Ordered Correlation Forest

Riccardo Di Francesco

April 22, 2025

Abstract

Empirical studies in various social sciences often involve categorical outcomes with inherent ordering, such as self-evaluations of subjective well-being and self-assessments in health domains. While ordered choice models, such as the ordered logit and ordered probit, are popular tools for analyzing these outcomes, they may impose restrictive parametric and distributional assumptions. This article introduces a novel estimator, the *ordered correlation forest*, that can naturally handle non linearities in the data and does not assume a specific error term distribution. The proposed estimator modifies a standard random forest splitting criterion to build a collection of forests, each estimating the conditional probability of a single class. Under an "honesty" condition, predictions are consistent and asymptotically normal. The weights induced by each forest are used to obtain standard errors for the predicted probabilities and the covariates' marginal effects. Evidence from synthetic data shows that the proposed estimator features a superior prediction performance than alternative forest-based estimators and demonstrates its ability to construct valid confidence intervals for the covariates' marginal effects. Comparisons using various real-world data sets further highlight the advantages of forest-based estimators over parametric models in larger samples while showing that the ordered correlation forest remains competitive in smaller samples.

Keywords: Ordered non-numeric outcomes, choice probabilities, machine learning.
JEL Codes: C14, C25, C55