The statistical map, high-dimensional models and concentration of measure

Sara van de Geer
Seminar for Statistics, ETH Zürich

Statistics is about trying to make sense out of data. Today we are stumbling over data. New ways for storage, visualisation, transfer are needed, and new measures for information privacy. But having a large amount of data is something else than knowing a lot. All these data need to be analysed. There are clever algorithms that do this, machine learning algorithms, deep learning, semi-supervised feature extraction .... The task of a (theoretical) statistician is to study the mathematical properties of these algorithms and access their accuracy.

Let me describe this task in more detail. Perhaps the best way to do this for a mathematical audience is by formulating statistics as the study of a *mapping* Q from a space of probability measures $\mathcal{P}$ to a target space $\mathcal{B}$ at a random point. Let $\mathcal{X}$ be the observation space, and let the data be $X_1, \ldots, X_n$, with $X_i \in \mathcal{X}$, $i = 1, \ldots, n$. Aim is to learn something from these data. For example, the X_i are paintings and the aim is to learn how to distinguish a Mondriaan painting from a Picasso. The data are modelled as being random. For simplicity, let us assume they are independent and all have the same distribution P. Let $Q(P)$ be the target, that what we want to learn. The map Q is known, the distribution P is unknown. Let $\hat{P}_n := \sum_{i=1}^n \delta_{X_i}/n$ be the observed distribution of the data. Note that $\hat{P}_n$ is a random element of the class of distributions on $\mathcal{X}$. The idea is now to estimate $Q(P)$ by the observed counterpart $Q(\hat{P}_n)$. The mapping Q is our algorithm and the task of the statistician is to study its "modulus of continuity". Of course, there are some issues here, for example, $Q(\hat{P}_n)$ may not be defined or produce nonsense.

A special case of mapping Q is the one defined by risk minimisation. Let $\beta^0 \in \mathcal{B}$ be the target and suppose it can be defined as the minimiser of some risk function $R : \mathcal{B} \to \mathbb{R}$. The risk function R is unknown. It is estimated from data by a empirical risk function $\hat{R}_n$. Hence, the empirical risk function is observed and is a random function. The target is estimated by the minimiser $\hat{\beta}_n$ of the empirical risk. So $\hat{\beta}_n$ is observed and random. Question: how close is the observed $\hat{\beta}_n$ to the unknown β^0?

With today's high-dimensional data, straightforward minimisation of the empirical risk may give nonsense. Suppose $\mathcal{B} \subset \mathbb{R}^p$, say. By high-dimensional we then mean that the number of parameters p is much larger than the number of observations n. In a sense, one has more unknowns than equations. To deal with this, one needs to "regularise" the empirical risk: parameters β with small empirical risk $\hat{R}_n(\beta)$ are made less attractive by adding a cost for certain properties of this value. Thus, instead of minimising $\hat{R}_n$ one minimises $\hat{R}_n + \text{pen}(\cdot)$ where $\text{pen}(\cdot)$ is a given penalty function. Instead of estimating $\beta^0 = Q(P)$ by $Q(\hat{P}_n)$ one changes Q to a more regular map $\tilde{Q}$ and uses the estimator $\hat{\beta}_n = \tilde{Q}(\hat{P}_n)$.

In my talk I will give several examples. The penalty in these examples is such that it reflects the idea that although the data may be complex, one needs essentially only a few parameters to describe them. This is termed *sparsity*. What we need to do in these examples is study the regularised map $\tilde{Q}$ at the random point $\hat{P}_n$. Here is a simple prototype example. Suppose (for $i = 1, \ldots, n$) that X_i is a random row-vector in $\mathbb{R}^p$. Each entry $X_{i,j}$ is a measurement (measurement i) of a variable j, $j = 1, \ldots, p$. Let Σ_0 be the $p \times p$ matrix of variances and covariances of these p variables. Suppose for simplicity that the mean of X_i is zero. Let X be the $n \times p$ matrix with rows X_i, $i = 1, \ldots, n$, and let $\hat{\Sigma}_n := X^T X/n$ be the $p \times p$ empirical covariance matrix. Then the random matrix $\hat{\Sigma}_n$ is an estimator of the unknown matrix Σ_0. But when $p \gg n$ it is not a good estimator at all. We need to regularise depending on what the target is. Say our target is the first principal component q^0 of Σ_0 together with the corresponding eigenvalue $\phi^2_{\max}$. The target is thus $\beta^0 := \phi_{\max} q^0 \in \mathbb{R}^p$. Suppose there are only a few variables with a relevant contribution to the first principal component. Then a reasonable penalty is the ℓ_1-penalty $\text{pen}(\beta) := \lambda \|\beta\|_1$. The constant $\lambda > 0$ is a tuning parameter: large values enforce more sparsity. As empirical risk function, one may take $\hat{R}_n(\beta) = -2\beta^T \hat{\Sigma}_n \beta + \|\beta\|_2^4$. The "sparse PCA" estimator then becomes

$$\hat{\beta}_n := \arg \min_{\beta \in \mathcal{B} \subset \mathbb{R}^p} \Big\{ \underbrace{-2\beta^T \hat{\Sigma}_n \beta + \|\beta\|_2^4}_{\hat{R}_n(\beta)} + \underbrace{\lambda \|\beta\|_1}_{\text{pen}(\beta)} \Big\}.$$

The theoretical version of the risk is $R(\beta) = -2\beta^T \Sigma_0 \beta + \|\beta\|_2^4$. It is indeed minimised at $\beta^0 = \phi_{\max} q^0$.

Now the work starts. The task is to show that $\hat{\beta}_n$ is close to β^0. In fact, we aim at showing that if β^0 has many zero entries, the estimator behaves as if it *knew* how many zero there are and where they are. If β^0 is not sparse in the sense of having many zeroes, we want to show that the estimator is very close to the best sparse approximation of β^0. We want to find lower bounds for the error of any estimator of β^0. Actually, the ℓ_2-error depends on β^0 which is unknown. Task is also to show negative results: a proof that it is not possible to estimate the error in ℓ_2.

A tool from probability theory which is very much used for studying high-dimensional problems is concentration of measure. I will give a few examples where one sees that the error $\|\hat{\beta} - \beta^0\|_2$ concentrates on a single point s_0 (that is, its deviation from s_0 is of much smaller order than s_0 itself). Further tools are from convex analysis, approximation theory, geometry, etc. And of course, statistics itself also develops new mathematical tools which (might) have their merits in other branches of mathematics.