Chapter 2
A Survey of Bayesian Statistical Approaches for Big Data

Farzana Jahan, Insha Ullah, and Kerrie L. Mengersen

Abstract The modern era is characterised as an era of information or Big Data. This has motivated a huge literature on new methods for extracting information and insights from these data. A natural question is how these approaches differ from those that were available prior to the advent of Big Data. We present a survey of published studies that present Bayesian statistical approaches specifically for Big Data and discuss the reported and perceived benefits of these approaches. We conclude by addressing the question of whether focusing only on improving computational algorithms and infrastructure will be enough to face the challenges of Big Data.

Keywords Bayesian statistics · Bayesian modelling · Bayesian computation · Scalable algorithms

2.1 Introduction

Although there are many variations on the definition of Big Data [51, 52, 91, 184], it is clear that it encompasses large and often diverse quantitative data obtained from increasing numerous sources at different individual, spatial and temporal scales, and with different levels of quality. Examples of Big Data include data generated from social media [22]; data collected in biomedical and healthcare informatics research such as DNA sequences and electronic health records [114]; geospatial data generated by remote sensing, laser scanning, mobile mapping, geo-located sensors, geo-tagged web contents, volunteered geographic information (VGI), global navigation satellite system (GNSS) tracking and so on [103]. The volume and complexity of Big Data often exceeds the capability of the

F. Jahan (✉) · I. Ullah · K. L. Mengersen
School of Mathematical Sciences, ARC Centre of Mathematical and Statistical Frontiers, Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD, Australia
e-mail: f.jahan@hdr.qut.edu.au

K. L. Mengersen et al. (eds.), *Case Studies in Applied Bayesian Data Science*,
Lecture Notes in Mathematics 2259, https://doi.org/10.1007/978-3-030-42553-1_2

standard analytics tools (software, hardware, methods and algorithms) [70, 92]. The concomitant challenges of managing, modelling, analysing and interpreting these data have motivated a large literature on potential solutions from a range of domains including statistics, machine learning and computer science. This literature can be grouped into four broad categories of articles. The first includes general articles about the concept of Big Data, including the features and challenges, and their application and importance in specific fields. The second includes literature concentrating on infrastructure and management, including parallel computing and specialised software. The third focuses on statistical and machine learning models and algorithms for Big Data. The final category includes articles on the application of these new techniques to complex real-world problems.

In this chapter, we classify the literature published on Big Data into finer classes than the four broad categories mentioned earlier and briefly reviewed the contents covered by those different categories. But the main focus of the chapter is around the third category, in particular on statistical contributions to Big Data. We examine the nature of these innovations and attempt to catalogue them as modelling, algorithmic or other contributions. We then drill further into this set and examine the more specific literature on Bayesian approaches. Although there is an increasing interest in this paradigm from a wide range of perspectives including statistics, machine learning, information science, computer science and the various application areas, to our knowledge there has not yet been a survey of Bayesian statistical approaches for Big Data. This is the primary contribution of this chapter.

This chapter provides a survey of the published studies that present Bayesian statistical models specifically for Big Data and discusses the reported and perceived benefits of these approaches. We conclude by addressing the question of whether focusing only on improving computational algorithms and infrastructure will be enough to face the challenges of Big Data.

The chapter proceeds as follows. In the next section, literature search and inclusion criteria for this chapter is outlined. A classification of Big Data literature along with brief survey of relevant literature in each class is presented in Sect. 2.3. Section 2.4 consists of a brief survey of articles discussing Big Data problems from statistical perspectives, followed by a survey of Bayesian approaches applied to Big Data. The final section includes a discussion of this survey with a view to answering the research question posed above.

2.2 Literature Search and Inclusion Criteria

The literature search for this survey paper was undertaken using different methods. The search methods implemented to find the relevant literature and the criteria for the inclusion of the literature in this chapter are briefly discussed in this section.

2.2.1 Inclusion Criteria

Acknowledging the fact that there has been a wide range of literature on Big Data, the specific focus in this chapter was on recent developments published in the last 5 years, 2013–2019.

For quality assurance reasons, of the literature only peer reviewed published articles, book chapters and conference proceedings were included in the chapter. Some articles were also included from arXiv and pre-print versions for those to be soon published and from well known researchers working in that particular area of interest.

2.2.2 Search Methods

Database Search The database "Scopus" was used to initiate the literature search. To identify the availability of literature and broadly learn about the broad areas of concentration, the following keywords were used: Big Data, Big Data Analysis, Big Data Analytics, Statistics and Big Data.

The huge range of literature obtained by this initial search was complemented by a search of "Google Scholar" using more specific key words as follows: Features and Challenges of Big Data, Big Data Infrastructure, Big Data and Machine Learning, Big Data and Cloud Computing, Statistical approaches/methods/models in Big Data, Bayesian Approaches/Methods/Models in Big Data, Big Data analysis using Bayesian Statistics, Bayesian Big Data, Bayesian Statistics and Big Data.

Expert Knowledge In addition to the literature found by the above Database search, we used expert knowledge and opinions in the field and reviewed the works of well known researchers in the field of Bayesian Statistics for their research works related to Bayesian approaches to Big Data and included the relevant publications for survey in this chapter.

Scanning References of Selected Literature Further studies and literature were found by searching the references of selected literature.

Searching with Specific Keywords Since the focus of this chapter is to survey the Bayesian approaches to Big Data, more literature was sourced by using specific Bayesian methods or approaches found to be applied to Big Data: Approximate Bayesian Computation and Big Data, Bayesian Networks in Big Data, Classification and regression trees/Bayesian Additive regression trees in Big Data, Naive Bayes Classifiers and Big Data, Sequential Monte Carlo and Big Data, Hamiltonian Monte Carlo and Big Data, Variational Bayes and Big Data, Bayesian Empirical Likelihood and Big Data, Bayesian Spatial modelling and Big Data, Non parametric Bayes and Big Data.

This last step was conducted in order to ensure that this chapter covers the important and emerging areas of Bayesian Statistics and their application to Big Data. These searches were conducted in "Google Scholar" and up to 30 pages of results were considered in order to find relevant literature.

2.3 Classification of Big Data Literature

The published articles on Big Data can be divided into finer classes than the four main categories described above. Of course, there are many ways to make these delineations. Table 2.1 shows one such delineation, with representative references from the last 5 years of published literature. The aim of this table is to indicate the wide ranging literature on Big Data and provide relevant references in different categories for interested readers.

The links between these classes of literature can be visualised as in Fig 2.1 and a brief description of each of the classes and the contents covered by the relevant references listed are provided in Table 2.2. The brief surveys presented in Table 2.2 can be helpful for interested readers to develop a broad idea about each of the classes mentioned in Table 2.1. However, Table 2.2 does not include brief surveys of the last two classes, namely, Statistical Methods and Bayesian Methods, since these classes are discussed in detail in Sects. 2.4 and 2.5. We would like to acknowledge the fact that Bayesian methods are essentially part of statistical methods, but in this chapter, the distinct classes are made intentionally to be able to identify and discuss the specific developments in Bayesian approaches.

Table 2.1 Classes of big data literature

Topic	Representative references
Features and challenges	[51, 52, 63, 65, 70, 140, 160, 168, 184, 200]
Infrastructure	[10, 50, 96, 108, 117, 132, 137, 142, 165, 183, 193, 206, 207]
Cloud computing	[11, 32, 38, 54, 113, 120, 130, 139, 148, 175, 201, 205]
Applications (3 examples)	Social science: [5, 22, 37, 39, 121, 164]
	Health/medicine/medical science: [8, 9, 16, 19, 21, 28, 34, 46, 82, 118, 153, 158, 159, 182, 202]
	Business: [2, 31, 36, 60, 66, 122, 154, 172]
Machine learning methods	[3, 4, 26, 27, 55, 64, 89, 97, 138, 173]
Statistical methods	[44, 45, 58, 61, 67, 84, 86, 87, 111, 136, 143, 162, 174, 188, 191, 192, 194, 198, 204]
Bayesian methods	[7, 80, 81, 100, 102, 105, 109, 110, 115, 128, 129, 151, 163, 170, 180, 199, 210, 211]

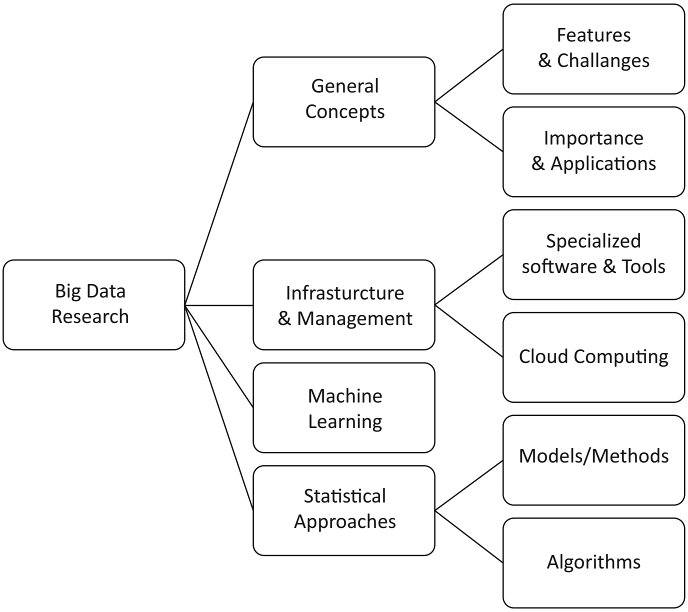

Fig. 2.1 Classification of big data literature

2.4 Statistical Approaches to Big Data

The importance of modelling and theoretical considerations for analysing Big Data are well stated in the literature [86, 198]. These authors pointed out that blind trust in algorithms without proper theoretical considerations will not result in valid outputs. The emerging challenges of Big Data are beyond the issues of processing, storing and management. The choice of suitable statistical methods is crucial in order to make the most of the Big Data [67, 87]. Dunson [61] highlighted the role of statistical methods for interpretability, uncertainty quantification, reducing selection bias in analysing Big Data.

In this section we present a brief survey of some of the published research on statistical perspectives, methods, models and algorithms that are targeted to Big Data. As above, the survey is confined to the last 5 years, commencing with the most recent contributions. Bayesian approaches are reserved for the next section.

Among the brief surveys of the relevant literature in Table 2.3, we include detailed surveys of three papers which are more generic in explaining the role of statistics and statistical methods in Big Data along with recent developments in this area.

Wang et al. [191] summarised the published literature on recent methodological developments for Big Data in three broad groups: subsampling, which calculates a statistic in many subsamples taken from the data and then combining the results [144]; divide and conquer, the principle of which is to break a dataset into

Table 2.2 Brief survey of relevant literature under identified classes

Features and challenges

- The general features of Big Data are volume, variety, velocity, veracity, value [52, 160] and some salient features include massive sample sizes and high dimensionality [160].
- Many challenges of Big Data regarding storage, processing, analysis and privacy are identified in the literature [52, 63, 65, 140, 160].

Infrastructure

- To manage and analyse Big Data, infrastructural support is needed such as sufficient storage technologies and data management systems. These are being continuously developed and improved. MangoDB, Terrastore and RethhinkDb are some examples of the storage technologies; more on evolution technologies with their strengths, weaknesses, opportunities and threats are available in [165].
- To analyse Big Data, parallel processing systems and scalable algorithms are needed. MapReduce is one of the pioneering data processing systems [206]. Some other useful and popular tools to handle Big Data are Apache, Hadoop, Spark [10].

Cloud computing

- Cloud computing, the practice of using a network of remote servers hosted on the Internet rather than a local server or a personal computer, plays a key role in Big Data analysis by providing required infrastructure needed to store, analyse, visualise and model Big Data using scalable and adaptive systems [11].
- Opportunities and challenges of cloud computing technologies, future trends and application areas are widely discussed in the literature [32, 175, 201] and new developments on cloud computing are proposed to overcome known challenges, such as collaborative anomaly detection [130], hybrid approach for scalable sub-tree anonymisation using MapReduce on cloud [205] etc.

Applications (3 examples)

- Big Data has made it possible to analyse social behaviour and an individual's interactions with social systems based on social media usage [5, 37, 164]. Discussions on challenges and future of social science research using Big Data have been made in the literature [39, 164].
- Research involving Big Data in medicine, public health, biomedical and health informatics has increased exponentially over the last decade [19, 28, 46, 114, 153, 158]. Some examples include infectious disease research [16, 82], developing personalised medicine and health care [9, 182] and improving cardiovascular care [159].
- Analysis of Big Data is used to solve many real world problems in business, in particular, using Big Data analytics for innovations in leading organisations [122], predictive analytics in retail [31], analysis of business risks and benefits [154], development of market strategies [60] and so on. The opportunities and challenges of Big Data in e-commerce and Big Data integration in business processes can be found in the survey articles by Akter and Wamba [2] and Wamba et al. [184].

(continued)

Table 2.2 (continued)

Machine learning methods
• Machine learning is an interdisciplinary field of research primarily focusing on theory, performance, properties of learning systems and algorithms [149]. Traditional machine learning is evolving to tackle the additional challenges of Big Data [4, 149]. • Some examples of developments in machine learning theories and algorithms for Big Data include high performance machine learning toolbox [3], scalable machine learning online services for Big Data real time analysis [14]. • There is a large and increasing research on specific applications of machine learning tools for Big Data in different disciplines. For example, [138] discussed the future of Big Data and machine learning in clinical medicine; [13] discussed a classifier specifically for medical Big Data and [26] reviewed the state of art and future prospects of machine learning and Big Data in radiation oncology.

smaller subsets to analyse these in parallel and combine the results at the end [169]; and online updating of streaming data [162], based on online recursive analytical processing. He summarised the following methods in the first two groups: subsampling based methods (bag of little bootstraps, leveraging, mean log likelihood, subsample based MCMC), divide and conquer (aggregated estimating equations, majority voting, screening with ultra high dimension, parallel MCMC). The authors, after reviewing existing online updating methods and algorithms, extended the online updating of stream data method by including criterion based variable selection with online updating. The authors also discussed the available software packages (open source R as well as commercial software) developed to handle computational complexity involving Big Data. For breaking the memory barrier using R, the authors cited and discussed several data management packages (sqldf, DBI, RSQLite, filehash, bigmemory, ff) and packages for numerical calculation (speedglm, biglm, biganalytics, ffbase, bigtabulate, bigalgebra, bigpca, bigrf, biglars, PopGenome). The R packages for breaking computing power were cited and discussed in two groups: packages for speeding up (compiler, inline, Rcpp, RcpEigen, RcppArmadilo, RInside, microbenchmark, proftools, aprof, lineprof, GUIprofiler) and packages for scaling up (Rmpi, snow, snowFT, snowfall, multicore, parallel, foreach, Rdsm, bigmemory, pdpMPI, pbdSLAP, pbdBASE, pbdMAT, pbdDEMO, Rhipe, segue, rhbase, rhdfs, rmr, plymr, ravroSparkR, pnmath, pnmath0, rsprng, rlecuyer, doRNG, gputools, bigvis). The authors also discussed the developments in Hadoop, Spark, OpenMP, API and using FORTRAN and C++ from R in order to create flexible programs for handling Big Data. The article also presented a brief summary about the commercial statistical software, e.g., SAS, SPSS, MATLAB. The study included a case study of fitting a logistic model to a massive data set on airline on-time performance data from the 2009 ASA Data Expo mentioning the use of some R packages discussed earlier to handle the problem with memory and computational capacity. Overall, this study provided a comprehensive

Table 2.3 Brief survey and classification of literature on statistical approaches to Big Data

Topic: Discussion article
Author: Dunson [61]
• Discussed the background of Big Data from the perspectives of the machine learning and statistics communities. • Listed the differences in the methods and inferences as replicability, uncertainty quantification, sampling, selection bias and measurement error drawn from statistical perspectives to those of machine learning. • Identified the statistical challenges for high dimensional complex data (Big Data) in quantifying uncertainty, scaling up sampling methods and selection of priors in Bayesian methods.
Topic: Survey
Author: Nongxa [136]
• Identified challenges of Big Data as: high dimensionality, heterogeneity and incompleteness, scale, timeliness, security and privacy. • Pointed out that mathematical and statistical challenges of Big Data require updating the core knowledge areas (i.e., linear algebra, multivariable calculus, elementary probability and statistics, coding or programming) to more advanced topics (i.e., randomised numerical linear algebra, topological data analysis, matrix and tensor decompositions, random graphs; random matrices and complex networks) in mathematical and statistical education.
Author: Franke et al. [67]
• Reviewed different strategies of analysis as: data wrangling, visualisation, dimension reduction, sparsity regularisation, optimisation, measuring distance, representation learning, sequential learning and provided detailed examples of applications.
Author: Chen et al. [45]
• Emphasised the importance of statistical knowledge and skills in Big Data Analytics using several examples. • Discussed some statistical methods that are useful in the context of Big Data as: confirmatory and exploratory data analysis tools, data mining methods including supervised learning (classification, regression/prediction) and unsupervised learning (cluster analysis, anomaly detection, association rule learning), visualisation techniques etc. • Elaborated on the computational skills needed for statisticians in data acquisition, data processing, data management and data analysis.
Author: Hoerl et al. [87]
• Provided a background of Big Data reviewing relevant articles. • Discussed the importance of statistical thinking in Big Data problems reviewing some misleading results produced by sophisticated analysis of Big Data without involving statistical principles. • Elaborated on the roles of statistical thinking for data quality, domain knowledge, analysis strategies in order to solve complex unstructured problems involving Big Data.

(continued)

Table 2.3 (continued)

Topic: Survey of methods & extension
Author: Wang et al. [191] • Reviewed statistical methods and software packages in R and recently developed tools to handle Big Data, focusing on three groups: sub-sampling, divide and conquer and online processing. • Extended the online updating approach by employing variable selection criteria.
Topic: Methods survey, new methods
Author: Genuer et al. [72] • Reviewed proposals dealing with scaling random forests to Big Data problems. • Discussed subsampling, parallel implementations, online processing of random forests in detail. • Proposed five variants of Random Forests for Big Data.
Author: Wang and Xu [185] • Reviewed different clustering methods applicable to Big Data situations. • Proposed a clustering procedure with adaptive density peak detection applying multivariate kernel estimation and demonstrated the performance through simulation studies and analysis of a few benchmark gene expression data sets. • Developed a R-package “ADPclust” to implement the proposed methods.
Author: Wang et al. [192] • Proposed a method and algorithm for online updating implementing bias corrections with extensions for application in a generalised linear model (GLM) setting. • Evaluated the proposed strategies in comparison with previous algorithms [162].
Topic: New methods and algorithms
Author: Liu et al. [111] • Proposed a novel sparse GLM with L0 approximation for feature selection and prediction in big omics data scenarios. • Provided novel algorithm and software in MATLAB (L0ADRIDGE) for performing L0 penalised GLM in ultra high dimensional Big Data. • Comparison of performance with other methods (SCAD, MC+) using simulation and real data analysis (mRNA, microRNA, methylation data from TGCA ovarian cancer).
Author: Schifano et al. [162] • Developed new statistical methods and iterative algorithms for analysing streaming data. • Proposed methods to enable update of the estimations and models with the arrival of new data.

(continued)

Table 2.3 (continued)

Author: Allen et al. [6] • Proposed generalisations to Principal Components Analysis (PCA) to take into account structural relationships in Big Data settings. • Developed fast computational algorithms using the proposed methods (GPCA, sparse GPCA and functional GPCA) for massive data sets.
Topic: New algorithms
Author: Wang and Samworth [188] • Proposed a new algorithm "inspect" (informative sparse projection for estimation of change points) to estimate the number and location of change points in high dimensional time series. • The algorithm, starting from a simple time series model, was extended to detect multiple change points and was also extended to have spatial or temporal dependence, assessed using simulation studies and real data application.
Author: Yu and Lin [203] • Extended the alternating direction method of multipliers (ADMM) to solve penalised quantile regression problems involving massive data sets having faster computation and no loss of estimation accuracy.
Author: Zhang and Yang [204] • Proposed new algorithms using ridge regression to make it efficient for handling Big Data.
Author: Doornik and Hendry [58] • Discussed the statistical model selection algorithm "autometrics" for econometric data [57] with its application to fat Big Data (having larger number of variables than the number of observations). • Extended algorithms for tackling computational issues of fat Big Data applying block searches and re-selection by lasso for correlated regressors.
Author: Sysoev et al. [174] • Presented efficient algorithms to estimate bootstrap or jackknife type confidence intervals for fitted Big Data sets by Multivariate Monotonic Regression. • Evaluated the performance of the proposed algorithms using a case study on death in coronary heart disease for a large population.
Author: Pehlivanlı [143] • Proposed a novel approach for feature selection from high dimensional data. • Tested the efficiency of the proposed method using sensitivity, specificity, accuracy and ROC curve. • Demonstrated the approach on micro-array data.

survey and discussion of state-of-the-art statistical methodologies and software development for handling Big Data.

Chen et al. [45] presented their views on the challenges and importance of Big Data and explained the role of statistics in Big Data Analytics based on a survey of relevant literature. This study emphasised the importance of statistical knowledge and skills in Big Data Analytics using several examples. As detailed in Table 2.3, the authors broadly discussed a range of statistical methods which can be really helpful in better analysis of Big Data, such as, the use of exploratory data analysis principle in Statistics to investigate correlations among the variables in the data or establish causal relationships between response and explanatory variables in the Big Data. The authors specifically mentioned hypothesis testing, predictive analysis using statistical models, statistical inference using uncertainty estimation to be some key tools to use in Big Data analysis. The authors also explained that the combination of statistical knowledge can be combined with the Data mining methods such as unsupervised learning (cluster analysis, Association rule learning, anomaly detection) and supervised learning (regression and classification) can be beneficial for Big Data analysis. The challenges for the statisticians in coping with Big Data were also described in this article, with particular emphasis on computational skills in data acquisition (knowledge of programming languages, knowledge of web and core communication protocols), data processing (skills to transform voice or image data to numeric data using appropriate software or programming), data management (knowledge about database management tools and technologies, such as NoSQL) and scalable computation (knowledge about parallel computing, which can be implemented using MapReduce, SQL etc.).

As indicated above, many of the papers provide a summary of the published literature which is not replicated here. Some of these surveys are based on large thematic programs that have been held on this topic. For example, the paper by Franke et al. [67] is based on presentations and discussions held as part of the program on Statistical Inference, Learning and Models for Big Data which was held in Canada in 2015. The authors discussed the four V's (volume, variety, veracity and velocity) of Big Data and mentioned some more challenges in Big Data analysis which are beyond the complexities associated with the four V's. The additional "V" mentioned in this article is veracity. Veracity refers to biases and noise in the data which may be the result of the heterogeneous structure of the data sources, which may make the sample non representative of the population. Veracity in Big Data is often referred to as the biggest challenge compared with the other V's. The paper reviewed the common strategies for Big Data analysis starting from data wrangling which consists of data manipulation techniques for making the data eligible for analysis; visualisation which is often an important tool to understand the underlying patterns in the data and is the first formal step in data analysis; reducing the dimension of data using different algorithms such as Principal Component Analysis (PCA) to make Big Data models tractable and interpretable; making models more robust by enforcing sparsity in the model by the use of regularisation techniques such as variable selection and model fitting criteria; using optimisation methods based on different distance measures proposed for high dimensional data and by using different learning algorithms such as representation learning and sequential learning. Different applications of Big Data were shown in public health, health policy, law and order, education, mobile application security, image recognition and labelling, digital humanities and materials science.

There are few other research articles focused on statistical methods tailored to specific problems, which are not included in Table 2.3. For example, Castruccio and Genton [40] proposed a statistics-based algorithm using a stochastic space-time model with more than 1 billion data points to reproduce some features of a climate model. Similarly, [123] used various statistical methods to obtain associations between drug-outcome pairs in a very big longitudinal medical experimental database (with information on millions of patients) with a detailed discussion on the big results problem by providing a comparison of statistical and machine learning approaches. Finally, Hensman et al. [84] proposed stochastic variational inference for Gaussian processes which makes the application of Gaussian process to huge data sets (having millions of data points).

From the survey of some relevant literature related to statistical perspectives for analysing Big Data, it can be seen that along with scaling up existing algorithms, new methodological developments are also in progress in order to face the challenges associated with Big Data.

2.5 Bayesian Approaches in Big Data

As described in the Introduction, the intention of this survey is to commence with a broad scope of the literature on Big Data, then focus on statistical methods for Big Data, and finally to focus in particular on Bayesian approaches for modelling and analysis of Big Data. This section consists of a survey of published literature on the last of these.

There are two defining features of Bayesian analysis: (1) the construction of the model and associated parameters and expectations of interest, and (2) the development of an algorithm to obtain posterior estimates of these quantities. In the context of Big Data, the resultant models can become complex and suffer from issues such as unavailability of a likelihood, hierarchical instability, parameter explosion and identifiability. Similarly, the algorithms can suffer from too much or too little data given the model structure, as well as problems of scalability and cost. These issues have motivated the development of new model structures, new methods that avoid the need for models, new Markov chain Monte Carlo (MCMC) sampling methods, and alternative algorithms and approximations that avoid these simulation-based approaches. We discuss some of the concomitant literature under two broad headings, namely computation and models realising that there is often overlap in cited papers.

2.5.1 Bayesian Computation

In Bayesian framework a main-stream computational tool has been the Markov chain Monte Carlo (MCMC). The traditional MCMC methods do not scale well because they need to iterate through the full data set at each iteration to evaluate the likelihood [199]. Recently several attempts have been made to scale MCMC methods up to massive data. A widely used strategy to overcome the computational cost is to distribute the computational burden across a number of machines. The strategy is generally referred to as divide-and-conquer sampling. This approach breaks a massive data set into a number of easier to handle subsets, obtains posterior samples based on each subset in parallel using multiple machines and finally combines the subset posterior inferences to obtain the full-posterior estimates [169]. The core challenge is the recombination of sub-posterior samples to obtain true posterior samples. A number of attempts have been made to address this challenge.

Neiswanger et al. [134] and White et al. [195] approximated the sub-posteriors using kernel density estimation and then aggregated the sub-posteriors by taking their product. Both algorithms provided consistent estimates of the posterior. Neiswanger et al. [134] provided faster MCMC processing since it allowed the machine to process the parallel MCMC chains independently. However, one limitation of the asymptotically embarrassing parallel MCMC algorithm [134] is that it only works for real and unconstrained posterior values, so there is still scope of works to make the algorithm work under more general settings.

Wang and Dunson [187] adopted a similar approach of parallel MCMC but used a Weierstrass transform to approximate the sub-posterior densities instead of a kernel density estimate. This provided better approximation accuracy, chain mixing rate and potentially faster speed for large scale Bayesian analysis.

Scott et al. [163] partitioned the data at random and performed MCMC independently on each subset to draw samples from posterior given the data subset. To obtain consensus posteriors they proposed to average samples across subsets and showed the exactness of the algorithm under a Gaussian assumption. This algorithm is scalable to a very large number of machines and works in cluster, single multi core or multiprocessor computers or any arbitrary collection of computers linked by a high speed network. The key weakness of consensus MCMC is it does not apply to non Gaussian posterior.

Minsker et al. [128] proposed dividing a large set of independent data into a number of non-overlapping subsets, making inferences on the subsets in parallel and then combining the inferences using the median of the subset posteriors. The median posterior (M-posterior) is constructed from the subset posteriors using Weiszfeld's algorithm, which provides a scalable algorithm for robust estimation.

Guhaniyogi and Banerjee [77] extended this notion to spatially dependent data, provided a scalable divide and conquer algorithm to analyse big spatial data sets named spatial meta kriging. The multivariate extension of spatial meta kriging has been addressed by Guhaniyogi and Banerjee [78]. These approaches of meta kriging are practical developments for Bayesian spatial inference for Big Data, specifically with "big-N" problems [98].

Wu and Robert [199] proposed a new and flexible divide and conquer framework by using re-scaled sub-posteriors to approximate the overall posterior. Unlike other parallel approaches of MCMC, this method creates artificial data for each subset, and applies the overall priors on the artificial data sets to get the subset posteriors. The sub-posteriors are then re-centred to their common mean and then averaged to approximate the overall posterior. The authors claimed this method to have statistical justification as well as mathematical validity along with sharing same computational cost with other classical parallel MCMC approaches such as consensus Monte Carlo, Weierstrass sampler. Bouchard-Côté et al. [30] proposed a non-reversible rejection-free MCMC method, which reportedly outperforms state-of-the-art methods such as: HMC, Firefly by having faster mixing rate and lower variances for the estimators for high dimensional models and large data sets. However, the automation of this method is still a challenge.

Another strategy for scalable Bayesian inference is the sub-sampling based approach. In this approach, a smaller subset of data is queried in the MCMC algorithm to evaluate the likelihood at every iteration. Maclaurin and Adams [116] proposed to use an auxiliary variable MCMC algorithm that evaluates the likelihood based on a small subset of the data at each iteration yet simulates from the exact posterior distribution. To improve the mixing speed, Korattikara et al. [95] used an approximate Metropolis Hastings (MH) test based on a subset of data. A similar approach is used in [17], where the accept/reject step of MH evaluates the likelihood of a random subset of the data. Bardenet et al. [18] extended this approach by replacing a number of likelihood evaluations by a Taylor expansion centred at the maximum of the likelihood and concluded that their method outperforms the previous algorithms [95].

The scalable MCMC approach was also improved by Quiroz et al. [150] using a difference estimator to estimate the log of the likelihood accurately using only a small fraction of the data. Quiroz et al. [151] introduced an unbiased estimator of the log likelihood based on weighted sub-sample which is used in the MH acceptance step in speeding up based on a weighted MCMC efficiently. Another scalable adaptation of MH algorithm was proposed by Maire et al. [119] to speed up Bayesian inference in Big Data namely informed subsampling MCMC which involves drawing of subsets according to a similarity measure (i.e., squared L2 distance between full data and maximum likelihood estimators of subsample) instead of using uniform distribution. The algorithm showed excellent performance in the case of a limited computational budget by approximating the posterior for a tall dataset.

Another variation of MCMC in Big Data has been made by Strathmann et al. [170]. These authors approximated the posterior expectation by a novel Bayesian inference framework for approximating the posterior expectation from a different perspective suitable for Big Data problems, which involves paths of partial posteriors. This is a parallelisable method which can easily be implemented using existing MCMC techniques. It does not require the simulation from full posterior, thus bypassing the complex convergence issues of kernel approximation. However, there is still scope for future work to look at computation-variance trade off and finite time bias produced by MCMC.

Hamiltonian Monte Carlo (HMC) sampling methods provide powerful and efficient algorithms for MCMC using high acceptance probabilities for distant proposals [44]. A conceptual introduction to HMC is presented by Betancourt [25]. Chen et al. [44] proposed a stochastic gradient HMC using second-order Langevin dynamics. Stochastic Gradient Langevin Dynamics (SGLD) have been proposed as a useful method for applying MCMC to Big Data where the accept-reject step is skipped and decreasing step size sequences are used [1]. For more detailed and rigorous mathematical framework, algorithms and recommendations, interested readers are referred to [178].

A popular method of scaling Bayesian inference, particularly in the case of analytically intractable distributions, is Sequential Monte Carlo (SMC) or particle filters [24, 48, 80]. SMC algorithms have recently become popular as a method to approximate integrals. The reasons behind their popularity include their easy implementation and parallelisation ability, much needed characteristics in Big Data implementations [100]. SMC can approximate a sequence of probability distributions on a sequence of spaces with an increasing dimension by applying resampling, propagation and weighting starting with the prior and eventually reaching to the posterior of interest of the cloud of particles. Gunawan et al. [80] proposed a subsampling SMC which is suitable for parallel computation in Big Data analysis, comprising two steps. First, the speed of the SMC is increased by using an unbiased and efficient estimator of the likelihood, followed by a Metropolis within Gibbs kernel. The kernel is updated by a HMC method for model parameters and a block-pseudo marginal proposal for the auxiliary variables [80]. Some novel approaches of SMC include: divide-and-conquer SMC [105], multilevel SMC [24], online SMC [75] and one pass SMC [104], among others.

Stochastic variational inference (VI, also called Variational Bayes, VB) is a faster alternative to MCMC [88]. It approximates probability densities using a deterministic optimisation method [109] and has seen widespread use to approximate posterior densities for Bayesian models in large-scale problems. The interested reader is referred to [29] for a detailed introduction to variational inference designed for statisticians, with applications. VI has been implemented in scaling up algorithms for Big Data. For example, a novel re-parameterisation of VI has been implemented for scaling latent variable models and sparse GP regression to Big Data [69].

There have been studies which combined the VI and SMC in order to take advantage from both strategies in finding the true posterior [56, 133, 152]. Naesseth et al. [133] employed a SMC approach to get an improved variational approximation, Rabinovich et al. [152] by splitting the data into block, applied SMC to compute partial posterior for each block and used a variational argument to get a proxy for the true posterior by the product of the partial posteriors. The combination of these two techniques in a Big Data context was made by Donnet and Robin [56]. Donnet and Robin [56] proposed a new sampling scheme called Shortened Bridge Sampler, which combines the strength of deterministic approximations of the posterior that is variational Bayes with those of SMC. This sampler resulted in reduced computational time for Big Data with huge numbers of parameters, such as data from genomics or network.

Guhaniyogi et al. [79] proposed a novel algorithm for Bayesian inference in the context of massive online streaming data, extending the Gibbs sampling mechanism for drawing samples from conditional distributions conditioned on sequential point estimates of other parameters. The authors compared the performance of this conditional density filtering algorithm in approximating the true posterior with SMC and VB, and reported good performance and strong convergence of the proposed algorithm.

Approximate Bayesian computation (ABC) is gaining popularity for statistical inference with high dimensional data and computationally intensive models where the likelihood is intractable [125]. A detailed overview of ABC can be found in [167] and asymptotic properties of ABC are explored in [68]. ABC is a likelihood free method that approximates the posterior distribution utilising imperfect matching of summary statistics [167]. Improvements on existing ABC methods for efficient estimation of posterior density with Big Data (complex and high dimensional data with costly simulations) have been proposed by Izbicki et al. [90]. The choice of summary statistics from high dimensional data is a topic of active discussion; see, for example, [90, 166]. Pudlo et al. [147] provided a reliable and robust method of model selection in ABC employing random forests which was shown to have a gain in computational efficiency.

There is another aspect of ABC recently in terms of approximating the likelihood using Bayesian Synthetic likelihood or empirical likelihood [59]. Bayesian synthetic likelihood arguably provides computationally efficient approximations of the likelihood with high dimensional summary statistics [126, 196]. Empirical likelihood, on the other hand is a non-parametric technique of approximating the likelihood empirically from the data considering the moment constraints; this has been suggested in the context of ABC [127], but has not been widely adopted. For further reading on empirical likelihood, see [141].

Classification and regression trees are also very useful tools in data mining and Big Data analysis [33]. There are Bayesian versions of regression trees such as Bayesian Additive Regression Trees (BART) [7, 47, 93]. The BART algorithm has also been applied to the Big Data context and sparse variable selection by Rocková and van der Pas [157], van der Pas and Rockova [181], and Linero [106].

Some other recommendations to speed up computations are to use graphics processing units (see, e.g., [101, 171]) and parallel programming approaches (see, e.g., [42, 71, 76, 197]).

2.5.2 Bayesian Modelling

The extensive development of Bayesian computational solutions has opened the door to further developments in Bayesian modelling. Many of these new methods are set in the context of application areas. For example, there have been applications of ABC for Big Data in many different fields [62, 102]. For example, Dutta et al. [62] developed a high performance computing ABC approach for estimation of parameters in platelets deposition, while Lee et al. [102] proposed ABC methods for inference in high dimensional multivariate spatial data from a large number of locations with a particular focus on model selection for application to spatial extremes analysis. Bayesian mixtures are a popular modelling tool. VB and ABC techniques have been used for fitting Bayesian mixture models to Big Data [29, 88, 124, 129, 177].

Variable selection in Big Data (wide in particular, having massive number of variables) is a demanding problem. Liquet et al. [107] proposed multivariate extensions of the Bayesian group lasso for variable selection in high dimensional data using Bayesian hierarchical models utilising spike and slab priors with application to gene expression data. The variable selection problem can also be solved employing ABC type algorithms. Liu et al. [112] proposed a sampling technique, ABC Bayesian forests, based on splitting the data, useful for high dimensional wide data, which turns out to be a robust method in identifying variables with larger marginal inclusion probability.

Bayesian non-parametrics [131] have unbounded capacity to adjust unseen data through activating additional parameters that were inactive before the emergence of new data. In other words, the new data are allowed to speak for themselves in non-parametric models rather than imposing an arguably restricted model (that was learned on an available data) to accommodate new data. The inherent flexibility of these models to adjust with new data by adapting in complexity makes them more suitable for Big Data as compared to their parametric counterparts. For a brief introduction to Bayesian non-parametric models and a nontechnical overview of some of the main tools in the area, the interested reader is referred to Ghahramani [73].

The popular tools in Bayesian non-parametrics include Gaussian processes (GP) [156], Dirichlet processes (DP) [155], Indian buffet process (IBP) [74] and infinite hidden Markov models (iHMM) [20]. GP have been used for a variety of applications [35, 41, 49] and attempts have been made to scale it to Big Data [53, 84, 85, 179]. DP have seen successes in clustering and faster computational algorithms are being adopted to scale them to Big Data [71, 104, 115, 186, 189]. IBP are used for latent feature modeling, where the number of features are determined in a data-driven fashion and have been scaled to Big Data through variational inference algorithms [211]. Being an alternative to classical HMM, one of the distinctive properties of iHMM is that it infers the number of hidden states in the system from the available data and has been scaled to Big Data using particle filtering algorithms [180].

Gaussian Processes are also employed in the analysis of high dimensional spatially dependent data [15]. Banerjee [15] provided model-based solutions employing low rank GP and nearest neighbour GP (NNGP) as scalable priors in a hierarchical framework to render full Bayesian inference for big spatial or spatio temporal data sets. Zhang et al. [208] extended the applicability of NNGP for inference of latent spatially dependent processes by developing a conjugate latent NNGP model as a practical alternative to onerous Bayesian computations. Use of variational optimisation with structured Bayesian GP latent variable model to analyse spatially dependent data is made in in Atkinson and Zabaras [12]. For a survey of methods of analysis of massive spatially dependent data including the Bayesian approaches, see Heaton et al. [83].

Another Bayesian modelling approach that has been used for big and complex data is Bayesian Networks (BN). This methodology has generated a substantial literature examining theoretical, methodological and computational approaches, as well as applications [176]. BN belong to the family of probabilistic graphical models and based on direct acyclic graphs which are very useful representation of causal relationship among variables [23]. BN are used as efficient learning tool in Big Data analysis integrated with scalable algorithms [190, 209]. For a more detailed understanding of BN learning from Big Data, please see Tang et al. [176].

Classification is also an important tool for extracting information from Big Data and Bayesian classifiers, including Naive Bayes classifier (NBC) are used in Big Data classification problems [94, 110]. Parallel implementation of NBC has been proposed by Katkar and Kulkarni [94]. Moreover, Liu et al. [110] evaluated the scalability of NBC in Big Data with application to sentiment classification of millions of movie survey and found NBC to have improved accuracy in Big Data. Ni et al. [135] proposed a scalable multi step clustering and classification algorithm using Bayesian nonparametrics for Big Data with large n and small p which can also run in parallel.

The past 15 years has also seen an increase in interest in Empirical Likelihood (EL) for Bayesian modelling. The idea of replacing the likelihood with an empirical analogue in a Bayesian framework was first explored in detail by Lazar [99]. The author demonstrated that this Bayesian Empirical Likelihood (BEL) approach increases the flexibility of EL approach by examining the length and coverage of BEL intervals. The paper tested the methods using simulated data sets. Later, Schennach [161] provided probabilistic interpretations of BEL exploring moment condition models with EL and provided a non parametric version of BEL, namely Bayesian Exponentially Tilted Empirical Likelihood (BETEL). The BEL methods have been applied in spatial data analysis in Chaudhuri and Ghosh [43] and Porter et al. [145, 146] for small area estimation.

We acknowledge that there are many more studies on the application of Bayesian approaches in different fields of interest which are not included in this survey. There are also other survey papers on overlapping and closely related topics. For example, Zhu et al. [210] describes Bayesian methods of machine learning and includes some of the Bayesian inference techniques reviewed in the present study. However, the scope and focus of this survey is different from that of Zhu et al. [210], which was focused around the methods applicable to machine learning.

2.6 Conclusions

We are living in the era of Big Data and continuous research is in progress to make most use of the available information. The current chapter has attempted to survey the recent developments made in Bayesian statistical approaches for handling Big Data along with a general overview and classification of the Big Data literature with brief survey in last 5 years. This survey chapter provides relevant references in Big Data categorised in finer classes, a brief description of statistical contributions to the field and a more detailed discussion of the Bayesian approaches developed and applied in the context of Big Data.

On the basis of the surveys made above, it is clear that there has been a huge amount of work on issues related to cloud computing, analytics infrastructure and so on. However, the amount of research conducted from statistical perspectives is also notable. In the last 5 years, there has been an exponential increase in published studies focused on developing new statistical methods and algorithms, as well as scaling existing methods. These have been summarised in Sect. 2.4, with particular focus on Bayesian approaches in Sect. 2.5. In some instances citations are made outside of the specific period (see Sect. 2.2) to refer the origin of the methods which are currently being applied or extended in Big Data scenarios.

With the advent of computational infrastructure and advances in programming and software, Bayesian approaches are no longer considered as being very computationally expensive and onerous to execute for large volumes of data, that is Big Data. Traditional Bayesian methods are now becoming much more scalable due to the advent of parallelisation of MCMC algorithms, divide and conquer and/or subsampling methods in MCMC, and advances in approximations such as HMC, SMC, ABC, VB and so on. With the increasing volume of data, non-parametric Bayesian methods are also gaining in popularity.

This survey chapter aimed to survey a range of methodological and computational advancement made in Bayesian Statistics for handling the difficulties arose by the advent of Big Data. By not focusing to any particular application, this chapter provided the readers with a general overview of the developments of Bayesian methodologies and computational algorithms for handling these issues. The survey has revealed that most of the advancements in Bayesian Statistics for Big Data have been around computational time and scalability of particular algorithms, concentrating on estimating the posterior by adopting different techniques. However the developments of Bayesian methods and models for Big Data in the recent literature cannot be overlooked. There are still many open problems for further research in the context of Big Data and Bayesian approaches, as highlighted in this chapter.

Based on the above discussion and the accompanying survey presented in this chapter, it is apparent that to address the challenges of Big Data along with the strength of Bayesian statistics, research on both algorithms and models are essential.

Acknowledgments This research was supported by an ARC Australian Laureate Fellowship for project, Bayesian Learning for Decision Making in the Big Data Era under Grant no. FL150100150. The authors also acknowledge the support of the Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

References

1. S. Ahn, B. Shahbaba, M. Welling, Distributed stochastic gradient MCMC, in *International Conference on Machine Learning* (2014), pp. 1044–1052
2. S Akter, S.F. Wamba, Big data analytics in e-commerce: a systematic review and agenda for future research. Electron. Mark. **26**(2), 173–194 (2016)
3. A. Akusok, K.M. Björk, Y. Miche, A. Lendasse, High-performance extreme learning machines: a complete toolbox for big data applications. IEEE Access **3**, 1011–1025 (2015)
4. O.Y. Al-Jarrah, P.D. Yoo, S. Muhaidat, G.K. Karagiannidis, K. Taha, Efficient machine learning for big data: a review. Big Data Res. **2**(3), 87–93 (2015)
5. K. Albury, J. Burgess, B. Light, K Race, R. Wilken, Data cultures of mobile dating and hook-up apps: emerging issues for critical social science research. Big Data Soc. **4**(2), 1–11 (2017)
6. G.I. Allen, L. Grosenick, J. Taylor, A generalized least-square matrix decomposition. J. Am. Stat. Assoc. **109**(505), 145–159 (2014)
7. G.M. Allenby, E.T. Bradlow, E.I. George, J. Liechty, R.E. McCulloch, Perspectives on Bayesian methods and big data. Cust. Needs Solut. **1**(3), 169–175 (2014)
8. S.G. Alonso, I. de la Torre Díez, J.J. Rodrigues, S. Hamrioui, M. López-Coronado, A systematic review of techniques and sources of big data in the healthcare sector. J. Med. Syst. **41**(11), 183 (2017)
9. A. Alyass, M. Turcotte, D. Meyre, From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med. Genomics **8**(1), 33 (2015)
10. D. Apiletti, E. Baralis, T. Cerquitelli, P. Garza, F. Pulvirenti, L. Venturini, (2017) Frequent itemsets mining for big data: a comparative analysis. Big Data Res. **9**, 67–83
11. M.D. Assunção, R.N. Calheiros, S. Bianchi, M.A. Netto, R. Buyya, Big data computing and clouds: trends and future directions. J. Parallel Distrib. Comput. **79**, 3–15 (2015)
12. S. Atkinson, N. Zabaras, Structured Bayesian Gaussian process latent variable model: applications to data-driven dimensionality reduction and high-dimensional inversion. J. Comput. Phys. **383**, 166–195 (2019)
13. A.T. Azar, A.E. Hassanien, Dimensionality reduction of medical big data using neural-fuzzy classifier. Soft Comput. **19**(4), 1115–1127 (2015)
14. A. Baldominos, E. Albacete, Y. Saez, P. Isasi, A scalable machine learning online service for big data real-time analysis, in *2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD)* (IEEE, Piscataway, 2014), pp. 1–8
15. S. Banerjee, High-dimensional Bayesian geostatistics. Bayesian Anal. **12**(2), 583 (2017)
16. S. Bansal, G. Chowell, L. Simonsen, A. Vespignani, C. Viboud, Big data for infectious disease surveillance and modeling. J. Infect. Dis. **214**(suppl_4), S375–S379 (2016)
17. R. Bardenet, A. Doucet, C. Holmes, Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach, in *International Conference on Machine Learning (ICML)* (2014), pp. 405–413
18. R. Bardenet, A. Doucet, C. Holmes, On Markov chain Monte Carlo methods for tall data. J. Mach. Learn. Res. **18**(1), 1515–1557 (2017)
19. D.W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. **33**(7), 1123–1131 (2014)
20. M.J. Beal, Z. Ghahramani, C.E. Rasmussen, The infinite hidden Markov model, in *Advances in Neural Information Processing Systems* (2002), pp. 577–584

21. A. Belle, R. Thiagarajan, S. Soroushmehr, F. Navidi, D.A. Beard, K. Najarian, Big data analytics in healthcare. BioMed. Res. Int. **2015**, 370194 (2015)
22. G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: recent achievements and new challenges. Inf. Fusion **28**, 45–59 (2016)
23. I. Ben-Gal, Bayesian Networks. Encycl. Stat. Qual. Reliab. **1**, 1–6 (2008)
24. A. Beskos, A. Jasra, E.A. Muzaffer, A.M. Stuart, Sequential Monte Carlo methods for Bayesian elliptic inverse problems. Stat. Comput. **25**(4), 727–737 (2015)
25. M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo. Preprint, arXiv: 170102434 (2017)
26. J.E. Bibault, P. Giraud, A. Burgun, Big data and machine learning in radiation oncology: state of the art and future prospects. Cancer Lett. **382**(1), 110–117 (2016)
27. A. Bifet, Morales GDF Big data stream learning with Samoa, in *2014 IEEE International Conference on Data Mining Workshop (ICDMW)*, IEEE, pp. 1199–1202 (2014)
28. H. Binder, M. Blettner, Big data in medical science–a biostatistical view: Part 21 of a series on evaluation of scientific publications. Dtsch. Ärztebl Int. **112**(9), 137 (2015)
29. D.M. Blei, A. Kucukelbir, J.D. McAuliffe, Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017)
30. A. Bouchard-Côté, S.J. Vollmer, A. Doucet, The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. J. Am. Stat. Assoc. **113**, 1–13 (2018)
31. E.T. Bradlow, M. Gangwar, P. Kopalle, S. Voleti, The role of big data and predictive analytics in retail. J. Retail. **93**(1), 79–95 (2017)
32. R. Branch, H. Tjeerdsma, C. Wilson, R. Hurley, S. McConnell, Cloud computing and big data: a review of current service models and hardware perspectives. J. Softw. Eng. Appl. **7**(08), 686 (2014)
33. L. Breiman, *Classification and Regression Trees* (Routledge, Abingdon, 2017)
34. P.F. Brennan, S. Bakken, Nursing needs big data and big data needs nursing. J. Nurs. Scholarsh. **47**(5), 477–484 (2015)
35. F. Buettner, K.N. Natarajan, F.P. Casale, V. Proserpio, A. Scialdone, F.J. Theis, S.A. Teichmann, J.C. Marioni, O. Stegle, Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. **33**(2), 155 (2015)
36. J. Bughin, Big data, big bang? J. Big Data **3**(1), 2 (2016)
37. R. Burrows, M. Savage, After the crisis? Big data and the methodological challenges of empirical sociology. Big Data Soc. **1**(1), 1–6 (2014)
38. H. Cai, B. Xu, L. Jiang, A.V. Vasilakos, Iot-based big data storage systems in cloud computing: perspectives and challenges. IEEE Internet Things J. **4**(1), 75–87 (2017)
39. J.N. Cappella, Vectors into the future of mass and interpersonal communication research: big data, social media, and computational social science. Hum. Commun. Res. **43**(4), 545–558 (2017)
40. S. Castruccio, M.G. Genton, Compressing an ensemble with statistical models: an algorithm for global 3d spatio-temporal temperature. Technometrics **58**(3), 319–328 (2016)
41. K. Chalupka, C.K. Williams, I. Murray, A framework for evaluating approximation methods for Gaussian process regression. J. Mach. Learn. Res. **14**(Feb), 333–350 (2013)
42. J. Chang, J.W. Fisher III, Parallel sampling of DP mixture models using sub-cluster splits, in *Advances in Neural Information Processing Systems* (2013), pp. 620–628
43. S. Chaudhuri, M. Ghosh, Empirical likelihood for small area estimation. Biometrika **98**, 473–480 (2011)
44. T. Chen, E. Fox, C. Guestrin, Stochastic gradient Hamiltonian Monte Carlo, in *Int. Conference on Machine Learning* (2014), pp. 1683–1691
45. J.J. Chen, E.E. Chen, W. Zhao, W. Zou, Statistics in big data. J. Chin. Stat. Assoc. **53**, 186–202 (2015)
46. A.S. Cheung, Moving beyond consent for citizen science in big data health and medical research. Northwest J. Technol. Intellect. Prop. **16**(1), 15 (2018)

47. H.A. Chipman, E.I. George, R.E. McCulloch et al., BART: Bayesian additive regression trees. Ann. Appl. Stat. **4**(1), 266–298 (2010)
48. N. Chopin, P.E. Jacob, O. Papaspiliopoulos, Smc2: an efficient algorithm for sequential analysis of state space models. J. R. Stat. Soc. Ser. B (Stat Methodol.) **75**(3), 397–426 (2013)
49. A. Damianou, N. Lawrence, Deep Gaussian processes, in *Artificial Intelligence and Statistics* (2013), pp. 207–215
50. T. Das, P.M. Kumar, Big data analytics: a framework for unstructured data analysis. Int. J. Eng. Sci. Technol. **5**(1), 153 (2013)
51. A. De Mauro, M. Greco, M. Grimaldi, What is big data? a consensual definition and a review of key research topics, in *AIP Conference Proceedings, AIP*, vol. 1644 (2015), pp. 97–104
52. A. De Mauro, M. Greco, M. Grimaldi A formal definition of big data based on its essential features. Libr. Rev. **65**(3), 122–135 (2016)
53. M.P. Deisenroth, J.W. Ng, Distributed Gaussian processes, in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, vol. 37, JMLR.org (2015), pp. 1481–1490
54. H. Demirkan, D. Delen Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud. Decis. Support Syst. **55**(1), 412–421 (2013)
55. K.S. Divya, P. Bhargavi, S. Jyothi Machine learning algorithms in big data analytics. Int. J. Comput. Sci. Eng. **6**(1), 63–70 (2018)
56. S. Donnet, S. Robin Shortened bridge sampler: using deterministic approximations to accelerate SMC for posterior sampling. Preprint, arXiv 170707971 (2017)
57. J.A. Doornik, Autometrics, in *The Methodology and Practice of Econometrics, A Festschrift in Honour of David F. Hendry*, University Press, pp. 88–121 (2009)
58. J.A. Doornik, D.F. Hendry, Statistical model selection with "big data". Cogent Econ. Finan. **3**(1), 1045216 (2015)
59. C.C. Drovandi, C. Grazian, K. Mengersen, C. Robert, Approximating the likelihood in ABC, in *Handbook of Approximate Bayesian Computation*, ed. by S.A. Sisson, Y. Fan, M. Beaumont (Chapman and Hall/CRC, Boca Raton, 2018), pp. 321–368
60. P. Ducange, R. Pecori, P. Mezzina, A glimpse on big data analytics in the framework of marketing strategies. Soft Comput. **22**(1), 325–342 (2018)
61. D.B. Dunson, Statistics in the big data era: failures of the machine. Stat. Probab. Lett. **136**, 4–9 (2018)
62. R. Dutta, M. Schoengens, J.P. Onnela, A. Mira, Abcpy, in *Proceedings of the Platform for Advanced Scientific Computing Conference on - PASC* (2017)
63. C.K. Emani, N. Cullot, C. Nicolle, Understandable big data: a survey. Comput. Sci. Rev. **17**, 70–81 (2015)
64. A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A.Y. Zomaya, S. Foufou, A. Bouras, A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans. Emerg. Top. Comput. **2**(3), 267–279 (2014)
65. J. Fan, F. Han, H. Liu, Challenges of big data analysis. Natl. Sci. Rev. **1**(2), 293–314 (2014)
66. S. Fosso Wamba, D. Mishra, Big data integration with business processes: a literature review. Bus. Process Manag. J. **23**(3), 477–492 (2017)
67. B. Franke, J.F. Plante, R. Roscher, A. Lee, C. Smyth, A. Hatefi, F. Chen, E. Gil, A. Schwing, A. Selvitella et al., Statistical inference, learning and models in big data. Int. Stat. Rev. **84**(3), 371–389 (2016)
68. D.T. Frazier, G.M. Martin, C.P. Robert, J. Rousseau, Asymptotic properties of approximate Bayesian computation. Biometrika **105**(3), 593–607 (2018)
69. Y. Gal, M. Van Der Wilk, C.E. Rasmussen, Distributed variational inference in sparse Gaussian process regression and latent variable models, in *Advances in Neural Information Processing Systems* (2014), pp. 3257–3265
70. A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics. Int. J. Inf. Manag. **35**(2), 137–144 (2015)
71. H. Ge, Y. Chen, M. Wan, Z. Ghahramani, Distributed inference for Dirichlet process mixture models, in *International Conference on Machine Learning* (2015), pp. 2276–2284

72. R. Genuer, J.M. Poggi, Tuleau-Malot C, N. Villa-Vialaneix, Random forests for big data. Big Data Res. **9**, 28–46 (2017)
73. Z. Ghahramani, Bayesian non-parametrics and the probabilistic approach to modelling. Phil. Trans. R. Soc. A. **371**(1984), 20110553 (2013)
74. Z. Ghahramani, T.L. Griffiths, Infinite latent feature models and the Indian buffet process, in *Advances in Neural Information Processing Systems* (2006), pp. 475–482
75. P. Gloaguen, M.P. Etienne, S. Le Corff Online sequential Monte Carlo smoother for partially observed diffusion processes. URASIP J. Adv. Signal Process. **2018**(1), 9 (2018)
76. S. Guha, R. Hafen, J. Rounds, J. Xia, J. Li, B. Xi, W.S. Cleveland, Large complex data: divide and recombine (D&R) with RHIPE. Stat **1**(1), 53–67 (2012)
77. R. Guhaniyogi, S. Banerjee, Meta-Kriging: scalable Bayesian modeling and inference for massive spatial datasets. Technometrics **60**(4), 430–444 (2018)
78. R. Guhaniyogi, S. Banerjee, Multivariate spatial meta kriging. Stat. Probab. Lett. **144**, 3–8 (2019)
79. R. Guhaniyogi, S. Qamar, D.B. Dunson, Bayesian conditional density filtering for big data. Stat **1050**, 15 (2014)
80. D. Gunawan, R. Kohn, M. Quiroz, K.D. Dang, M.N. Tran, Subsampling Sequential Monte Carlo for Static Bayesian Models. Preprint, arXiv:180503317 (2018)
81. H. Hassani, E.S. Silva, Forecasting with big data: a review. Ann. Data Sci. **2**(1), 5–19 (2015)
82. S.I. Hay, D.B. George, C.L. Moyes, J.S. Brownstein, Big data opportunities for global infectious disease surveillance. PLoS Med. **10**(4), e1001413 (2013)
83. M.J. Heaton, A. Datta, A. Finley, R. Furrer, R. Guhaniyogi, F. Gerber, R.B. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren et al., Methods for analyzing large spatial data: a review and comparison. Preprint, arXiv:171005013 (2017)
84. J. Hensman, N. Fusi, N.D. Lawrence, Gaussian processes for big data. Preprint, arXiv:13096835 (2013)
85. J. Hensman, A.G.d.G. Matthews, Z. Ghahramani, Scalable variational Gaussian process classification, in *18th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2015), pp. 351–360
86. M. Hilbert, Big data for development: a review of promises and challenges. Dev. Policy Rev. **34**(1), 135–174 (2016)
87. R.W. Hoerl, R.D. Snee, R.D. De Veaux, Applying statistical thinking to “Big Data” problems. Wiley Interdiscip. Rev. Comput. Stat. **6**(4), 222–232 (2014)
88. M.D. Hoffman, D.M. Blei, C. Wang, J. Paisley, Stochastic variational inference. J. Mach. Learn. Res. **14**(1), 1303–1347 (2013)
89. H.H. Huang, H. Liu, Big data machine learning and graph analytics: Current state and future challenges, in *2014 IEEE International Conference on Big Data (Big Data)* (IEEE, Piscataway, 2014), pp. 16–17
90. R. Izbicki, A.B. Lee, T. Pospisil, ABC–CDE: toward approximate Bayesian computation with complex high-dimensional data and limited simulations. J. Comput. Graph. Stat. **28**, 1–20 (2019)
91. G. Jifa, Z. Lingling, Data, DIKW, big data and data science. Procedia Comput. Sci. **31**, 814–821 (2014)
92. S. Kaisler, F. Armour, J.A. Espinosa, W. Money, Big data: issues and challenges moving forward, in *2013 46th Hawaii International Conference on System Sciences* (IEEE, Piscataway, 2013), pp. 995–1004
93. A. Kapelner, J. Bleich bartMachine: machine learning with Bayesian additive regression trees. Preprint, arXiv:13122171 (2013)
94. V.D. Katkar, S.V. Kulkarni, A novel parallel implementation of Naive Bayesian classifier for big data, in *2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)* (IEEE, Piscataway, 2013), pp. 847–852
95. A. Korattikara, Y. Chen, M. Welling, Austerity in MCMC land: Cutting the Metropolis-Hastings budget, in *International Conference on Machine Learning* (2014), pp. 181–189

96. H. Kousar, B.P. Babu, Multi-Agent based MapReduce Model for Efficient Utilization of System Resources. Indones. J. Electr. Eng. Comput. Sci. **11**(2), 504–514 (2018)
97. S. Landset, T.M. Khoshgoftaar, A.N. Richter, T. Hasanin, A survey of open source tools for machine learning with big data in the hadoop ecosystem. J. Big Data **2**(1), 24 (2015)
98. G.J. Lasinio, G. Mastrantonio, A. Pollice, Discussing the "big n problem". Stat. Methods Appt. **22**(1), 97–112 (2013)
99. N.A. Lazar, Bayesian empirical likelihood. Biometrika **90**(2), 319–326 (2003)
100. A. Lee, N. Whiteley, Forest resampling for distributed sequential Monte Carlo. Stat. Anal. Data Min. **9**(4), 230–248 (2016)
101. A. Lee, C. Yau, M.B. Giles, A. Doucet, C.C. Holmes, On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. J. Comput. Graph. Stat. **19**(4), 769–789 (2010)
102. X.J. Lee, M. Hainy, McKeone JP, C.C. Drovandi, A.N. Pettitt, ABC model selection for spatial extremes models applied to South Australian maximum temperature data. Comput. Stat. Data Anal. **128**, 128–144 (2018)
103. S. Li, S. Dragicevic, F.A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein et al., Geospatial big data handling theory and methods: a review and research challenges. ISPRS J. Photogramm. Remote Sens. **115**, 119–133 (2016)
104. D. Lin, Online learning of nonparametric mixture models via sequential variational approximation, in *Advances in Neural Information Processing Systems* (2013), pp. 395–403
105. F. Lindsten, A.M. Johansen, C.A. Naesseth, B. Kirkpatrick, T.B. Schön, J. Aston, A. Bouchard-Côté, Divide-and-conquer with sequential Monte Carlo. J. Comput. Graph. Stat. **26**(2), 445–458 (2017)
106. A.R. Linero, Bayesian regression trees for high-dimensional prediction and variable selection. J. Am. Stat. Assoc. **113**, 1–11 (2018)
107. B. Liquet, K. Mengersen, A. Pettitt, M. Sutton et al., Bayesian variable selection regression of multivariate responses for group data. Bayesian Anal. **12**(4), 1039–1067 (2017)
108. L. Liu, Computing infrastructure for big data processing. Front. Comput. Sci. **7**(2), 165–170 (2013)
109. Q. Liu, D. Wang, Stein variational gradient descent: a general purpose Bayesian inference algorithm, in *Advances In Neural Information Processing Systems* (2016), pp. 2378–2386
110. B. Liu, E. Blasch, Y. Chen, D. Shen, G. Chen, Scalable sentiment classification for big data analysis using Naive Bayes classifier, in *2013 IEEE International Conference on Big Data* (IEEE, Piscataway, 2013), pp. 99–104
111. Z. Liu, F. Sun, D.P. McGovern, Sparse generalized linear model with L0 approximation for feature selection and prediction with big omics data. BioData Min. **10**(1), 39 (2017)
112. Y. Liu, V. Ročková, Y. Wang, ABC variable selection with Bayesian forests. Preprint, arXiv:180602304 (2018)
113. C. Loebbecke, A. Picot, Reflections on societal and business model transformation arising from digitization and big data analytics: a research agenda. J. Strategic Inf. Syst. **24**(3), 149–157 (2015)
114. J. Luo, M. Wu, D. Gopukumar, Y. Zhao, Big data application in biomedical research and health care: a literature review. Biomed. Inform. Insights **8**, BII–S31559 (2016)
115. Z. Ma, P.K. Rana, J. Taghia, M. Flierl, A. Leijon, Bayesian estimation of Dirichlet mixture model with variational inference. Pattern Recognit. **47**(9), 3143–3157 (2014)
116. D. Maclaurin, R.P. Adams, Firefly Monte Carlo: exact MCMC with subsets of data, in *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2014), pp. 543–552
117. T. Magdon-Ismail, C. Narasimhadevara, D. Jaffe, R. Nambiar, Tpcx-hs v2: transforming with technology changes, in *Technology Conference on Performance Evaluation and Benchmarking* (Springer, Berlin, 2017), pp. 120–130
118. L. Mählmann, M. Reumann, N. Evangelatos, A. Brand, Big data for public health policy-making: policy empowerment. Public Health Genomics **20**(6), 312–320 (2017)
119. F. Maire, N. Friel, P. Alquier, Informed sub-sampling MCMC: approximate Bayesian inference for large datasets. Stat. Comput. 1–34 (2017). https://doi.org/10.1007/s11222-018-9817-3

120. R. Manibharathi, R. Dinesh, Survey of challenges in encrypted data storage in cloud computing and big data. J. Netw. Commun. Emerg. Technol. **8**(2) (2018). ISSN:2395-5317
121. R.F. Mansour, Understanding how big data leads to social networking vulnerability. Comput. Hum. Behav. **57**, 348–351 (2016)
122. A. Marshall, S. Mueck, R. Shockley, How leading organizations use big data and analytics to innovate. Strateg. Leadersh. **43**(5), 32–39 (2015)
123. T.H. McCormick, R. Ferrell, A.F. Karr, P.B. Ryan, Big data, big results: knowledge discovery in output from large-scale analytics. Stat. Anal. Data Min. **7**(5), 404–412 (2014)
124. C.A. McGrory, D. Titterington, Variational approximations in Bayesian model selection for finite mixture distributions. Comput. Stat. Data Anal. **51**(11), 5352–5367 (2007)
125. T.J. McKinley, I. Vernon, I. Andrianakis, N. McCreesh, J.E. Oakley, R.N. Nsubuga, M. Goldstein, R.G. White et al., Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. Stat. Sci. **33**(1), 4–18 (2018)
126. E. Meeds, M. Welling, GPS-ABC: Gaussian process surrogate approximate Bayesian computation. Preprint, arXiv:14012838 (2014)
127. K.L. Mengersen, P. Pudlo, C.P. Robert, Bayesian computation via empirical likelihood. Proc. Natl. Acad. Sci. **110**(4), 1321–1326 (2013)
128. S. Minsker, S. Srivastava, L. Lin, D.B. Dunson, Robust and scalable Bayes via a median of subset posterior measures. J. Mach. Learn. Res. **18**(1), 4488–4527 (2017)
129. M.T. Moores, C.C. Drovandi, K. Mengersen, C.P. Robert, Pre-processing for approximate Bayesian computation in image analysis. Stat. Comput. **25**(1), 23–33 (2015)
130. N. Moustafa, G. Creech, E. Sitnikova, M. Keshk, Collaborative anomaly detection framework for handling big data of cloud computing, in *Military Communications and Information Systems Conference (MilCIS), 2017* (IEEE, Piscataway, 2017), pp. 1–6
131. P. Müller, F.A. Quintana, A. Jara, T. Hanson, *Bayesian Nonparametric Data Analysis* (Springer, Berlin, 2015)
132. O. Müller, I. Junglas, J.v. Brocke, S. Debortoli, Utilizing big data analytics for information systems research: challenges, promises and guidelines. Eur. J. Inf. Syst. **25**(4), 289–302 (2016)
133. C.A. Naesseth, S.W. Linderman, R. Ranganath, D.M. Blei, Variational sequential Monte Carlo. Preprint, arXiv:170511140 (2017)
134. W. Neiswanger, C. Wang, E. Xing, Asymptotically exact, embarrassingly parallel MCMC. Preprint, arXiv:13114780 (2013)
135. Y. Ni, P. Müller, M. Diesendruck, S. Williamson, Y. Zhu, Y. Ji Scalable Bayesian nonparametric clustering and classification. J. Comput. Graph. Stat. 1–45 (2019). https://doi.org/10.1080/10618600.2019.1624366
136. L.G. Nongxa, Mathematical and statistical foundations and challenges of (big) data sciences. S. Afr. J. Sci. **113**(3–4), 1–4 (2017)
137. B. Oancea, R.M. Dragoescu et al., Integrating R and hadoop for big data analysis. Romanian Stat. Rev. **62**(2), 83–94 (2014)
138. Z. Obermeyer, E.J. Emanuel, Predicting the future—big data, machine learning, and clinical medicine. N. Engl. J. Med. **375**(13), 1216 (2016)
139. A. O'Driscoll, J. Daugelaite, R.D. Sleator, 'Big data', Hadoop and cloud computing in genomics. J. Biomed. Inform. **46**(5), 774–781 (2013)
140. D. Oprea, Big questions on big data. Rev. Cercet. Interv. Soc. **55**, 112 (2016)
141. A.B. Owen, *Empirical Likelihood* (Chapman and Hall/CRC, Boca Raton, 2001)
142. S. Pandey, V. Tokekar, Prominence of mapreduce in big data processing, in *2014 Fourth International Conference on Communication Systems and Network Technologies (CSNT)* (IEEE, Piscataway, 2014), pp. 555–560
143. A.Ç. Pehlivanlı, A novel feature selection scheme for high-dimensional data sets: four-staged feature selection. J. Appl. Stat. **43**(6), 1140–1154 (2015)
144. D.N. Politis, J.P. Romano, M. Wolf, *Subsampling* (Springer Science & Business Media, New York, 1999)

145. A.T. Porter, S.H. Holan, C.K. Wikle, Bayesian semiparametric hierarchical empirical likelihood spatial models. J. Stat. Plan. Inference **165**, 78–90 (2015)
146. A.T. Porter, S.H. Holan, C.K. Wikle, Multivariate spatial hierarchical Bayesian empirical likelihood methods for small area estimation. Stat **4**(1), 108–116 (2015)
147. P. Pudlo, J.M. Marin, A. Estoup, J.M. Cornuet, M. Gautier, C.P. Robert, Reliable ABC model choice via random forests. Bioinformatics **32**(6), 859–866 (2015)
148. F. Qi, F. Yang, Analysis of large data mining platform based on cloud computing, in *2018 4th World Conference on Control Electronics and Computer Engineering* (2018)
149. J. Qiu, Q. Wu, G. Ding, Y. Xu, S. Feng, A survey of machine learning for big data processing. EURASIP J. Adv. Signal Process. **2016**(1), 67 (2016)
150. M. Quiroz, M. Villani, R. Kohn, Scalable MCMC for large data problems using data subsampling and the difference estimator. SSRN Electron. J. (2015). arXiv:1507.02971
151. M. Quiroz, R. Kohn, M. Villani, M.N. Tran, Speeding up MCMC by efficient data subsampling. J. Am. Stat. Assoc. 1–13 (2018). https://doi.org/10.1080/01621459.2018.1448827
152. M. Rabinovich, E. Angelino, M.I. Jordan, Variational consensus Monte Carlo, in *Advances in Neural Information Processing Systems* (2015), pp. 1207–1215
153. W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(1), 3 (2014)
154. E. Raguseo, Big data technologies: an empirical investigation on their adoption, benefits and risks for companies. Int. J. Inf. Manag. **38**(1), 187–195 (2018)
155. C.E. Rasmussen, The infinite Gaussian mixture model, in *Advances in Neural Information Processing Systems* (2000), pp. 554–560
156. C.E. Rasmussen, Gaussian processes in machine learning, in *Advanced Lectures on Machine Learning* (Springer, Berlin, 2004), pp. 63–71
157. V. Rocková, S. van der Pas, Posterior concentration for Bayesian regression trees and forests. Ann. Stat. (in revision) 1–40 (2017). arXiv:1708.08734
158. J. Roski, G.W. Bo-Linn, T.A. Andrews, Creating value in health care through big data: opportunities and policy implications. Health Aff. **33**(7), 1115–1122 (2014)
159. J.S. Rumsfeld, K.E. Joynt, T.M. Maddox, Big data analytics to improve cardiovascular care: promise and challenges. Nat. Rev. Cardiol. **13**(6), 350–359 (2016)
160. S. Sagiroglu, D. Sinanc, Big data: a review, in *2013 International Conference on Collaboration Technologies and Systems (CTS)* (IEEE, Piscataway, 2013), pp. 42–47
161. S.M. Schennach, Bayesian exponentially tilted empirical likelihood. Biometrika **92**(1), 31–46 (2005)
162. E.D. Schifano, J. Wu, C. Wang, J. Yan, M.H. Chen, Online updating of statistical inference in the big data setting. Technometrics **58**(3), 393–403 (2016)
163. S.L. Scott, A.W. Blocker, F.V. Bonassi, H.A. Chipman, E.I. George, R.E. McCulloch (2016) Bayes and big data: The consensus Monte Carlo algorithm. Int. J. Manag. Sci. Eng. Manag. **11**(2), 78–88
164. D.V. Shah, J.N. Cappella, W.R. Neuman, Big data, digital media, and computational social science: possibilities and perils. Ann. Am. Acad. Pol. Soc. Sci. **659**(1), 6–13 (2015)
165. A. Siddiqa, A. Karim, A. Gani, Big data storage technologies: a survey. Front. Inf. Technol. Electron. Eng. **18**(8), 1040–1070 (2017)
166. P. Singh, A. Hellander, Multi-statistic Approximate Bayesian Computation with multi-armed bandits. Preprint, arXiv:180508647 (2018)
167. S. Sisson, Y. Fan, M. Beaumont, Overview of ABC, in *Handbook of Approximate Bayesian Computation* (Chapman and Hall/CRC, New York, 2018), pp. 3–54
168. U. Sivarajah, M.M. Kamal, Z. Irani, V. Weerakkody, Critical analysis of big data challenges and analytical methods. J. Bus. Res. **70**, 263–286 (2017)
169. S. Srivastava, C. Li, D.B. Dunson, Scalable Bayes via barycenter in Wasserstein space. J. Mach. Learn. Res. **19**(1), 312–346 (2018)
170. H. Strathmann, D. Sejdinovic, M. Girolami, Unbiased Bayes for big data: paths of partial posteriors. Preprint, arXiv:150103326 (2015)

171. M.A. Suchard, Q. Wang, C. Chan, J. Frelinger, A. Cron, M. West, Understanding GPU programming for statistical computation: studies in massively parallel massive mixtures. J. Comput. Graph. Stat. **19**(2), 419–438 (2010)
172. Z. Sun, L. Sun, K. Strang, Big data analytics services for enhancing business intelligence. J. Comput. Inf. Syst. **58**(2), 162–169 (2018)
173. S. Suthaharan, Big data classification: problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Perform. Eval. Rev. **41**(4), 70–73 (2014)
174. O. Sysoev, A. Grimvall, O. Burdakov, Bootstrap confidence intervals for large-scale multivariate monotonic regression problems. Commun. Stat. Simul. Comput. **45**(3), 1025–1040 (2014)
175. D. Talia, Clouds for scalable big data analytics. Computer **46**(5), 98–101 (2013)
176. Y. Tang, Z. Xu, Y. Zhuang, Bayesian network structure learning from big data: a reservoir sampling based ensemble method, in *International Conference on Database Systems for Advanced Applications* (Springer, Berlin, 2016), pp. 209–222
177. A. Tank, N. Foti, E. Fox, Streaming variational inference for Bayesian nonparametric mixture models, in *Artificial Intelligence and Statistics* (2015), pp. 968–976
178. Y.W. Teh, A.H. Thiery, S.J. Vollmer, Consistency and fluctuations for stochastic gradient Langevin dynamics. J. Mach. Learn. Res. **17**(1), 193–225 (2016)
179. D. Tran, R. Ranganath, D.M. Blei, The variational Gaussian process. Preprint, arXiv:151106499 (2015)
180. N. Tripuraneni, S. Gu, H. Ge, Z. Ghahramani, Particle Gibbs for infinite hidden Markov models, in *Advances in Neural Information Processing Systems* (2015), pp. 2395–2403
181. S. van der Pas, V. Rockova, Bayesian dyadic trees and histograms for regression, in *Advances in Neural Information Processing Systems* (2017), pp. 2089–2099
182. M. Viceconti, P. Hunter, R. Hose, Big data, big knowledge: big data for personalized healthcare. IEEE J. Biomed. Health Inform. **19**(4), 1209–1215 (2015)
183. A. Vyas, S. Ram, Comparative study of MapReduce frameworks in big data analytics. Int. J. Mod. Comput. Sci. **5**(Special Issue), 5–13 (2017)
184. S.F. Wamba, S. Akter, A. Edwards, G. Chopin, D. Gnanzou, How “big data” can make big impact: findings from a systematic review and a longitudinal case study. Int. J. Prod. Econ. **165**, 234–246 (2015)
185. X.F. Wang, Fast clustering using adaptive density peak detection. Stat. Methods Med. Res. **26**(6), 2800–2811 (2015)
186. L. Wang, D.B. Dunson, Fast Bayesian inference in Dirichlet process mixture models. J. Comput. Graph. Stat. **20**(1), 196–216 (2011)
187. X. Wang, D.B. Dunson, Parallelizing MCMC via weierstrass sampler. Preprint, arXiv:13124605 (2013)
188. T. Wang, R.J. Samworth, High dimensional change point estimation via sparse projection. J. R. Stat. Soc. Ser. B (Stat Methodol.) **80**(1), 57–83 (2017)
189. C. Wang, J. Paisley, D. Blei, Online variational inference for the hierarchical Dirichlet process, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (2011), pp. 752–760
190. J. Wang, Y. Tang, M. Nguyen, I. Altintas, A scalable data science workflow approach for big data Bayesian network learning, in *2014 IEEE/ACM Int Symp. Big Data Comput.* (IEEE, Piscataway, 2014), pp. 16–25
191. C. Wang, M.H. Chen, E. Schifano, J. Wu, J. Yan, Statistical methods and computing for big data. Stat. Interface **9**(4), 399–414 (2016)
192. C. Wang, M.H. Chen, J. Wu, J. Yan, Y. Zhang, E. Schifano, Online updating method with new variables for big data streams. Can. J. Stat. **46**(1), 123–146 (2017)
193. H.J. Watson, Tutorial: big data analytics: concepts, technologies, and applications. Commun. Assoc. Inf. Syst. **34**, 65 (2014)
194. Y. Webb-Vargas, S. Chen, A. Fisher, A. Mejia, Y. Xu, C. Crainiceanu, B. Caffo, M.A. Lindquist, Big data and neuroimaging. Stat. Biosci. **9**(2), 543–558 (2017)

195. S. White, T. Kypraios, S.P. Preston, Piecewise Approximate Bayesian Computation: fast inference for discretely observed Markov models using a factorised posterior distribution. Stat. Comput. **25**(2), 289–301 (2015)
196. R. Wilkinson, Accelerating ABC methods using Gaussian processes, in *Artificial Intelligence and Statistics* (2014), pp. 1015–1023
197. S. Williamson, A. Dubey, E.P. Xing, Parallel Markov chain Monte Carlo for nonparametric mixture models, in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (2013), pp. 98–106
198. A.F. Wise, D.W. Shaffer, Why theory matters more than ever in the age of big data. J. Learn. Anal. **2**(2), 5–13 (2015)
199. C. Wu, C.P. Robert, Average of recentered parallel MCMC for big data. Preprint, arXiv:170604780 (2017)
200. X.G. Xia, Small data, mid data, and big data versus algebra, analysis, and topology. IEEE Signal Process. Mag. **34**(1), 48–51 (2017)
201. C. Yang, Q. Huang, Z. Li, K. Liu, F. Hu, Big data and cloud computing: innovation opportunities and challenges. Int. J. Digit Earth **10**(1), 13–53 (2017)
202. C. Yoo, L. Ramirez, J. Liuzzi, Big data analysis using modern statistical and machine learning methods in medicine. Int. Neurourol. J. **18**(2), 50 (2014)
203. L. Yu, N. Lin, ADMM for penalized quantile regression in big data. Int. Stat. Rev. **85**(3), 494–518 (2017)
204. T. Zhang, B. Yang, An exact approach to ridge regression for big data. Comput. Stat. **32**, 1–20 (2017)
205. X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, J. Chen, A hybrid approach for scalable subtree anonymization over big data using MapReduce on cloud. J. Comput. Syst. Sci. **80**(5), 1008–1020 (2014)
206. Y. Zhang, T. Cao, S. Li, X. Tian, L. Yuan, H. Jia, A.V. Vasilakos, Parallel processing systems for big data: a survey. Proc. IEEE **104**(11), 2114–2136 (2016)
207. Z. Zhang, K.K.R. Choo, B.B. Gupta, The convergence of new computing paradigms and big data analytics methodologies for online social networks. J. Comput. Sci. **26**, 453–455 (2018)
208. L. Zhang, A. Datta, S. Banerjee, Practical Bayesian modeling and inference for massive spatial data sets on modest computing environments. Stat. Anal. Data Min. **12**(3), 197–209 (2019)
209. L. Zhou, S. Pan, J. Wang, A.V. Vasilakos, Machine learning on big data: Opportunities and challenges. Neurocomputing **237**, 350–361 (2017)
210. J. Zhu, J. Chen, W. Hu, B. Zhang, Big learning with Bayesian methods. Natl. Sci. Rev. **4**(4), 627–651 (2017)
211. G. Zoubin, Scaling the Indian Buffet process via submodular maximization, in *International Conference on Machine Learning* (2013), pp. 1013–1021