Release summary

- Release Version: 0.03
- Release Type: beta
- Product Type: R library
- Date: 20 December 2022
- Contact: Charles Shaw

1 Motivation

To draw scientifically meaningful conclusions and build reliable models of quantitative phenomena, uncertainty quantification must be taken into consideration (either implicitly or explicitly). This is particularly challenging when the measurements are not from controlled experimental (interventional) settings, since estimates can be obscured by spurious, indirect influences. Modern predictive techniques from machine learning are capable of capturing high-dimensional, nonlinear relationships between variables while relying on few parametric or probabilistic model assumptions.

However, since these techniques are associational, applied to observational data they are prone to picking up spurious influences from non-experimental (observational) data, making their predictions unreliable. thms since they typically require explicit probabilistic models. We augment Robyn with nonparametric bootstrap resampling to estimate the uncertainty about the population. In this way, we can use modern statistical learning algorithms unaltered to make statistically powerful, yet causally-robust, predictions.

2 What is dryad MMM

dryad MMM is a fork of Robyn, the semi-automated and open-sourced Marketing Mix Modeling (MMM) package from Meta Marketing Science. Robyn uses various machine learning techniques (Ridge regression, multi-objective evolutionary algorithm for hyperparameter optimization, time-series decomposition for trend and season, gradient-based optimization for budget allocation etc.) to define media channel efficiency and effectivity, explore adstock rates and saturation curves.

Dryad extends Robyn by performing bootstrapped ridge regressions for the same models to estimate penalized (added bias to reduce variance) least squares (PLS) model coefficients, which are more optimal, stable estimates that are more likely to validate in new samples.

Dryad further extends Robyn by adding diagnostics, visuals, augmented numerical outputs, and a forecasting function. To help uncover causal interdependencies from observational data (one of the great challenges of nonlinear time series analysis!), we use the popular Shannon-entropy based Transfer Entropy, which represents a prominent tool for assessing directed information flow between joint processes.

3 Features

- Transfer entropy : a non-parametric measure of directed, asymmetric information transfer between depVar and paid media spends
- KPSS statistic to test the null of a unit root against stationary alternatives
- Traditional in-sample diagnostic measures such as confidence intervals, p-values, and the concept of statistical significance.
- QQ-plot
- Bootstrapped standard errors
- Cramer von Mises test for normality
- Other visual and statistical features.

4 What's new?

4.1 Extended module: model.r

Dryad extends Robyn by performing bootstrapped ridge regressions for the same models to estimate penalized model coefficients. The bootstrap merges simulation with formal model-based statistical inference. A statistical model for a sample X_n of size n is a family of distributions $\{P_{\theta,n}: \theta \in \Theta\}$. The parameter space Θ is typically metric, possibly infinite-dimensional. The value of θ that identifies the true distribution from which X_n is drawn is unknown. Suppose that $\hat{\theta}_n = \hat{\theta}_n(X_n)$ is a consistent estimator of θ. The bootstrap idea is:

(a) Create an artificial *bootstrap world* in which the true parameter value is $\hat{\theta}_n$ and the sample X_n^* is generated from the fitted model $P_{\hat{\theta}_n,n}$. That is, the conditional distribution of X_n^*, given the data X_n, is $P_{\hat{\theta}_n,n}$.

(b) Act as if a sampling distribution computed in the fully known bootstrap world is a trustworthy approximation to the corresponding, but unknown, sampling distribution in the model world.

4.2 New module: exposure_spend.r

Shows output of regression analysis that represents the relationship between spend and exposure variables in a given data set.

4.3 New module: forecast.r

We use Prophet to generate a n-step forecast. Facebook's Prophet is an additive regression model designed for making forecasts for uni-variate time series datasets and to automatically find an optimal set of hyperparameters for the model with trends and seasonal structure by default. It belongs to the family of General Additive models (GAM) which fit a set of smooth functions that describe trend, seasonality and predictable special events or cycles to the data. The GAM is the sum of its smooth functions. A GAM treats a time series as a curve-fitting exercise.

Its core comprises of the sum of three functions of time plus an error term: growth or trend g(t), seasonality s(t), holidays h(t) and error e_t. The growth function incorporates "changepoints", which are moments in the data where the data shifts direction, to model the overall trend of the data. The seasonality function is a Fourier Series as a function of time. Prophet automatically detects the Fourier order which is the optimal number of terms in the series. The holiday function allows Prophet to adjust accordingly to an anomaly that may change the forecast. The error term stands for random fluctuations that cannot be explained by the model.

4.4 New module: diagnostics.r

4.4.1 Transfer Entropy

The transfer entropy (TE) is a nonlinear measure that quantifies the amount of information explained in Y at h time steps ahead from the state of X accounting for the concurrent state of Y. Let x_t, y_t be two stationary time series and $\mathbf{x}_t = [x_t, x_{t-1}, \ldots, x_{(m-1)\tau}]'$ and $\mathbf{y}_t = [y_t, y_{t-1}, \ldots, y_{(m-1)\tau}]'$ the reconstructed vectors of the state space of each system, where τ is the delay time and m is the embedding dimension. The TE from X to Y can be defined based on entropy terms as

$$TE_{X \to Y} = -H(y_{t+h} \mid \mathbf{x}_t, \mathbf{y}_t) + H(y_{t+h} \mid \mathbf{y}_t)$$

where $H(x)$ is the Shannon entropy of the variable X.

4.4.2 KPSS Statistic

The KPSS test assesses the null hypothesis that a univariate time series is trend stationary against the alternative that it is a nonstationary unit root process.

4.4.3 Cramer von Mises statistic

Cramér-von Mises's test is an empirical distribution function omnibus test for the composite hypothesis of normality.

4.4.4 QQ Plot

The quantile-quantile (q-q) plot is a graphical technique for determining if the distribution of the data against the expected normal distribution. For normally distributed data, observations should lie approximately on a straight line.