Royal Statistical Society Series B: Statistical Methodology (1993-09-01)
https://doi.org/gmfvmb
DOI: 10.1111/j.2517-6161.1993.tb01939.x

2. **Bayesian Edge Regression in Undirected Graphical Models to Characterize Interpatient Heterogeneity in Cancer**
Zeya Wang, Veerabhadran Baladandayuthapani, Ahmed O Kaseb, Hesham M Amin, Manal M Hassan, Wenyi Wang, Jeffrey S Morris
Journal of the American Statistical Association (2022-01-05) https://doi.org/gt68hr
DOI: 10.1080/01621459.2021.2000866 · PMID: 36090952 · PMCID: PMC9454401

3. **Statistical estimation in varying coefficient models**
Jianqing Fan, Wenyang Zhang
The Annals of Statistics (1999-10-01) https://doi.org/dsxd4s
DOI: 10.1214/aos/1017939139

4. **Time-Varying Coefficient Model Estimation Through Radial Basis Functions**
Juan Sosa, Lina Buitrago
arXiv (2021-03-02) https://arxiv.org/abs/2103.00315

5. **Contextual Explanation Networks**
Maruan Al-Shedivat, Avinava Dubey, Eric P Xing
arXiv (2017) https://doi.org/gt68h9
DOI: 10.48550/arxiv.1705.10301

6. **Contextualized Machine Learning**
Benjamin Lengerich, Caleb N Ellington, Andrea Rubbi, Manolis Kellis, Eric P Xing
arXiv (2023) https://doi.org/gt68jg
DOI: 10.48550/arxiv.2310.11340

7. **NOTMAD: Estimating Bayesian Networks with Sample-Specific Structures and Parameters**
Ben Lengerich, Caleb Ellington, Bryon Aragam, Eric P Xing, Manolis Kellis
arXiv (2021) https://doi.org/gt68jc
DOI: 10.48550/arxiv.2111.01104

8. **Contextualized: Heterogeneous Modeling Toolbox**
Caleb N Ellington, Benjamin J Lengerich, Wesley Lo, Aaron Alvarez, Andrea Rubbi, Manolis Kellis, Eric P Xing
Journal of Open Source Software (2024-05-08) https://doi.org/gt68h8
DOI: 10.21105/joss.06469

9. **Contextualized Policy Recovery: Modeling and Interpreting Medical Decisions with Adaptive Imitation Learning**
Jannik Deuschel, Caleb N Ellington, Yingtao Luo, Benjamin J Lengerich, Pascal Friederich, Eric P Xing
arXiv (2023) https://doi.org/gt68jf
DOI: 10.48550/arxiv.2310.07918

10. **Automated interpretable discovery of heterogeneous treatment effectiveness: A COVID-19 case study**
Benjamin J Lengerich, Mark E Nunnally, Yin Aphinyanaphongs, Caleb Ellington, Rich Caruana
Journal of Biomedical Informatics (2022-06) https://doi.org/gt68h5
DOI: 10.1016/j.jbi.2022.104086 · PMID: 35504543 · PMCID: PMC9055753

11. **Discriminative Subtyping of Lung Cancers from Histopathology Images via Contextual Deep Learning**
Benjamin J Lengerich, Maruan Al-Shedivat, Amir Alavi, Jennifer Williams, Sami Labbaki, Eric P Xing
Cold Spring Harbor Laboratory (2020-06-26) https://doi.org/gt68h6
DOI: 10.1101/2020.06.25.20140053

12. **Contextualized Networks Reveal Heterogeneous Transcriptomic Regulation in Tumors at Sample-Specific Resolution**
Caleb N Ellington, Benjamin J Lengerich, Thomas BK Watkins, Jiekun Yang, Hanxi Xiao, Manolis Kellis, Eric P Xing
Cold Spring Harbor Laboratory (2023-12-04) https://doi.org/gt68h7
DOI: 10.1101/2023.12.01.569658

13. **Contextual Feature Selection with Conditional Stochastic Gates**
Ram Dyuthi Sristi, Ofir Lindenbaum, Shira Lifshitz, Maria Lavzin, Jackie Schiller, Gal Mishne, Hadas Benisty
arXiv (2023) https://doi.org/gt68jh
DOI: 10.48550/arxiv.2312.14254

14. **Estimating time-varying networks**
Mladen Kolar, Le Song, Amr Ahmed, Eric P Xing
The Annals of Applied Statistics (2010-03-01) https://doi.org/b3rn6q
DOI: 10.1214/09-aoas308

15. **When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction**
Vinith M Suriyakumar, Marzyeh Ghassemi, Berk Ustun
arXiv (2022) https://doi.org/gt68jd
DOI: 10.48550/arxiv.2206.02058

16. **Learning Sample-Specific Models with Low-Rank Personalized Regression**
Benjamin Lengerich, Bryon Aragam, Eric P Xing
arXiv (2019) https://doi.org/gt68jb
DOI: 10.48550/arxiv.1910.06939