Least squares under convex constraint

Sourav Chatterjee

Stanford University

Questions

- Let Z be an n-dimensional standard Gaussian random vector.
- Let μ be a point in $\mathbb{R}^n$ and let $Y = Z + \mu$.
- We are interested in estimating μ from the data vector Y, under the assumption that μ belongs to a given closed convex set K.
- Many problems in modern statistics are special cases of this general setup, including the Lasso and other high-dimensional regression techniques, function estimation problems, matrix estimation and completion, shape-restricted regression, etc.
- The least squares estimator (LSE) of μ, which is also the MLE, is $\hat{\mu} := P_K(Y)$, where P_K denotes projection on to K.
- Question 1: What can we say about the magnitude and behavior of the error $\|\hat{\mu} - \mu\|$, where $\|\cdot\|$ denotes Euclidean norm?
- Question 2: Does the LSE have any general optimality property that holds for any K?

Brief survey of literature

- Standard approach is to get an upper bound on $\mathbb{E}\|\hat{\mu} - \mu\|^2$ using empirical process theory.
- Statistical theory developed by Birgé, Tsirelson, Pollard, van de Geer, Massart, van der Vaart, Wellner, Koltchinskii, Wegkamp, Tsybakov and many others over a period of more than thirty years.
- Recently gained prominence in the statistical signal processing literature. LSE problem is equivalent to the problem of constrained denoising in signal processing and is connected with linear inverse problems.
- In the signal processing domain, important contributions from Rudelson, Vershynin, Stojnic, Oymak, Hassibi, Chandrasekaran, Tropp, and others.
- The expected squared error $\mathbb{E}\|\hat{\mu} - \mu\|^2$ is closely related to the concept of "statistical dimension" of convex sets introduced by Tropp and coauthors in 2013.

Classical approach

- All existing approaches to this problem are refinements and generalizations of the following idea.
- Since $\hat{\mu}$ is the projection of $Y = Z + \mu$ on to the convex set K and $\mu \in K$, a simple geometric argument shows that the angle between the vectors $Y - \hat{\mu}$ and $\mu - \hat{\mu}$ must be ≥ 90 degrees.
- In other words, $(Y - \hat{\mu}) \cdot (\mu - \hat{\mu}) \leq 0$.
- This may be rewritten as $\|\hat{\mu} - \mu\|^2 \leq Z \cdot (\hat{\mu} - \mu)$.
- Consequently,

$$\|\hat{\mu} - \mu\|^2 \leq \sup_{\nu \in K} Z \cdot (\nu - \mu)\,.$$

- The expected value of the right-hand side is bounded using upper bounds on expected maxima of Gaussian processes.
- Clearly, this approach can only give upper bounds.

Main results

In this talk, I will ...

- give a formula for the exact behavior of $\|\hat{\mu} - \mu\|$, instead of an upper bound;
- give an application of this formula to Lasso;
- show that $\hat{\mu}$ may not be minimax optimal, not even minimax rate-optimal, but ...
- $\hat{\mu}$ is always admissible up to a universal constant factor;
- as a corollary, prove an optimality property for Lasso.

In particular, this solves the two questions posed in the first slide.

Estimation error

Recall: $K \subseteq \mathbb{R}^n$ is a closed convex set, P_K is projection on to K, Z is a standard Gaussian vector, and $\hat{\mu} = P_K(Z + \mu)$.

Theorem (C., 2014)

Given any $\mu \in \mathbb{R}^n$ and $t \geq 0$, define

$$f_\mu(t) := \mathbb{E}\left(\sup_{\nu \in K\,:\,\|\nu-\mu\| \leq t} Z \cdot (\nu - \mu)\right) - \frac{t^2}{2}\,.$$

Then f_μ is a strictly concave function and there is a unique point $t_\mu \in [0, \infty)$ where f_μ is maximized. For any $x \geq 0$,

$$\mathbb{P}\left(\left|\|\hat{\mu} - \mu\| - t_\mu\right| \geq x\sqrt{t_\mu}\right) \leq 3\exp\left(-\frac{x^4}{32(1 + \frac{x}{\sqrt{t_\mu}})^2}\right).$$

Implication: $\|\hat{\mu} - \mu\| = t_\mu + O(\max\{\sqrt{t_\mu}, 1\})$ with high probability.

Steps in the proof

- Define a random function

$$F_\mu(t) = \sup_{\nu \in K \,:\, \|\nu - \mu\| \le t} Z \cdot (\nu - \mu) - \frac{t^2}{2},$$

  so that $f_\mu(t) = \mathbb{E}(F_\mu(t))$.
- Using the convexity of K, prove that F_μ and f_μ are both strictly concave functions. Let t^* be the unique point at which F_μ is maximized.
- Again, use convexity of K to prove the identity $\|\hat{\mu} - \mu\| = t^*$.
- Using the concentration of Gaussian maxima, show that $F_\mu(t)$, although random, is close to $f_\mu(t)$ with high probability.
- Since F_μ and f_μ are two strictly concave functions that are close to each other (with high probability), their points of maxima must also be close. That is, $t_\mu \approx t^* = \|\hat{\mu} - \mu\|$ with high probability.

Remarks

- To compute t_μ, one needs to compute expected maxima of certain Gaussian empirical processes. This is similar to the requirements of existing theory.
- However, unlike existing theory, this theorem gives an asymptotically exact formula instead of an upper bound.
- The theorem also shows that $\|\hat{\mu} - \mu\|$ is concentrated in a small window around its expected value, whenever that expected value is large. This is a universal law of large numbers for the global error in this class of problems.
- One may go as far as to claim that this theorem completes the quest for a precise connection between empirical process theory and estimation errors of least squares estimators under convex constraints (in the Gaussian setting).

But ... is this any good for actual computations?

- Yes.
- Recall that $\|\hat{\mu} - \mu\| \approx t_\mu$ with high probability, where t_μ is the point at which the function $f_\mu(t)$ is maximized.
- Recall also that the function f_μ is strictly concave.
- Therefore, to get upper and lower bounds for t_μ, it suffices to get upper and lower bounds for $f_\mu(t)$, which, in turn, involves getting upper and lower bounds for expected maxima of Gaussian processes. There are a lot of tools for doing that.
- The following proposition is useful in applications:

Proposition (C., 2014)

If $r_1 < r_2 < r_3$ are such that $f_\mu(r_1) \leq f_\mu(r_2)$ and $f_\mu(r_2) \geq f_\mu(r_3)$, then $r_1 \leq t_\mu \leq r_3$.

How to apply this: An illustration

- Isotonic regression: Here K is the set of all $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ such that $\mu_1 \leq \cdots \leq \mu_n$.
- Fix μ. Let C and c denote any constants that depend on μ only through the quantities $\max_i n(\mu_{i+1} - \mu_i)$ and $\min_i n(\mu_{i+1} - \mu_i)$.
- Using standard results for the entropy of monotone functions and maxima of Gaussian fields, one can show that

$$c\sqrt{t}n^{1/4} \leq \mathbb{E}\left(\sup_{\nu \in K\,:\,\|\nu-\mu\| \leq t} Z \cdot (\nu - \mu)\right) \leq C\sqrt{t}n^{1/4}\,.$$

- Consequently, $c\sqrt{t}n^{1/4} - t^2/2 \leq f_\mu(t) \leq C\sqrt{t}n^{1/4} - t^2/2$.
- Choosing $r_1 = c_1 n^{1/6}$, $r_2 = c_2 n^{1/6}$ and $r_3 = c_3 n^{1/6}$ with appropriately chosen constants c_1, c_2 and c_3, one can use the above inequalities to arrange that $r_1 < r_2 < r_3$ and $f_\mu(r_1) \leq f_\mu(r_2)$ and $f_\mu(r_2) \geq f_\mu(r_3)$. Therefore by the main theorem, $cn^{1/6} \leq \|\hat{\mu} - \mu\| \leq Cn^{1/6}$ with high probability.

Lasso

- Let X be an $n \times p$ matrix. Let

$$Y = X\beta + Z\,,$$

  where $\beta \in \mathbb{R}^p$ and Z is a standard Gaussian vector.
- Tibshirani's Lasso estimator of β is the vector $\hat{\beta}$ that minimizes $\|Y - X\hat{\beta}\|$ subject to $|\hat{\beta}|_1 \le L$, where $|\hat{\beta}|_1$ is the ℓ^1 norm of $\hat{\beta}$ and L is a number chosen by the statistician.
- Extensive theoretical investigations by Donoho, Johnstone, Zou, Wainwright, Candès, Tao, Meinshausen, Bühlmann, Yu, Koltchinskii, van de Geer, Greenshtein, Ritov, Bickel, Bartlett, Rigollet, Tsybakov, Tibshirani, Taylor, and many others.
- Research focused on two things: (a) What is the magnitude of the prediction error $\|X\beta - X\hat{\beta}\|$, and (b) whether the procedure is good at choosing the correct nonzero coordinates of β (model consistency).
- The theory presented here is good for only part (a).

Generally...

- If β_0 is the true value of the parameter, let

$$f(t) := \mathbb{E}\left(\sup_{\substack{\beta : |\beta|_1 \le L \\ \|X(\beta-\beta_0)\| \le t}} Z \cdot (X\beta - X\beta_0) \right) - \frac{t^2}{2}.$$

- Let t^* be the point at which $f(t)$ maximized.
- Then our main theorem implies that the prediction error $\|X\hat{\beta} - X\beta_0\|$ is equal to $t^* + O(\sqrt{t^*})$ with high probability.
- In the next slide, I will show a special situation in which an explicit quantification of the error may be deduced from the general result given above.

Error in Lasso

Theorem (C., 2014)

Let $\Sigma := X^T X/n$, and let a and b be the smallest and largest eigenvalues of Σ. Assume that $a > 0$, and that all the diagonal entries of Σ are equal to 1. Take any $\beta \in \mathbb{R}^p$. Let s be the number of nonzero entries of β and $r = p/n$. Imagine a sequence of problems where $|\beta|_1$, L, a, b, s and r remain fixed and $n \to \infty$. With probability tending to one, the following happens.

- *If $L > |\beta|_1$, then $\|X\beta - X\hat{\beta}\|^2 = n^{1/2+o(1)}$.*
- *If $L = |\beta|_1$, then $\|X\beta - X\hat{\beta}\|^2 \le C \log n$.*
- *If $L < |\beta|_1$, then $\|X\beta - X\hat{\beta}\|^2 \asymp n$.*

Remarks

- This theorem emphasizes the need for choosing the "correct" value of the penalty parameter in Lasso.
- The theorem, however, assumes that L is a fixed number, independent of the data.
- In practice, the penalty parameter is usually chosen in a data-dependent manner, often using cross-validation.
- It will require more work to show that cross-validated Lasso works far better than Lasso without cross-validation, in terms of prediction error (I think that can be proved by extending the techniques presented here). Incidentally, there's very little work on cross validation in Lasso. See recent papers of Homrighausen and Macdonald for some progress.
- I think this is the first theorem that gives a *lower bound* on the error of the Lasso.
- In principle, the theory should work for $p \gg n$ also. Working out the details is challenging but seems doable.

Back to the original problem

- Recall: $K \subseteq \mathbb{R}^n$ is a closed convex set, P_K is projection on to K, Z is a standard Gaussian vector, and $\hat{\mu} = P_K(Z + \mu)$.
- Question: Is $\hat{\mu}$ always minimax rate-optimal?
- More precise question: Does there exist a positive universal constant C such that for any n, any choice of K, and any other estimator $\tilde{\mu}$, the maximum risk of $\tilde{\mu}$ is at least as large as C times the maximum risk of $\hat{\mu}$?
- Here "risk" means expected squared error in Euclidean norm, as always.
- Answer: No. Given any positive ϵ, one can construct a scenario where there exists an estimator $\tilde{\mu}$ whose maximum risk is less than ϵ times the maximum risk of $\hat{\mu}$. Example in the next slide.

Counterexample to minimaxity

- Take any n. Define a closed convex set $K \subseteq \mathbb{R}^n$ as follows.
- Take any $\alpha \in [0,1]$, $\theta_1, \ldots, \theta_n \in [-1,1]$, and let

$$\mu_i := \alpha n^{-1/4} + \alpha \theta_i n^{-1/2}, \quad i = 1, \ldots, n.$$

  Let K be the set of all $\mu = (\mu_1, \ldots, \mu_n)$ obtained as above.
- This is a regression problem with unknown parameters α and $\theta_1, \ldots, \theta_n$.
- As usual, let $\hat{\mu} = P_K(Z + \mu)$.
- Let $\tilde{\mu}$ be the estimate whose coordinates are all equal to the average of the coordinates of $Z + \mu$.

Proposition (C., 2014)

The maximum risk of $\hat{\mu}$ is bounded below by $C_1 n^{1/2}$ whereas the maximum risk of $\tilde{\mu}$ is bounded above by C_2, where C_1 and C_2 are positive constants that do not depend on n. (Here "risk" means expected total squared error.)

(Proof involves an application of the main theorem.)

So, what kind of optimality can we hope for?

- The preceding counterexample shows that in some cases the LSE under convex constraint can have significantly bad worst-case performance compared to another estimator.
- But it's hard to believe that the LSE under convex constraint will have no general optimality property at all.
- The next slide contains a theorem which shows that the LSE under convex constraint is always guaranteed to be admissible up to a universal constant factor.
- That is, given any convex constraint and any other estimator, there always exists *some* region in the parameter space where its performance is comparable to that of the LSE.

Admissibility up to universal constant

Theorem (C., 2014)

There is a positive universal constant C such that for any n, any nonempty closed convex set $K \subseteq \mathbb{R}^n$, and any estimator $\tilde{\mu}$, there exists $\mu \in K$ such that the risk of $\tilde{\mu}$ is at least as large as C times the risk of $\hat{\mu}$ when μ is the true mean.

Remarks

- Recall that by Stein's paradox, $\hat{\mu}$ may not be admissible. Therefore in general, admissibility up to a universal constant factor is the best that one can hope for.
- It would be interesting to figure out the optimal value of C. The value given by the current proof is too embarrassingly small to report.
- Proof sketch, very briefly: Main idea is to choose an appropriate $\mu \in K$, and use the distribution of $\hat{\mu}$ for this μ as a prior distribution in a Bayesian problem. Choosing this μ is a tricky problem.

Optimality of Lasso

- Linear regression model: $Y = X\beta + \varepsilon$, where ε is normal with i.i.d. components.
- Lasso estimator with penalty parameter L is $\hat{\beta}$ that minimizes $\|Y - X\beta\|$ among all β with $|\beta|_1 \le L$.
- Let $\tilde{\beta}$ be any other estimator of β.
- In this setup, we have the following corollary of the admissibility theorem:

Corollary (C., 2014)

There exists β^* with $|\beta^*|_1 \le L$, such that if β^* is the true value of β, then

$$\mathbb{E}\|X\tilde{\beta} - X\beta^*\|^2 \ge C\,\mathbb{E}\|X\hat{\beta} - X\beta^*\|^2\,,$$

where C is a universal constant.

Note: No hidden assumptions.

Reference

- "A new perspective on least squares under convex constraint", recently uploaded on arXiv.