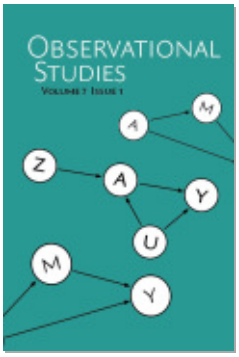
Modern Data Modeling: Cross-Fertilization of the Two Cultures

Jianqing Fan, Cong Ma, Kaizheng Wang, Ziwei Zhu



Observational Studies, Volume 7, Issue 1, 2021, pp. 65-76 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2021.0023*

Observational Studies Vol. 7.1 (2021): Published July 2021

Modern data modeling: Cross-fertilization of the two cultures

Jianqing Fan **jqfan@princeton.edu**
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08540, USA

Cong Ma **congm@berkeley.edu**
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720, USA

Kaizheng Wang **kw2934@columbia.edu**
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027, USA

Ziwei Zhu **ziweiz@umich.edu**
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

Abstract

The past two decades have witnessed deep cross-fertilization between the two cultures—statistics (data/generative modeling) and machine learning (algorithmic modeling), which is in stark contrast to the scene pictured in Breiman's inspiring work. In light of this major confluence, we find it helpful to single out a few salient examples showcasing the impacts of one to the other, and the research progress out of them. We point out in the end that the current big data era especially requires joint efforts from both cultures in order to address some common challenges including decentralized data analysis, privacy, fairness, etc.

Keywords: generative modeling, algorithmic modeling, computational thinking, inferential thinking, distributed learning

1. Introduction

It has been two decades since the publication of Breiman's thought-provoking work, and we find his views still incisive and inspiring. Back then, the gap between the two cultures (i.e., the generative[1] and algorithmic modeling) was perceived to be huge and Breiman urged the embrace of algorithmic modeling from the generative modeling side. However, the past two decades have witnessed deep cross-fertilization of the two cultures of generative and algorithmic modeling on data analytics: they absorb each other's advantage and together underpin the practice of modern data science. In what follows, we first discuss how machine

1. Breiman used the term "data modeling" for this, whereas we refer data modeling to both generative and algorithmic modeling.

learning, the modern term for algorithmic modeling, and statistics, with a longstanding culture of generative modeling, fundamentally impact each other; we then conclude with a few common challenges for both statistics and machine learning that require joint efforts from both cultures.

2. How machine learning impacts statistics

2.1 Assumption-lean statistical inference

"With data gathered from uncontrolled observations on complex systems involving unknown physical, chemical, or biological mechanisms, the a priori assumption that nature would generate the data through a parametric model selected by the statistician can result in questionable conclusions that cannot be substantiated by appeal to goodness-of-fit tests and residual analysis." — page 204

"The one assumption made in the theory (of machine learning) is that the data is drawn i.i.d. from an unknown multivariate distribution." — page 205

Breiman was rightfully skeptical about the (simple) parametric assumptions imposed by statisticians on the data generating process. By contrast, machine learners strive to make minimal assumptions on the data at hand. In fact, the desire for assumption lean statistical inference has existed for a long time in the statistics community; see the transformation from parametric to nonparametric/semiparametric inference. With that being said, it is fair to say that the advance of machine learning speeds up this process and has resulted in immensely useful and scalable methodologies that are capable of handling high-dimensional big data we are seeing every day. In what follows, we single out two salient examples.

Conformal prediction. Let $\{(X_i, Y_i)\}_{i=1}^n$ be n independent and identically distributed (i.i.d.) copies of (X, Y) with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. One crucial task in statistics is to construct a prediction region $C(X) \subseteq \mathbb{R}$ such that, say,

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \approx 0.9. \tag{1}$$

Here (X_{n+1}, Y_{n+1}) denotes a new data point under the same law as (X, Y). To achieve this goal, conventional statistics relies on parametric models such as $Y = \beta^\top X + \varepsilon$ to make inference on the linear coefficients β, and hence on the outcome Y_{n+1}. However, without the help of the parametric assumptions, how shall we construct a valid and tight prediction band?

The method of *split conformal prediction* (Vovk et al., 2005; Lei et al., 2018) starts by randomly splitting the sample into two parts, say $\mathcal{D}_1 := \{(X_i, Y_i)\}_{i=1}^{n/2}$ and $\mathcal{D}_2 := \{(X_i, Y_i)\}_{i=n/2+1}^n$. Then on $\mathcal{D}_1$, one fits a regression function $\hat{\mu}(X)$ using any off-of-shelf methods (e.g., linear regression, LASSO, random forest, neural network, etc.). Next, one applies the estimated regression function to the hold-out sample $\mathcal{D}_2$ and calculates a particular quantile q of the residuals $\{|\hat{\mu}(X_i) - Y_i|\}_{i=n/2+1}^n$. In the end, a prediction band of the form

$$C(X) = [\hat{\mu}(X) - q, \hat{\mu}(X) + q] \tag{2}$$

can be constructed. The validity of such a prediction band can be easily verified via an exchangeability argument, regardless of the actual relationship between X and Y and the

regression method being used. In light of the assumption-free property of conformal prediction, it has been deployed to perform valid uncertainty quantification for a variety of "black boxes" including random forest and neural networks.

Double machine learning. Treatment effect estimation is of great importance in observational studies. To accommodate the effect of confounders, a strong parametric assumption (e.g., a linear and additive form) is normally imposed on the functional norm of the model. However, nowadays one can measure an increasingly large number of covariates, which renders a precise parametric form improbable.

Without specifying the exact form of the model, it is tempting to apply machine learning methods to nonparametrically learn the functional form. To make things concrete, let us consider a partially linear regression model

$$Y = \theta_0 T + f_0(\mathbf{X}) + \varepsilon, \quad \mathbb{E}[\varepsilon \mid \mathbf{X}, T] = 0, \tag{3a}$$

$$T = g_0(\mathbf{X}) + \eta, \quad \mathbb{E}[\eta \mid \mathbf{X}] = 0. \tag{3b}$$

Here T denotes the treatment (e.g., years of education) whose effect θ_0 on the response Y (e.g., income) is to be evaluated, but the response depends also on the measured covariates $\mathbf{X}$ (e.g., age, gender, SES) which might also have impact on the treatment T. Let $Z \coloneqq (T, Y, \mathbf{X})$. One can apply machine learning in a plain fashion to the data $\{Z_i\}_{i=1}^n$ and obtain estimates $\hat{\theta}^{\mathsf{ML}}$ and $\hat{f}^{\mathsf{ML}}$ for θ_0 and f_0, respectively. However, due to the nonparametric nature of f_0, the estimation rate is typically slower than $n^{-1/2}$, which fails to achieve $\sqrt{n}$-consistency for θ_0. As a remedy, double machine learning—proposed in the work Chernozhukov et al. (2018)—splits the data $\{Z_i\}_{i=1}^n$ into two parts $\{Z_i\}_{i=1}^{n/2}$ and $\{Z_i\}_{i=n/2+1}^n$. One uses the first sample $\{Z_i\}_{i=1}^{n/2}$ to estimate the nuisance parameters, in this case the functions f_0 and g_0, and obtain $\hat{f}$ and $\hat{g}$. This step of estimation can be performed using any nonparametric or high-dimensional regression methods including random forest and neural networks. Then, the second sample $\{Z_i\}_{i=n/2+1}^n$ is used to perform Neyman orthogonalization. More specifically, one regresses $Y_i - \hat{f}(\mathbf{X}_i)$ on $T_i - \hat{g}(\mathbf{X}_i)$ to secure an estimate $\hat{\theta}^{\mathsf{DML}}$ of the treatment effect θ_0.

To appreciate the idea, we note by substituting (3b) into (3a) that

$$Y - f(\mathbf{X}) = \theta_0 \eta + \varepsilon, \qquad \text{with} \quad f(\mathbf{X}) = f_0(\mathbf{X}) + \theta_0 g_0(\mathbf{X}).$$

Hence $f(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$ and is estimated by $\hat{f}(\mathbf{X})$ via a machine learning algorithm and η is analogously estimated by $T - \hat{g}(\mathbf{X})$. Now a simple regression of the residuals gives an estimate of the target parameter θ_0. The idea of this kind of residual-regression dates back to the work by Robinson (1988). It has been shown in Chernozhukov et al. (2018) that the estimator $\hat{\theta}^{\mathsf{DML}}$ resulting from the double machine learning procedure is $\sqrt{n}$-consistent whenever both f_0 and g_0 are estimated with an $n^{-1/4}$ rate.

2.2 Discriminative modeling

According to Breiman, algorithmic modelers preferred discriminative models targeting the response given the input while statisticians preferred generative models to explain the whole data generation mechanism. Things have changed dramatically over the past two decades.

Statistics has indeed benefited a lot from the incorporation of discriminative modeling. As an example, consider the binary classification problem. Suppose that our training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are i.i.d. pairs of features and labels from a joint distribution over $\mathbb{R}^d \times \{-1, 1\}$. The optimal Bayes classifier

$$f^*(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) \geq 1/2 \\ -1, & \text{if } \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) < 1/2 \end{cases}$$

achieves the minimum misclassification error (Friedman et al., 2001). A canonical statistical approach usually starts with a probabilistic model of the joint distribution $(\mathbf{X}, Y)$ or the conditional distribution $Y|\mathbf{X}$, followed by parameter estimation or density estimation in order to approximate the optimal Bayes classifier. Traditional statistics provides numerous consistency results given a well-specified model class. Yet the true model is rarely known in practice and different model assumptions may yield very different estimators. One wonders whether there exist methods that achieve consistency under general conditions.

In algorithmic modeling, one looks for a function f from a function class $\mathcal{F}$ to make the misclassification error $\mathbb{P}[\text{sgn}(f(\mathbf{X})) \neq Y]$ small. While it is natural to estimate the optimal classifier f^* by minimizing the empirical misclassification rate

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}[\text{sgn}(f(\mathbf{X}_i)) \neq Y_i] = \frac{1}{n}\sum_{i=1}^n \mathbf{1}[Y_i f(\mathbf{X}_i) < 0] \tag{4}$$

over a function class $\mathcal{F}$, the discontinuity of the step function makes optimization hard. Instead, a common practice is to choose a continuous surrogate loss function $\ell : \mathbb{R} \to \mathbb{R}$ and then minimize

$$\frac{1}{n}\sum_{i=1}^n \ell[Y_i f(\mathbf{X}_i)].$$

Popular choices of ℓ include the hinge function $\max\{1 - x, 0\}$, exponential loss e^{-x}, the logistic loss $\log(1 + e^{-x})$, among others (Boucheron et al., 2005; Fan et al., 2020). The new loss reduces to the empirical misclassification rate (4) when $\ell(x) = \mathbf{1}(x < 0)$. If $\mathcal{F}$ is the family of all linear functions and ℓ is convex, then the new loss function is convex in the model parameters and it can be minimized efficiently. One may wonder whether the estimation procedure enjoys Fisher consistency even with ℓ not chosen according to the true model. A series of works (Zhang, 2004; Bartlett et al., 2006) by statisticians answer the question in the affirmative under minimal assumptions. These algorithms manage to approximate the Bayes optimal classifier without fitting the model. They greatly enriched the arsenal of statisticians.

2.3 Computational thinking

"Some ingenious algorithms make finding the optimal separating hyperplane computationally feasible. These devices reduce the search to a solution of a quadratic programming problem with linear inequality constraints that are of the order of the number N of cases, independent of the dimension of the feature space." — page 209

Though not the main focus of Breiman's paper, computational efficiency indeed lies at at the heart of algorithmic modeling in modern machine learning, due to the ever-increasing

Figure 1: The pursuit of computational efficiency in low-rank matrix completion.

scale of the data set we encounter in practice. In sharp contrast, conventional statistical theory and practice do not include the notion of computational complexity, i.e., the runtime of statistical procedures. The last two decades witnessed statisticians' quest for computationally-efficient statistical methods and desire to establish statistical properties for computable estimators. Below, we use two vignettes to showcase the impact of computational considerations in statistics.

Matrix completion: a journey towards computational efficiency. Given partial entries of a low-rank matrix, can one fill in the missing entries faithfully? This problem, often dubbed as low-rank matrix completion (Candès and Recht, 2009), finds numerous applications in recommendation systems, sensor localization, causal inference, to name a few. To fix ideas, let $\mathbf{M}^\star \in \mathbb{R}^{m\times n}$ be the unknown rank-r matrix of interest, and $\Omega \subseteq [m]\times[n]$ be the index set for the observed entries. A natural statistical solution is given by the rank minimization approach:

$$\text{minimize}_{\mathbf{M}} \quad \mathsf{rank}(\mathbf{M}), \qquad \text{subject to} \;\; \mathbf{M}_{i,j} = \mathbf{M}^\star_{i,j}, \;\text{for}\; (i,j)\in\Omega. \tag{5}$$

Under certain incoherence assumptions, $\mathbf{M}^\star$ can be shown to be the unique minimizer of the above rank minimization problem. Consequently, from a statistical perspective, the problem of matrix completion is "solved". However, a major drawback of this approach is its computational efficiency: the above program is NP hard to solve in the worst case. With this computational considerations in mind, statisticians started the pursuit of computationally-efficient methods for matrix completion.

Inspired by the success story of ℓ_1 minimization in compressed sensing and sparse regression, the following nuclear norm minimization approach arises:

$$\text{minimize}_{\mathbf{M}} \quad \|\mathbf{M}\|_*, \qquad \text{subject to} \;\; \mathbf{M}_{i,j} = \mathbf{M}^\star_{i,j}, \;\text{for}\; (i,j)\in\Omega. \tag{6}$$

It turns out that this convex relaxation method retains the desired statistical property—$\mathbf{M}^\star$ is still the unique minimizer of the problem (6). What is more important is that now the nuclear norm minimization problem is a convex problem that is solvable in polynomial times. Though polynomial-time solvable, the convex relaxation approach (6) oftentimes requires super-linear computational and memory footprints, which makes it challenging to scale to huge data matrices.

To further improve the computational efficiency, one can apply the least-squares principle to the factorized parametrization $\mathbf{M} = \mathbf{X}\mathbf{Y}^\top$ and arrive at the following nonconvex optimization problem

$$\text{minimize}_{\mathbf{X}\in\mathbb{R}^{m\times r},\mathbf{Y}\in\mathbb{R}^{n\times r}} \quad \sum_{i,j\in\Omega}\left\{(\mathbf{X}\mathbf{Y}^\top)_{i,j} - \mathbf{M}^\star_{i,j}\right\}^2. \tag{7}$$

q Though nonconvex, the above problem can be solved in linear time—in time proportional to read the data matrix—via simple gradient descent method; see e.g., the work Ma et al. (2019). Figure 1 illustrates the evolution of computational efforts.

The pursuit of computationally efficient statistical methods is not unique to matrix completion. Similar endeavor has been made in problems including folded-concave penalized likelihood, mixed linear regression, tensor recovery and community detection.

Sparse principal component analysis: a computational-statistical gap. In addition to the search for computationally-efficient statistical methods, the introduction of computational complexity to statistics also brings about an interesting phenomenon called the computational-statistical gap. Informally, there exists a range of signal-to-noise ratios (SNRs) such that statistical inference is information-theoretically possible, while no computationally-efficient methods exist.

To illustrate the ideas, we use the example of sparse principal component analysis from the paper Berthet et al. (2013). Let $X_1, X_2, \ldots, X_n$ be n i.i.d. copies of $X \in \mathbb{R}^d$. One would like to perform the following hypothesis testing:

$$H_0 : X \sim \mathcal{N}(0, \mathbf{I}_d), \qquad H_1 : X \sim \mathcal{N}(0, \mathbf{I}_d + \theta \mathbf{v}\mathbf{v}^\top) \text{ with } \|\mathbf{v}\|_2 = 1, \|\mathbf{v}\|_0 \le k, \tag{8}$$

with $\theta > 0$ a measure of SNR. In words, one wishes to detect whether the data arise from an isotropic Gaussian model (cf. H_0) or a spiked one (cf. H_1). It turns out that, the statistical limit for the detection threshold is $\theta_{\mathsf{stat}} \asymp \sqrt{k \log d/n}$, namely, nontrivial power of testing is possible if and only if $\theta \gtrsim \sqrt{k \log d/n}$. In comparison, the computational limit is $\theta_{\mathsf{comp}} \asymp \sqrt{k^2 \log d/n}$, i.e., when $\lambda \lesssim \sqrt{k^2 \log d/n}$, there is no computationally efficient algorithm that can faithfully detect the sparse eigenvector $\mathbf{v}$. Figure 2 depicts the results.

Such an intriguing computation-statistics gap has also been observed in problems including community detection, spiked tensor models, sparse phase retrieval.

3. How statistics impacts machine learning

3.1 Generative models

Generative models, perhaps a namecard of statistics, has had a huge impact on machine learning. Physicist Richard Feynman once said, "what I cannot create, I do not understand". Being able to generate data as well as the nature does is a crucial step towards full understanding of it. Generative models also come in handy when handling incomplete (e.g.

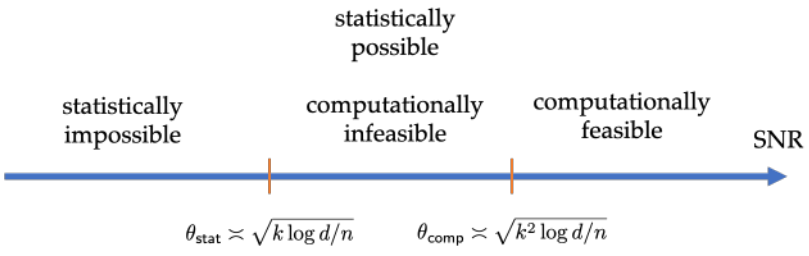


Figure 2: The computational-statistical gap in sparse principle component detection.

missing, unlabeled) data and quantifying uncertainty. To generate high-quality data, the model needs to see through the raw training data and capture their underlying structure. A common belief is that many high-dimensional datasets are intrinsically low-dimensional: the overall variability can be explained by a few key factors. Statisticians have been studying such phenomena for over a century, resulting in powerful tools such as Principal Component Analysis (PCA) (Pearson, 1901) and factor analysis (Thurstone, 1931). Nonlinear generalizations of these classical methods (Kramer, 1991) enjoy better flexibility. Recent developments of deep neural networks have furthered their success.

Two of the most popular deep generative models are autoencoders and Generative Adversarial Networks (GANs). As the name suggests, an autoencoder first "encodes" raw inputs $\mathbf{x} \in \mathbb{R}^d$ to a low-dimensional vector $f(\mathbf{x}) \in \mathbb{R}^K$ and then "decodes" that to get a data point $g[f(\mathbf{x})] \in \mathbb{R}^d$ back in the original space. Given a training sample $\{\mathbf{X}_i\}_{i=1}^n$ from a distribution $\mathcal{P}$, a natural loss function is the mean squared reconstruction error

$$L(f, g) = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{X}_i - g[f(\mathbf{X}_i)]\|_2^2.$$

The functions f and g are parametrized by neural networks. In particular, if both of them are linear, then PCA gives the optimal solution. Extensions include variational autoencoders (Kingma and Welling, 2019) and sparse autoencoders (Ranzato et al., 2007).

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014a), on the other hand, mostly focus on generating data from low-dimensional representations (the decoding step above). Suppose that the data distribution $\mathcal{P}$ concentrates around a K-dimensional manifold. Let $\mathcal{P}_Z$ be a source distribution over $\mathbb{R}^K$, usually set to be $N(\mathbf{0}, \mathbf{I}_K)$. Then, one aims for a function $g : \mathbb{R}^K \to \mathbb{R}^d$ (called a generator) such that the distribution of a random fake sample $g(\mathbf{Z})$ with $\mathbf{Z} \sim \mathcal{P}_Z$ is close to $\mathcal{P}$. In particular, the vanilla GAN seeks to minimize their Jensen-Shannon (JS) divergence via a two-player game between the generator g and a discriminator $d : \mathbb{R}^d \to [0, 1]$ that evaluates the likelihood of a sample $\mathbf{x} \in \mathbb{R}^d$ being real (coming from $\mathcal{P}$) or fake. Eventually, the discriminator d becomes sharp and the generator g produces samples that are hard to distinguish from the real ones. Again, both g and d are neural networks. One can also choose other discrepancy measures between the learned distribution and the truth, such as the Wasserstein distance (Arjovsky et al., 2017).

3.2 Robust statistics

Robust statistics (Box, 1953; Tukey, 1960; Huber, 2004) have also made profound impact in the culture of algorithmic modelling, especially the machine learning community. Since Tukey (1960) observed the extreme sensitivity of some conventional statistical methods to model deviation, there have been massive endeavors in developing stable statistical procedures in the presence of outliers or model misspecification. Robust statistics substantially broaden the capacity of traditional data models so that they can embrace real-world examples. For instance, a recent series of works such as Catoni (2012); Minsker (2015); Devroye et al. (2016); Fan et al. (2017, 2020+); Lugosi and Mendelson (2019) study how to handle heavy-tailed data in point estimation and regression analysis. These works do not assume any parametric form of the data distribution but a bounded moment, and they show that the estimator based on median-of-means or robustified risk minimization exhibits sub-Gaussian

behavior around the truth. Ironically, modern sophisticated supervised learning architectures may still suffer from similar sensitivity problems as conventional statistical procedures. Goodfellow et al. (2014b) found that an imperceptibly small perturbation of the pixel values can change GoogLeNet's (Szegedy et al., 2015) classification of an image (see Figure 3). To achieve robustness against adversarial perturbation, Goodfellow et al. (2014b) proposed to train the architecture by minimizing an adversarial objective that combines both clean and adversarial instances. This technique reduced the original error rate of 89.4% on adversarial examples to 17.9%.

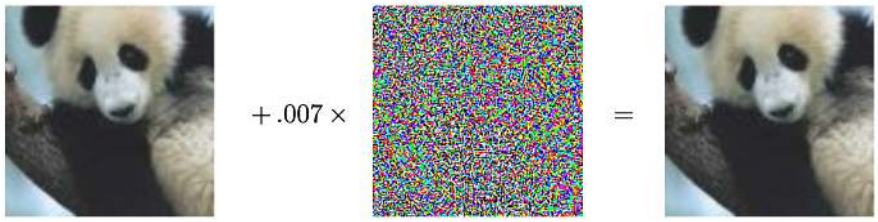


Figure 3: Adversarial perturbation: GoogLeNet classifies the left panda correctly, but misclassifies the right panda as a gibbon.

3.3 Interpretability, causal inference and uncertainty quantification

Interpretability, cause inference, and uncertainty quantification have always been a top priority in the generative modeling culture. In high-dimensional setups, pursuit of simple and interpretable models has inspired a myriad of dimension reduction tools, such as principal component analysis (PCA) and kernel PCA, and variable selection techniques such as LASSO, SCAD and MCP (Fan et al., 2020). Variable selection and its associated inference also provide useful tools for causal inference. Nevertheless, as Breiman said in the beginning of Section 9, "Accuracy and simplicity (interpretability) are in conflict." The past two decades have witnessed revolutionary success of black-box algorithmic models, in particular deep neural nets, in computer vision, natural language processing, among others. A natural question thus arises: Is the sacrifice on interpretability necessary to achieve high prediction accuracy? Recently the so-called interpretable AI has become an increasingly heated topic; the goal is seeking for an interpretable model that yields the state-of-the-art prediction accuracy.

Statistics traditionally places more emphasis on causal inference and uncertainty quantifications via generative models; a modern example is the statistical inference for predicted entries in matrix completion (Chen et al., 2019). Whereas machine learning focuses more on predictive tasks via algorithmic modeling. Bootstrap and cross-validations are frequently used in machine learning to quantify uncertainties and to increase stabilities. The community nowadays has frequently employed statistical models for understanding causal relation in disciplinary science with uncertainty quantification.

4. Common challenges in the big data era

As we have emphasized throughout the manuscript, the gap between the two cultures is much smaller than that in Breiman's time. Both cultures have learned from each other and the cross-fertilization greatly enhanced our ability to tackle more challenging data science problems. However, it is worth pointing out that in the current big data era, there are still a variety of interesting and important challenges that await joint efforts from the two cultures. In addition to the aforementioned desiderata (e.g., statistical/computational efficiency, interpretability, causal inference, uncertainty quantifications), privacy, fairness and also decentralized data analysis become increasingly important when applying data analytical tools in practice.

To conclude, we detail the challenge on decentralized data analysis. Besides having immense volume, modern data sets are often decentralized in the sense that they are scattered across different places across which the communication is highly restrictive. Consider international IT companies that collect data worldwide. Constraints due to communication budget, network bandwidth and legal policies stifle the hope of aggregating and maintaining global data in a single data center. Another example is the health data that are generated in many hospitals or labs. Fusing them in a central location is often prohibited by privacy and ownership concerns. It is imperative to design distributed statistical and machine learning algorithms that enjoy sharp statistical accuracy, strong privacy guarantee and cheap communication overhead.

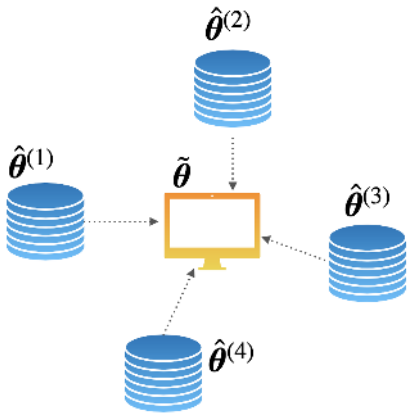


Figure 4: Decentralized data and distributed statistical learning. $\{\hat{\boldsymbol{\theta}}^{(k)}\}_{k=1,\ldots,4}$ are local estimators. They are transferred to a central server to generate an aggregated estimator $\tilde{\boldsymbol{\theta}}$, which provably enjoys full-sample accuracy as long as the bias of the local estimators is small.

Averaging model parameters turn out to be a powerful approach (Figure 4). It only requires transferring local models instead of local data, thereby dramatically reducing communication cost. The past decade has witnessed success of this strategy in a wide range of statistical tasks (Zhang et al., 2013; Chen and Xie, 2014; Lee et al., 2017; Rosenblatt and

Nadler, 2016; Fan et al., 2019; Battey et al., 2018). The main message is that averaging achieves full-sample accuracy as long as the bias of local estimators is negligible. As regard to modern complex algorithmic models, a recent work by McMahan et al. (2017) from Google studies how to leave the training data on the mobile devices and learn a shared model by aggregating local updates. They propose a federated averaging (FedAve) algorithm where the central server averages the model parameters from clients every time they perform few epochs of local training. They show that FedAve requires much less communication than naive average of stochastic gradient to achieve the same testing performance.

Decentralized data setups inspire both generative and algorithmic modeling to consider new criteria for methodologies and theory. Besides prediction accuracy and model interpretability, many practical concerns need to be taken into account: privacy, fairness, communication, computation, storage, among others. Conquering these challenges will undoubtedly inspire more joint effort and mutual influence of the two cultures.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3): 1352–1382, 2018.

Quentin Berthet, Philippe Rigollet, et al. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(4):1780–1815, 2013.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré*, 48(4):1148–1185, 2012.

Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24(4):1655–1684, 2014.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116 (46):22931–22937, 2019.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Jianqing Fan, Quefeng Li, and Yuyan Wang. Robust estimation of high-dimensional mean regression. *Journal of Royal Statistical Society, Series B*, 79(1):247–265, 2017.

Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47:3009–3031, 2019.

Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical foundations of data science*. CRC press, 2020.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *The Annals of Statistics*, page To appear, 2020+.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Peter J Huber. *Robust Statistics*, volume 523. John Wiley & Sons, 2004.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901.

Marc Ranzato, Christopher Poultney, Sumit Chopra, and Yann LeCun. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, 19:1137, 2007.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

Jonathan D Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Louis Leon Thurstone. Multiple factor analysis. *Psychological review*, 38(5):406, 1931.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Divide and conquer kernel ridge regression. In *COLT*, pages 592–617, 2013.