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

This manuscript (permalink) was automatically generated from AdaptInfer/context-review@6368038 on July 24, 2025.

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 extends classical statistical modeling by allowing parameters to vary across individuals, environments, or tasks. This adaptation may be explicit—through parameterized functions of context—or implicit, via interactions between context and input features. In this review, we survey recent advances in modeling sample-specific variation, including varying-coefficient models, transfer learning, and in-context learning. We also examine the emerging role of foundation models as flexible context encoders. Finally, we outline key challenges and open questions for the development of principled, scalable, and interpretable context-adaptive methods.

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.

Relevant references:

- Can Subpopulation Shifts Explain Disagreement in Model Generalization? [2]

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.

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

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): [3] Markov networks: [4] 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: [5] 2-step estimation with RBF kernels: [6]

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 [7] Contextualized ML generalization and applications: [8], [9], [10], [11], [12], [13], [14], [15]

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: [16] 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: [17] Noted in foundational literature for linear varying coefficient models [5]

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: [18]

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

What does it mean for a model to be adaptive? When is it good for a model to be adaptive? While the appeal of adaptivity lies in flexibility and personalized inference, not all adaptivity is good adaptivity. In

this section, we formalize the core principles that underlie adaptive modeling.

1. Adaptivity requires flexibility

A model cannot adapt unless it has the capacity to represent multiple behaviors. Flexibility may take the form of nonlinearity, hierarchical structure, or modular components that allow different responses in different settings.

- Interaction effects in regression models [19]
- Hierarchical models that allow for varying effects across groups
- Meta-learning and mixtures-of-experts models that learn to adapt based on context
- Varying-coefficient models that allow coefficients to change with context [3]

2. Adaptivity requires a signal of heterogeneity

- Varying-coefficient models adapt parameters based on observed context [3]
- Contextual bandits adapt actions to context features [20]
- Multi-domain models adapt across known environments or inferred partitions [21]

3. Modularity improves adaptivity

Adaptive systems are easier to design, debug, and interpret when built from modular parts. Modularity supports targeted adaptation, transferability, and disentanglement.

- []

4. Adaptivity implies selectivity

Adaptation must be earned. Overreacting to limited data leads to overfitting. The best adaptive methods include mechanisms for deciding when not to adapt. - Lepski's method [22] - Aggregation of classifiers [23]

5. Adaptivity is bounded by data efficiency

[24]

When Adaptivity Fails: Common Failure Modes

Even when all the ingredients are present, adaptivity can backfire. Common failure modes include:

- Spurious Adaptation: Adapting to unstable or confounded features [25]
- Overfitting in Low-Data Contexts: Attempting fine-grained adaptation with insufficient signal
- Modularity Mis-specification: Adapting in the wrong units or groupings [26]
- Feedback Loops: Models that change the data distribution they rely on [27]

Advances in Varying-Coefficient Models

TODO: Outlining key theoretical and methodological breakthroughs.

Relevant references:

- [28]

Flexible Functional Forms

Relevant references:

- [29]

Integration with State-of-the-Art Machine Learning

TODO: Enhancing VC models with modern ML technologies (e.g. deep learning, boosted trees, etc).

Relevant references:

- [30]
- [31]
- [32]

Structured data (Spatio-Temporal, Graphs, etc.)

Related references:

- [33]
- [34]
- [35]
- [36]
- [37]
- [38]

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.

Relevant references:

- Invariant Risk Minimization [39]
- Out-of-Distribution Generalization via Risk Extrapolation [40]
- The Risks of Invariant Risk Minimization [25]
- Conditional Variance Penalties and Domain Adaptation [41]
- Can Subpopulation Shifts Explain Disagreement in Model Generalization? [2]

Adversarial Robustness as Context-Invariant Training

Related references:

- Towards Deep Learning Models Resistant to Adversarial Attacks [42]
- Robustness May Be at Odds with Accuracy [43]

Training methods for Context-Invariant Models

- Just Train Twice: Improving Group Robustness without Training Group Information [44]
- Environment Inference for Invariant Learning [45]
- Distributionally Robust Neural Networks for Group Shifts [26]

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.

Relevant references:

- [46]

Opportunities for Foundation Models

Expanding Frameworks

Foundation models refer to large-scale, general-purpose neural networks, predominantly transformer-based architectures, trained on vast datasets using self-supervised learning [47]. These models have significantly transformed modern statistical modeling and machine learning due to their flexibility, adaptability, and strong performance across diverse domains. Notably, large language models (LLMs) such as GPT-4 [48] and LLaMA-3.1 [49] have achieved substantial advancements in natural language processing (NLP), demonstrating proficiency in tasks ranging from text generation and summarization to question-answering and dialogue systems. Beyond NLP, foundation models also excel in multimodal (text-vision) tasks [50], text embedding generation [51], and structured tabular data analysis [52], highlighting their broad applicability.

A key strength of foundation models lies in their capacity to dynamically adapt to different contexts provided by inputs. This adaptability is primarily achieved through techniques such as prompting, which involves designing queries to guide the model's behavior implicitly, allowing task-specific responses without additional fine-tuning [53]. Furthermore, mixture-of-experts (MoE) architectures amplify this contextual adaptability by employing routing mechanisms that select specialized sub-models or "experts" tailored to specific input data, thus optimizing computational efficiency and performance [54].

Foundation Models as Context

Foundation models offer significant opportunities by supplying context-aware information that enhances various stages of statistical modeling and inference:

Feature Extraction and Interpretation: Foundation models transform raw, unstructured data into structured and interpretable representations. For example, targeted prompts enable LLMs to extract insightful features from text, providing meaningful insights and facilitating interpretability [57]. This allows statistical models to operate directly on semantically meaningful features rather than on raw, less interpretable data.

Contextualized Representations for Downstream Modeling: Foundation models produce adaptable embeddings and intermediate representations useful as inputs for downstream models, such as decision trees or linear models [58]. These embeddings significantly enhance the training of both complex, black-box models [59] and simpler statistical methods like n-gram-based analyses [60], thereby broadening the application scope and effectiveness of statistical approaches.

Post-hoc Interpretability: Foundation models support interpretability by generating natural-language explanations for decisions made by complex models. This capability enhances transparency and trust in statistical inference, providing clear insights into how and why certain predictions or decisions are made [61].

Recent innovations underscore the role of foundation models in context-sensitive inference and enhanced interpretability:

FLAN-MoE (Fine-tuned Language Model with Mixture of Experts) [62] combines instruction tuning with expert selection, dynamically activating relevant sub-models based on the context. This method significantly improves performance across diverse NLP tasks, offering superior few-shot and zero-shot capabilities. It also facilitates interpretability through explicit expert activations. Future directions may explore advanced expert-selection techniques and multilingual capabilities.

LMPriors (Pre-Trained Language Models as Task-Specific Priors) [63] leverages semantic insights from pre-trained models like GPT-3 to guide tasks such as causal inference, feature selection, and reinforcement learning. This method markedly enhances decision accuracy and efficiency without requiring extensive supervised datasets. However, it necessitates careful prompt engineering to mitigate biases and ethical concerns.

Mixture of In-Context Experts (MoICE) [63] introduces a dynamic routing mechanism within attention heads, utilizing multiple Rotary Position Embeddings (RoPE) angles to effectively capture token positions in sequences. MoICE significantly enhances performance on long-context sequences and retrieval-augmented generation tasks by ensuring complete contextual coverage. Efficiency is achieved through selective router training, and interpretability is improved by explicitly visualizing attention distributions, providing detailed insights into the model's reasoning process.

Applications, Case Studies, and Evaluations

Implementation Across Sectors

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

Relevant references:

- [64]
- [65]

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. **Sketch-Based Anomaly Detection in Streaming Graphs**
Siddharth Bhatia, Mohit Wadhwa, Kenji Kawaguchi, Neil Shah, Philip S Yu, Bryan Hooi
arXiv (2023-07-18) https://arxiv.org/abs/2106.04486

3. **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

4. **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

5. **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

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

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

8. **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

9. **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

10. **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

11. **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

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

19. **Intelligible Models for HealthCare**
Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad
Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015-08-10) https://doi.org/gftgxk
DOI: 10.1145/2783258.2788613

20. **Adapting multi-armed bandits policies to contextual bandits scenarios**
David Cortes

arXiv (2019-11-26) https://arxiv.org/abs/1811.04383

21. **Environment Inference for Invariant Learning**
Elliot Creager, JÃ¶rn-Henrik Jacobsen, Richard Zemel
arXiv (2021-07-16) https://arxiv.org/abs/2010.07249

22. **Lepski's Method and Adaptive Estimation of Nonlinear Integral Functionals of Density**
Rajarshi Mukherjee, Eric Tchetgen Tchetgen, James Robins
arXiv (2016-01-12) https://arxiv.org/abs/1508.00249

23. **Optimal Rates of Aggregation**
Alexandre B Tsybakov
Lecture Notes in Computer Science (2003) https://doi.org/czntw5
DOI: 10.1007/978-3-540-45167-9_23

24. **Optimal Estimation of Change in a Population of Parameters**
Ramya Korlakai Vinayak, Weihao Kong, Sham M Kakade
arXiv (2019-12-02) https://arxiv.org/abs/1911.12568

25. **The Risks of Invariant Risk Minimization**
Elan Rosenfeld, Pradeep Ravikumar, Andrej Risteski
arXiv (2021-03-30) https://arxiv.org/abs/2010.05761

26. **Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization**
Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, Percy Liang
arXiv (2020-04-03) https://arxiv.org/abs/1911.08731

27. **The Selective Labels Problem**
Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan
Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017-08-04) https://doi.org/ggd7hz
DOI: 10.1145/3097983.3098066 · PMID: 29780658 · PMCID: PMC5958915

28. **Publication Trends on the Varying Coefficients Model: Estimating the Actual (Under)Utilization of a Highly Acclaimed Method for Studying Statistical Interactions**
Assaf Botzer
Publications (2025-04-07) https://doi.org/g9t2rq
DOI: 10.3390/publications13020019

29. **Semi-nonparametric Varying Coefficients Models**
Ting Li, Yang Yu, Xiao Wang, JS Marron, Hongtu Zhu
Statistica Sinica (2027) https://doi.org/g9t2rr
DOI: 10.5705/ss.202024.0118

30. **A tree-based varying coefficient model**
Henning Zakrisson, Mathias Lindholm
Computational Statistics (2025-02-04) https://doi.org/g869k6
DOI: 10.1007/s00180-025-01603-8

31. **VCBART: Bayesian trees for varying coefficients**
Sameer K Deshpande, Ray Bai, Cecilia Balocchi, Jennifer E Starling, Jordan Weiss
arXiv (2024-09-26) https://arxiv.org/abs/2003.06416

32. **Neural Additive Models: Interpretable Machine Learning with Neural Nets**

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, Geoffrey Hinton
arXiv (2021-10-26) https://arxiv.org/abs/2004.13912

33. **Network Varying Coefficient Model**
Xinyan Fan, Kuangnan Fang, Wei Lan, Chih-Ling Tsai
Journal of the American Statistical Association (2025-04-11) https://doi.org/g9t2rm
DOI: 10.1080/01621459.2025.2470481

34. **Spatially Varying Coefficient Models for Estimating Heterogeneous Mixture Effects**
Jacob Englert, Howard Chang
arXiv (2025-02-21) https://arxiv.org/abs/2502.14651

35. **Fast Spatio-Temporally Varying Coefficient Modeling With Reluctant Interaction Selection**
Daisuke Murakami, Shinichiro Shirota, Seiji Kajita, Mami Kajita
Geographical Analysis (2025-04-15) https://doi.org/g9t2rn
DOI: 10.1111/gean.70005

36. **Varying-coefficient spatial dynamic panel data models with fixed effects: Theory and application**
Han Hong, Gaosheng Ju, Qi Li, Karen X Yan
Journal of Econometrics (2024-10) https://doi.org/g9t2rj
DOI: 10.1016/j.jeconom.2024.105883

37. **Varying coefficient panel data models and methods under correlated error components: Application to disparities in mental health services in England**
Pipat Wongsa-art, Namhyun Kim, Yingcun Xia, Francesco Moscone
Regional Science and Urban Economics (2024-05) https://doi.org/g9t2rk
DOI: 10.1016/j.regsciurbeco.2024.104009

38. **NOTMAD: Estimating Bayesian Networks with Sample-Specific Structures and Parameters**
Ben Lengerich, Caleb Ellington, Bryon Aragam, Eric P Xing, Manolis Kellis
arXiv (2021-11-02) https://arxiv.org/abs/2111.01104

39. **Invariant Risk Minimization**
Martin Arjovsky, LÃ©on Bottou, Ishaan Gulrajani, David Lopez-Paz
arXiv (2020-03-31) https://arxiv.org/abs/1907.02893

40. **Out-of-Distribution Generalization via Risk Extrapolation (REx)**
David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, Aaron Courville
arXiv (2021-02-26) https://arxiv.org/abs/2003.00688

41. **Conditional Variance Penalties and Domain Shift Robustness**
Christina Heinze-Deml, Nicolai Meinshausen
arXiv (2019-04-16) https://arxiv.org/abs/1710.11469

42. **Towards Deep Learning Models Resistant to Adversarial Attacks**
Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu
arXiv (2019-09-06) https://arxiv.org/abs/1706.06083

43. **Robustness May Be at Odds with Accuracy**
Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, Aleksander Madry
arXiv (2019-09-10) https://arxiv.org/abs/1805.12152

44. **On the Sample Complexity of Adversarial Multi-Source PAC Learning**
Nikola Konstantinov, Elias Frantar, Dan Alistarh, Christoph H Lampert
arXiv (2020-07-01) https://arxiv.org/abs/2002.10384

45. **Conflict-Averse Gradient Descent for Multi-task Learning**
Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, Qiang Liu
arXiv (2024-02-22) https://arxiv.org/abs/2110.14048

46. **In-Context Explainers: Harnessing LLMs for Explaining Black Box Models**
Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju
arXiv (2024-07-12) https://arxiv.org/abs/2310.05797

47. **On the Opportunities and Risks of Foundation Models**
Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, … Percy Liang
arXiv (2021) https://doi.org/hw3v
DOI: 10.48550/arxiv.2108.07258

48. **GPT-4 Technical Report**
OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, … Barret Zoph
arXiv (2023) https://doi.org/grx4cb
DOI: 10.48550/arxiv.2303.08774

49. **The Llama 3 Herd of Models**
Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, … Zhiyu Ma
arXiv (2024) https://doi.org/ndw6
DOI: 10.48550/arxiv.2407.21783

50. **Learning Transferable Visual Models From Natural Language Supervision**
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, … Ilya Sutskever
arXiv (2021) https://doi.org/hs7z
DOI: 10.48550/arxiv.2103.00020

51. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
arXiv (2018) https://doi.org/hm65
DOI: 10.48550/arxiv.1810.04805

52. **TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second**
Noah Hollmann, Samuel Müller, Katharina Eggensperger, Frank Hutter
arXiv (2022) https://doi.org/g9t22b
DOI: 10.48550/arxiv.2207.01848

53. **Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing**
Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, Graham Neubig
ACM Computing Surveys (2023-01-16) https://doi.org/gq5fh2
DOI: 10.1145/3560815

54. **Mixture of experts: a literature survey**
Saeed Masoudnia, Reza Ebrahimpour
Artificial Intelligence Review (2012-05-12) https://doi.org/f59sxs
DOI: 10.1007/s10462-012-9338-y

55. **CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models**
Denis Jered McInerney, Geoffrey Young, Jan-Willem van de Meent, Byron C Wallace
arXiv (2023) https://doi.org/g9t22g
DOI: 10.48550/arxiv.2302.12343

56. **Learning Interpretable Style Embeddings via Prompting LLMs**
Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, Chris Callison-Burch
arXiv (2023) https://doi.org/g9t22h
DOI: 10.48550/arxiv.2305.12696

57. **Tree Prompting: Efficient Task Adaptation without Fine-Tuning**
Chandan Singh, John Morris, Alexander Rush, Jianfeng Gao, Yuntian Deng
Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023) https://doi.org/gtgrkq
DOI: 10.18653/v1/2023.emnlp-main.384

58. **What Can Transformers Learn In-Context? A Case Study of Simple Function Classes**
Shivam Garg, Dimitris Tsipras, Percy Liang, Gregory Valiant
arXiv (2022) https://doi.org/g9t22c
DOI: 10.48550/arxiv.2208.01066

59. **One Embedder, Any Task: Instruction-Finetuned Text Embeddings**
Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu
arXiv (2022) https://doi.org/g9t22f
DOI: 10.48550/arxiv.2212.09741

60. **Augmenting interpretable models with large language models during training**
Chandan Singh, Armin Askari, Rich Caruana, Jianfeng Gao
Nature Communications (2023-11-30) https://doi.org/g9t2z9
DOI: 10.1038/s41467-023-43713-1 · PMID: 38036543 · PMCID: PMC10689442

61. **Explaining Datasets in Words: Statistical Models with Natural Language Parameters**
Ruiqi Zhong, Heng Wang, Dan Klein, Jacob Steinhardt
arXiv (2024) https://doi.org/g9t22k
DOI: 10.48550/arxiv.2409.08466

62. **Mixture-of-Experts Meets Instruction Tuning:A Winning Combination for Large Language Models**
Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, … Denny Zhou
arXiv (2023) https://doi.org/g9t22j
DOI: 10.48550/arxiv.2305.14705

63. **LMPriors: Pre-Trained Language Models as Task-Specific Priors**
Kristy Choi, Chris Cundy, Sanjari Srivastava, Stefano Ermon
arXiv (2022) https://doi.org/g9t22d
DOI: 10.48550/arxiv.2210.12530

64. **Exact Inference for Transformed Large-Scale Varying Coefficient Models with Applications**
Tianyu Chen, Robert Habans, Thomas Douthat, Jenna Losh, Lida Chalangar Jalili Dehkharghani, Li-Hsiang Lin
Journal of Data Science (2025-01-01) https://doi.org/g9t2rs

DOI: 10.6339/25-jds1181

65. **Variable Selection for Generalized Single-Index Varying-Coefficient Models with Applications to Synergistic G × E Interactions**
Shunjie Guan, Xu Liu, Yuehua Cui
Mathematics (2025-01-31) https://doi.org/g9t2rp
DOI: 10.3390/math13030469