Reliable Uncertainties for Stellar Properties from Gaia XP Spectra

Joshua D. Ingram[1], **James Carzon**[1], **Joshua S. Speagle**[2], **and Ann B. Lee**[1]

[1]Department of Statistics and Data Science, Carnegie Mellon University; [2]David A. Dunlap Department of Astronomy & Astrophysics, University of Toronto, Canada

GitHub Repository

Overview

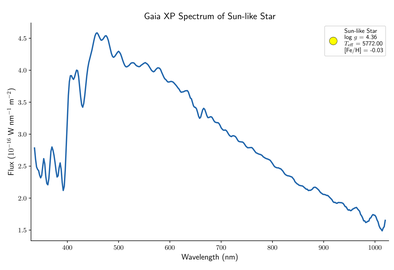
- **Gaia Data Release 3** (DR3) has low-resolution XP spectra for over 220 million stars
- Cross-matching with high-quality measurements from **APOGEE Data Release 17** (DR17) enables estimation of stellar properties (T_{eff}, $\log g$, [Fe/H]) at scale
- Cross-matched data are often **biased due to systematics**: instrument limitations and mission priorities result in **distribution shift**, where training and target data distributions do not match
- Conventional approaches often provide point estimates with no uncertainties [1]

Goal: Given a Gaia XP spectrum $x \in \mathbb{R}^{110}$, estimate stellar properties $\theta = (\log g,\ T_{\text{eff}},\ [\text{Fe/H}])$ with **reliable uncertainties having guaranteed local coverage** at any desired confidence level $1-\alpha$

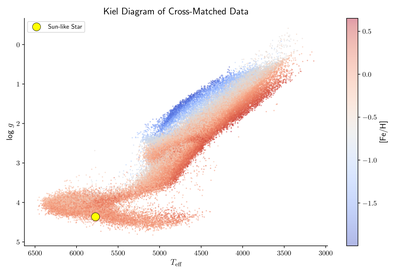
Gaia-APOGEE Cross-Matched Catalog

Data: **202,970** Gaia XP spectra cross-matched with APOGEE DR17 stellar properties [2]

- **XP Spectra ($\mathbf{X}$)**: 110-dimensional Hermite polynomial coefficients for constructing BP and RP optical spectra
- **Properties (θ)**: $\log g$, T_{eff}, [Fe/H] derived from high-resolution spectra



XP spectrum of a Sun-like star.



Kiel diagram of cross-matched data, colored by metallicity.

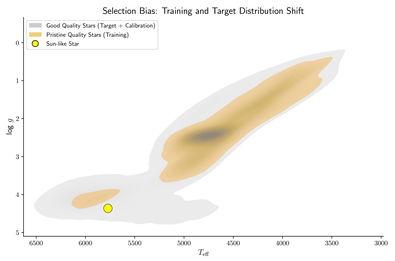
Inference Setting

Challenge: Construct uncertainties with desired level of local coverage for stellar properties that are reliable in the presence of selection bias

We study performance under two data settings:

1. **No selection bias:** Training data match target distribution
2. **Selection bias:** Train on only "pristine" quality data, predominantly giant-branch (GB) stars, with targets including main sequence (MS) and metal-poor stars; held-out "good" quality data used for calibration

1. No Selection bias: training distribution spans the target distribution of stars (no distribution shift).



2. Selection bias: "pristine" quality cuts concentrate training data on the giant branch (distribution shift).

Example Sun-like star: We illustrate performance with a held-out Sun-like star:

Gaia DR3 Source ID	Distance [pc]	$\log g$ [dex]	T_{eff} [K]	[Fe/H] [dex]
4660210013529490176	334.15	4.36	5772	−0.03

Reliable Uncertainties Via Confidence Distributions

Key Idea: Learn confidence distribution that encodes the confidence sets for all levels α simultaneously

Let $F(t \mid \theta) = P_{X|\theta}\big(\lambda(\theta, X) \le t\big)$ be the CDF of the test statistic under θ. Then the $1-\alpha$ confidence set

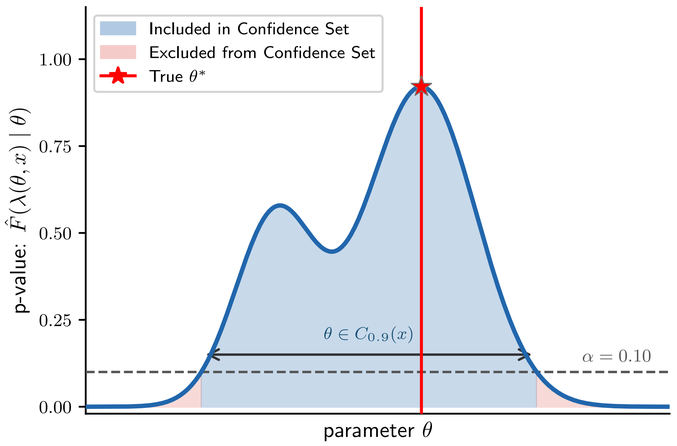
$$C_{1-\alpha}(x) \;=\; \{\theta : F(\lambda(\theta, x) \mid \theta) \ge \alpha\}$$

satisfies $P_{\theta^*}\big(\theta^* \in C_{1-\alpha}(X)\big) \ge 1-\alpha$ for all θ^* and all $\alpha \in (0,1)$.

Framework: [3] [4]

1. Train a **neural posterior estimator** (normalizing flow) $\hat{p}(\theta \mid x)$ on labeled spectra
2. Define test statistic $\lambda(\theta, x) = \log \hat{p}(\theta \mid x)$
3. For each θ, evaluate the **learned CDF** of λ on a held-out calibration set to obtain the p-value:
$$\hat{F}(\lambda(\theta, x) \mid \theta) \;=\; P_{X|\theta}\big(\lambda(\theta, X) \le \lambda(\theta, x)\big)$$
4. Confidence sets at any level $1-\alpha$ follow by thresholding:
$$C_{1-\alpha}(x) = \{\theta : \hat{F}(\lambda(\theta, x) \mid \theta) \ge \alpha\}$$

Key Advantage: Unlike posterior credible sets, confidence distributions achieve **guaranteed frequentist coverage** by construction, even when the posterior is trained under distribution shift.



1-dimensional confidence distribution sliced at level α; the set of θ with $\hat{F}\Big(\log \hat{p}(\theta \mid x_{\text{obs}}) \,\Big|\, \theta\Big) > \alpha$ are included in the confidence set.

Learning the CDF of the Test Statistic

We impose a **parametric model on the conditional CDF of the test statistic** $\hat{F}(\lambda \mid \theta)$, learned using calibration data $\{(\theta_i, \lambda_i)\}_{i=1}^n$

Sigmoid CDF Model: We model $F(\lambda \mid \theta)$ as a logistic function with θ-dependent location and scale parameters:

$$\hat{F}(\lambda \mid \theta) \;=\; \sigma\big(\kappa(\theta)\,(\lambda - \mu(\theta))\big), \qquad \sigma(t) = \tfrac{1}{1+e^{-t}}$$

where $\mu(\theta)$ is the location (median) and $\kappa(\theta) > 0$ controls the spread/sharpness.

Shallow MLP: A shallow multi-layer perceptron (MLP) maps θ to the CDF parameters:

$$\text{MLP}(\theta) \;=\; \big(\mu(\theta),\ \log \kappa(\theta)\big), \qquad \theta \in \mathbb{R}^4 \to \mathbb{R}^2$$

- Architecture: 2 hidden layers $\times$ 64 units, `tanh` activations; θ inputs normalized

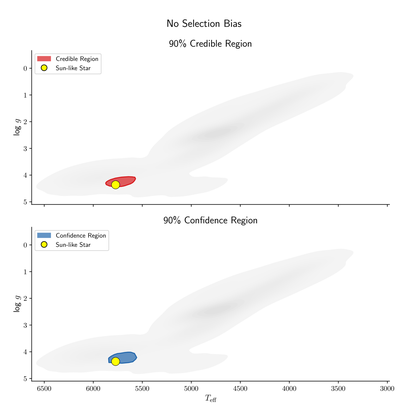
Inference: For a new observation x_{obs}, evaluate the learned CDF over a grid of θ, and construct $1-\alpha$ confidence sets by thresholding at level α:

$$C(\theta \mid x_{\text{obs}}) \;=\; \hat{F}\Big(\log \hat{p}(\theta \mid x_{\text{obs}}) \,\Big|\, \theta\Big) > \alpha$$

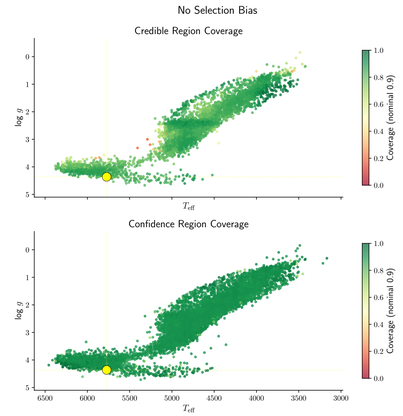
Implementation

GitHub: Available in the open-source `lf2i` package: `github.com/lee-group-cmu/lf2i`

Results: No Selection Bias

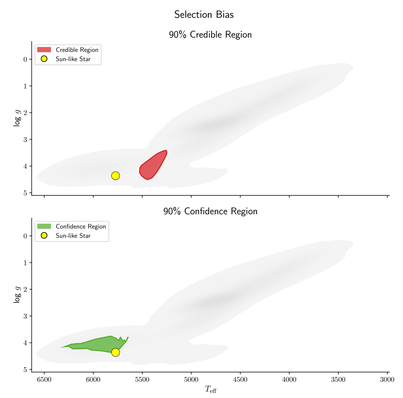


90% confidence region (blue) and HPD credible region (red) for the Sun-like star under selection bias.

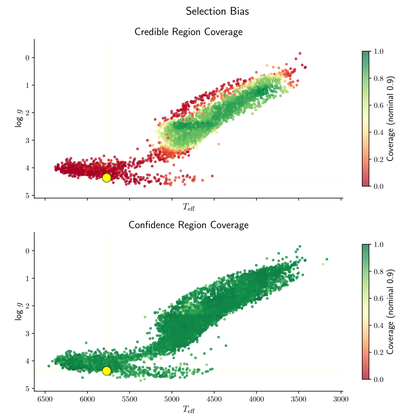


Local coverage at 90% level. Confidence regions achieves desired coverage while credible regions may under-cover.

Results: Selection Bias



90% confidence region (green) and credible region (red) for the Sun-like star under selection bias.



Local coverage at 90% level. Confidence regions achieve desired coverage while credible regions severely under-cover.

Takeaways

- **Confidence sets achieve nominal local coverage** in both settings
- **Credible regions under-cover**, especially under selection bias where the posterior fails to cover main-sequence and metal-poor stars
- **Calibration is modular and amortized**: any pre-trained posterior estimator can serve as the test statistic; only a held-out calibration set is required, no retraining needed

Next Steps

- Extend to estimate additional stellar properties, such as [α/M]
- Explore richer parametric CDF families beyond the logistic model and alternative regressors
- Apply to larger set of the Gaia DR3 catalog (up to 220M+ stars) for large-scale inference

References

[1] Andrae et al. (2023) [2] Laroche & Speagle (2025) [3] Carzon et al. (2026) [4] Dalmasso et al. (2021)

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No DGE-2140739. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Contact: jdingram@andrew.cmu.edu