MODEL-FREE BOOTSTRAP

Dimitris N. Politis
Department of Mathematics and Halicioglu Data Science Institute
University of California–San Diego, La Jolla, CA 92093-0112, USA

1 Model-based regression and bootstrap

With the advent of widely accessible powerful computing in the late 1970s, computer-intensive methods such as resampling and cross-validation created a revolution in modern statistics. Using computers, statisticians became able to analyze big datasets for the first time, paving the way towards the 'big data' era of the 21st century. But perhaps more important was the realization that the way we do the analysis could/should be changed as well, as practitioners were gradually freed from the limitations of parametric models. For instance, the great success of the bootstrap was in providing a complete framework for statistical inference under a nonparametric setting much like Maximum Likelihood Estimation had done half a century earlier under the restrictive parametric setup.

The original bootstrap of Efron (1979) was designed for data $Y_1, \ldots, Y_n$ that are independent and identically distributed (i.i.d.). However, it was soon realized that one can resample residuals from a model as if they were i.i.d. To fix ideas, let us first focus on regression, i.e., data that are pairs: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ where Y_i is the measured response associated with a (deterministic) regressor value of X_i. The standard homoscedastic additive model in this situation reads:

$$Y_i = \mu(X_i) + \epsilon_i \tag{1}$$

where the ϵ_i are i.i.d. from a distribution $F(\cdot)$ with mean zero; consider three cases:

Parametric model: Both $\mu(\cdot)$ and $F(\cdot)$ belong to parametric families of functions, e.g., $\mu(x) = \beta_0 + \beta_1 x$ and $F(\cdot)$ is $N(0, \sigma^2)$.
Semiparametric model: $\mu(\cdot)$ belongs to a parametric family but $F(\cdot)$ does not.
Nonparametric model: Neither $\mu(\cdot)$ nor $F(\cdot)$ can be assumed to belong to parametric families of functions. Instead, it is typically assumed that $\mu(\cdot)$ belongs to some smoothness class, e.g., it is twice differentiable.

Whether the model is parametric, semiparametric or nonparametric, the practitioner will typically estimate $\mu(\cdot)$ by a consistent estimator $\hat{\mu}(\cdot)$, and construct the residuals $e_i = Y_i - \hat{\mu}(X_i)$. Consistency will then imply that the residual e_i is a good proxy for the unobserved error ϵ_i when n is large. Since $E\epsilon_i = 0$, it is advisable to center the residuals, i.e., define $r_i = e_i - n^{-1} \sum_j e_j$, and use r_i as a proxy for ϵ_i. We can then resample the r_is as if they were i.i.d. yielding $r_1^*, \ldots, r_n^*$; we then create a bootstrap dataset $(Y_1^*, X_1), (Y_2^*, X_2), \ldots, (Y_n^*, X_n)$ by letting $Y_i^* = \hat{\mu}(X_i) + r_i^*$. Finally, we re-compute the estimator $\hat{\mu}(\cdot)$ on many such bootstrap datasets, and

witness how it varies across different datasets in order to gauge its accuracy; this procedure is called a *residual* bootstrap—see e.g. Efron and Tibshirani (1993).

2 Model-free regression and bootstrap

Even under the flexible nonparametric setup, eq. (1) constitutes a model, and can thus be rather restrictive. For example, the well known `cps71` dataset from the `np` package of R has been a workhorse for nonparametric function estimation. As it turns out, it can not be modelled by eq. (1) even after allowing for heteroscedasticity of the errors— see Ch. 4.2 of Politis (2015).

Nevertheless, it is possible to shun eq. (1) altogether and instead adopt a *model-free* regression setup. The deterministic design case is described below but a random design is also possible—see Wang and Politis (2021).

Model-free regression: The variables $X_1, \ldots, X_n$ are deterministic, and the random variables $Y_1, \ldots, Y_n$ are independent with common conditional distribution, i.e., $P\{Y_j \leq y | X_j = x\} = D_x(y)$ not depending on j.

Inference for features, i.e. functionals, of the common conditional distribution $D_x(\cdot)$ is still possible under some regularity conditions, e.g. smoothness. Arguably, the most important such feature is the conditional mean $E(Y|X = x)$ that can be denoted $\mu(x)$. When $\mu(x)$ can be assumed smooth, it can be consistently estimated by a nonparametric estimator $\hat{\mu}(x)$; the procedure is completely analogous to nonparametric estimation of $\mu(x)$ under model (1).

The question is: *how to gauge the accuracy of $\hat{\mu}(x)$ without assuming eq. (1)?* The **Model-free bootstrap** comes to the rescue; to describe it, we briefly move away from the regression setup, and consider a data vector $\underline{Y}_n = (Y_1, \ldots, Y_n)'$ whose elements are not i.i.d.; this can be either because they are not identically distributed (a regression effect), or not independent (a time series effect), or both. The practitioner now uses the structure of the problem in order to **find an invertible transformation H_n that can map the non-i.i.d. vector $\underline{Y}_n$ to a vector $\underline{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)'$ that has i.i.d. components.** Letting H_n^{-1} denote the inverse transformation, we have $\underline{\epsilon}_n = H_n(\underline{Y}_n)$ and $\underline{Y}_n = H_n^{-1}(\underline{\epsilon}_n)$, i.e.,

$$\underline{Y}_n \overset{H_n}{\longmapsto} \underline{\epsilon}_n \quad \text{and} \quad \underline{\epsilon}_n \overset{H_n^{-1}}{\longmapsto} \underline{Y}_n. \tag{2}$$

Under regularity condition such a transformation always exists but is not unique—see Ch. 2.3.3 of Politis (2015). It is up to the ingenuity of the practitioner to employ a transformation H_n whose form is easily estimable from the data at hand.

Let $\widehat{H}_n$ denote the data-based estimate of H_n, and define $\underline{e}_n = \widehat{H}_n(\underline{Y}_n)$; we will use e_i as a proxy for ϵ_i. The Model-Free bootstrap procedure goes as follows: resample the e_is as if they were i.i.d. yielding $e_1^*, \ldots, e_n^*$; let $\underline{e}_n^* = (e_1^*, \ldots, e_n^*)'$, and construct a bootstrap data vector by $\underline{Y}_n^* = \widehat{H}_n^{-1}(\underline{e}_n^*)$. We can now re-compute our

estimator of interest on many such bootstrap datasets, and witness how it varies across different datasets in order to gauge its accuracy.

Going back to the regression setup, it is apparent that the transformation H_n will depend on the regressor values $X_1, \ldots, X_n$. To fix ideas, assume the Model-free regression setup with the distribution $D_x(y)$ assumed continuous in both x and y. Letting $\epsilon_i = D_{X_i}(Y_i)$, it is apparent that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Uniform (0,1) by the probability integral transform. Assuming that for each x, the function $D_x(\cdot)$ is one-to-one, we may then write $Y_i = D_{X_i}^{-1}(\epsilon_i)$, thus verifying both parts of eq. (2). Although $D_x(y)$ will generally be unknown, it is straightforward to employ a nonparametric estimator $\widehat{D}_x(y)$ to construct $\widehat{H}_n$; see Politis (2013). Consequently, $e_i = \widehat{D}_{X_i}(Y_i)$ will be our proxy for ϵ_i.

Variations to this theme are possible. For example, the bootstrap data vector $\underline{Y}_n^* = \widehat{H}_n^{-1}(\underline{e}_n^*)$ where $\underline{e}_n^* = (e_1^*, \ldots, e_n^*)'$ can be obtained either by resampling the proxies $(e_1, \ldots, e_n)$ as already discussed *or* by i.i.d. sampling from a Uniform (0,1). The latter was termed a Limit Model-free (LMF) bootstrap by Politis (2015) since it employs the theoretical distribution of the ϵ_is which is the limiting distribution of the e_is. Interestingly, the LMF approach retains its validity even when $D_x(\cdot)$ and/or $\widehat{D}_x(\cdot)$ are not invertible; in this case, $Y_i^* = \widehat{D}_{X_i}^{-1}(e_i^*)$ with $e_1^*, \ldots, e_n^*$ drawn i.i.d. Uniform (0,1), and $\widehat{D}_x^{-1}(\cdot)$ denoting the quantile inverse of distribution $\widehat{D}_x(\cdot)$.

Further examples of transformations applicable to diverse settings with regression and/or time series data are discussed in Wang and Politis (2022). Beyond inference on parameters, e.g. confidence intervals, hypothesis tests, etc., the Model-free bootstrap is also applicable to predictive inference, e.g. prediction intervals; the latter would then follow the *Model-free Prediction Principle* of Politis (2015).

References

[1] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann. Statist.*, vol. 7, pp. 1–26.

[2] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

[3] Politis, D.N. (2013). Model–free model–fitting and predictive distributions, *Test*, vol. 22, no. 2, pp. 183–250.

[4] Politis, D.N. (2015). *Model–free Prediction and Regression: a Transformation-based Approach to Inference*, Springer, New York.

[5] Wang, Y. and Politis, D.N. (2021). Model-free bootstrap and conformal prediction in regression: conditionality, conjecture testing, and prediction intervals, Preprint arXiv:2109.12156.

[6] Wang, Y. and Politis, D.N. (2022). Model-free bootstrap for a general class of stationary time series, *Bernoulli*, vol. 22, no. 2, pp. 744-770.