Quantifying the Uncertainty of Complex Models: Three Vignettes

Lucas Janson
Harvard University Department of Statistics

AI Day for Federal Statistics
May 2, 2024

Overview

AI and machine learning:

- great at maximizing prediction accuracy
- bad at interpretability and uncertainty quantification (UQ)

Overview

AI and machine learning:

- great at maximizing prediction accuracy
- bad at interpretability and uncertainty quantification (UQ)

Three settings where UQ for complex models is **critical (and how)**

Overview

AI and machine learning:

- great at maximizing prediction accuracy
- bad at interpretability and uncertainty quantification (UQ)

Three settings where UQ for complex models is **critical (and how)**

- Variable selection

Overview

AI and machine learning:

- great at maximizing prediction accuracy
- bad at interpretability and uncertainty quantification (UQ)

Three settings where UQ for complex models is **critical (and how)**

- Variable selection
- Computational modeling

Overview

AI and machine learning:

- great at maximizing prediction accuracy
- bad at interpretability and uncertainty quantification (UQ)

Three settings where UQ for complex models is **critical (and how)**

- Variable selection
- Computational modeling
- Predictive decision-making

Variable selection

The data: Observations of features and an outcome variable, but with many more features than observations

Variable selection

The data: Observations of features and an outcome variable, but with many more features than observations

E.g., people's genetic features and disease state

Variable selection

The data: Observations of features and an outcome variable, but with many more features than observations

E.g., people's genetic features and disease state

The goal: Find important features (genetic features related to disease)

Variable selection

The data: Observations of features and an outcome variable, but with many more features than observations

E.g., people's genetic features and disease state

The goal: Find important features (genetic features related to disease)

The model: Linear model with shrinkage and sparsity

Variable selection

The data: Observations of features and an outcome variable, but with many more features than observations

E.g., people's genetic features and disease state

The goal: Find important features (genetic features related to disease)

The model: Linear model with shrinkage and sparsity

Why we need UQ: How big does a coefficient need to be to be confident it represents a real relationship?

Genomics: what can we do?

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for Gold: **Model-X Knockoffs** for High- dimensional Controlled Variable Selection. Journal of the Royal Statistical Society: Series B, 80(3):551–577, 2018.

Genomics: what can we do?

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for Gold: **Model-X Knockoffs** for High- dimensional Controlled Variable Selection. Journal of the Royal Statistical Society: Series B, 80(3):551–577, 2018.

- Can't do UQ of fitted linear model coefficients directly:
  - shinkage induces bias
  - linear model only an approximation

Genomics: what can we do?

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for Gold: **Model-X Knockoffs** for High- dimensional Controlled Variable Selection. Journal of the Royal Statistical Society: Series B, 80(3):551–577, 2018.

- Can't do UQ of fitted linear model coefficients directly:
  - shinkage induces bias
  - linear model only an approximation
- **Key idea**: Create a knockoff for each feature:
  - statistically mimics the real feature
  - but has no real relationship with outcome

Genomics: what can we do?

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for Gold: **Model-X Knockoffs** for High- dimensional Controlled Variable Selection. Journal of the Royal Statistical Society: Series B, 80(3):551–577, 2018.

- Can't do UQ of fitted linear model coefficients directly:
  - shinkage induces bias
  - linear model only an approximation
- **Key idea**: Create a knockoff for each feature:
  - statistically mimics the real feature
  - but has no real relationship with outcome
- Compare fitted coefficient on real feature with fitted coefficient on knockoff (or many such)

Genomics: what can we do?

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for Gold: **Model-X Knockoffs** for High- dimensional Controlled Variable Selection. Journal of the Royal Statistical Society: Series B, 80(3):551–577, 2018.

- Can't do UQ of fitted linear model coefficients directly:
  - shinkage induces bias
  - linear model only an approximation
- **Key idea**: Create a knockoff for each feature:
  - statistically mimics the real feature
  - but has no real relationship with outcome
- Compare fitted coefficient on real feature with fitted coefficient on knockoff (or many such)
- If real coefficient bigger than most/all knockoff coefficients, conclude relationship is real with confidence

Computational modeling

The data: Data from a complex/expensive computational model

Computational modeling

The data: Data from a complex/expensive computational model

E.g., climate model inputs and outputs

Computational modeling

The data: Data from a complex/expensive computational model

E.g., climate model inputs and outputs

The goal: Estimate model output's sensitivity to an input

Computational modeling

The data: Data from a complex/expensive computational model

E.g., climate model inputs and outputs

The goal: Estimate model output's sensitivity to an input

E.g., impact of a country's CO2 emissions on future climate

Computational modeling

The data: Data from a complex/expensive computational model

E.g., climate model inputs and outputs

The goal: Estimate model output's sensitivity to an input

E.g., impact of a country's CO2 emissions on future climate

The model: A *surrogate* model (deep learning, or less complex physics)

- approximates the full model
- but much faster to run

Computational modeling

The data: Data from a complex/expensive computational model

E.g., climate model inputs and outputs

The goal: Estimate model output's sensitivity to an input

E.g., impact of a country's CO2 emissions on future climate

The model: A *surrogate* model (deep learning, or less complex physics)

- approximates the full model
- but much faster to run

Why we need UQ: Can't directly estimate full model's sensitivity well: can for surrogate, but *how different are they?*

Computational modeling: what can we do?

M. Aufiero and L. Janson. Surrogate-Based Global Sensitivity Analysis with Statistical Guarantees via **Floodgate**. 2022. [https://arxiv.org/abs/2208.05885]

Computational modeling: what can we do?

M. Aufiero and L. Janson. Surrogate-Based Global Sensitivity Analysis with Statistical Guarantees via **Floodgate**. 2022. [https://arxiv.org/abs/2208.05885]

- Full model too expensive to accurately estimate its sensitivity

Computational modeling: what can we do?

M. Aufiero and L. Janson. Surrogate-Based Global Sensitivity Analysis with Statistical Guarantees via **Floodgate**. 2022. [https://arxiv.org/abs/2208.05885]

- Full model too expensive to accurately estimate its sensitivity
- **Key idea**: Design a special proxy sensitivity quantity:
  - depends on full and surrogate models
  - if surrogate close to full model, proxy sensitivity close to full sensitivity
  - guaranteed to be at most full sensitivity

Computational modeling: what can we do?

M. Aufiero and L. Janson. Surrogate-Based Global Sensitivity Analysis with Statistical Guarantees via **Floodgate**. 2022. [https://arxiv.org/abs/2208.05885]

- Full model too expensive to accurately estimate its sensitivity
- **Key idea**: Design a special proxy sensitivity quantity:
  - depends on full and surrogate models
  - if surrogate close to full model, proxy sensitivity close to full sensitivity
  - guaranteed to be at most full sensitivity
- Generate a lot of data from the surrogate model to estimate proxy sensitivity:
  - proxy sensitivity estimate much more accurate
  - comes with UQ for proxy sensitivity
  - lower confidence bound (LCB) for proxy sensitivity immediately a LCB for full sensitivity

Computational modeling: what can we do?

M. Aufiero and L. Janson. Surrogate-Based Global Sensitivity Analysis with Statistical Guarantees via **Floodgate**. 2022. [https://arxiv.org/abs/2208.05885]

- Full model too expensive to accurately estimate its sensitivity
- **Key idea**: Design a special proxy sensitivity quantity:
  - depends on full and surrogate models
  - if surrogate close to full model, proxy sensitivity close to full sensitivity
  - guaranteed to be at most full sensitivity
- Generate a lot of data from the surrogate model to estimate proxy sensitivity:
  - proxy sensitivity estimate much more accurate
  - comes with UQ for proxy sensitivity
  - lower confidence bound (LCB) for proxy sensitivity immediately a LCB for full sensitivity
  - same idea for upper-bound $\longrightarrow$ confidence interval for full sensitivity
  - **valid regardless of surrogate, very narrow if surrogate accurate**

Predictive uncertainty

The data: Observations of features and an outcome variable, with many observations

Predictive uncertainty

The data: Observations of features and an outcome variable, with many observations

E.g., patient electronic health records and survival data

Predictive uncertainty

The data: Observations of features and an outcome variable, with many observations

E.g., patient electronic health records and survival data

The goal: Provide prediction interval for outcome

Predictive uncertainty

The data: Observations of features and an outcome variable, with many observations

E.g., patient electronic health records and survival data

The goal: Provide prediction interval for outcome

E.g., range of survival times for patient

Predictive uncertainty

The data: Observations of features and an outcome variable, with many observations

E.g., patient electronic health records and survival data

The goal: Provide prediction interval for outcome

E.g., range of survival times for patient

The model: Complex (black box) predictive model (machine learning/AI)

Predictive uncertainty

The data: Observations of features and an outcome variable, with many observations

E.g., patient electronic health records and survival data

The goal: Provide prediction interval for outcome

E.g., range of survival times for patient

The model: Complex (black box) predictive model (machine learning/AI)

Why we need UQ: Important decisions made based on predictions:

- point predictions insufficient
- e.g., don't want to give patient false sense of certainty

Predictive uncertainty: what can we do?

A. Angelopoulos, S. Bates. **Conformal prediction**: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494-591, 2023.

Predictive uncertainty: what can we do?

A. Angelopoulos, S. Bates. **Conformal prediction**: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494-591, 2023.

- Complexity of predictive model prevents direct UQ

Predictive uncertainty: what can we do?

A. Angelopoulos, S. Bates. **Conformal prediction**: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494-591, 2023.

- Complexity of predictive model prevents direct UQ
- **Key idea**: use interchangeability of observations:
  - holding out a validation set, train on subset of observations
  - compute residuals on validation set
  - residuals on validation set distributed same as residual on a new data point

Predictive uncertainty: what can we do?

A. Angelopoulos, S. Bates. **Conformal prediction**: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494-591, 2023.

- Complexity of predictive model prevents direct UQ
- **Key idea**: use interchangeability of observations:
  - holding out a validation set, train on subset of observations
  - compute residuals on validation set
  - residuals on validation set distributed same as residual on a new data point
- Prediction $\pm$ high quantile of (absolute) validation residuals forms prediction interval:
  - exact coverage regardless of machine learning model
  - better predictions mean narrower intervals

Conclusion

Modern nonparametric statistical inference methods:

- rigorously **quantify uncertainty** for arbitrarily complex models
- require no knowledge of model—just need to be able to compute it

Conclusion

Modern nonparametric statistical inference methods:

- rigorously **quantify uncertainty** for arbitrarily complex models
- require no knowledge of model—just need to be able to compute it

Would love to engage with federal statistical applications!
ljanson@fas.harvard.edu

http://lucasjanson.fas.harvard.edu