Statistical Learning: Distribution-free Prediction and Confidence Intervals for Linear Regression Problems

Babolcsay, Barbara

1. Introduction

One of the most fundamental methods in statistics and therefore in machine learning is linear regression. It provides us a model - a point estimate of the parameters - fitting our noisy observations well and so a prediction in the not yet visited points. A natural expectation is to guarantee a confidence region around these estimated values. There are methods that work under assumptions about the distribution but in real life we hardly have appropriate prior knowledge. There are also asymptotic results but these are not exact on a finite sample. In this report I describe an algorithm introduced by B. Cs. Csáji, M. C. Campi and E. Weyer in (2) which gives non-asymptotic distribution-free confidence regions for the parameters of our model. I summarize Sign-Perturbed Sums (SPS) method and show how we can use it through numerical experiencies.

2. SPS method

2.1. Exact Confidence Regions

Consider a linear regression system:

$$Y_t = \phi_t^T \theta^* + N_t, \tag{1}$$

where Y_t is the output, ϕ_t (d dimensional) is the regressor, θ^* (d dimensional) is the real parameter and N_t is the noise. We assume that our observations of size n are in the upper form. It is known that the Least-Squares Estimate (LSE) of θ^* is:

$$\hat{\theta_n} = (\Phi_n^T \Phi_n)^{-1} \Phi_n^T Y,$$

where Φ_n is the matrix of the n regressors and Y is the vector of the n outputs. The first goal of SPS is to decide if a given θ is in the region that contains the real parameter with a given probability or not.

The only asumptions we need:

- $\{N_t\}$ must be independent and have symmetric distribution around zero
- If we define R_n as: $R_n = \frac{1}{n}\Phi_n^T\Phi_n = \frac{1}{n}\sum_{t=1}^n \phi_t\phi_t^T$, R_n must be invertible

The main idea of the algorithm is to perturb the sign of the prediction errors of our LSE in order to evaluate the uncertainty of the estimate. $\hat{\theta}$ was the root of the following equation:

$$0 = \sum_{t=1}^n \phi_t(Y_t - \phi_t^T\theta) = \sum_{t=1}^n \phi_t\phi_t^T(\theta^* - \theta) + \sum_{t=1}^n \phi_t N_t,$$

from which we can generate:

$$H_0(\theta) = \sum_{t=1}^n \phi_t\phi_t^T(\theta^* - \theta) + \sum_{t=1}^n \phi_t N_t$$

and

$$H_i(\theta) = \sum_{t=1}^n \alpha_{i,t}\phi_t\phi_t^T(\theta^* - \theta) + \sum_{t=1}^n \alpha_{i,t}\phi_t N_t,$$

where $\{\alpha_{i,t}\}$, $i = 1, .., m-1, t = 1, ..., n$ are random signs ($P(\alpha_{i,t} = 1) = P(\alpha_{i,t} = -1) = \frac{1}{2}$).

In (2) it is shown that if $\|\theta^* - \theta\|_2$ is large - so the parameter in question is far from the real parameter - then $\|H_0(\theta)\|_2$ dominates in the ordering of $\{\|H_i(\theta)\|_2\}$. So if we want to build a confidence set $\hat{\Theta}$ with probability $p = 1 - \frac{q}{m}$, we shall compare the following values:

$$\|S_0(\theta)\|_2 = \|R_n^{1/2}\frac{1}{n}H_0(\theta)\|_2$$
$$\|S_i(\theta)\|_2 = \|R_n^{1/2}\frac{1}{n}H_i(\theta)\|_2$$

for $i = 1, ..., m-1$ and if $\|S_0(\theta)\|_2$ is at most the $(m-q)th$ smallest then $\theta \in \hat{\Theta}$.

With this method we can check every θ in the parameter space and define confidence regions. These are star-convex around $\hat{\theta}$ and contains θ^* with the desired certainty. The exact pseudocode of the algorithm may be found in (2).

2.2. Ellipsoidal Outer Approximation

In practice we might need confidence regions that are easier to calculate than the exact ones. For this reason we can define ellipsoids around $\hat{\theta}$ that is an over-bound of $\hat{\Theta}$. Considering a larger set we quarentee that the probability of the real parameter being in the ellipsoid is greater than what we want.

If $\theta \in \hat{\Theta}$ than:

$$\|S_0(\theta)\|_2^2 = (\theta - \hat{\theta})R_n^{-1}(\theta - \hat{\theta}) \leq r(\theta),$$

where $r(\theta)$ is the qth largest value in $\{\|S_i(\theta)\|_2^2\}$. We want to give a θ-free upper bound which leads to a maximization and in dual a *convex minimization* problem for all $i = 1, ..., m-1$:

$$\min \gamma$$
$$\text{s.t.} \lambda \geq 0$$
$$\begin{bmatrix} -I + \lambda A_i & \lambda b_i \\ \lambda b_i^T & \lambda c_i + \lambda \end{bmatrix} \geq 0,$$

where A_i, b_i, c_i come from $\Phi_n, Y, \hat{\theta}$ and $\{\alpha_{i,t}\}$. After solving the $m-1$ convex minimization problem, the qth largest optimum will be the proper radius of our ellipsoid.

2.3. Modifying the method to Ridge Regression

Ridge Regression (RR) is a commonly used regularized version of the Ordinary Least-Squares (OLS) problem where the function to minimize is:

$$\|Y - \Phi\Theta\|_2^2 - \lambda\|_2\Theta\|_2.$$

It can be reformulated as an OLS problem with $\Phi_{RR} = [\Phi, \sqrt{\lambda}I]^T$ and $Y_{RR} = [Y, 0]^T$ so we can easily determine the analytic solution:

$$\hat{\theta}_{n,RR} = (\Phi_n^T\Phi_n + \lambda I)^{-1}\Phi_n^T Y.$$

In (1) it is discussed how we can use SPS to generate confidence regions around RR-estimates.

As we have the OLS version of the RR-problem, we could apply SPS with Φ_{RR} and Y_{RR}. The only change we have to make is in the perturbation part: because the last d rows in Φ_{RR} and zeros in Y_{RR} are only responsible for encoding the regularization, we shall not perturb the signs of these residuals.

3. From the Parameter Space to the Function Space

Until now we were in the space of parameters: $\theta \in \mathbb{R}^d$, but in the case of $d > 3$ we can represent the confidence regions better in the function space: $f : \mathbb{R} \to \mathbb{R}$. We would like to have an interval in each point around the estimated function value we have from the LSE. We can transport our results here with the help of the ellipsoidal outer approximation of our parameter. If we fix a $t_0 \in \mathbb{R}$, that gives us a linear optimization proplem on an ellipsoid (Figure 1).

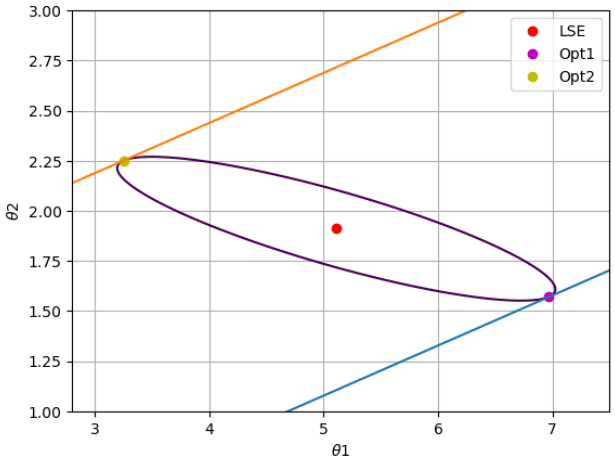

Figure 1: Example for a linear optimization problem on a confidence ellipsoid in the parameter space ($\theta = (\theta 1, \theta 2) \in \mathbb{R}^2$)

This problem has an exact analytic solution which one can find e.g. using Lagrangian relaxation what I would like to show.

The convex minimization problem:

$$\min c^T\theta$$
$$\text{s.t. } (\theta - \hat{\theta})R_n(\theta - \hat{\theta}) < r,$$

where $c = \phi_{t_0}$, $\hat{\theta}$ is the LSE, $R_n = \frac{1}{n}\Phi_n^T\Phi_n$ and r is the radius of the confidence ellipsoid.

If we take the Lagrangian form:

$$l(\theta, \lambda) = c^T\theta + \lambda(\theta - \hat{\theta})R_n(\theta - \hat{\theta})$$
$$\frac{d}{d\theta}l(\theta, \lambda) = c + \lambda 2R_n(\theta - \hat{\theta}) = 0 \Rightarrow \theta_{\text{inf}} = \hat{\theta} - \frac{1}{2\lambda}R_n^{-1}c$$
$$\inf_\theta l(\theta, \lambda) = -\frac{1}{4\lambda}c^T R_n^{-1}c + c^T\hat{\theta} - \lambda r$$
$$\frac{d}{d\lambda}\inf_\theta l(\theta, \lambda) = \frac{1}{4\lambda^2}c^T R_n^{-1}c - r = 0 \Rightarrow \lambda_{\text{sup}} = \sqrt{\frac{c^T R_n^{-1}c}{4r}}$$
$$\sup_{\lambda\geq 0}\inf_\theta l(\theta, \lambda) = c^T\hat{\theta} - \sqrt{c^T R_n^{-1}cr}, \operatorname{argmin}_\theta l(\theta, \lambda) = \hat{\theta} - \frac{\sqrt{r}R_n^{-1}c}{\sqrt{c^T R_n^{-1}c}}$$

One can find the maximum at t_0 similarly:

$$\sup_{\lambda\geq 0}\inf_\theta \tilde{l}(\theta, \lambda) = c^T\hat{\theta} + \sqrt{c^T R_n^{-1}cr}, \operatorname{argmin}_\theta \tilde{l}(\theta, \lambda) = \hat{\theta} + \frac{\sqrt{r}R_n^{-1}c}{\sqrt{c^T R_n^{-1}c}}$$

4. Numerical Experiments

I made experiments using Python programming language, Numpy, CVXPy and Matplotlib.Pyplot packages. The noise in the simulated observations were uncorrelated Gaussian so I could compare the outer confidence ellipsoid with the asymptotic ones given by F-distribution:

$$\{\theta : (\theta - \hat{\theta})R_n^{-1}(\theta - \hat{\theta}) \leq \frac{q\hat{\sigma_n}^2}{n}\},$$

where $\hat{\sigma_n}^2$ is the estimated variation of the noise from the sample and $F_{\chi^2}(q) = p$. In the figure below (Figure 2) we can see an experiment where SPS gives us a narrower ellipsoid than the asymptotic result.

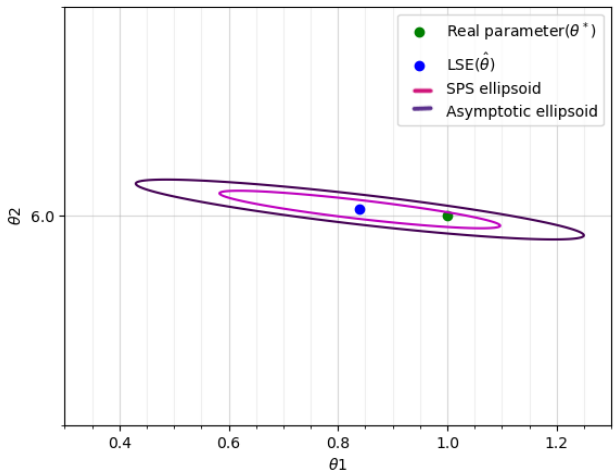

Figure 2: Example for confidence ellipsoids given by SPS and F-distribution for $n = 100$, $p = 0.9$ (for SPS $m = 10$)

I experimented with different regressors, polynomial, exponential and trigonometrial and plotted the results in the function space.

A typical result with polynomial regressors is plotted in Figure 3. We can observe that as we move away from zero the confidence region adheres to the function. The explanation can be that the larger exponents start dominate as the absolute value of x grows.

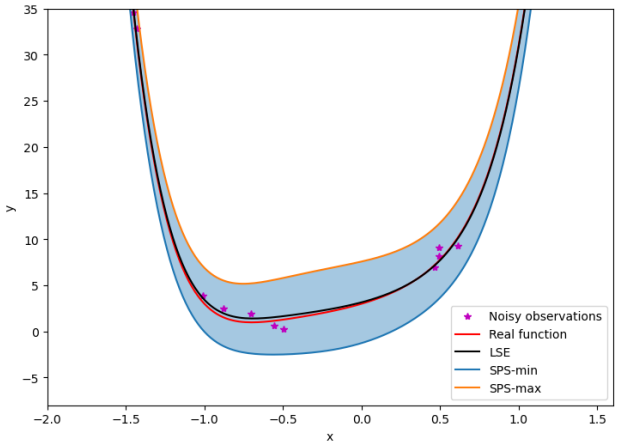

Figure 3: Example for a confidence region estimate in the function space with polynomial regressors, $n = 30$, $d = 7$, $p = 0.8$

5. Summary and future work

In this semester my main goal was to get to know the SPS method and its modifications in outer ellipsoid approximation and ridge regression. I was able to implement them in Python and examine them in work. As it is discussed in (3), SPS can be generalized to define the uncertainty for models given by kernel methods (4). In the future I would like to continue my work in this direction.

References

[1] B. C. Csáji. Non-asymptotic confidence regions for regularized linear regression estimates. In *Progress in industrial mathematics at ECMI 2018*, volume 30 of *Math. Ind.*, pages 605–611. Springer, Cham, [2019] ©2019.

[2] B. C. Csáji, M. C. Campi, and E. Weyer. Sign-perturbed sums: a new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models. *IEEE Trans. Signal Process.*, 63(1):169–181, 2015.

[3] B. C. Csáji and K. B. Kis. Distribution-free uncertainty quantification for kernel methods by gradient perturbations. *Mach. Learn.*, 108(8-9):1677–1699, 2019.

[4] B. Schölkopf and A. J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.