Data Science, Statistical Modeling, and Financial and Health Care Reforms

by Tze Leung Lai

Department of Statistics, Stanford University

This paper discusses some new trends in the field of Statistics, in response to technological innovations leading to "big data" as well as opportunities and challenges in the wake of health care and financial reform legislation.

Introduction

The past decade witnessed exciting developments and new trends in statistical modeling and its applications. These trends are related to the "explosion of digital data," as pointed out in the article "For Today's Graduate, Just One Word: Statistics" of *The New York Times* (August 6, 2009):

- "In field after field, computing and the Web are creating new realms of data to explore — sensor signals, surveillance tapes, social network chatter, public records and more. And the digital data surge only promises to accelerate, rising five-fold by 2012, according to a projection by IDC, a research firm."
- "Even the recently ended Netflix contest, which offered $1 million to anyone who could significantly improve the company's movie recommendation system, was a battle waged with the weapons of modern statistics."
- "I.B.M., seeing an opportunity in data-hunting services, created a Business Analytics and Optimization Services group in April. The unit will tap the expertise of the more than 200 mathematicians, statisticians and other data analysts in its research labs — but the number is not enough."
- "If the data explosion magnifies longstanding issues in statistics, it also opens up new frontiers."

Closely related to these developments is "big data," which is one of the most promising business trends in the past decade and which has also led to the development of a new interdisciplinary program called "data science." In the next section we give an overview of data science and also describe some active areas of research on statistical modeling and analysis of big data.

The past few years also witnessed the beginning of a new era in financial markets and in the US health care system. In March 2010, landmark health care reform was passed through two federal statutes: the Patient Protection and Affordable Care Act and subsequently amended by the Health Care and Education Reconciliation Act. A few months later, the Dodd-Frank Wall Street Reform and Consumer Protection Act was signed into law in the US on July 21, 2010, in response to widespread calls for regulatory reform following the 2007–2008 financial crisis. The last three sections of the paper discuss the challenges and opportunities for innovative study design, data analysis, and statistical modeling in this new era for finance and health care.

Data Science and Statistical Methods for Big Data

The term "data science" arose in the field of computer science. In 1974, Naur [55] freely used this term in his survey of contemporary data processing methods for a wide range of applications. In the field of statistics, Cleveland [17] introduced this term as a new direction for the field in 2001. Two years later, the *Journal of Data Science* began its publication under the founding editor Min-Te Chao of the Institute of Statistical Science at Academia Sinica. It currently has editorial offices in Taipei (Fu Jen University), Beijing (Renmin University), and New York (Columbia University). Its emphasis is on "statistical methods at large for collecting, analyzing, modeling" data in applications. Its scope is very different from that of the *Data Science Journal*, which started in 2002, and which "publishes data or data compilations if the quality of data is excellent or if significant efforts are required in compilation."

Data science, therefore, includes high performance computing, data processing, development and management of databases, data warehousing, mathematical representations, statistical modeling and analysis, and visualization with the goal of extracting information from the data collected for domain-specific applications. Interdisciplinary graduate programs to train data scientists are being established at a number of universities, including Columbia, Stanford, New York University, and North Carolina State University. At Stanford, the MS programs in Statistics and in Computational & Mathematical Engineering offer a joint Data Science track, beginning in the 2013–2014 academic year. The track offers courses in advanced statistical methods and models, machine learning and data mining, high-performance computing, numerical analysis and optimization, and applied and computational

mathematics. In addition, PhD programs in the Departments of Statistics and Computer Science and in the Institute of Computational & Mathematical Engineering cover different aspects of research in data science.

One of the most active areas of data science research is related to very large data sets, or "big data," which pose computational and statistical challenges. As an illustration of the statistical challenges, consider the linear regression model

(1) $\quad y_t = \alpha + \beta_1 x_{t1} + \cdots + \beta_p x_{tp} + \epsilon_t, \quad (t = 1, \ldots, n),$

in which ϵ_t represent random, unobserved disturbances with $E(\epsilon_t) = 0$. Estimation of the regression parameters $\alpha, \beta_1, \ldots, \beta_p$ is an age-old topic now often taught in introductory statistics courses. It became a "hot" topic again in the last decade, in response to data explosion that results in p (# of input variables) considerably larger than n (sample size). The $p \gg n$ problem appears hopeless at first sight since it involves many more parameters than the sample size and therefore the parameters cannot be well estimated, resulting in unacceptably large variances of the estimates. On the other hand, it was recognized that the regression function $f(x_1, \ldots, x_p) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$ is still estimable if the regression model is "sparse" and that many applications indeed involve sparse regression models. Such problems are of increasing importance in genomics, in which n is the number of subjects in a clinical study that requires informed consent of human subjects and p is the number of locations in a genome at which measurements are taken. Advances in high-throughput microarray technology have resulted in data explosion concerning p, but it is difficult to recruit patients into clinical trials, resulting in $n \ll p$. There are two major issues with estimating $\beta = (\beta_1, \ldots, \beta_p)'$ when $p \gg n$. The first is singularity of $X'X$, where $X = (x_{tj})_{1 \le t \le n, 1 \le j \le p}$. Denoting the row vectors of X by X_t, note that the n values $X_1\beta, \ldots, X_n\beta$ cannot determine the p-dimensional vector β uniquely for $p > n$. Assuming the ϵ_t to be independent and identically distributed (i.i.d.) normal and using a normal prior $\beta \sim N(0, \lambda I)$ can remove such singularity since the posterior mean of β is the ridge regression estimator [29, Sect. 3.4.1]:

(2) $\quad \hat{\beta}^{\text{ridge}} = (X'X + \lambda I)^{-1} X'Y,$

where $Y = (y_1, \ldots, y_n)'$. The posterior mean minimizes the penalized residual sum of squares $\|Y - X\beta\|^2 + \lambda\|\beta\|^2$, with the L_2-penalty $\|\beta\|^2 = \sum_{j=1}^p \beta_j^2$ and regularization parameter λ.

The second issue with estimating β when $p \gg n$ is sparsity. Although the number of parameters is much larger than the sample size, one expects for the problem at hand that most of them are small and can be shrunk to 0. While the L_2-penalty does not lead to a sparse estimator $\hat{\beta}^{\text{ridge}}$, the L_1-penalty $\sum_{j=1}^p |\beta_j|$ does. The minimizer $\hat{\beta}^{\text{lasso}}$ of $\|Y - X\beta\|^2 + \lambda \sum_{j=1}^p |\beta_j|$ introduced by Tibshirani [68], is called lasso (*least absolute shrinkage and selection operator*) because it sets some coefficients to 0 and shrinks the others toward 0, thus performing both subset section and

shrinkage. Unlike ridge regression, $\hat{\beta}^{\text{lasso}}$ does not have an explicit solution unless X has orthogonal columns, but can be computed by convex optimization algorithms. Oracle inequalities for $\sum_{t=1}^n (\hat{y}^M(x_t) - y(x_t))^2$ have been developed recently, when M is lasso or the closely related Dantzig selector, by Candés and Tao [13], Bickel et al. [10], and Candés and Plan [14]. Zhao and Yu [73] and Zhang and Huang [72] have also shown that lasso is variable-selection consistent under certain conditions.

Zou and Hastie [75] introduced the elastic net estimator β^{enet} that minimizes a linear combination of L_1- and L_2-penalties:

(3)
$$\beta^{\text{enet}} = (1 + \lambda_2) \arg\min_{\beta} \left\{ \|Y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2 \right\},$$

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. The factor $(1 + \lambda_2)$ in (3) is used to correct the "double shrinkage" effect of the naive elastic net estimator, which is Bayes with respect to the prior density proportional to $\exp\{-\lambda_2\|\beta\|^2 - \lambda_1\|\beta\|_1\}$, a compromise between the Gaussian prior (for ridge regression) and the double exponential prior (for lasso). Note that (3) is still a convex optimization problem. The choice of the regularization parameters λ_1 and λ_2 in (3), and λ in ridge regression or lasso, is carried out by k-fold cross-validation [29, Sect. 7.10.1].

Since the non-convex optimization problem of minimizing

(4) $\qquad \|Y - X\beta\|^2 + \lambda \sum_{j=1}^p I_{\{\beta_j \ne 0\}},$

which corresponds to the L_0-penalty, is infeasible for large p, lasso is sometimes regarded as an approximation of (4) by a convex optimization problem. Ing and Lai [30] recently introduced a fast stepwise regression method, called the *orthogonal greedy algorithm* (OGA), following the terminology introduced by Temlyakov [66], and used it in conjunction with a *high-dimensional information criterion* (HDIC) for variable selection along the OGA path. The method, which provides an approximate solution of the L_0-regularization problem, has three components. The first is the forward selection of input variables in a greedy manner so that the selected variables at each step minimize the residual sum of squares after ordinary least squares (OLS) is performed on it together with previously selected variables. This is carried out by OGA that orthogonalizes the included input variables sequentially so that OLS can be computed by component-wise linear regression, thereby circumventing matrix inversion. The second component of the procedures is a stopping rule to terminate forward inclusion after K_n variables are included. The choice $K_n = O((n/\log p_n)^{1/2})$ is based on an oracle-type inequality, similar to those for lasso, and a convergence rate result reflecting the bias-variance trade-off in the OGA iterations. The third component of the procedure is variable selection along the OGA path according to

(5) $\qquad \text{HDIC}(J) = n \log \hat{\sigma}_J^2 + \sharp(J) w_n \log p,$

where $J \subset \{1,\ldots,p\}$ represents a set of selected variables, $\hat{\sigma}_J^2 = n^{-1}\sum_{t=1}^{n}(y_t - \hat{y}_{t;J})^2$ in which $\hat{y}_{t;J}$ denotes the fitted value of y_t when Y is projected into the linear space spanned by the column vectors X_j of X, with $j \in J$, and w_n characterizes the information criterion used, e.g., $w_n = \log n$ for HDBIC, and $w_n = c$ for some constant c that does not change with n for HDAIC. Letting $\hat{J}_k = \{\hat{j}_1,\ldots,\hat{j}_k\}$ denote the set of selected variables up to the kth step of OGA iterations, Ing and Lai [30] choose $\hat{k}_n = \arg\min_{1 \le k \le K_n} \text{HDIC}(\hat{J}_k)$ and eliminate irrelevant variables $\hat{j}_l$ along the OGA path if $\text{HDIC}(\hat{J}_{\hat{k}_n} - \{\hat{j}_l\}) > \text{HDIC}(\hat{J}_{\hat{k}_n})$, $1 \le l \le k_n$. The procedure is denoted by the acronym OGA+HDIC+Trim, which they showed to have the oracle property of being equivalent to OLS on an asymptotically minimal set of relevant regression under a strong sparsity assumption.

A closely related topic in introductory courses on multivariate statistics, which was also "revolutionized" by big data with $p > n$, is estimation of the covariance matrix V of a p-dimensional distribution, based on a sample of i.i.d. vectors $X_1,\ldots,X_n$. A basic result often taught in those courses is that the sample covariance matrix is a consistent estimator of V and its eigenvalues are also consistent estimators of those of V, for fixed p and $n \to \infty$. In this case, the joint asymptotic distribution of the eigenvalues is normal if those of V are distinct, and is given by Anderson's formula [2] for general V. Johnstone [32] pointed out the inadequacy of this asymptotic theory in modern applications such as climate studies where n is the number of time points and p is the number of observation stations; information retrieval and search engines for which n and p are at least in the thousands; and functional data analysis in which each data point is a curve and therefore typically high-dimensional. He studied the asymptotic distribution of the largest eigenvalue $\hat{\lambda}_{\max}$ of the sample covariance matrix in the case where $p/n \to c \le 1$ and the components of X_i are i.i.d. standard normal. He showed that $\{\hat{\lambda}_{\max} - (\sqrt{n-1} + \sqrt{p})^2\}/\sigma_{n,p}$ converges in distribution to the Tracy–Widom law of order 1, where $\sigma_{n,p} = \{\sqrt{n-1} + \sqrt{p}\}\{(n-1)^{-1/2} + p^{-1/2}\}^{1/3}$. This limiting distribution was found by Tracy and Widom [69] and its distribution function is defined by Painlevé's type II differential equation. The numerical studies in [32] and [33] show that the Tracy–Widom distribution can provide very good approximations to the distribution of $\hat{\lambda}_{\max}$ even for p as small as 2.

The asymptotic distribution of the eigenvalues of a $p \times p$ real symmetric or complex Hermitian matrix is a basic theme in random matrix theory (RMT). Random matrices can be used to model certain behavior of physical systems. For example, Wigner [71] used the empirical measure of the eigenvalues of a symmetric $p \times p$ matrix such that the entries above the diagonal are i.i.d. and are independent of those on the diagonal that are also i.i.d., with all entries being zero-mean random variables, to model the spectrum of a heavy atom. He established weak convergence of the empirical measure to a semi-circle law, which has been named after him, as $p \to \infty$. RMT is an active

area of research in probability theory and mathematical physics. One topic in RMT that witnessed major recent advances, by Erdös et al. [24] and Tao and Vu [65], is related to the "universality principle" for certain local behavior of eigenvalues of matrix ensembles. Another is the application of RMT to solve the longstanding problem concerning the asymptotic distribution of the length of the longest increasing subsequence in a random permutation by Baik et al. [6] and Okounkov [56].

In 1967, Marčenko and Pastur [54] published a foundational paper in RMT, in which they generalized random Hermitian matrices to random self-adjoint linear operators and introduced a powerful method involving Stieltjes transforms to derive the limiting distribution of the eigenvalues of the operator. For the special case of sample covariance matrices, this technique was further developed by Bai [3], Bai and Silverstein [4, 5], and others. El Karoui [23] and Ledoit and Wolf [51] recently made use of these refinements, together with the original Marčenko–Pastur equation that links the Stieltjes transform of the empirical spectral distribution to an integral with respect to the population spectral distribution, to derive consistent estimators of large-dimensional covariance matrices and to show that they have certain oracle properties, without *a priori* assumptions such as sparsity [9, 59], or graphical model structure [57], or factor model structure [25].

Another example of how big data have changed statistical inference is multiple testing, which is a topic often taught in a first course on analysis of variance and inference in linear regression models. Efron [22, pp. ix and 15–16] points out that classical multiple testing meant "somewhere between two and ten (simultaneous testing) problems, not thousands" as in what the era of "scientific mass production excels at, e.g., expression levels comparing sick and healthy subjects for thousands of genes at the same time by means of microarrays." He gives an example of testing the null hypothesis H_{0i} that the expression levels for gene i have the same distribution for normal and cancer patients, $i = 1,\ldots,p$, with $p = 6033$ genes and $n_1 = 50$ normal men and $n_2 = 52$ prostate cancer patients. The principal goal of the study was to discover a small number of genes that might have a causal link to prostate cancer. New statistical methodologies, based on empirical Bayes arguments to "combine frequentist and Bayesian elements in analyzing problems of repeated structure," are introduced in [22] to handle these "large-scale" multiple testing problems. Dudoit and van der Laan [20] describe other approaches to a wide range of genomic testing problems, including identification of differentially expressed and co-expressed genes in high-throughput microarray experiments, associations between gene expression measures and biological annotation metadata, protein sequence analysis and genetic mapping of complex traits.

The monograph by Hastie et al. [29] describes new methods and important ideas for "statistical learning" from vast amounts of data that arise in this information age. Its coverage is broad, ranging from supervised learning, which includes classification, regression, and prediction, to unsupervised learning that includes association

rules, cluster analysis, principal components, independent component analysis, multidimensional scaling, and dimension reduction. Chapter 18 of the book considers the case $p \gg n$ that has "become of increasing importance, especially in genomics and other areas of computational biology," focusing on classification problems, supervised principal components, and multiple testing for assessing the significance of each of p features (predictors).

In the preceding paragraphs of this section we have focused on massive data in genomics and other areas of computational biology. Vast amounts of data also arise naturally in measurements in engineering applications and experiments in the physical sciences. A case in point is the data from the Large Hadron Collider at CERN (European Organization for Nuclear Research) to create and study Higgs bosons and related questions to test the Standard Model of particle physics. Careful and thorough analysis of these massive data and assessing the statistical significance of the experimental results led CERN to confirm in March 2013 that the particle generated "strongly indicates that it is a Higgs boson." A major difference between the big data in the engineering and physical sciences and those in the biomedical sciences is the high signal-to-noise ratio of the former and the large between-subject and within-subject variabilities of the latter. Thus the former focuses on signal recovery whereas the latter's focus is on prediction and inference, using statistical models to address these variabilities. As an illustration, consider the regression model (1) for which the random disturbances ϵ_t are typically assumed to have substantial variances in the applications and methods described in the three paragraphs that follow (1). In particular, the regularization methods in the case $p \gg n$ attempt to strike a good balance between bias and variance. In engineering applications, the $p \times 1$ vector β represents a signal and $X_1\beta, \ldots, X_n\beta$ are n measurements of the signal. The basic questions of signal recovery are how many measurements are necessary to reconstruct the signal and what algorithms can be used for the reconstruction. Suppose the signal is m-sparse in the sense that β has m nonzero components. Then one requires $n \geq m$ but sparse signals can be reconstructed with far fewer than p measurements.

The *basis pursuit* algorithm for sparse signal recovery [16] solves an optimization problem of the form

(6) $$\min_{\beta} \|\beta\|_1 \text{ subject to } Y = X\beta,$$

where Y is an $n \times 1$ vector of measurements and X is an $n \times p$ matrix with row vectors X_t. Instead of exact recovery, one may want a more sparse solution by imposing a constraint on $\|Y - X\beta\|^2$, hence minimizing $\|\beta\|_1$ subject to $\|Y - X\beta\|^2 \leq \delta$. An equivalent minimization problem is *basis pursuit denoising*:

(7) $$\min_{\beta} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\},$$

which is the same as the regularized least squares criterion used by lasso. Concerning the sample size n required

to construct m-sparse signals, Candés and Tao [15] have shown that for $n \geq Km\log(p/m)$ and i.i.d. X_t drawn from the standard p-variate normal distribution, it is possible to reconstruct every m-sparse signal β with probability exceeding $1 - e^{-kN}$, where K and k are positive universal constants. Donoho [18] found important applications of this kind of results to compressed sensing, which reconstructs a signal by efficiently acquiring a minimum amount of data needed. While lasso is closely related to the basis pursuit denoising algorithm for signal recovery, the orthogonal greedy algorithm described earlier in connection with (5) is the same as *orthogonal matching pursuit* (OMP) in the signal processing literature. Tropp and Gilbert [70] have shown that similar to the aforementioned result of Candés and Tao, OMP can recover an m-sparse signal with probability exceeding $1 - 2\delta$ if $n \geq Km\log(p/\delta)$ and X_t are i.i.d. standard p-variate normal. The constant K is ≤ 20 and can be reduced to around 4 for large values of m. OMP is a refinement of *matching pursuit*, introduced by Mallat and Zhang [53], that uses a greedy algorithm to search over an over-complete dictionary ($p > n$) of basis functions to match n measurements of a signal. It is also called the *pure greedy algorithm* by Temlyakov [66] and is the same as the population version of the L_2-boosting procedure of Bühlmann [12], who studied the sample version of L_2-boosting for the regression model (1) with $p \gg n$. Earlier, Friedman [28] introduced a general gradient boosting procedure for classification and regression, particularly in the context of stagewise additive modeling.

In the second paragraph of this section, we have noted the interdisciplinary nature of data science that involves statistics, computer science, and application areas such as computational biology and signal processing. In 2004, Jordan [34] commented: "The fields of statistics and computer science have generally followed separate paths for the past several decades, each field providing useful services to the other, but with the core concerns of the two fields rarely appearing to intersect. In recent years, however, it has become increasingly evident that the long-term goals of the two fields are closely aligned." He pointed out that graphical models, which consist of probability distributions defined in terms of directed or undirected graphs, as an example where the two fields have shared goals and methods, "including the algorithmic ideas that allow graphical models to be deployed in large-scale data analysis problems." In particular, the sampling algorithms such as Markov chain Monte Carlo widely used in Bayesian inference can be better designed by exploiting the graphical structure of the underlying statistical model. Another example is ensemble learning that combines many computationally convenient "weak learners" to produce a "strong learner," as in the AdaBoost classification algorithm of Freund and Schapire [27], which is equivalent to forward stagewise additive modeling using the exponential loss function; see Chapter 10 of [29]. Chapters 15 and 16 of [29] describe important developments of ensemble learning in the fields of statistics and computer science.

Health Care Reform and Innovative Study Design and Analysis

One of the provisions of the Patient Protection and Affordable Care Act mentioned in Introduction is the establishment of a non-profit Patient-Centered Outcomes Research Institute to undertake comparative effectiveness research (CER), examining the "relative health outcomes, clinical effectiveness, and appropriateness" of different medical treatments. This involves the design of comparative studies of the treatments and their statistical analysis. One approach to CER is to use observational studies, including analysis of claims or registry data [64]. As pointed out by Shih and Lavori [62], such an approach involves "confounding by indication," the tendency for clinicians and patients to choose treatments with their anticipated effects in mind. This leads to bias in estimating the effectiveness, which has to be handled by statistical adjustments and modeling techniques, or instrumental variables methods, or some combination. An obvious way to remove confounding is a randomized trial. However, conventional randomized trial designs are not only too costly but also ineffective in changing medical practice. An example is the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT), which was a randomized, double-blind, multi-center clinical trial designed to recruit 40,000 hypertensive patients to be randomized to a diuretic treatment (chlorthalidone) and three alternative antihypertensive pharmacologic treatments. Patients were followed every three months for the first year and every four months thereafter for an average of six years of follow-up. This large-scale CER trial cost over $100 million, and the results showed no difference in the prevention of heart attack and the superiority of chlorthalidone in preventing one or more forms of cardiovascular disease [1]. Yet, a few years later, the impact of the trial was found to be disappointing because of difficulty in pursuading doctors to change, scientific disagreement about the interpretation of the results, and heavy marketing by the pharmaceutical companies of their own drugs, as pointed by Lai and Lavori [38].

Section 4 of [38] describes some innovative approaches that are promising to meet the challenges of designing CER clinical trials. One is sequential multiple-assignment randomization for dynamic treatment strategies in the management of patients with chronic diseases, as in Thall et al. [67] who describe a two-stage randomized trial of twelve different strategies of first-line and second-line treatments of androgen-independent prostate cancer. Another is equipoise-stratified randomization as in the STAR*D trial [60] that compares seven treatment options in patients who did not attain a satisfactory response with citalopram, an inhibitor antidepressant. A third approach is to design point-of-care (POC) clinical trials which can be regarded as experiments embedded into clinical care, as described by Fiore et al. [26] for a Veterans Administration-sponsored trial that compares the effectiveness of two insulin dosing regimens for hospitalized diabetic patients. In POC trials, subjects are randomized at the health care encounter, clinician equipoise defines the reference population, and baseline and/or outcome data are captured through electronic medical records. By using outcome-adaptive randomization, POC trials integrate experimentation into implementation and learn sequentially the superior treatment(s). This is similar to the classical multi-arm bandit problem, a review of which is given in Section 6 of [37] except that POC uses adaptive randomization to implement it in a clinical setting with clinician equipoise and patient's informed consent. Moreover, recent advances in group sequential generalized likelihood ratio tests with efficient outcome-adaptive randomization to multiple arms, which will be reviewed in the penultimate paragraph of this section, are particularly suited to these POC trials.

Besides CER studies, Lai and Lavori [38] describe novel design methods for clinical studies of personalized treatments and targeted cancer therapies in translational medicine. "From bench to bedside" is a maxim of translational medicine. "Bench" refers to laboratory experiments to study new biochemical principles and discover novel treatments. The experiments with promising results are followed by pre-clinical animal studies. After understanding the effect of the treatment on animals (e.g., rodents), the next stage of drug development consists of clinical trials involving human subjects, starting with Phase I studies to determine a safe dose or dosage regimen and/or to collect information on the pharmacokinetics and pharmacodynamics of the drug. Phase II trials are precursors of Phase III trials. Their goal is to demonstrate the effectiveness of the drug for its approval by the regulatory agency (FDA) and to provide adequate evidence for its labeling. The evidence of effectiveness culminates in the regulatory approval of the new drug that can be used at the "bedside," that is, to treat patients. Phase III trials also collect safety information from the relatively large samples of patients accrued to the trial. Safety of the drug is evaluated from data obtained from all three phases of clinical trials prior to marketing approval of the drug, and continues to be evaluated through post-marketing Phase IV trials. Despite the sequential nature of Phase I–III trials, trials are often planned separately, treating each trial as an independent study whose design depends on results from studies in previous phases. The need for innovative study designs to better integrate the phases is now widely recognized, as are statistical methodologies for (a) group sequential and adaptive designs of clinical trials and (b) analysis of data from these trials. The monograph by Bartroff et al. [8] gives an overview of the existing methodologies and describes some major advances in the past decade. In particular, since the sample sizes of the trials are often inadequate because of separate planning, an alternative strategy is to expand a trial seamlessly from one phase into the next phase, as in the Phase II-III cancer trial designs of Inoue et al. [31] and Lai et al. [39].

We now return to personalized treatments and targeted cancer therapies mentioned at the beginning of the

preceding paragraph. The development of imatinib, the first drug to target the genetic defects of a particular cancer while leaving healthy cells unharmed, exemplifies this maxim and has revolutionized the treatment of cancer. A Phase I clinical trial treating CML (chronic myeloid leukemia) patients with the drug began in June 1998, and within six months remissions had occurred in all patients as determined by their white blood cell counts returning to normal. In a subsequent five-year study on survival, which followed 553 CML patients who had received imatinib as their primary therapy, only 5% of the patients died from CML and 11% died from all causes during the five-year period. Moreover, there were few significant side effects; see [19]. Such remarkable success of targeted therapies has led to hundreds of kinase inhibitors and other targeted drugs that are in various stages of development in the present anticancer drug pipeline. Most new targeted treatments, however, have resulted in only modest clinical benefit, with less than 50% remission rates and less than one year of progression-free survival, unlike a few cases such as trastuzumab in HER2-positive breast cancer, imatinib in CML, and gefitinib and erlotinib in non-small cell lung cancer. While the targeted treatments are devised to attack specific targets, the "one size fits all" treatment regimens commonly used may have diminished their effectiveness, and genomic-guided and risk-adapted personalized therapies that are tailored for individual patients are expected to substantially improve the effectiveness of these treatments. To achieve this potential for personalized therapies, the first step is to identify and measure the relevant biomarkers. The markers can be individual genes or proteins or gene expression signatures. The next step is to select drugs (standard cytotoxins, monoclonal antibodies, kinase inhibitors and other targeted drugs) based on the genetics of the disease in individual patients and biomarkers of drug sensitivity and resistance. The third step is to design clinical trials to provide data for the development and verification of personalized therapies. This is an active area of research and several important developments are reviewed in [40].

In particular, whereas traditional designs require large sample sizes and cannot adapt to evolving knowledge about biomarkers, an editorial in *Nature* (April 2010, vol. 464), in which two clinical trials of biomarker-guided personalized therapies are reported, highlights the new possibilities offered by innovative clinical trial design "which allows researchers to avoid being locked into a single, static protocol of the trial," and can "yield breakthroughs, but must be handled with care" to ensure that "the risk of reaching a false positive conclusion" is not inflated. These two clinical trials, BATTLE [74, 52] and I-SPY2 [7], use Bayesian adaptive randomization designs and perform Bayesian inference from the posterior distributions, although the frequentist operating characteristics of the Bayes tests are also assessed by simulation studies under certain assumed models. In the same issue of the journal, Ledford [50] comments on these Bayesian designs: "The approach has been controversial, but is catching on

with both researchers and regulators as companies struggle to combat the nearly 50% failure rate of (cancer) drugs in large, late-stage trials." Since the posterior distributions can be defined irrespective of whether the observations are generated adaptively or by independent sampling from some population, Bayesian inference can be carried out in the same way for sequential/adaptive samples as in the conventional fixed samples. The controversy lies in whether the frequentist type I error rate is inflated and in the assumptions of prior distributions and parametric models in the Bayesian approach. Moreover, the complexity of the design requires computationally intensive Markov chain Monte Carlo methods to implement the Bayesian approach.

The results of the BATTLE trial are reported by Kim et al. [35]. Despite applying the Bayesian approach to adaptive randomization, "standard statistical methods (used in the Results section) included the Fisher's exact test for contingency tables and log-rank test for survival data" together with standard confidence intervals based on normal approximations, without adjustments for Bayesian adaptive randomization (AR) and possible treatment suspension, even though Zhou et al. [74] noted earlier that "one known ramification of the AR design is that it results in biased estimates due to dependent samples." The overall 8-week disease control rate (DCR) using the biomarker-guided AR scheme was 46%, compared to "the historical 30% DCR estimate in similar patients," showing that the "learn-as-we-go" approach in Bayesian AR can indeed "leverage accumulating patient data to improve the treatment outcome" by "allowing more patients to be assigned to more effective therapies and fewer patients to be assigned to less effective therapies;" see [35, pp. 46–48, 52]. This shares the philosophy of the classical *multi-arm bandit problem* mentioned in the second paragraph of this section. Suppose there are K treatments of unknown efficacy to be chosen sequentially to treat a large class of n patients. How should we allocate the treatments to maximize the mean treatment effect? Lai and Robbins [42] and Lai [36] consider the problem in the setting where the treatment effect has a density function $f(x; \theta_k)$ for the kth treatment, where the θ_k are unknown parameters. There is an apparent dilemma between the need to learn the unknown parameters and the objective of allocating patients to the best treatment to maximize the total treatment effect $S_n = X_1 + \cdots + X_n$ for the n patients. If the θ_k were known, then the optimal rule would use the treatment with parameter $\theta^* = \arg\max_{1 \leq k \leq K} \mu(\theta_k)$, where $\mu(\theta) = \mathbb{E}_\theta(X)$. In ignorance of θ_k, Lai and Robbins [42] define the *regret* of an allocation rule by

$$
\begin{aligned}
R_n(\theta_1, \ldots, \theta_K) &= n\mu(\theta^*) - \mathbb{E}_{\theta_1,\ldots,\theta_K}(S_n) \\
&= \sum_{k:\mu(\theta_k)<\mu(\theta^*)} (\mu(\theta^*) - \mu(\theta_k)) \mathbb{E}_{\theta_1,\ldots,\theta_K} T_n(k),
\end{aligned}
$$

where $T_n(k)$ is the number of patients receiving treatment k. They show that adaptive allocation rules can be constructed to attain the asymptotically minimal order of

$\log n$ for the regret, in contrast to the regret of order n for the traditional equal randomization rule that assigns patients to each treatment with equal probability $1/K$. A subsequent refinement by Lai [36] shows the relatively simple rule that chooses the treatment with the largest upper confidence bound $U_k^{(n)}$ for θ_k to be asymptotically optimal. The upper confidence bound at stage n, with $n > k$, is defined by

$$U_k^{(n)} = \inf \left\{ \theta \in A : \theta \geq \hat{\theta}_k \text{ and } 2T_n(k)I\left(\hat{\theta}_k, \theta\right) \geq h^2\left(T_n(k)/n\right) \right\},$$

where $\inf \emptyset = \infty$, A is some open interval known to contain θ, $\hat{\theta}_k$ is the maximum likelihood estimate of θ_k, $I(\theta, \lambda)$ is the Kullback–Leibler information number, and the function h has a closed-form approximation. For the first K stages, the K treatments are assigned successively. It is noted in [11, p. 97] that the upper confidence bound $U_k^{(n)}$ corresponds to inverting a generalized likelihood ratio (GLR) test based on the GLR statistic $T_n(k)I(\hat{\theta}_k, \theta)$ for testing $\theta_k = \theta$.

The multi-arm bandit problem has the same "learn-as-we-go" spirit of the BATTLE trial and focuses on attaining the best response rate for patients in the trial. However, such a trial does not establish, with a guaranteed probability of correct selection, which treatment is the best for future patients. Lai et al. [41] have recently developed a group sequential design for jointly developing and testing treatment recommendations for biomarker classes, while using multi-armed bandit ideas to provide sequentially optimizing treatments to patients in the trial. Thus, the design has to fulfill multiple objectives, which include (a) treating accrued patients with the best (yet unknown) available treatment, (b) developing a treatment strategy for future patients, and (c) demonstrating that the strategy developed indeed has better treatment effect than the historical mean effect of standard of care plus a predetermined margin. In a group sequential trial, sequential decisions are made only at times of interim analysis. Let n_i denote the total sample size up to the time of the ith analysis, $i = 1, \ldots, I$, so that n_I is the total sample size by the scheduled end of the trial, and let n_{ij} be the total sample size from biomarker class j up to the time of the ith analysis, hence $n_i = \sum_{j=1}^J n_{ij}$. Because of the need for informed consent, the treatment allocation that uses the aforementioned upper confidence bound rule is no longer appropriate. It is unlikely for patients to consent to being assigned to a seemingly inferior treatment for the sake of collecting more information to ensure that it is significantly inferior (as measured by the upper confidence bounds). Instead, randomization in a double blind setting is required, and the randomization probability $\pi_{jk}^{(i)}$, determined at the ith interim analysis, of assigning a patient in group j to treatment k cannot be too small to suggest obvious inferiority of the treatments being tried, that is, $\pi_{jk}^{(i)} \geq \epsilon$ for some $0 < \epsilon < 1/K$. The unknown mean treatment effect μ_{jk} of treatment k in biomarker class j can be estimated by the sample mean $\hat{\mu}_{ijk}$ at interim analysis i. Let $k_j = \arg\max_k \mu_{jk}$,

which can be estimated by $\hat{k}_{ij} = \arg\max_k \hat{\mu}_{ijk}$ at the ith interim analysis. Analogy with multi-arm bandit theory suggests assigning the highest randomization probability to treatment $\hat{k}_{ij}$ and randomizing to the other available treatments in biomarker class j with probability ϵ. Because the randomization probabilities are only updated at interim analyses in a group sequential design and because $\hat{k}_{ij}$ may fluctuate over i among treatments whose treatment effects do not differ by more than δ_{ij}, it is more stable to lump these "nearby" treatments into the set $\mathcal{H}_{ij} = \{k \in \mathcal{K}_{ij} : |\hat{\mu}_{ij}^* - \hat{\mu}_{ijk}| \leq \delta_{ij}\}$, where $\hat{\mu}_{ij}^* = \hat{\mu}_{ij\hat{k}_{ij}}$ and $\mathcal{K}_{ij}$ is the set of available treatments in biomarker class j at interim analysis i. The randomization probabilities $\pi_{jk}^{(i)}$ in [41] are determined at the ith interim analysis by

$$\pi_{jk}^{(i)} = \epsilon \text{ for } k \in \mathcal{K}_{ij} \backslash \mathcal{H}_{ij},$$
$$\pi_{jk}^{(i)} = (1 - |\mathcal{K}_{ij} \backslash \mathcal{H}_{ij}|\epsilon)/|\mathcal{H}_{ij}| \text{ for } k \in \mathcal{H}_{ij},$$

where $|A|$ denotes the number of elements of a finite set A. Equal randomization is used up to the first interim analysis. Besides aiming at treating patients in the trial with the best available treatment, the group sequential design can also be used to address testing and inference questions, with guaranteed error probabilities, that are of basic interest to personalized treatment selection for future patients based on their biomarkers. Lai et al. [41] use GLR statistics and modified Haybittle–Peto stopping rules introduced by Lai and Shih [43] to include early elimination of significantly inferior treatments from a biomarker class.

In the case of approved drugs, pharmaceutical companies would not sponsor clinical trials for developing and testing biomarker-guided personalized treatment selection strategies. Funding for such trials can come from private foundations and government agencies as in the case of the BATTLE trial, or from the Patient-Centered Outcomes Research Institute, established after the 2010 Patient Protection and Affordable Care Act to undertake comparative effectiveness research (CER). Fiore et al. [26] and Shih and Lavori [62] have recently proposed to use (a) the infrastructure of clinical experiments in natural clinical settings, such as POC (point of care) clinical trials, and (b) group sequential designs to conduct CER trials more easily and at a much lower cost than the traditional randomized clinical trial approach. The innovative designs introduced in [41] are a continuation of that line of work, incorporating biomarkers into CER for personalized treatment selection.

Statistics and Data Science in a New Era for Financial Markets

We begin this section with a review of the 2007–2008 financial crisis and the 2010 Dodd-Frank Act. The financial crisis was associated with unpreparedly high default rates of subprime mortgage loans in 2007 and culminated in the collapse of large financial institutions such as Bear Stearns and Lehman Brothers in 2008. Parallel to the increasing volume of subprime mortgage loans whose value was es-

timated to be \$1.3 trillion by March 2007, an important development in financial markets from 2000 to 2007 was the rapid growth of credit derivatives, culminating in \$32 trillion worth of notional principal for outstanding credit derivatives by December 2009. These derivative contracts are used to hedge against credit loss of either a single corporate bond, as in a "credit default swap" (CDS), or a portfolio of corporate bonds, as in a cash CDO ("collateralized debt obligation"), or variant thereof called synthetic CDO. In July 2007, Bear Stearns disclosed that two of its subprime hedge funds which were invested in CDOs had lost nearly all their value following a rapid decline in the subprime mortgage market, and Standard & Poor's (S&P) downgraded the company's credit rating. In March 2008, the Federal Reserve Bank of New York initially agreed to provide a \$25 billion collateralized 28-day loan to Bear Stearns, but subsequently changed the deal to make a \$30 billion loan to JPMorgan Chase to purchase Bear Stearns. Lehman Brothers also suffered unprecedented losses for its large positions in subprime and other lower-rated mortgage-backed securities in 2008. After attempts to sell it to Korea Development Bank and then to Bank of America and to Barclays failed, it filed for Chapter 11 bankruptcy protection on September 15, 2008, making the largest bankruptcy filing, with over \$600 billion in assets, in US history. A day after Lehman's collapse, American International Group (AIG) needed bailout by the Federal Reserve Bank, which gave the insurance company a secured credit facility of up to \$85 billion to enable it to meet collateral obligations after its credit ratings were downgraded below AA, in exchange for a stock warrant for 79.9% of its equity. AIG's London unit had sold credit protection in the form of CDS and CDO to insure \$44 billion worth of securities originally rated AAA. As Lehman's stock price was plummeting, investors found that AIG had valued its subprime mortgage-backed securities at 1.7 to 2 times the values used by Lehman and lost confidence in AIG. Its share prices had fallen over 95% by September 16, 2008. The "contagion" phenomenon, from increased default probabilities of subprime mortgages to those of counterparties in credit derivative contracts whose values vary with credit ratings, was mostly neglected in the models of joint default intensities that were used to price CDOs and mortgage-backed securities. These models also failed to predict well the "frailty" traits of latent macroeconomic variables that underlie mortgages and mortgage-backed securities.

For a multiname credit derivative such as CDO involving k firms, it is important to model not only the individual default intensity process but also the joint distribution of these processes. Finding tractable models that can capture the key features of the interrelationships of the firms' default intensities has been an active area of research since intensity-based (also called reduced-form) models have become a standard approach to pricing the default risk of a corporate bond; see [21] and [49]. Let Φ denote the standard normal distribution function, and let G_i be the distribution function of the default time τ_i for the ith firm, $1 \le i \le M$. Then $Z_i = \Phi^{-1}(G_i(\tau_i))$ is standard normal. A widely used tool to price CDOs and other multi-name credit derivatives was the Gaussian copula model, which assumes that $(Z_1, \ldots, Z_M)$ is multivariate normal and specifies its correlation matrix Γ by using the correlations of the stock returns of the M firms. Because it provides a simple way to model default correlations, it became popular despite the lack of convincing argument to connect the stock return correlations to the correlations of the normally distributed transformed default times. In a commentary on "the biggest financial meltdown since the Great Depression," Salmon [61] mentioned that the Gaussian copula approach, which "looked like an unambiguously positive breakthrough", was used uncritically by "everybody from bond investors and Wall Street banks to rating agencies and regulators" and "became so deeply entrenched — and was making people so much money — that warnings about its limitations were largely ignored."

One of the provisions of the Dodd-Frank Act is the establishment of the Financial Stability Oversight Council (FSOC) and the Office of Financial Research (OFR, under the US Treasury Department). Their tasks are to monitor systemic risk, to identify risks to the financial stability of the US from both financial and non-financial organizations, to conduct supporting research that includes data collection from bank holding companies and non-bank financial companies and development of tools for monitoring risk, and to promote market discipline and maintain investor confidence. The OFR has a Data Center and a Research and Analysis Center, and provides technical support for FSOC that makes general regulatory recommendations and advises Congress on financial stability. The Dodd-Frank Act also has provisions for credit derivatives, asset-backed securities, and insurance, with the Federal Insurance Office established within the Department of the Treasury to monitor all aspects of the insurance industry (except health insurance that is reformed by the 2010 Affordable Health Care Act).

In the wake of the financial crisis, it was recognized that better albeit less tractable models of correlated default intensities are needed for pricing CDOs and risk management of credit portfolios. It was also recognized that such models should include relevant firm-level and macroeconomic variables for default prediction and also incorporate frailty and contagion. The monograph by Lai and Xing [48] reviews recent works on dynamic frailty and contagion models in the finance literature and describes a new approach involving dynamic empirical Bayes and generalized linear mixed models (GLMM), which have been shown by Lai et al. [44] to compare favorably with the considerably more complicated hidden Markov models for the latent frailty processes or the additive intensity models for contagion. The empirical Bayes (EB) methodology, introduced by Robbins [58] and Stein [63], considers n independent and structurally similar problems of inference on the parameters θ_i from observed data Y_i $(i = 1, \ldots, n)$, where Y_i has probability density $f(y|\theta_i)$. The θ_i are assumed to have a common prior distribution G that has unspecified hyperparameters. Letting $d_G(y)$

be the Bayes decision rule (with respect to some loss function and assuming known hyperparameters) when $Y_i = y$ is observed, the basic principle underlying EB is that a parametric form of G (as in [63]) or even G itself (as in [58]) can be consistently estimated from $Y_1, \ldots, Y_n$, leading to the EB rule $d_{\hat{G}}$. Dynamic EB extends this idea to longitudinal data Y_{it}; see [44]. In the context of insurance claims over time for n contracts belonging to the same risk class, the conventional approach to insurance rate-making (called "evolutionary credibility" in actuarial science) assumes a linear state-space for the longitudinal claims data so that the Kalman filter can be used to estimate the claims' expected values, which are assumed to form an autoregressive time series. Applying the EB principle to the longitudinal claims from the n insurance contracts, Lai and Sun [45] have developed a class of linear mixed models as an alternative to linear state-space models for evolutionary credibility and have shown that the predictive performance is comparable to that of the Kalman filter when the claims are generated by a linear state-space model. This approach can be readily extended to GLMMs not only for longitudinal claims data but also for default probabilities of n firms, incorporating frailty, contagion, and regime switching. Details are given in [44] and [48].

The Dodd-Frank Act and recent financial reforms in the European Union and other countries have led to new financial regulations that enforce transparency and accountability and enhance consumer financial protection. The need for good and timely data for risk management and regulatory supervision is well recognized, but how to analyze these massive datasets and use them to give early warning and develop adaptive risk control strategies is a challenging statistical problem that requires domain knowledge and interdisciplinary collaboration. Closely related areas of research in statistical modeling and data science include sequential surveillance and early warning, particularly for systemic risk which is the risk of a broad-based breakdown in the financial system as experienced in the recent financial crisis; core-periphery network models for mathematical representation of the critical financial market infrastructure; transmission of risk and liquidity in these networks; representation and standardization of financial data and non-numeric information such as contracts and regulations; algorithms and models for measuring threats to financial stability; storage and monitoring tools for financial networks; on-line detection, estimation, and stochastic adaptive control; and simulation methods for stress testing and risk analytics.

Statistics and data science are assuming an increasingly important role in financial markets after the financial crisis, which exposed the weakness and limitations of traditional financial models, pricing and hedging theories, risk measures and management. Technological advances have also dramatically changed the investment landscape. Electronic exchanges/venues have been developed to cover every asset class: equities, derivatives, fixed income securities, currencies, commodities, and exchange-traded funds (ETFs). Statistical methods and trading strategies that have been developed for daily open-close/high-low data as described in [47] have to be modified to model and analyze the tick-by-tick transactions data. The forthcoming monograph by Lai et al. [46] on algorithmic trading and quantitative strategies describes these modifications, together with developments in data science and statistical methodologies that can be used to address the challenges in modeling, analysis, computation, decision, and execution in this new era for financial markets.

Conclusion

There are some common threads linking data science and statistical modeling in finance and health care for the new era. One is related to "big data" for regulatory supervision, risk management and algorithmic trading, and for emerging health care systems that involve electronic medical records, genomic and proteomic biomarkers, and computer-assisted support for patient care. Another is related to the need for collaborative research that can integrate the methodological research with domain knowledge and subject-matter issues. A third thread is related to innovations and major advances in statistical modeling that are inspired by, and in turn are applicable to, the challenges in this new era for finance and health care.

References

[1] ALLHAT Collaborative Research Group (2002). Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *J. Amer. Med. Assoc.* 288: 2981–2997.

[2] Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* 34: 122–148.

[3] Bai, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica* 9: 611–677, with comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.

[4] Bai, Z. D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.* 26: 316–345.

[5] Bai, Z. D. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices.* New York: Springer, 2nd ed.

[6] Baik, J., Deift, P. and Johansson, K. (1999). On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.* 12: 1119–1178.

[7] Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry, D. A. and Esserman, L. J. (2009). I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin. Pharmacol. Ther.* 86: 97–100.

[8] Bartroff, J., Lai, T. L. and Shih, M.-C. (2012). *Sequential Experimentation in Clinical Trials: Design and Analysis.* New York: Springer.

[9] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* 36: 199–227.

[10] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* 37: 1705–1732.

[11] Brezzi, M. and Lai, T. L. (2002). Optimal learning and experimentation in bandit problems. *J. Econ. Dyn. Control* 27: 87–108.

[12] Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* 34: 559–583.

[13] Candés, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* 35: 2313–2351.

[14] Candés, E. J. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* 37: 2145–2177.

[15] Candés, E. J. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory* 52: 5406–5425.

[16] Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20: 33–61.

[17] Cleveland, W. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *Int. Statist. Rev.* 69: 21–26.

[18] Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* 52: 1289–1306.

[19] Druker, B. J., Guilhot, F., O'Brien, S. G., Gathmann, I., Kantarjian, H., Gattermann, N., Deininger, M. W. N., Silver, R. T., Goldman, J. M., Stone, R. M., Cervantes, F., Hochhaus, A., Powell, B. L., Gabrilove, J. L., Rousselot, P., Reiffers, J., Cornelissen, J. J., Hughes, T., Agis, H., Fischer, T., Verhoef, G., Shepherd, J., Saglio, G., Gratwohl, A., Nielsen, J. L., Radich, J. P., Simonsson, B., Taylor, K., Baccarani, M., So, C., Letvak, L., Larson, R. A. and IRIS Investigators (2006). Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *N. Engl. J. Med.* 355: 2408–2417.

[20] Dudoit, S. and van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics.* New York: Springer.

[21] Duffie, D. and Singleton, K. J. (2003). *Credit Risk: Pricing, Measurement, and Management.* Princeton University Press.

[22] Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge: Cambridge University Press.

[23] El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* 36: 2757–2790.

[24] Erdös, L., Péché, S., Ramírez, J. A., Schlein, B. and Yau, H.-T. (2010). Bulk universality for Wigner matrices. *Comm. Pure Appl. Math.* 63: 895–925.

[25] Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* 147: 186–197.

[26] Fiore, L. D., Brophy, M., Ferguson, R. E., D'Avolio, L., Hermos, J. A., Lew, R. A., Doros, G., Conrad, C. H., O'Neil, J. A. ., Sabin, T. P., Kaufman, J., Swartz, S. L., Lawler, E., Liang, M. H., Gaziano, J. M. and Lavori, P. W. (2011). A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. *Clin. Trials* 8: 183–195.

[27] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* 55: 119–139.

[28] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 29: 1189–1232.

[29] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer, 2nd ed.

[30] Ing, C.-K. and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica* 21: 1473–1513.

[31] Inoue, L. Y. T., Thall, P. F. and Berry, D. A. (2002). Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* 58: 823–831.

[32] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* 29: 295–327.

[33] Johnstone, I. M. (2009). Approximate null distribution of the largest root in multivariate analysis. *Ann. Appl. Stat.* 3: 1616–1633.

[34] Jordan, M. I. (2004). Graphical models. *Statist. Sci.* 19: 140–155.

[35] Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J.,

Gupta, S. et al. (2011). The BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discov.* 1: 44–53.

[36] Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* 15: 1091–1114.

[37] Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statist. Sinica* 11: 303–408.

[38] Lai, T. L. and Lavori, P. W. (2011). Innovative clinical trial designs: Toward a 21st-century health care system. *Stat. Biosci.* 3: 145–168.

[39] Lai, T. L., Lavori, P. W. and Shih, M.-C. (2012). Sequential design of Phase II-III cancer trials. *Stat. Med.* 31: 1944–1960.

[40] Lai, T. L., Lavori, P. W., Shih, M.-C. and Sikic, B. I. (2012). Clinical trial designs for testing biomarker-based personalized therapies. *Clin. Trials* 9: 141–154.

[41] Lai, T. L., Liao, O. Y.-W. and Kim, D. W. (2013). Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research, *Contemp. Clin. Trials* (to appear).

[42] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6: 4–22.

[43] Lai, T. L. and Shih, M.-C. (2004). Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika* 91: 507–528.

[44] Lai, T. L., Su, Y. and Sun, K. H. (2013). Dynamic empirical Bayes models and their applications to longitudinal data analysis and prediction, *Statist. Sinica* (to appear).

[45] Lai, T. L. and Sun, K. H. (2012). Evolutionary credibility theory: A generalized linear mixed modeling approach. *N. Amer. Actuarial J.* 16: 273–284.

[46] Lai, T. L., Velu, R. P. and Wong, S. P. (2014). *Algorithmic Trading and Quantitative Strategies.* Chapman and Hall/CRC (to appear).

[47] Lai, T. L. and Xing, H. (2008). *Statistical Models and Methods for Financial Markets.* New York: Springer.

[48] Lai, T. L. and Xing, H. (2014). *Active Risk Management: Financial Models and Statistical Methods.* Chapman & Hall/CRC (to appear).

[49] Lando, D. (2004). *Credit Risk Modeling: Theory and Applications.* Princeton University Press.

[50] Ledford, H. (2010). Clinical drug tests adapted for speed. *Nature* 464: 1258.

[51] Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* 40: 1024–1060.

[52] Lee, J., Gu, X. and Liu, S. (2010). Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials* 7, in press.

[53] Mallat, S. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* 41: 3397–3415.

[54] Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)* 72: 507–536.

[55] Naur, P. (1974). *Concise Survey of Computer Methods.* New York: Petrocelli/Charter.

[56] Okounkov, A. (2000). Random matrices and random permutations. *Internat. Math. Res. Notices*: 1043–1095.

[57] Rajaratnam, B., Massam, H. and Carvalho, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.* 36: 2818–2849.

[58] Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I.* Berkeley and Los Angeles: University of California Press, 157–163.

[59] Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* 39: 887–930.

[60] Rush, A., Fava, M., Wisniewski, S., Lavori, P., Trivedi, M., Sackeim, H., Thase, M., Nierenberg, A., Quitkin, F., Kashner, T., Kupfer, D., Rosenbaum, J., Alpert, J., Stewart, J., McGrath, P., Biggs, M., Shores-Wilson, K., Lebowitz, B., Ritz, L., Niederehe, G. and STAR D Investigators Group (2004). Sequenced treatment alternatives to relieve depression

(STAR*D): Rationale and design. *Contr. Clin. Trial.* 25: 119-142.

[61] Salmon, F. (2012). The formula that killed Wall Street. *Significance* 9: 16-20.

[62] Shih, M.-C. and Lavori, P. W. (2013). Sequential methods for comparative effectiveness experiments: Point of care clinical trials, *Statist. Sinica* (to appear).

[63] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954-1955, vol. I.* Berkeley and Los Angeles: University of California Press, 197-206.

[64] Stukel, T. A., Fisher, E. S., Wennberg, D. E., Alter, D. A., Gottlieb, D. J. and Vermeulen, M. J. (2007). Analysis of observational studies in the presence of treatment selection bias: Effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *J. Amer. Med. Assoc.* 297: 278-285.

[65] Tao, T. and Vu, V. (2010). Random matrices: Universality of local eigenvalue statistics up to the edge. *Comm. Math. Phys.* 298: 549-572.

[66] Temlyakov, V. N. (2000). Weak greedy algorithms. *Adv. Comput. Math.* 12: 213-227.

[67] Thall, P. F., Logothetis, C., Pagliaro, L. C., Wen, S., Brown, M. A., Williams, D. and Millikan, R. E. (2007). Adaptive therapy for androgen-independent prostate cancer: A randomized selection trial of four regimens. *J. Nat. Cancer Inst.* 99: 1613-1622.

[68] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58: 267-288.

[69] Tracy, C. A. and Widom, H. (1998). Correlation functions, cluster functions, and spacing distributions for random matrices. *J. Statist. Phys.* 92: 809-835.

[70] Tropp, J. A. and Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory* 53: 4655-4666.

[71] Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Ann. Math. (2)* 62: 548-564.

[72] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* 36: 1567-1594.

[73] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* 7: 2541-2563.

[74] Zhou, X., Liu, S., Kim, E. S., Herbst, R. S. and Lee, J. L. (2008). Bayesian adaptive design for targeted therapy development in lung cancer - A step toward personalized medicine. *Clin. Trials* 5: 181-193.

[75] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* 67: 301-320.