Biometrics & Biostatistics

Young et al., J Biomet Biostat 2015, S12
http://dx.doi.org/10.472/2155-6180.S12-e001

Editorial Open Access

Some Flexible Modeling Paradigms for Analyzing Big Data

Derek S Young[1]*, Limin Feng[1,2] and Richard J Charnigo[1,3]

[1]Department of Statistics, University of Kentucky, Lexington, KY, USA
[2]Intel Corporation, Hillsboro, OR, USA
[3]Department of Biostatistics, University of Kentucky, Lexington, KY, USA

Introduction

Data analysts — whether in industry, government, or academia – are faced with increasingly large datasets, or *big data*. Some examples of discourse on big data include its impact on marketing analytics [1], official statistics [2], and biomedical research [3]. However, when data actually becomes *big* is subjective. This can refer to a large number of records, a large number of measured variables, or both. Nonetheless, the growing number of data collection strategies, as well as increases in computational efficiency and storage, have resulted in big data being relatively cheap to obtain and manage. However, it is fundamentally more important to assess how well that data is being used, whether for classification, prediction, or modeling. We focus on the last of these goals and highlight some flexible modeling paradigms that can be helpful for analyzing big data.

Nonnormal Parametric Models

Many classical statistical modeling paradigms for continuous data (e.g., linear regression modeling) assume normality, either because practitioners characterize the randomness of their observed data using a normal distribution or because they appeal to the Central Limit Theorem. However, such an assumption may not be strictly or even approximately valid. A pertinent issue in this regard is the deviation from normality that can be tolerated before one must abandon such a classical paradigm; this may vary from one scientific application to another. A complicating factor is that, as the sample size grows, the power of almost any test for detecting non-normality will increase; thus, smaller deviations from normality will be perceptible, and one must consider whether such deviations truly warrant abandoning a classical paradigm.

In some cases, they will. And, in these cases, the nature of the scientific application from which the data arise may suggest other parametric distributions to consider. For example, right-skewed distributions are often used for lifetime data, which in turn are usually parameterized via hazard rate functions. One such right-skewed distribution is the Rayleigh distribution, which has been used to model background data from magnetic resonance images [4]. Another example is the generalized extreme value (GEV) distribution, which has a cumulative distribution function of the form

$$F_X(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \quad (1)$$

and is defined on the set $\left\{x : 1+\frac{\xi(x-\mu)}{\sigma} > 0\right\}$ with parameter space $\left\{\mu, \sigma, \xi : \mu, \xi \in \mathbb{R}, \sigma \in \mathbb{R}^+\right\}$, though the above formula must be modified for $\xi = 0$. This distribution is often used to characterize the stochastic behavior of a process at unusually large (or small) values, such as for novelty detection in medical screening [5]. Note that the GEV distribution function also includes the Gumbel, Fréchet, and Weibull families as special cases; see the text by [6] for a thorough discussion.

For multivariate data also, not constraining oneself to a normality assumption may be potentially advantageous. Indeed, while random generation, estimation, and testing under multivariate normality are well-studied, greater flexibility and fidelity to underlying scientific phenomena may be achieved without such an assumption. Related to this, [7] presents an approach for joint random generation of binary and nonnormal continuous variables in this special issue of the journal. The author partitions the correlation matrix of such variables into three components: one for the binary variates only, one for the nonnormal continuous variates only, and one governing pairs in which one variable is binary and the other is nonnormal continuous. The author derives modified versions of each component and reassembles them into a second correlation matrix, from which multivariate normal data are then randomly generated. Finally, these multivariate normal data are converted to binary variables and nonnormal continuous variates respecting the original correlation matrix via thresholding and a power polynomials procedure rooted in the work of [8]. [7] demonstrates the efficacy of this strategy through a numerical study based on data from the National Institute of Mental Health Schizophrenia Collaborative Study [9]. The reported biases for correlation measures, moment-based quantities, and regression coefficients are all relatively small under this strategy.

Finite Mixture Models

When the sampled data arise from a population that consists of several homogenous subpopulations, then one can use finite mixture models to characterize the data (e.g., [10]). Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ denote a random sample of size n such that $\mathbf{Y}_i \in \mathbb{R}^d$, i = 1, . . . , n, and let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ denote the corresponding realizations. Assume further that these data are drawn from a population consisting of k<∞ subpopulations. The *k-component mixture density* for these data is written as

$$f(\mathbf{y}_i) = \sum_{j=1}^{k} \lambda_j g_j(\mathbf{y}_i) \quad (2)$$

where $\lambda_j = 0$ and $\sum_{j=1}^{k} \lambda_j = 1$ are *mixing proportions* and the $g_j(\cdot)$ are *component densities*. Usually the component densities are taken from a known parametric family, with the value of the parameter regarded

***Corresponding author:** Derek S Young, Assistant Professor, Department of Statistics, 323 Multidisciplinary Science Building, University of Kentucky, USA, Tel: (859)-218-3408; E-mail: derek.young@uky.edu

Received October 30, 2014; **Accepted** December 09, 2014; **Published** January 21, 2015

Citation: Young DS, Feng L, Charnigo RJ (2015) Some Flexible Modeling Paradigms for Analyzing Big Data. J Biomet Biostat S12: e001. doi:10.4172/2155-6180.S12-e001

as unknown in at least one (and perhaps every) component. There are many mixture model applications to biomedical research problems, such as studies involving phenylthiocarbamide (PTC) sensitivity in people [11], screening for hemochromatosis [12], and assessing prostate cancer clinical trial data [13].

Note that a k'-component mixture can be re-expressed as a k-component mixture for any $k' \in \{1, \ldots, k-1\}$, simply by adding more components with zero mixing proportions and/or equal parameter values in the component densities. Thus, if one is uncertain that a k-component mixture provides the most parsimonious characterization of the data among all that would be deemed acceptable, one may be interested in testing a null hypothesis that the k-component mixture could be reduced to a k'-component mixture. In this special issue [14] study a modified likelihood ratio test for two-component binomial mixture models, in which k'=1, k=2, $g_1 = Bin(m, 0.5)$ for known m, and $g_2 = Bin(m, \theta)$ for unknown $\theta \in [0, 0.5]$. More specifically, they derive the limiting distribution of the test statistic under two local alternatives, one in which θ is close to 0.5 and one in which λ_2 is close to 0. Their work is motivated by a genetic linkage study for schizophrenia, in which the local alternatives correspond to weakness and rarity of the linkage, respectively.

For the remainder of this section, let d=1 so that $Y_1, \ldots, Y_n$ are univariate. In this special issue [15] discuss testing procedures for the *bilaterally contaminated normal with nuisance parameter (BCN+NP) model*. The density for this model is

$$f(y) = (1-\lambda_1-\lambda_2)\frac{1}{\sigma}\phi\left(\frac{y}{\sigma}\right) + \sum_{j=1}^{2}\frac{1}{\sigma}\lambda_j\phi\left(\frac{y-\mu_j}{\sigma}\right) \quad (3)$$

where $\phi(\cdot)$ is the standard normal density, $\mu_1 \geq 0 \geq \mu_2$, $\sigma^2>0$, and all parameters are unknown. More specifically [15] provide asymptotic and simulation results regarding two hypothesis tests:

$$H_0: \lambda_1\mu_1=0 \text{ and } \lambda_2\mu_2=0 \text{ versus } H_1: \lambda_1\mu_1 \neq 0 \text{ or } \lambda_2\mu_2 \neq 0 \quad (4)$$

and

$$H_0: \lambda_1\mu_1=0 \text{ or } \lambda_2\mu_2=0 \text{ versus } H_1: \lambda_1\mu_1 \neq 0 \text{ and } \lambda_2\mu_2 \neq 0 \quad (5)$$

These are referred to as testing the *omnibus null hypothesis* and *unliateral null hypothesis*, respectively. The former procedure employs sample moments in a union-intersection test, while the latter procedure is based on sample moments and an auxiliary estimator of the nuisance parameter σ^2. The procedures are demonstrated on *logarithm of the odds (LOD) scores* in a whole genome linkage analysis from an autism study.

When each observation Y_i is measured along with a vector of covariates, say $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})^T$, the conditional distribution for $Y_i|\mathbf{X}_i$ may be characterized by a *mixtures-of-regressions model*. Mixtures of regressions have been applied to diverse data problems involving music perception [16], viral propagation by aphids [17], and protein structures of DNA [18]. There are also extensions to the mixtures-of-regressions model that aim at greater flexibility. For example [19] developed a mixtures-of-regressions model in which the components have changepoints. Maximum likelihood estimation for this model is accomplished using an expectation-conditional maximization (ECM) algorithm [20] where the parameter vector of interest is partitioned and optimized in a series of conditional maximization steps. In this special issue [21] investigate mixtures of self-modeling regressions for flexibility in describing functional data. The starting point for their work is the *shape invariant model*, in which functions $f_1, \ldots, f_n$ are defined by affine transformations of a common shape function g,

$$f_i(x) = a_i g(c_i x + d_i) + b_i, \quad (6)$$

where $\boldsymbol{\theta}_i=(a_i, b_i, c_i, d_i)^T$ is a vector of self-modeling coefficients. [21] induce a mixture structure by proposing that each f_i transform one of k possible shape functions $g_1, \ldots, g_k$ instead of a common shape function. Their approach to inference entails Bayesian adaptive regression splines [22] and is illustrated in an application to synaptic transmission data, in which components in the mixture structure may correspond to different active zones in a synapse.

Local Models

Greater flexibility in describing data can also be achieved using local models. Here we refer to models in which the probabilistic structure is not described a priori by a finite specified set of parameters but rather is estimated at a given location using mainly those observations in proximity to the location. For example, consider estimating the probability density function f(·) underlying d–dimensional realizations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. A kernel density estimate (e.g., [23]) puts

$$\hat{f}_{\mathbf{H}}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} |\mathbf{H}|^{-1/2} K\left(\mathbf{H}^{-1/2}(\mathbf{y}_i - \mathbf{y})\right) \quad (7)$$

where $\mathbf{H}$ is a d×d bandwidth matrix, symmetric and positive definite, and the non-negative kernel function K satisfies $\int_{\mathbb{R}^d} K(\mathbf{t})\,d\mathbf{t} = 1$ and $K(\mathbf{t}) = K(-\mathbf{t})$ for all $\mathbf{t}$. One chooses $\mathbf{H}$ so that, at the location $\mathbf{y}$, observations with $\mathbf{y}_i$ close to $\mathbf{y}$ play a greater role in estimating f($\mathbf{y}$) than observations with $\mathbf{y}_i$ far from $\mathbf{y}$.

Local models can also be used in regression settings [24], in which case interest lies in locally estimating a mean response function and perhaps its derivatives as well. A nonparametric regression model has a form such as

$$Y_i = \mu(\mathbf{x}_i) + \varepsilon_i \quad (8)$$

where μ(·) is an unknown mean response function and we have again assumed that d=1, so that the responses are univariate and are measured along with a (vector of) covariate(s). There are various local procedures that one can use to estimate μ(·). For example, local averaging moves a window continuously over the data and averages the observations that fall within that window. Locally-weighted averaging (or kernel regression) modifies (7) to

$$\hat{\mu}(\mathbf{x}) = \frac{\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}_i, \mathbf{x})\mathbf{y}_i}{\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}_i, \mathbf{x})} \quad (9)$$

where $K_{\mathbf{H}}(\mathbf{t}, \mathbf{t}_0) = |\mathbf{H}|^{-\frac{1}{2}} K\left(|\mathbf{H}|^{-\frac{1}{2}}(\mathbf{t} - \mathbf{t}_0)\right)$.

One can gain even more flexibility by employing a local polynomial structure [25] and solving a minimization problem such as

$$\arg\min_{\beta_0, \boldsymbol{\beta}_1} \sum_{i=1}^{n} \left\{y_i - \beta_0 - \boldsymbol{\beta}_1^T(\mathbf{x}_i - \mathbf{x})\right\}^2 K_{\mathbf{H}}(\mathbf{x}_i, \mathbf{x}) \quad (10)$$

The minimizing β_0 becomes the estimate of μ($\mathbf{x}$), while the minimizing $\boldsymbol{\beta}_1$ becomes the estimate of the (vector of partial) derivative(s) of μ(·) evaluated at $\mathbf{x}$. Using a local polynomial of degree one may suffice for some applications, while a local polynomial of higher degree may be desirable for other applications due to a reduction in estimation bias.

There also exist other nonparametric regression procedures.

For example [26] discussed a paradigm that allows simultaneous estimation of the mean response function and its derivatives when there is a single covariate. This is accomplished using a compound estimator, which is *self-consistent* (i.e., the estimates of the derivatives equal the derivatives of the estimated mean response function) and achieves essentially optimal convergence rates in consistency. On the other hand, minimizing (10) does not yield self-consistent estimates.

One area that relies on local modeling is *image analysis*, which includes big data problems requiring rapid or even real-time solutions. For example [27] proposed a semi-local paradigm that divides a volumetric image into blocks and then applies wavelet denoising to the blocks individually before re-assembling them. Their approach was illustrated using a noisy phantom positron emission tomography (PET) image and found to outperform a competing method for image processing. Wang and Ye [28] developed a nonparametric test for comparing a group of images or multivariate local regression surfaces. The authors illustrated their procedure on medical rehabilitation data from a neuro-muscular electrical stimulation experiment.

In this special issue [29] discusses nonrigid image registration, which maps each pixel from one image to the corresponding pixel of another image, in such a way that local distortions can be accommodated; the basic idea is to alter a *template image* so that it more closely conforms to a reference image. (The artistic-minded reader may try to visualize what Salvador Dali's timepieces might have looked like before they melted; the template image could be of one of the melted timepieces, and the reference image could be of a similar but non-melted instrument.) More specifically [29] presents fluid registration methodology, so named because the local distortions may resemble fluid flow and potentially useful in medical applications in which, for instance, different people's brains are imaged. [29] also assesses two image similarity measures: the sum of squared intensity differences (SSD) and mutual information (MI). The computation of MI uses a univariate version of (7) with a truncated normal kernel function.

Incorporating a kernel structure into mixture models (thus making them local models) is also possible. For example, local approaches in mixture models appear in the literature as nonparametric mixture models [30] and even as mixtures of regressions ([31-33]). While local models are often termed nonparametric (or semiparametric), a sort of parametric form may be induced by the estimation method. For example, the compound estimator of [26] is a normalized Gaussian convolution of polynomials and so can be represented in terms of a finite number of parameters, mainly the polynomial coefficients; however, the parameters are a feature of the estimation method rather than of the model itself, and as such the typical results for parametric statistical inference (e.g., $\sqrt{n}$ -consistency) are not available. Be that as it may, local models offer tremendous flexibility in describing big data. However, local models can also break down in high-dimensional settings due to the *curse of dimensionality* [34], from which big data is not immune.

Conclusion

We discussed three flexible modeling paradigms that can be helpful for analyzing big data: nonnormal parametric models, finite mixture models, and local models. As highlighted above, all three paradigms have enjoyed successes in various biomedical applications. However, the utility of these paradigms is by no means relegated to big data settings; on the other hand, they may be particularly valuable when attempting to characterize the sorts of complex relationships that are often present in big data. While we addressed only a few flexible modeling paradigms, the researcher's recognition of assumptions for and limitations of the chosen approach to data analysis should not be restricted to these particular paradigms. Indeed, even with flexible modeling, misspecifications are still possible and may be more detrimental with big data than with small data.

References

1. Ratner B (2005) Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data. Chapman & Hall/CRC, Boca Raton, FL, USA.
2. Capps C, Wright T (2013) Toward a vision: Official statistics and big data. AMSTAT News, 434: 9-13.
3. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, et al. (2008) Big data: The future of biocuration. Nature 455: 47-50.
4. Sijbers J, den Dekker AJ, Raman E, Van Dyck D (1999) Parameter estimation from magnitude mr images. International Journal of Imaging Systems and Technology 10: 109-114.
5. Roberts SJ (2000) Extreme value statistics for novelty detection in biomedical data processing. IEE Proceedings–Science Measurement and Technology 147: 363-367.
6. Coles S (2001) An Introduction to Statistical Modeling of Extreme Values. Springer, London.
7. Demirtas H (2014) Joint generation of binary and nonnormal continuous data. Journal of Biometrics and Biostatistics 5:199.
8. Fleishman AI (1978) A method for simulating non-normal distributions. Psychometrika 43: 521-532.
9. Hedeker D, Gibbons RD (1997) Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychological Methods, 2: 64-78.
10. McLachlan GJ, Peel D (2000) Finite Mixture Models. Wiley, New York.
11. Jones PN, McLachlan GJ (1991) Fitting mixture distributions to phenylthiocarbamide (ptc) sensitivity. American Journal of Human Genetics 48: 117-120.
12. McLaren CE, McLachlan GJ, Halliday JW, Webb SI, Leggett BA, et al. (1998) The distribution of transferrin saturation in hereditary haemochromatosis in Australians. Gastroenterology 114: 543-549.
13. Hunt LA, Jorgensen MA (1999) Mixture model clustering: A brief introduction to the MULTMIX program. Australian and New Zealand Journal of Statistics 41: 153-171.
14. Fu Y, Li P, Chung S (2014) Sample size calculation for the modified likelihood ratio test in genetic link analysis. Journal of Biometrics and Biostatistics, 5: 205
15. Fan Q, Charnigo RJ, Talebizadeh Z, Dai H (2014) Hypothesis testing in normal admixture models to detect heterogeneous genetic signals. Journal of Biometrics and Biostatistics. 5: 213.
16. DeVeaux RD (1989) Mixtures of linear regressions. Computational Statistics and Data Analysis, 8: 227-245.
17. Turner TR (2000) Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. Applied Statistics 49: 371-384.
18. Martin-Magniette ML, Mary-Huard T, B´erard C, Robin S (2008) ChIPmix: Mixture model of regressions for two-color ChIP-chip analysis. Bioinformatics 24: 181-186.
19. Young DS (2014) Mixtures of regressions with changepoints. Statistics and Computing, 24: 265-281.
20. Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80: 267-278.
21. Szczesniak RD, Viele K, Cooper RL (2014) Mixtures of self-modeling regressions. Journal of Biometrics and Biostatistics, 5: 208.
22. DiMatteo I, Genovese C, Kass R (2001) Bayesian curve fitting with free knot splines. Biometrika, 88: 1055-1071.

23. Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall/CRC, London, UK.

24. Härdle W (1990) Applied Nonparametric Regression. Cambridge University Press, Cambridge, UK.

25. Loader C (1999) Local Regression and Likelihood. Springer-Verlag.

26. Charnigo RJ, Srinivasan C (2011) Self-consistent estimation of mean response functions and their derivatives. The Canadian Journal of Statistics 39: 280-299.

27. Charnigo RJ, Sun J, Muzic R (2006) A semi-local paradigm for wavelet denoising. IEEE Trans-actions on Image Processing 15: 666-677.

28. Wang XF, Ye D (2010) On nonparametric comparison of images and regression surfaces. Journal of Statistical Planning and Inference 140: 2875-2884.

29. Huang X (2014) Nonrigid image registration problem using fluid dynamics and mutual information. Journal of Biometrics and Biostatistics. 5: 212.

30. Lindsay BG (1995) Mixture Models: Theory, Geometry and Applications, volume 5 of NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics and the American Statistical Association.

31. Hunter DR, Young DS (2012) Semiparametric mixtures of regressions. Journal of Nonparametric Statistics 24: 19-38.

32. Huang M, Li R, Wang S (2013) Nonparametric mixture of regression models. Journal of the American Statistical Association 108: 929-941.

33. Bordes L, Kojadinovic I, Vandekerkhove P (2013) Semiparametric estimation of a two-component mixture of linear regressions in which one component is known. Electronic Journal of Statistics 7: 2603-2644.

34. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY.

Citation: Young DS, Feng L, Charnigo RJ (2015) Some Flexible Modeling Paradigms for Analyzing Big Data. J Biomet Biostat S12: e001. doi:10.4172/2155-6180.S12-e001

This article was originally published in a special issue, **Big Data and Flexible Modeling** handled by Editor. Richard Charingo, University of Kentucky, USA