RMM Vol. 2, 2011, 201–209
Special Topic: Statistical Science and Philosophy of Science
Edited by Deborah G. Mayo, Aris Spanos and Kent W. Staley
http://www.rmm-journal.de/

Larry Wasserman

Low Assumptions, High Dimensions

Abstract:

These days, statisticians often deal with complex, high dimensional datasets. Researchers in statistics and machine learning have responded by creating many new methods for analyzing high dimensional data. However, many of these new methods depend on strong assumptions. The challenge of bringing low assumption inference to high dimensional settings requires new ways to think about the foundations of statistics. Traditional foundational concerns, such as the Bayesian versus frequentist debate, have become less important.

1. In the Olden Days

There is a joke about media bias from the comedian Al Franken:

> "To make the argument that the media has a left- or right-wing, or a liberal or a conservative bias, is like asking if the problem with Al-Qaeda is: do they use too much oil in their hummus?"

I think a similar comment could be applied to the usual debates in the foundations of statistical inference. The important foundation questions are not 'Bayes versus Frequentist' or 'Objective Bayesian versus Subjective Bayesian'.

To me, the most pressing foundational question is: how do we reconcile the two most powerful needs in modern statistics: the need to make methods assumption free and the need to make methods work in high dimensions. Methods that hinge on weak assumptions are always valuable. But this is especially so in high dimensional problems. The Bayes-Frequentist debate is not irrelevant but it is not as central as it once was. I'll discuss Bayesian inference in *section 4*.

Our search for low assumption, high dimension methods is complicated by the fact that our intuition in high dimensions is often misguided. In the olden days, statistical models had low dimension d and large sample size n. These models guided our intuition but this intuition is inadequate for modern data where $d > n$.

An analogy from physics is helpful. Physics was initially guided by simple thought (and real) experiments about falling apples, balls rolling down inclined planes and moving objects bumping into each other. This approach guided physics successfully for a while. But modern physics (quantum mechanics, fields

theories, string theory etc.) shows that the intuition built from simple physical scenarios is misleading. Research fields get more complex as they mature and we have to give up our simple models. Similarly, the foundations of statistics was once guided by simple thought experiments where the models were: correct, had a few, interpretable parameters and the biggest issue of the day was Bayes versus frequentist. But this is no longer true. Modern problems involve many of the following characteristics:

1. The number of parameters is larger then the number of data points.
2. Data can be numbers, images, text, video, manifolds, geometric objects, etc.
3. The model is always wrong. We use models, and they lead to useful insights but the parameters in the model are not meaningful.

2. Low Assumptions Inference

Before discussing high-dimensional statistics, let us begin with a discussion of low assumptions inference.

I think most statisticians would agree that methods based on weak assumptions are generally preferable to methods based on strong assumptions. For simple, low dimensional problems, low assumption inference is well-studied. But the extensions to high-dimensions are far from obvious. I'll now describe three approaches to low assumption inference which make increasingly weaker assumptions.

2.1 Completely Nonparametric Inference

Let $X_1,\ldots,X_n$ be random variables on $\mathbb{R}$. We take our model to be the set $\mathscr{P}$ of all distributions on the real line. Let $\theta = T(P)$ be a function of P. A confidence set $C = C(X_1,\ldots,X_n)$ is completely nonparametric if $P^n(T(P) \in C) \geq 1-\alpha$ for all $P \in \mathscr{P}$. Bahadur and Savage (1956) showed that there is no non-trivial completely nonparametric confidence set when $T(P)$ is the mean of P. Donoho (1988) and Tibshirani and Wasserman (1988) extended the Bahadur and Savage result to other functionals T.

On a more positive note, Donoho (1988) showed that completely nonparametric one-sided inference is sometimes possible. For example, let $M(P)$ be the number of modes of the density of P. (If P has no density, then define $M(P)$ to be the limit of $M(P \star K_h)$ as $h \to 0$ where K_h is a kernel with bandwidth h.)

Let $\mathscr{P}$ be a $1-\alpha$ confidence set for P. (For example, invert the Kolomogorov-Smirnov test.) Then $C = [\min\{m(P) : P \in \mathscr{P}\}, \infty)$ is a completely nonparametric one-sided confidence interval for $M(P)$. That is, for all P, $P^n(M(P) \notin C)$. Donoho gives examples of one-sided intervals for even more complex functionals.

However, things break down as soon as we increase the dimension. Even when $X_i \in \mathbb{R}^2$, it can be shown that there is not even a one-sided nonparametric confidence interval for $M(P)$. A basic question which has not been answered (as

far as I know) is: for which quantities do there exist non-trivial nonparametric confidence intervals when the dimension $d > 2$? A more interesting question is: for which quantities do there exist non-trivial nonparametric confidence intervals when when the dimension $d > n$?

2.2 Inference without Models

P. Laurie Davies (and his co-workers) have written several interesting papers where probability models, at least in the sense that we usually use them, are eliminated. Data are treated as deterministic. One then looks for *adequate models* rather than *true models*. His basic idea is that a distribution P is an adequate approximation for $x_1, \ldots, x_n$, if typical data sets of size n, generated under P 'look like' $x_1, \ldots, x_n$. In other words, he asks whether we can approximate the deterministic data with a stochastic model.

For nonparametric regression, the idea is implemented as follows (Davies and Kovac 2001; Davies, Kovac and Meise 2009). We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$. Given an regression function f, we can define the residuals

$$\epsilon(f) = (\epsilon_1, \ldots, \epsilon_n)$$

where $\epsilon_i = Y_i - f(X_i)$. Now we apply a test for randomness to $\epsilon(f)$. The particular test is not important in our discussion. Write $R(\epsilon(f)) = 1$ if randomness is rejected and $R(\epsilon(f)) = 0$ if randomness is not rejected.

Next define a measure of complexity $\psi(f)$. For example, $\psi(f)$ might be the number of maxima and minima of f. Finally, we define $\hat{f}$ to minimize $\psi(f)$ subject to $R(\epsilon(f)) = 0$. Thus, $\hat{f}$ is the simplest function such that the residuals 'look random'.

Davies' approach is intriguing but again, I have not seen extensions to high-dimensional problems.

2.3 Individual Sequences

A more extreme example of using weak assumptions is to abandon probability completely. This is the idea behind *individual sequence prediction*. This is a large and well-researched area, yet most statisticians (and I suspect most philosophers) have never heard of it.

We might say that, to a Bayesian, everything is random. To a frequentist, some things are random. To researchers in individual sequence prediction, nothing is random. Abandoning the idea that data are generated from some random process is very appealing. After all, when we have data we really just have a bunch of numbers. The idea that they were generated from a distribution is usually just a fiction.

Here is a brief description of individual sequence prediction, taken from Cesa-Bianchi and Lugosi (2006). We observe a sequence $y_1, y_2, \ldots,$. For simplicity assume these are binary. After observing $y_1, \ldots, y_{t-1}$ we issue a prediction

p_t of y_t. We suffer a loss $\ell(y_t, p_t)$. Again, to keep it simple, let us suppose that $\ell(y_t, p_t) = |y_t - p_t|$.[1]

We construct N algorithms $A_1, \ldots, A_N$ (called experts in this research area). The algorithm A_j takes the data $y_1, \ldots, y_{t-1}$ and yields a prediction $A_{j,t}$. To summarize: at time t:

1. You see y^{t-1} and $(A_{1,t}, \ldots, A_{N,t})$.
2. You predict p_t.
3. y_t is revealed.
4. You suffer loss $\ell(p_t, y_t)$.

Define the cumulative loss of algorithm A_j by $L_j(y^n) = \frac{1}{n}\sum_{i=1}^n |A_{j,t} - y_t|$. Define the *maximum regret*

$$R_n = \max_{y^t \in \{0,1\}^t} \left(L_p(y^n) - \min_j L_j(y^n) \right)$$

and the *minimax regret*

$$V_n = \inf_P \max_{y^t \in \{0,1\}^t} \left(L_P(y^n) - \min_j L_j(y^n) \right).$$

A number of authors (Vovk, Littelstone and Warmuth, Cesa-Bianchi and Lugosi) have shown the following. If we define

$$P_t(y^{t-1}) = \sum_{j=1}^N w_{j,t-1}\, F_{j,t}$$

where $w_{j,t-1} \propto \exp\{-\gamma n L_{j,t-1}\}$ and $\gamma = \sqrt{8 \log N / n}$ then

$$L_P(y^n) - \min_{1 \le j \le N} L_j(y^n) \le \sqrt{\frac{\log N}{2n}}.$$

Moreover, this bound is tight. Note that there is no assumption about randomness (subjective or frequentist).

In summary, we can do sensible inference without invoking probability at all. Why are scholars in foundations ignoring this?

3. Low Assumptions in High Dimensions

Can we apply some of this thinking to high dimensional problems? Let us consider a few specific cases.

[1] See the book by Cesa-Bianchi and Lugosi (2006) and various papers by: Cesa-Bianchi, Lugosi, Rakhlin, Bartlett, Freund, Feder, Merhav, Gutman, Vovk, Shafer, Littlestone, Warmuth, Schapire and others.

3.1 Prediction

Consider linear prediction. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^d$. We want to predict the next Y given a new X.

A linear predictor has the form $L(x;\beta) = \beta^T x$. Traditional, low dimensional thinking involves (i) estimating β by least squares, (ii) treating the linear model as if it is correct and (iii) treating the β's as meaningful quantities. None of these ideas make sense when $d \gg n$. Instead we (i) estimate β using sparse regression, (ii) assume that the linear model is incorrect and (iii) we make no attempt to interpret the β's.

For example, we can estimate β using the lasso (Tibshirani 1996) where $\hat{\beta}$ is chosen to minimize

$$\sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1$$

where $||\beta||_1 = \sum_{j=1}^d |\beta_j|$. This is equivalent to minimizing $\sum_i (Y_i - \beta^T X_i)^2$ subject to $||\beta||_1 \leq C$. The estimator $\hat{\beta}$ can be found quickly since this is a convex minimization. Furthermore, $\hat{\beta}$ is sparse: most of its entries are 0. (The regularization parameter is typically chosen by cross-validation.) Our predictor is then $L(x, \hat{\beta}) = \hat{\beta}^T x$.

The lasso (and its variants) has been used successfully in so many application areas that one cannot doubt its usefulness. This is interesting because the linear model is certainly wrong and is also uncheckable. For example, if Y is a disease outcome and X represents the expression levels of 50,000 genes, it is inconceivable that the mean of Y given $X = x$ would be linear.

Greenshtein and Ritov (2004) provide a low assumption justification for the lasso: it approximates the best sparse linear predictor. Let $R(\beta) = \mathbb{E}(Y - \beta^T X)^2$ denote the risk of predicting a new Y from a new X. Define the best, sparse linear predictor to be $L_*(x) = \beta_*^T x$ where $\beta_* = \text{argmin}_{||\beta||_1 \leq L} R(\beta)$ is the best, sparse linear predictor. Let $\hat{\beta}$ be the lasso estimator. Under very weak assumptions, it can be shown that

$$R(\hat{\beta}) - R(\beta_*) \leq O_P\left(L^2 \sqrt{\frac{\log d}{n}}\right).$$

Thus, the lasso 'works' under very weak conditions. This reasoning can even be extended to nonparametric versions of the lasso (Ravikumar, Lafferty, Liu and Wasserman 2009).

3.2 Salient Structure

A more difficult challenge is to give a low assumption interpretation to the many 'structure finding' algorithms that are now common for estimating high dimensional data.

As an example, we consider forest density estimation (Liu, Xu, Gu, Gupta, Lafferty and Wasserman 2011). Let $X^{(1)}, \ldots, X^{(n)}$ be n vectors, drawn from a distribution P where $X^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$. Imagine measuring d genes on n subjects, for example. Let $X = (X_1, \ldots, X_d)$ denote a generic random variable.

One way to explore the structure of P is to consider the undirected graph G whose d vertices $X_1,\ldots,X_d$ correspond to the d elements of the vector X. An edge is omitted between X_i and X_j if and only if X_i and X_j are independent conditional on the other variables. Graphs of this form are now routinely used to find gene networks, for example.

The preferred method for estimating the graph G is to assume that P is multivariate Normal with mean vector μ and covariance matrix Σ. In this case, an edge between X_i and X_j is missing if and only if $\Omega_{ij} = 0$ where $\Omega = \Sigma^{-1}$. Estimating Ω when d is large can be done using the graphical lasso. We maximize the Gaussian likelihood subject to the sparsity penalty $\sum_{i\neq j} |\Omega_{ij}| \leq L$.

There are many apparent successes using the graphical lasso. But we have moved far from the world of low assumptions. The forest approach attempts to remedy this. Suppose we require the graph to be a forest, which means that the graph has no cycles. Under this assumption, the density of P can be written as

$$p(x) = \prod_{j=1}^{d} p_j(x_j) \prod_{j\sim k} \frac{p_{jk}(x_j,x_k)}{p_j(x_j)p_k(x_k)}$$

where $j \sim k$ means there is an edge between X_j and X_k. Low assumption inference is now possible because, despite the fact that d might be large, we need only estimate one and two-dimensional marginals of P which can be done very nonparametrically. An example is shown in *figure 1*.

What have we done here? We have traded a strong distributional assumption on P for a strong structural assumption on the graph G. My sense is that the latter is preferable. But currently, we have no way to make that intuition precise. More importantly, I doubt very much that any of the variables are truly conditionally independent of any other variables. Yet I don't think that renders the graph useless. Rather, I think that the graph is capturing a salient structure. Again, I don't know how to make this precise or give it a secure foundational justification. More importantly, I don't know how to make sensible (and rigorous) uncertainty statements in a problem like this.

What's happening here is that statisticians are reacting to challenging problems by creating new methods. The methods appear to be reasonable and effective. But they lack philosophical foundations.

4. Bayes?

What is the role of Bayesian inference in our high dimensional world? In principle, low assumption Bayesian inference is possible. We simply put a prior π on the set of all distributions $\mathscr{P}$. The rest follows from Bayes theorem. But this is clearly unsatisfactory. The resulting priors have no guarantees, except the solipsistic guarantee that the answer is consistent with the assumed prior.

There are some successes with nonparametric Bayesian methods in difficult machine learning problems. But the answers are usually checked against held out data. This is quite sensible but then this is Bayesian in form not substance.

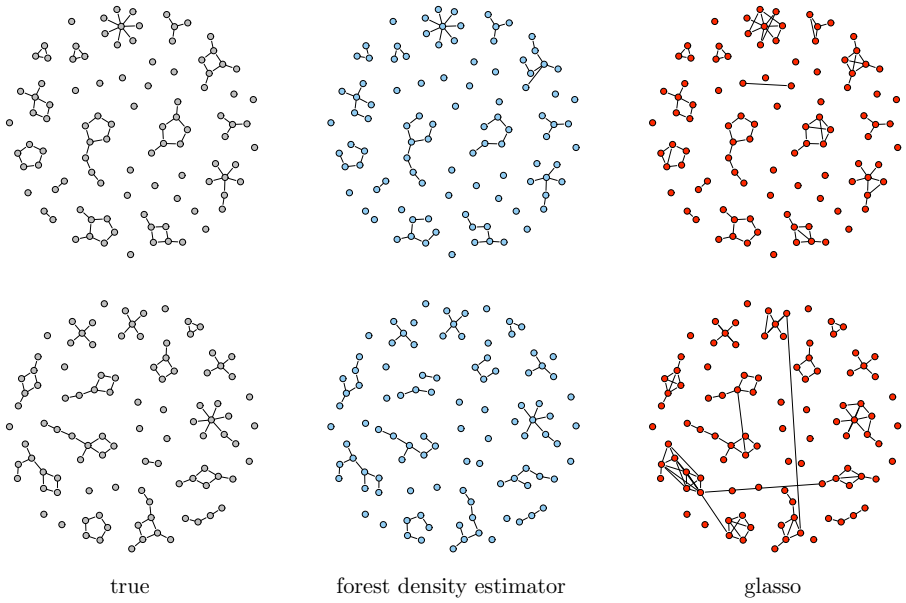

Figure 1: Graph estimation (from Liu, Xu, Gu, Gupta, Lafferty and Wasserman 2011). Top row: Gaussian data. Bottom row: Non-Gaussian data. First column: true graph. Middle column: forest estimator. Right column: graphical lasso.

It is worth recalling the main result from Freedman (1965) which shows how hard it is to construct good priors on large spaces. Let Λ be the set of distributions on $\{1,2,3,\ldots,\}$. Let Π be the set of priors on Λ. The pair $(\lambda,\mu) \in \Lambda \times \Pi$ is consistent if the posterior under μ converges to a point mass at λ a.s. with respect to λ. Endow Λ with the weak* topology: $\lambda_n \to \lambda$ if $\lambda_n(i) \to \lambda(i)$ for all i. Endow Π with the weak* topology: $\mu_n \to \mu$ if $\int f d\mu_n \to \int f d\mu$ for all bounded, continuous functions f. Put the product topology on $\Lambda \times \Pi$. A set is *nowhere dense* if the interior of its closure is empty. A set is *meager* if it is a countable union of nowhere dense sets. A meager set is the topological version of a null set.

Theorem 1 (Freedman 1965)
The set of consistent pairs (λ,μ) is meager.

A corollary of the theorem is that most Bayesians will disagree with each other. As Freedman puts it:

> "[...] it is easy to prove that for essentially any pair of Bayesians, each thinks the other is crazy."

The lesson is that it is hard to choose a good prior even on the set of distributions for a countable set. Constructing sensible, low assumption priors for high dimensional problems can only be more vexing.

5. Conclusion

According to Wikipedia:

> "Foundations of statistics is the usual name for the epistemological debate in statistics over how one should conduct inductive inference from data. Among the issues considered in statistical inference are the question of Bayesian inference versus frequentist inference, the distinction between Fisher's 'significance testing' and Neyman-Pearson 'hypothesis testing', and whether the likelihood principle should be followed. Some of these issues have been debated for up to 200 years without resolution."

Wikipedia references Efron (1978) for this definition. It is telling that this definition is from a paper written over 20 years ago. Perhaps it is time to rethink what we mean by 'the foundations of statistics'. The best way to do this is to look at the wide array of new problems that statisticians (and computer scientists) are facing with the deluge of high dimensional complex data that is now so common.

References

Bahadur, R. and L. Savage (1956), "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems", *The Annals of Mathematical Statistics* 27, 1115–1122.

Cesa-Bianchi, N. and G. Lugosi (2006), *Prediction, Learning, and Games*, Cambridge: Cambridge University Press.

Davies, P. L. and A. Kovac (2001), "Local Extremes, Runs, Strings and Multiresolution", *The Annals of Statistics* 29, 1–65.

—, — and M. Meise (2009), "Nonparametric Regression, Confidence Regions and Regularization", *The Annals of Statistics* 37, 2597–2625.

Donoho, D. (1988), "One-Sided Inference about Functionals of a Density", *The Annals of Statistics* 16, 1390–1420.

Efron, B. (1978), "Controversies in the Foundations of Statistics", *American Mathematical Monthly* 85, 231–246.

Freedman, D. (1965), "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case II", *The Annals of Mathematical Statistics* 36, 454–456.

Greenshtein, E. and Y. Ritov (2004), "Persistence in High-Dimensional Linear Predictor Selection and the Virtue of Overparametrization", *Bernoulli* 10, 971–988.

Liu, H., M. Xu, H. Gu, A. Gupta, J. Lafferty and L. Wasserman (2011), "Forest Density Estimation", *Journal of Machine Learning Research* 12, 907–951.

Ravikumar, P., J. Lafferty, H. Liu and L. Wasserman (2009), "Sparse Additive Models", *Journal Of The Royal Statistical Society Series B* 71, 1009–1030.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society Series B* 58, 267–288.

— and L. Wasserman (1988), "Sensitive Parameters", *The Canadian Journal of Statistics* 16, 185–192.