Conformal Prediction

When doing estimation, we usually provide confidence intervals in addition to point estimates. Is there a similar notion for predictions? The answer is yes: we provide *prediction* sets or *set-valued predictions*. Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ we construct a set-valued function C_n, depending on $(X_1, Y_1), \ldots, (X_n, Y_n)$ such that

$$P(Y_{n+1} \in C_n(X_{n+1})) \geq 1 - \alpha.$$

The approach we consider in these notes is *conformal prediction*. The idea is due to Vovk, Gammerman and Shafer (2005). The statistical theory for conformal prediction was developed in Lei, Robins and Wasserman (2013), Lei and Wasserman (2014), Lei, G'Sell, Rinaldo, Tibshirani and Wasserman (2017), Sadinle, Lei and Wasserman (2018).

The Unsupervised Case. We begin with the following problem. We observe $Y_1, \ldots, Y_n$ and we want to predict Y_{n+1}. The basic algorithm is as follows:

1. Observe $Y_1, \ldots, Y_n$.
2. Define a permutation invariant *residual function* (or *conformity score*) $R_i = \phi(y, \mathcal{A})$ where $\mathcal{A}$ is any dataset of size $n + 1$.
3. For each y:
   (a) Set $Y_{n+1} = y$ and form the augmented dataset $\mathcal{A} = \{Y_1, \ldots, Y_{n+1}\}$.
   (b) Let $R_i = \phi(Y_i, \mathcal{A})$ for $i = 1, \ldots, n + 1$.
   (c) Test the hypothesis $H_0 : Y_{n+1} = y$ by computing the p-value

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1}).$$

   (d) Invert the test: set

$$C_n = \{y : \ \pi(y) \geq \alpha\}.$$

Note that when H_0 is true, the residuals are exchangeable and the p-value is uniform. Therefore, we have:

Theorem 1 *For every P,*

$$P(Y_{n+1} \in C_n) \geq 1 - \alpha.$$

If P is absolutely continuous, we also have $P(Y_{n+1} \in C_n) \leq 1 - \alpha + \frac{1}{n+1}$.

Note that this result is distribution-free and holds for all finite samples.

A simple example of a residual function is

$$R_i = \left| Y_i - \frac{Y_1 + \cdots + Y_{n+1}}{n+1} \right|.$$

A more complicated residual is

$$R_i = \frac{1}{\widehat{p}_h(Y_i)}$$

where $\widehat{p}_h$ is a kernel density estimator constructed from the augmented data.

The coverage validity of the prediction set does not depend on the choice of residual. But a poor choice can lead to large prediction sets. A careful choice can lead to minimax optimal sets. For example, suppose that P has a density p. Let t_α be such that $P(Y \in C_*) = 1 - \alpha$ where $C_* = \{y : \ p(y) \geq t_\alpha\}$. Note that C_* is the smallest set such that $P(Y \in C) = 1 - \alpha$. Suppose that $p \in \text{Holder}(\beta)$ and that there exist c_1, c_2 and γ such that

$$c_1|\epsilon|^\gamma \leq |P(p(Y) \leq t_\alpha + \epsilon) - \epsilon| \leq c_2|\epsilon|^\gamma$$

for all small ϵ. In this case, any prediction set must satisfy $\mu(C_* \Delta C_n) \geq r_n$ with high probability, where μ is Lebesgue, Δ is Lebesgue measure and

$$r_n = \left(\frac{\log n}{n} \right)^{\frac{\beta\gamma}{2\beta+d}}.$$

Theorem 2 *The conformal set C_n based on the kernel density estimator (with appropriate bandwidth) satisfies*

$$P\left(\mu(C_n \Delta C_\alpha) \geq r_n\right) \leq \left(\frac{1}{n}\right)^\lambda$$

for any $\lambda > 0$.

For a proof, see Lei, Robins and Wasserman (2013). Thus, in this case, C_n is minimax under the stated conditions. But C_n still has $1 - \alpha$ coverage even if the conditions fail. In fact, C_n has $1 - \alpha$ coverage even if P does not have a density.

The algorithm above requires that we test $H_0 : Y_{n+1} = y$ for every y. In practice, we only consider a grid of values for y. But this can be slow. The *split conformal method* is much faster. The steps are:

1. Split the data into two sets $\mathcal{D}_1$ and $\mathcal{D}_2$.
2. Compute the residuals $R_i = \phi(Y_i, \mathcal{D}_1)$ for $Y_i \in \mathcal{D}_1$.

3. Let q be the $1-\alpha$ quantile of the residuals.
4. Return $C_n = \{y:\ \phi(y, \mathcal{D}_1) \leq q$.

It is not hard to show that, once again we have

$$P(Y_{n+1} \in C_n) \geq 1 - \alpha$$

for all P. The split conformal method is fast but can result in larger prediction sets. Also, it depends on the particular split of the data. We might consider combining several splits. Suppose that we split the data N times. For each split we construct a prediction set C_j at level $1 - \alpha/N$. It follows that $C = \bigcap_{j=1}^{N} C_j$. It follows from the union bound that this is valid at level $1 - \alpha$. There are two effects: replacing α with α/N makes each set larger. But taking the intersection makes the set smaller. Unfortunately it can be shown that, under fairly general conditions, that the Lebesgue measure of C is larger than the set constructed with one split, with probability tending to 1. So there seems to be no advantage to suing several splits.

Regression. The extension to regression is straightforward. The data are $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. We augment the data with a new point (x, y). Again we define a residual $R_i = \phi((X_i, Y_i), \mathcal{A})$ and we define

$$\pi(x, y) = \frac{1}{n+1} \sum_i I(R_i \geq R_{n+1}).$$

Then we set $C_n(x) = \{y:\ \pi(x, y) \geq \alpha\}$. We then have

$$P(Y_{n+1} \in C_n(X_{n+1})) \geq 1 - \alpha$$

for every P.

An example of a residual is

$$R_i = |Y_i - \widehat{m}(X_i)|$$

where $\widehat{m}$ is based on the augmented data. The validity holds even if the model is wrong. Again we can use splitting to speed up the calculations.

Note that the coverage guarantees are marginal. Under regularity conditions it can bge shown that we get asymptotic conditional covage, that is,

$$P(Y_{n+1} \in C_n(x) | X_{n+1} = x) \to 1 - \alpha.$$

It is not possible to get finhite sample, distribution-free conditional coverage as shown on Lei and Wasserman (2014).

We can apply this method to high dimensional and nonparametric regression. The nice thing is that we do not need the model to be correct. To see how well it works, see Figures 1, 2 and 3. (These are from Lei, G'Sell, Rinaldo, Tibshirani and Wasserman 2017.)

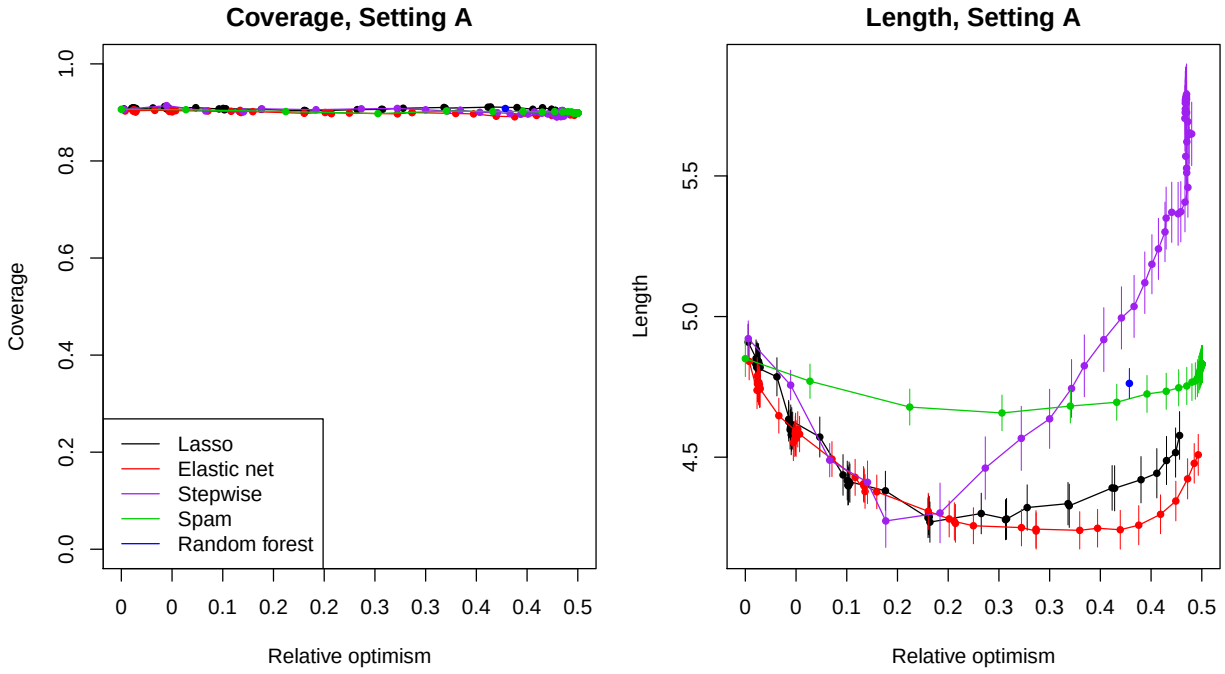

Figure 1: Example: $n = 200, d = 2,000$; linear and Normal

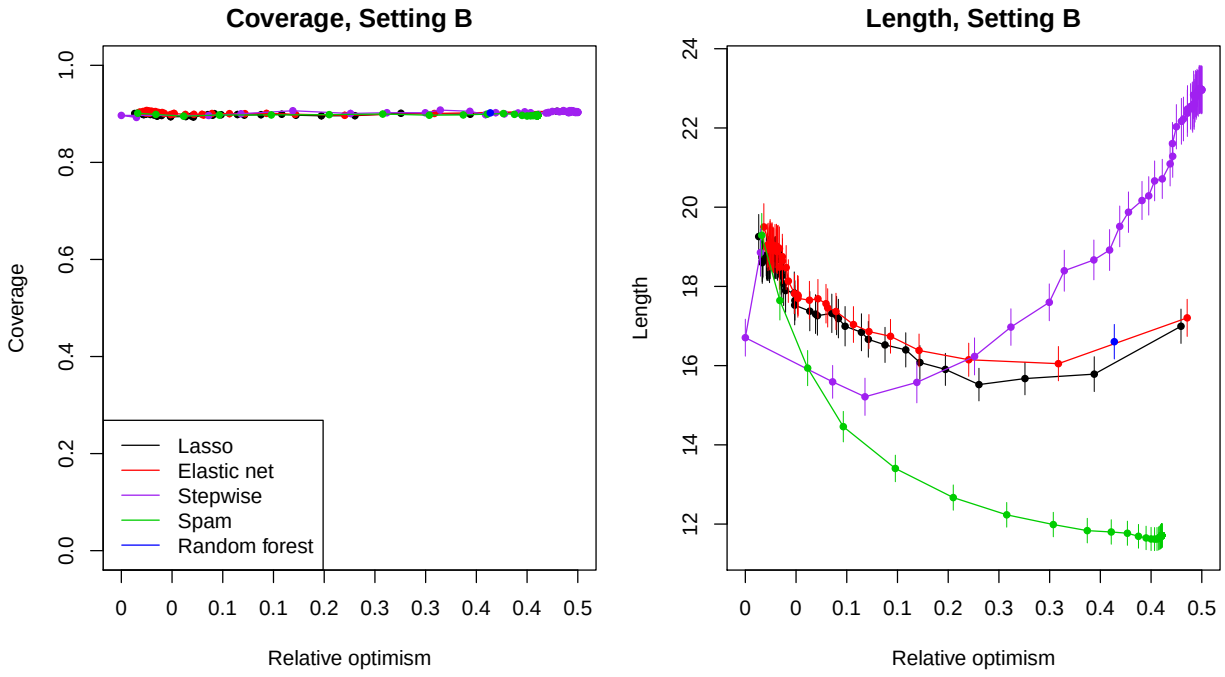

Figure 2: Example: $n = 200, d = 2,000$; nonlinear and heavy-tailed

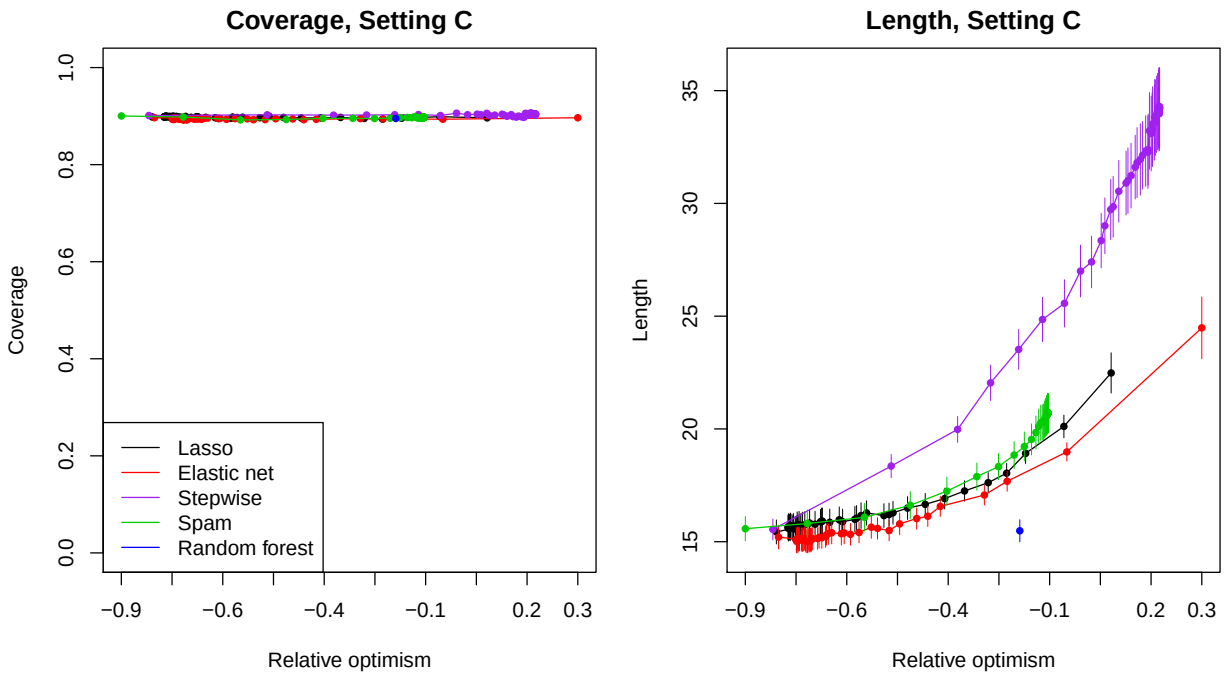

Figure 3: Example: $n = 200, d = 2,000$; linear, correlated, heteroskedastic, heavy-tailed

Classification. The extension to classification is straightforward. The only change is the choice of residual. An example of such a score is $1/\widehat{p}(Y_i|X_i)$. Another example is the nearest neighbor score

$$R_i = \frac{\min_{i:\ Y_i=y} ||x - X_i||}{\min_{i:\ Y_i\neq y} ||x - X_i||}.$$

One complication is that sometimes $C_n(x) = \emptyset$. Some methods for fixing this are discussed in Sadinle, Lei and Wasserman (2018). On the other hand, if one uses the score $1/\widehat{p}(X_i|Y_i)$ then $C_n(x) = \emptyset$ when X_i is an outlier i.e. we have not seen a datapoint like X_i before.