Non-Asymptotic Confidence Regions for Regularized Linear Regression Estimates

Balázs Csanád Csáji

Abstract Building *confidence regions* for regression models is of high importance, for example, they can be used for uncertainty quantification and are also fundamental for robust optimization. In practice, these regions are often computed from the asymptotic distributions, which however only lead to heuristic confidence sets. Sign-Perturbed Sums (SPS) is a resampling method which can construct *exact*, *non-asymptotic*, *distribution-free* confidence regions under very mild statistical assumptions. In its standard form, the SPS regions are built around the least-squares estimate of linear regression problems, and have favorable properties, such as they are star convex, strongly consistent, and have efficient ellipsoidal outer-approximations. In this paper, we extend the SPS method to *regularized* estimates, particularly, we present variants of SPS for ridge regression, LASSO and elastic net regularization.

1 Introduction

Estimating models based on noisy measurements is a fundamental problem for many scientific, engineering and economic applications. A very important issue in practice is to quantify the uncertainty of the obtained models. This is often done by building confidence regions for the models. While these regions are frequently built using the limiting distribution of the used point-estimate [6], such regions are not guaranteed for finite samples, and can only be seen as heuristics. It is of high importance to construct confidence regions with non-asymptotic guarantees, using minimal statistical assumptions. Resampling methods, such as bootstrap and Monte Carlo approaches, typically use some regularity of the noise to build such regions.

Sign-Perturbed Sums (SPS) is a recently developed resampling method with favorable properties. SPS can construct *exact*, distribution-free confidence regions for

Balázs Csanád Csáji
EPIC Centre of Excellence, MTA SZTAKI: Institute for Computer Science and Control, Hungarian Academy of Sciences, Budapest, Hungary, e-mail: balazs.csaji@sztaki.mta.hu

finite samples [2, 7]. Its standard form constructs (star convex, strongly consistent) confidence sets around the least-squares estimate of linear regression problems.

Regularization is an important tool in regression which can help, for example, to handle ill-posed and ill-conditioned problems, reduce over-fitting, enforce sparsity, and in general to control the shape and smoothness of the regression function. The paper extends SPS to various regularized linear regression problems, particularly, to ridge regression (Tikhonov regularization), LASSO and elastic net regularization.

2 Preliminaries: Asymptotic Confidence Ellipsoids

We start by recalling the standard "textbook" approach to build (asymptotic) confidence ellipsoids around the least-squares estimate of linear regression problems.

Assume we are given a data sample, $\mathscr{D}_n \doteq \{(\varphi_1, y_1), \ldots, (\varphi_n, y_n)\}$, with

$$y_t \doteq \varphi_t^{\mathrm{T}} \theta^* + \varepsilon_t, \qquad \text{for} \qquad t = 1, \ldots, n \tag{1}$$

where y_t is the *output*, φ_t is the *input* or *regressor* and ε_t is the (non-observable) *noise* for measurement t. We aim at estimating the (constant) "true" parameter, $\theta^* \in \mathbb{R}^d$. We assume that $\{\varphi_t\} \subset \mathbb{R}^d$ are *deterministic* and the noise $\{\varepsilon_t\}$ is an *independent* sequence of random variables, each having a *symmetric* distribution about zero, that is the distribution of ε_t is the same as that of $-\varepsilon_t$. Finally, for simplicity, we assume that the *regressor matrix*, $\Phi \doteq [\varphi_1, \ldots, \varphi_n]^{\mathrm{T}}$, is skinny ($n > d$) and full rank.

One of the standard estimators is the well-known *least-squares* (LS) method

$$\hat{\theta}_n \doteq \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} V(\theta \,|\, \mathscr{D}_n) = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \frac{1}{2} \| y - \Phi\theta \|_2^2, \tag{2}$$

where $y \doteq [y_1, \ldots, y_n]^{\mathrm{T}}$; $\hat{\theta}_n$, can be obtained from the *normal equation*, that is

$$\nabla_\theta V(\hat{\theta}_n \,|\, \mathscr{D}_n) = \Phi^{\mathrm{T}}\Phi\,\hat{\theta}_n - \Phi^{\mathrm{T}} y = 0, \tag{3}$$

which has a unique analyitical solution, famously given by $\hat{\theta}_n = (\Phi^{\mathrm{T}}\Phi)^{-1}(\Phi^{\mathrm{T}} y)$.

A crucial question is that how can we *quantify the uncertainty* of the so obtained estimate? This question can be answered, e.g., by constructing *confidence regions* around the point-estimate. More precisely, given a confidence probability $p \in (0,1)$, we aim at finding a region, $\widehat{\Theta}_{\mathscr{D}_n,p}$ around $\hat{\theta}_n$, such that $\mathbb{P}\big(\theta^* \in \widehat{\Theta}_{\mathscr{D}_n,p}\big) \geq p$.

The standard method to build such regions is to use the *asymptotic distribution* of the estimate [6]. It is known that the (scaled) error of LS is asymptotically Gaussian,

$$\sqrt{n}\,(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathscr{N}(0, \sigma^2 R^{-1}), \qquad \text{as} \qquad n \to \infty, \tag{4}$$

where $\mathscr{N}(\mu, \Sigma)$ is the (multivariate) Gaussian distribution with mean μ and covariance Σ. This property holds under various conditions, e.g., if the regressors are bounded, there exits a positive definite matrix R as the limit of matrices $R_n \doteq \frac{1}{n}\Phi_n^{\mathrm{T}}\Phi_n$, and $\{\varepsilon_t\}$ are i.i.d. as well as $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{E}[\varepsilon_t^2] = \sigma^2$, with $0 < \sigma^2 < \infty$.

Using the limiting distribution, a (heuristic) *confidence ellipsoid* can be built by

$$\widetilde{\Theta}_{n,p} \doteq \left\{ \theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq \frac{q\,\hat{\sigma}_n^2}{n} \right\}, \tag{5}$$

where $p = F_{\chi^2(d)}(q)$, with $F_{\chi^2(d)}$ being the CDF of the χ^2 distribution with d degrees of freedom; and $\hat{\sigma}_n^2$ is an (unbiased) estimate of the noise variance, that is

$$\hat{\sigma}_n^2 \doteq \frac{1}{n-d} \sum_{t=1}^{n} (y_t - \varphi_t^{\mathrm{T}} \hat{\theta}_n)^2. \tag{6}$$

Then, we *approximately* have $\mathbb{P}(\, \theta^* \in \widetilde{\Theta}_{n,p}\,) \approx p$ (and, obviously, $\hat{\theta}_n \in \widetilde{\Theta}_{n,p}$).

However, the confidence regions constructed using the asymptotic distribution are *not guaranteed* for finite samples, and are typically *imprecise* if the sample size is small. Another drawback of the asymptotic approach is that it presupposes the *existence* of a limiting distribution, which cannot be guaranteed in certain cases.

3 Sign-Perturbed Sums: Non-Asymptotic Confidence Regions

Now, we overview the *Sign-Perturbed Sums* (SPS) method [2, 7] that can construct *exact*, *non-asymptotic*, *distribution-free* confidence regions around the LS estimate.

As first glance, SPS can be seen as a *hypothesis testing* method. It tests the null hypothesis $\theta = \theta^*$, against the alternative hypothesis $\theta \neq \theta^*$. SPS is based on the idea that if $\theta = \theta^*$, then (1) we can compute the exact realization of the noise vector, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, by "inverting" the system, and (2) using some *regularity* of the noise (e.g., symmetry), alternative noise realizations can be generated, leading to alternative samples and estimates, which behave "similarly" (in the statistical sense) to the original ones. On the other hand, if $\theta \neq \theta^*$, then the residuals will be biased and the alternative samples and estimates based on them will behave statistically differently than the original ones. SPS applies a rank-test to decide whether the perturbed objects are similar to the original ones. Unlike other resampling based approahces, SPS avoids actually constructing the alternative samples and fitting surrogate models to them, as it directly perturbs the *gradient* of the objective function.

The principal building blocks of SPS are the following *evaluation functions*,

$$Z_i(\theta) \doteq \| \Psi^{1/2} \Phi^{\mathrm{T}} G_i \big(y - \Phi\theta\big) \|_2^2, \tag{7}$$

for $i \in \{0, 1, \ldots, m-1\}$, where $\Psi = (\Phi^{\mathrm{T}}\Phi)^{-1}$, $m > 0$ is a user-chosen integer, $G_0 \doteq I$, the identity matrix, and for $i \neq 0$, $G_i \doteq \mathrm{diag}(\alpha_{i,1}, \ldots, \alpha_{i,n})$; $\{\alpha_{i,j}\}$ are i.i.d. Rademacher variables[1]; and $\mathrm{diag}(\cdot)$ builds a diagonal matrix from its argument.

Notice that, apart from an (optional) linear transformation, $\Psi^{1/2}$, whose role is to make a covariance correction, $Z_0(\theta)$ is basically the norm of the (negative) *gradient* of the least-squares objective. The difference between $Z_0(\theta)$ and $Z_i(\theta)$, $i \neq 0$, is that in latter functions the signs of the residuals $(y - \Phi\theta)$ are perturbed in the gradient.

[1] Random variables which take values $+1$ and -1 with probability $1/2$ each.

In case $\theta = \theta^*$, the residuals are the true noises, $y - \Phi\theta^* = \varepsilon$, and we know from the symmetry assumption that for all i, ε and $G_i\varepsilon$ have the same distribution, where G_i is a diagonal matrix containing random signs as defined above. Then,

$$Z_0(\theta^*) \;=\; \| \Psi^{1/2}\Phi^{\mathrm{T}}\varepsilon \|_2^2 \;\overset{d}{=}\; \| \Psi^{1/2}\Phi^{\mathrm{T}}G_i\,\varepsilon \|_2^2 \;=\; Z_i(\theta^*), \tag{8}$$

for $i = 1, \ldots, m-1$, where "$\overset{d}{=}$" denotes equality in distribution. Nevertheless, variables $\{Z_i(\theta^*)\}$ are of course not independent. On the other hand, it can be proved [2] that they are *conditionally i.i.d.*, conditioned on the σ-algebra generated by $\{|\varepsilon_t|\}$. Consequently, they are also *exchangeable* and hence each ordering[2] of them, $Z_{i_0}(\theta^*) \prec \cdots \prec Z_{i_{m-1}}(\theta^*)$, has the same probability, namely, $1/m!$.

If however, $\theta \neq \theta^*$, then this exchangeability argument does not hold, moreover, $Z_0(\theta)$ will eventually dominate $\{Z_i(\theta)\}_{i\neq 0}$ with high probability as $\|\theta - \theta^*\| \to \infty$.

To make these ideas more precise, let us define the *normalized rank* of $Z_0(\theta)$ as

$$\mathscr{R}(\theta) \doteq \frac{1}{m}\left[1 + \sum_{i=1}^{m-1} \mathbb{I}(Z_0(\theta) \prec Z_i(\theta))\right], \tag{9}$$

where $\mathbb{I}(\cdot)$ is an indicator (its value is 1 if its argument is true and 0 otherwise).

Assume that the target confidence probability can be written as $p = 1 - q/m$ where $0 < q < m$ are user-chosen integers. Then, SPS accepts the null hypothesis, $\theta = \theta^*$, if $\mathscr{R}(\theta) \leq p$, and rejects it if $\mathscr{R}(\theta) > p$. As m and q are free-parameters, they are under our control, hence any (rational) probability can be achieved.

Based on these observations, the *SPS confidence regions* can be defined as

$$\widehat{\Theta}_{n,p} \;\doteq\; \Big\{\, \theta \in \mathbb{R}^d : \mathscr{R}(\theta) \leq p \,\Big\}. \tag{10}$$

It can be proved [2] that these regions have *exact* confidence $\mathbb{P}\big(\,\theta^* \in \widehat{\Theta}_{n,p}\,\big) = p$. Note that the exact confidence of the regions is guaranteed for *finite samples* despite no knowledge about the particular noise distributions is assumed, moreover, each noise term may have a different distribution with arbitrarily large variance.

There are several important properties of SPS confidence regions [2, 7]. For example, (1) they are *star convex* with the LS estimate as a star center; (2) they are uniformly *strongly consistent*; (3) they have asymptotically the *same size and shape* as the classical confidence ellipsoids; finally (4) they have *ellipsoidal outer approximation* that can be efficiently computed via semidefinite programming problems.

SPS has several generalizations, for example, it can be extended to general stochastic linear (dynamical) systems, even if they are operating in closed-loop [3], and to various non-linear dynamical systems, such as GARCH models [1].

Finally, we note that working with symmetric noises is not crucial for SPS as the theory can be extended to other noise distributions, as long as we know a group of transformations that leave the (joint) distribution of the noises unchanged. For example, one can assume that the noises are *exchangeable* and use random *permutation* matrices as $\{G_i\}$, see [5]. We refer to these generalized variants as (G)SPS.

[2] Relation "$\prec$" is a total order which we get from "$<$" by random tie-breaking, see [2].

4 Non-Asymptotic Confidence Sets for Regularized Estimates

In this section we are going to extend the theory of (G)SPS, in order to construct non-asymptotic, distribution-free confidence regions around regularized estimates.

First, we consider *ridge regression* (RR) which has the objective function

$$V_{\mathrm{R}}(\theta) \doteq \frac{1}{2}\|y-\Phi\theta\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2, \tag{11}$$

for a $\lambda \geq 0$ hyper-parameter. It is well-known that RR can be reformulated as LS,

$$\widetilde{\Phi} = \begin{bmatrix} \Phi \\ \sqrt{\lambda}\, I \end{bmatrix}, \qquad \text{and} \qquad \widetilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}, \tag{12}$$

where I is the identity matrix, after which we have $V_{\mathrm{R}}(\theta) = 1/2\,\|\widetilde{y}-\widetilde{\Phi}\theta\|^2$.

Then, one might be tempted to apply standard SPS to the obtained (ordinary) LS formulation. However, we should proceed with caution, as the new problem has some auxiliary output terms, the zero part of $\widetilde{y}$, to which there are no real noise terms in the original problem. Therefore, the last d terms of the residual vector, $\widetilde{y}-\widetilde{\Phi}\theta$, should not be perturbed, as the distributional invariance was only assumed for the original noise vector. Consequently, the $\{G_i\}$ matrices should be extended by

$$\widetilde{G}_i \doteq \begin{bmatrix} G_i & 0 \\ 0 & I \end{bmatrix}, \tag{13}$$

for $i = 1,\ldots,m-1$. Then, using an analogue of (7) to the new LS system with $\{\widetilde{G}_i\}$ perturbations, we arrive at the *(G)SPS evaluation function for ridge regression*,

$$Z_i(\theta) \doteq \left\| \Psi_{\mathrm{R}}^{1/2} \left[\Phi^{\mathrm{T}} G_i (y-\Phi\theta) - \lambda\theta \right] \right\|_2^2, \tag{14}$$

where $\Psi_{\mathrm{R}} = (\Phi^{\mathrm{T}}\Phi+\lambda I)^{-1}(\Phi^{\mathrm{T}}\Phi)(\Phi^{\mathrm{T}}\Phi+\lambda I)^{-1}$ is a correction term from the covariance of RR. Based on this evaluation function, *exact* confidence regions can be built around the RR estimate, using the same steps as we had for standard SPS.

Now, let us consider *LASSO* (least absolute shrinkage and selection operator) which applies L1 regularization to enforce *sparsity*. It has the objective function

$$V_{\mathrm{L}}(\theta) \doteq \frac{1}{2}\|y-\Phi\theta\|_2^2 + \lambda\|\theta\|_1, \tag{15}$$

for $\lambda \geq 0$. This objective is no more quadratic and it cannot be traced back to LS. However, the underlying idea of SPS, i.e., to perturb the residuals in the (negative) gradient of the objective, can still be applied. A (sub-) gradient[3] of (15) is

$$\nabla_\theta V_{\mathrm{L}}(\theta) = \Phi^{\mathrm{T}}\Phi\,\theta - \Phi^{\mathrm{T}} y + \lambda\,\mathrm{sign}(\theta), \tag{16}$$

where the sign function is understood component-wise.

[3] For our purposes, one of the subgradients is sufficient, thus we do not treat $\nabla_\theta V$ set-valued.

Then, we can proceed in the same way as before and perturb the residuals in (16) with $\{G_i\}$, leading to the *(G)SPS evaluation function for LASSO*,

$$Z_i(\theta) \doteq \left\| \Psi_{\text{L}}^{1/2} \left[\Phi^{\text{T}} G_i (y - \Phi\theta) - \lambda \operatorname{sign}(\theta) \right] \right\|_2^2, \tag{17}$$

where Ψ_{L} is an (optional) correction term, e.g., using the (asymptotic) results of [4], we may use $\Psi_{\text{L}} = (\Phi^{\text{T}}\Phi)^{-1}$. The correction matrix can be interpreted as the square-root of the (estimated) covariance of LASSO (modulo the variance of the noise, as multiplying each Z_i with the same positive scalar does not affect their ordering).

The last method that we discuss is the *elastic net* regularization with objective

$$V_{\text{E}}(\theta) \doteq \frac{1}{2} \| y - \Phi\theta \|_2^2 + \lambda_1 \| \theta \|_1 + \frac{\lambda_2}{2} \| \theta \|_2^2, \tag{18}$$

were $\lambda_1, \lambda_2 \geq 0$ are hyper-parameters. As the objective is the combination of the ridge regression and LASSO objectives, it can be handled using similar ideas. That is we can compute a subgradient of the objective and perturb the residuals based on the transformations $\{G_i\}$ which leave the (joint) distribution of the true noise terms invariant. Then, the *(G)SPS evaluation function for elastic net regularization* is

$$Z_i(\theta) \doteq \left\| \Psi_{\text{E}}^{1/2} \left[\Phi^{\text{T}} G_i (y - \Phi\theta) - \lambda_1 \operatorname{sign}(\theta) - \lambda_2 \theta \right] \right\|_2^2, \tag{19}$$

where Ψ_{E} can again be an (optional) covariance estimate for the elastic net solution.

The exact confidence of the constructed regions easily follows from the related results for SPS. We leave the investigation of their other properties for further work.

Acknowledgements This research was partially supported by the National Research, Development and Innovation Office (NKFIH), grant numbers ED_18-2-2018-0006 and KH_17 125698, and by the János Bolyai Research Fellowship of the Hung. Academy of Sciences, BO/00217/16/6.

References

1. B. Cs. Csáji. Score permutation based finite sample inference for generalized autoregressive conditional heteroskedasticity (GARCH) models. In *19th International Conference on Artificial Intelligence and Statistics (AISTATS), Cadiz, Spain*, pages 296–304, 2016.
2. B. Cs. 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 Transactions on Signal Processing*, 63(1):169–181, 2015.
3. B. Cs. Csáji and E. Weyer. Closed-loop applicability of the Sign-Perturbed Sums method. In *54th IEEE Conference on Decision and Control, Osaka, Japan*, pages 1441–1446, 2015.
4. K. Knight and W. Fu. Asymptotics for LASSO-type estimators. *Annals of Statistics*, (5):1356–1378, 2000.
5. S. Kolumbán. *System Identification in Highly Non-Informative Environment*. PhD thesis, Budapest Uni. of Techn. and Econ., Hungary, and Vrije Universiteit Brussels, Belgium, 2016.
6. L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, 2nd edition, 1999.
7. E. Weyer, M. C. Campi, and B. Cs. Csáji. Asymptotic properties of SPS confidence regions. *Automatica*, 82:287–294, 2017.