Comments on "Visualizing Statistical Models": Visualizing Modern Statistical Methods for Big Data

Genevera I. Allen[1,2,3], Frederick Campbell[1], and Yue Hu[1]

[1] Department of Statistics, Rice University

[2] Department of Electrical and Computer Engineering, Rice University

[3] Department of Pediatrics-Neurology, Baylor College of Medicine
& Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital

April 30, 2015

The authors are to be congratulated on this interesting and salient guide and review of visualization techniques and their role in the process of fitting statistical models. The introduction and illustrations of three main strategies, (i) visualizing the model in data space, (ii) visualizing the model fitting process, and (iii) visualizing collections, are important guiding principles. The manuscript, however, illustrates these principles on small or moderate sized data sets and for the purpose of model fitting and data analysis. These visualization strategies can be important in other contexts. Here, we discuss how the three visualization strategies introduced can be used for modern statistical methods developed for Big Data. Additionally we argue that when it comes to large and complex data, visualization should be an important part of the research process, and research developments are integral to improving the visualization process. As such, we outline many open research questions related to visualizing statistical machine learning methods and their application to Big Data.

Visualizing the Process of Model Fitting. When working with Big Data, some form of model sparsity or regularization is often necessary. Sparsity comes in many forms including sparsity in feature space as with sparse regression models, sparsity in observation space as with support vector machines, or sparsity in model dimension as with reduced-rank models. In order to fit any of these sparse statistical learning models, large-scale iterative optimization techniques must be employed. Visualization can be an important tool for understanding the properties of these sparse models and optimization techniques, and the insights gained from visualization can stimulate new research directions. We provide an example from our own research of how visualizing the iterates of sparse optimization techniques, as suggested in the manuscript, led to interesting insights and ultimately the development of a novel framework for sparse statistical learning.

Most sparse statistical machine learning methods seek to optimize a trade-off between a loss function $\mathcal{L}$ and a sparse penalty P: $\text{minimize}_{\boldsymbol{\beta}}\ \mathcal{L}(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}) +\ \lambda P(\boldsymbol{\beta})$. The regularization parameter, λ, governs this trade-off; typically, one applies iterative optimization techniques to solve for $\boldsymbol{\beta}$ over a range of λ values. A popular way to visualize this collection of sparse models is via regularization paths (Efron et al., 2004) which plot $\hat{\boldsymbol{\beta}}_j(\lambda_k)$ for all $\lambda_1 < \lambda_2 <$

$\ldots < \lambda_{max}$ and all coefficients $j = 1, \ldots p$. A couple of years ago when we were playing with a particular optimization method, the Alternating Direction Methods of Multipliers (ADMM) algorithm, we tried plotting the coefficient paths of the algorithm iterates (k) for a fixed value of λ, $\hat{\boldsymbol{\beta}}_j^k(\lambda)$, as opposed to the solutions (final iterates) over a range of λ values. In visualizing this model fitting process, we noticed something most interesting: the coefficient paths of ADMM algorithm iterates for one value of λ were very similar to regularization paths over a range of λ values!

In this case, visualizing the model fitting process led to an appealing research proposition: Instead of running the sparse optimization algorithm many times over a range of λ values, could we run the optimization algorithm (possibly tweaked) once and obtain the same or a similar sequence of sparse models? Surprisingly, the answer to this question turns out to be yes! We term this strategy the *Algorithmic Regularization Path* as we trace the coefficient paths of optimization algorithm iterates to achieve the effect of regularization paths. First, Hu and Allen (2015a) uses this strategy for two-way smoothly regularized regression showing empirically (we have later proven this theoretically; manuscript in preparation) that the Algorithmic Regularization Path is *exactly* equivalent to the regularization path for this problem. Hu et al. (2015) outlines a general strategy for generating Algorithmic Regularization Paths for sparse statistical machine learning problems, while Hu and Allen (2015b) expands on this for the well-studied but important problem of sparse regression. For the later, we show that our Algorithm Path strategy not only leads to major computational savings, but it also yields dramatic improvements in variable selection accuracy for the challenging case of highly-correlated high-dimensional data. Further, our strategy is general in that it can be applied to many sparse optimization algorithms and many statistical learning procedures, thus, paving the way for a novel framework with many potential benefits and fruitful research directions.

Overall, this example from our own work nicely illustrates the importance of visualization as a catalyst for theory and methods research and the particular strategy of visualizing the model fitting process for gaining further insight into statistical methods. Beyond this and other examples discussed in the manuscript, visualizing the iterates of optimization algorithms commonly used in statistical machine learning could be particularly useful for understanding online learning algorithms and distributed optimization. In the former, the fitted model evolves over time as data streams; visualizing algorithm iterates would give insights into how the data and model change with time. For distributed optimization, typically optimization algorithms begin by fitting models to each distributed data source and as the algorithm progresses, it obtains a consensus of the distributed models; visualizing algorithm iterates would give insights into both the commonalities and differences across the distributed data sources.

Visualizing Members of a Collection. Many examples of modern statistical methods used to analyze Big Data form a collection of models. Regularization methods, discussed previously, form a collection of models indexed by the regularization parameter, λ. Visualizing this collection by plotting the model coefficients over the entire range of λ, or regularization paths as previously discussed, are a popular way to visualize sparse regression models. Visualizing regularization paths for more complex penalties that enforce structured sparsity or low-rank structure, for example, is less common and an open area of research.

Perhaps more interesting, however, is the question of how to visualize members of a collection of random ensembles. Many statistical machine learning procedures use some form of randomization to deal with complex and large data sets. These include random forests, random penalization, stability selection, bootstrapping, random projections, random corruptions, and some forms of hashing, among many others. Often these random ensemble methods are used as black-box methods with only the mean summary statistic reported; it is yet uncommon to investigate the individual members of the ensemble. This can be extremely important, however, for improving prediction accuracy, model interpretability, as well as yielding possible insights into how to further extend and improve these methods. While the manuscript discusses summary statistics for visualizing random forests, it is often unclear how or what to visualize for each of the random ensembles as the collection of models as well as each model dimension can be large.

To motivate the importance of visualizing the random collection and especially reporting more than just the mean of an ensemble, consider the example of sparse regression via the Lasso, or ℓ_1-penalized regression. Recently, several have proposed to use random perturbations to improve variable selection accuracy (Meinshausen and Bühlmann, 2010). Either bootstrapping or sub-sampling is used to perturb the observation space while random penalties can be used to perturb the feature space. For each of these perturbations, the Lasso is fit and the selected variables are recorded. Then, the variables most stable to perturbations, or the variables most frequently selected in the ensemble, are most likely the true variables. Typically, one simply reports the proportion of times each variable is selected and takes the variables with the highest aggregated proportion to be the selected set. This however, throws away important information about each model in the collection, and can lead to erroneous results when there is a high degree of colinearity. Consider for example, two very important predictors that are highly correlated. Then each Lasso fit in the collection will tend to pick only one of these important predictors, which will each then have stability proportions around 50%. Hence, these two very important predictors would likely not be selected by simply looking at the mean of the ensemble. Further, reporting the most often selected variables does not ensure that these set of variables form the best or even a good multivariate model for predicting the response. Because of this, visualization strategies are needed to understand the relationships between variables selected in each member of the random collection. Perhaps, as suggested by the manuscript creating a set of summary statistics beyond the mean to visualize in a linked manner would be more useful. This example represents just one of many possible new frontiers for visualizing complex statistical machine learning methods that form an ensemble based on randomization.

Visualizing the Model in Data Space. While the manuscript makes a case for visualizing the model in data space as a means to better understand the model fit, this strategy is perhaps the most difficult and unrealistic to apply in Big Data settings. With huge samples sizes, scatterplots become too dense to see anything of interest and how to best subsample the observations is an open question. With large numbers of features, the data dimension is so large that finding "interesting" projections of the data to plot is a major open challenge. It would thus be a challenge to use the grand tour or a guided tour to visualize the data in these scenarios. But even beyond the traditional "large n" or "large p" settings, there are a host of Big Data problems that arise in which direct data visualization is all but impossible:

How do we visualize data stored in a distributed manner? How do we visualize mixed data in which features may be of different types (e.g. continuous, categorical, ordinal, counts, text, geographic coordinates, time-stamps etc.)? How do we visualize multi-modal data, or disparate but coupled data sets collected at the same time or over the same subjects? How do we visualize tensor data that comes in the form of a higher-order data cube? How do we visualize streaming data that is so big, that the raw data cannot be stored over time? How do we visualize private data or data with differential privacy levels?

In these situations, we would suggest that models may be the key to visualize and explore these types of complex Big Data sets. In particular, if one can apply and visualize many different exploratory data models and glean insights from each, then one could gain a "picture" of the data that would be otherwise impossible to visualize directly. Perhaps techniques introduced in the manuscript such as linked brushing could be used to understand and visualize connections between many of these exploratory models. But even basic exploratory models such as dimension reduction and clustering techniques are non-trivial in many Big Data settings. Thus, further methodological research is needed to develop new statistical machine learning models for these more complex types of Big Data.

Overall, there are many challenging open research problems related to visualizing Big Data and visualizing modern statistical models used to analyze Big Data. There also needs to be a close link between visualization and theory and methods development: As new statistical machine learning techniques are developed, visualization can aid in understanding properties of existing techniques and suggest new techniques. Further, newly developed statistical methods should serve as models to guide visualization of data that is otherwise difficult or impossible to see.

References

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

Hu, Y. and G. I. Allen (2015a). Local-aggregate modeling for big-data via distributed optimization: Applications to neuroimaging. *(Accepted pending minor revisions) Biometrics, arXiv:1405.0629*.

Hu, Y. and G. I. Allen (2015b). A new approach to variable selection via algorithmic regularization paths. In preparation.

Hu, Y., E. C. Chi, and G. I. Allen (2015). Admm algorithmic regularization paths for sparse statistical machine learning. *arXiv:1504.06637*.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.