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

Authors

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

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

✉ — Correspondence possible via GitHub Issues

Abstract

Context-adaptive inference enhances statistical methods by allowing model parameters to shift with context, either explicitly through parameter learning or implicitly via interactions between context and input features. In this review, we outline recent progress in integrating context into statistical models and explore the potential of foundation models to serve as context providers. We conclude by discussing future trends, challenges, and opportunities in context-adaptive statistical inference.

Introduction

A growing number of methods across statistics and machine learning aim to model how data distributions vary across individuals, environments, or tasks. This interest in context-adaptive inference reflects a shift from population-level models toward those that account for sample-specific variation.

In statistics, **varying-coefficient models** allow model parameters to change smoothly with covariates. In machine learning, **meta-learning** and **transfer learning** enable models to adapt across tasks or domains. More recently, **in-context learning** – by which foundation models adapt behavior based on support examples without parameter updates – has emerged as a powerful mechanism for personalization in large language models.

These approaches originate from different traditions but share a common goal: to use *context* in the form of covariates, support data, or task descriptors to guide inference about sample-specific *parameters*.

We formalize the setting by assuming each observation X_i is drawn from a sample-specific distribution:

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

where θ_i denotes the parameters governing the distribution of the ith observation. In the most general case, this formulation allows for arbitrary heterogeneity. However, estimating N distinct parameters from N observations is ill-posed without further structure.

To make the problem tractable, context-adaptive methods introduce structure by assuming that parameters vary systematically with context:
$$\theta_i = f(c_i).$$
This deterministic formulation is common in varying-coefficient models and many supervised personalization settings.

More generally, θ_i may be drawn from a context-dependent distribution:
$$\theta_i \sim P(\theta \mid c_i),$$
as in hierarchical Bayesian models or amortized inference frameworks. This stochastic formulation captures residual uncertainty or unmodeled variation beyond what is encoded in c_i.

The function f encodes how parameters vary with context, and may be linear, smooth, or nonparametric, depending on the modeling assumptions. In this view, the challenge of context-adaptive inference reduces to estimating or constraining f given data $\{(x_i, c_i)\}_{i=1}^{N}$.

Viewed this way, context-adaptive inference spans a spectrum—from models that seek **invariance** across environments to models that enable **personalization** at the level of individual samples. For example:

- **Population models** assume $\theta_i = \theta$ for all i.
- **Invariant risk minimization** [1] identifies components of θ that remain stable across distributions.
- **Transfer learning** assumes partial invariance, learning domain-specific shifts around a shared representation.
- **Varying-coefficient models** allow θ_i to vary smoothly with observed context.
- **In-context learning** treats parameters as an implicit function of support examples.

In this review, we survey methods across this spectrum. We highlight their shared foundations, clarify the assumptions they make about θ_i, and explore the emerging connections between classical approaches such as varying-coefficient models and modern inference mechanisms like in-context learning.

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

Without further assumptions, sample-specific parameter estimation is ill-defined. Single sample estimation is prohibitively high variance. We can begin to make this problem tractable by taking note from previous work and imposing assumptions on the topology of θ, or the relationship between θ and contextual variables.

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.

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

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): [2] Markov networks: [3] 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: [4] 2-step estimation with RBF kernels: [5]

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 [6] Contextualized ML generalization and applications: [7], [8], [9], [10], [11], [12], [13], [14]

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: [15] 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: [16] Noted in foundational literature for linear varying coefficient models [4]

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: [17]

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

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

Theoretical Foundations and Advances in Varying-Coefficient Models

Principles of Adaptivity

TODO: Analyzing the core principles that underpin adaptivity in statistical modeling.

Advances in Varying-Coefficient Models

TODO: Outlining key theoretical and methodological breakthroughs.

Integration with State-of-the-Art Machine Learning

TODO: Assessing the enhancement of VC models through modern ML technologies (e.g. deep learning, boosted trees, etc).

Context-Invariant Training

TODO: The converse of VC models, exploring the implications of training context-invariant models. e.g. out-of-distribution generalization, robustness to adversarial attacks.

Context-Adaptive Interpretations of Context-Invariant Models

In the previous section, we discussed the importance of context in model parameters. Such context-adaptive models can be learned by explicitly modeling the impact of contextual variables on model parameters, or learned implicitly in a model containing interaction effects between the context and the input features. In this section, we will focus on recent progress in understanding how context influences interpretations of statistical models, even when the model was not originally designed to incorporate context.

TODO: Discussing the implications of context-adaptive interpretations for traditional models. Related work including LIME/DeepLift/DeepSHAP.

Opportunities for Foundation Models

Expanding Frameworks

TODO: Define foundation models, Explore how foundation models are redefining possibilities within statistical models.

Foundation models as context

TODO: Show recent progress and ongoing directions in using foundation models as context.

Applications, Case Studies, and Evaluations

Implementation Across Sectors

TODO: Detailed examination of context-adaptive models in sectors like healthcare and finance.

Performance Evaluation

TODO: Successes, failures, and comparative analyses of context-adaptive models across applications.

Technological and Software Tools

Survey of Tools

TODO: Reviewing current technological supports for context-adaptive models.

Selection and Usage Guidance

TODO: Offering practical advice on tool selection and use for optimal outcomes.

Future Trends and Predictions

Emerging Technologies

TODO: Identifying upcoming technologies and predicting their impact on context-adaptive learning.

Advances in Methodologies

TODO: Speculating on potential future methodological enhancements.

Open Problems

Theoretical Challenges

TODO: Critically examining unresolved theoretical issues like identifiability, etc.

Ethical and Regulatory Considerations

TODO: Discussing the ethical landscape and regulatory challenges, with focus on benefits of interpretability and regulatability.

Complexity in Implementation

TODO: Addressing obstacles in practical applications and gathering insights from real-world data.

TODO: Other open problems?

Conclusion

Overview of Insights

TODO: Summarizing the main findings and contributions of this review.

Future Directions

TODO: Discussing potential developments and innovations in context-adaptive statistical inference.

References

1. **Invariant Risk Minimization**
   Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, David Lopez-Paz
   arXiv (2019) https://doi.org/gz355c
   DOI: 10.48550/arxiv.1907.02893

2. **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

3. **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

4. **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

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

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

7. **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

8. **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

9. **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

10. **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

11. **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

12. **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

13. **Learning to Estimate Sample-specific Transcriptional Networks for 7000 Tumors**
Caleb N Ellington, Benjamin J Lengerich, Thomas BK Watkins, Jiekun Yang, Abhinav Adduri, Sazan Mahbub, 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

14. **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

15. **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

16. **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

17. **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