Bayesian Analysis (2014) **9**, Number 1, pp. 29–38

Comment on Article by Rubio and Steel

Robert E. Weiss [*] and Marc A. Suchard [†]

Abstract. We discuss Rubio and Steel (2014). We discuss whether Jeffreys priors are worth the attention given to them, then move on to discuss the concepts of valid Bayesian inference and benchmark Bayesian inference. We briefly investigate the skew-normal and skew-$t(4)$ models for variables in the Australian Institute of Sport (AIS) data to investigate the range of estimates that occur for the skewness parameter. The discussion closes by wondering whether we shouldn't just use a Dirichlet Process Mixture instead of a skew-normal or skew-t.

Keywords: Jeffreys prior, Valid Bayesian inference, Benchmark Bayesian inference, Skew normal, Split normal

1 Jeffreys Prior

The Jeffreys prior is certainly model dependent, is definitely design dependent, may depend on predictors, and is often horribly improper. The Jeffreys prior can lead to improper posteriors; when producing a proper posterior, the posterior may not have any finite moments (Ibrahim and Laud 1991), the resulting posterior mean can be inadmissable. Eaton and Sudderth (1998) show that the predictive distribution in multivariate normal regression based on a Jeffreys prior (Geisser 1965; Keyes and Levy 1996) is strongly inconsistent where among other things, strongly inconsistent means that any prediction based on the Jeffreys prior is 'far away' from predictions based on every possible Bayesian analysis using a proper prior (Eaton and Sudderth 1998).

One has to wonder, other than the beauty of the invariance argument, why we continue to spend time with the Jeffreys prior. We start having problems in models with two parameters. In the independent and identically distributed (iid) normal model, $y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, already Jeffreys appeared to not like the Jeffreys prior $p(\mu, \sigma^2) \propto 1/\sigma^2$. In multivariate parameter models, Jeffreys definitely didn't like the Jeffreys prior. His comments in Jeffreys (1998, page 182) about the one way analysis of variance (ANOVA) model with more than two independent means make this clear.

The trivariate normal $y_i|\mu \sim N_3(\mu, I)$, $i = 1, \ldots, n$ with unknown mean μ and identity covariance matrix has Jeffreys prior $p(\mu) \propto 1$ but gives us $\mu|Y \sim N(\bar{Y}, I/n)$, whose posterior mean is not admissible (Stein 1956). Or the random intercept model, $y_{ij}|\mu, \beta_i, \sigma^2 \sim N(\mu + \beta_i, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, \ldots, J_i$ with $\beta_i|D \sim N(0, D)$, where Jeffreys as applied by Tiao and Tan (1965) gives a prior that is awkwardly dependent on the experimental design. The Jeffreys prior for the negative binomial and the binomial

[*] Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA 90095-1772, robweiss@ucla.edu

[†] Departments of Biomathematics, Biostatistics and Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, Los Angeles, CA 90095, msuchard@ucla.edu

DOI:10.1214/14-BA870

give different priors for the probability parameter, hence the Jeffreys prior is dependent on the investigator's intentions prior to actually performing the experiment.

If funders of medical research required that we present our priors as part of the body of the grant, long before data collection, would routine use of a Jeffreys prior even be possible?

Rubio and Steel (2014)'s (hereafter RS14) nicely laid out and entertaining analysis of the Jeffreys priors for the split/skew/two-piece/two-sided location scale model adds further gist to our shill. We're glad to be made aware of the problem here with Jeffreys, and especially the explanation in section 3.3. This problem of 1/variance being non-integrable on intervals with 0 as one endpoint is familiar to statisticians who have studied the random intercept model with its two variances (for example, Hobert and Casella 1996; Geyer 1992). RS14's example demonstrates the same problem in a non-hierarchical univariate sampling model.

2 Valid Bayesian Inferences

RS14 define "valid Bayesian inference" as do many others, that the resulting posterior is proper. This definition suffers from lack of ambition. We would prefer a stronger definition, perhaps that the resulting posterior leads to admissible estimators, or that the resulting posterior actually approximates the posterior based on a vague proper prior that might represent a naive analyst's actual proper prior. Given that mathematical statisticians need something to do to keep them out of trouble, this last option could provide plenty of interesting projects to chew on, as in Edwards et al. (1963). By a 'naive analyst' we mean data-naive and subject-naive, not statistically-naive or data-analytically-naive. A naive data analyst is one working on her first data analysis in the particular subject area (but definitely not on her first analysis) and therefore having no or at most minimal experience, (i.e. minimal prior information) in the particular subject area.

A better definition of a "valid Bayesian inference" might be a practical analysis using finite time and resources that approximates a hypothetical "true Bayesian analysis" that a data analyst would perform given infinite time, infinite resources, infinite introspective power and zero opportunity cost. We acknowledge that a "true Bayesian analysis" may not exist, or if it does, it is likely a subjective concept, but it is an interesting concept to contemplate.

3 Benchmark Bayesian Inferences

A claim by RS14:

> The availability of a "benchmark" Bayesian analysis is thus of particular importance for practitioners.

We understand early statements like this are for quickly positioning a paper so that the authors can get on to the fun and important material in the remainder of the paper. We have those statements in our papers too. But what is the point of a discussion if not to question obvious statements that on second thought are not so obvious? A benchmark analysis sounds desirable in principle, but like a "true Bayesian analysis", a claim that an analysis (or model, likelihood or prior) is a benchmark analysis (or model, likelihood or prior) is a deeply subjective claim.

In a world calling for a benchmark prior for the skewness parameter, the benchmark prior should be used at most once. Having estimated the skewness parameter one time, the information garnered in the first analysis can and should be used to help set the prior in the second analysis.

An analysis we might consider as a candidate for benchmark analysis is to report "0" as our estimate with standard deviation 0. This might seem a crappy benchmark analysis if you think (subjectively!) there is an effect, but it is quick and cheap and therefore has high utility, not to mention good reproducibility. Given current worries about published research (Ioannidis 2005; Johnson 2013), perhaps it should be considered more often. In the non-scientific sphere, when friends make claims about the latest health food fad, that such and such (American) politician is a [insert favorite insult here], what some celebrity did, or how good a diet is, we use the zero benchmark analysis as a better description of the world: the fad doesn't improve health, the politician isn't a [favorite insult here], the celebrity didn't do what was reported, and the diet doesn't work.

One assumption underlying statements about the need for a "benchmark prior" is that the likelihood/sampling density has been agreed upon and all that remains is to select a prior that can slip past readers without raising alarms. If the likelihood were in fact agreed upon, would statisticians be writing papers on skew-normal distributions? Entire disciplines engage in discussions on what covariates need to be mandatorily included in regression analyses of particular outcomes; these discussions evolve as the science develops. Many subject matter journals have roadblocks before accepting unfamiliar analyses such as Bayesian or hierarchical models that are not well represented in their archives. Likelihoods, sampling densities and statistical analyses are not agreed upon, and analysts can not assume that a model, likelihood, or sampling density will be acceptable to readers. And this applies to benchmark priors too.

More palatable as a benchmark prior is an informative hierarchical prior that could be agreeable by many people as being sensible. This takes time and data to construct. Suppose we have data sets and analyses indexed by $h = 1, \ldots, H$ with data/information Y_h, parameters θ_h and model $f(Y_h|\theta_h)$ with capital H increasing steadily with time. Interest lies in θ_H, the latest parameter. One can and should develop the hierarchical prior $p(\theta_h|\phi)$, a proper prior given hyperparameters ϕ with prior $p(\phi)$. When H is 1 or small, the hierarchical model may be difficult to implement and the extra work may not benefit the analysis or the analyst, (but see Liang et al. 2009). Instead one fits a model $f(Y_h|\theta_h)p_h(\theta_h)$, the prior $p_h(\theta_h)$ will typically be a simple, often overly diffuse or even improper prior, possibly even one of RS14's priors. Overly diffuse might mean a prior with prior variance much larger than a carefully elicited subjective prior.

The analyst reports posterior MCMC samples $\{\theta_h^\ell\}_{\ell=1}^L$ from each analysis h. Eventually H becomes large enough to contemplate a more complex hierarchical model $\prod_{h=1}^H [f(Y_h|\theta_h)p(\theta_h|\phi)]p(\phi)$. Liang and Weiss (2007) propose a nonparametric model for $p(\theta_h|\phi)$ and explain how to compute $p(\theta_h|\phi)$ given the posterior MCMC samples $\{\theta_h^\ell\}_{\ell=1}^L$, and $h = 1, \ldots, H$. Then the analyst can fit a model

$$f(Y_{H+1}|\theta_{H+1})p(\theta_{H+1}|\phi, Y_1, \ldots, Y_H)p(\phi|Y_1, \ldots, Y_H).$$

Even in the situation where $H = 1$ and H can't get larger, it can be possible to execute this approach to provide a utile, proper and actively informative prior (Liang et al. 2009).

4 Data and Priors

In thinking about priors for our next analysis, it is helpful to have fit the model to previous data sets. How else can one develop intuition, understanding and prior knowledge about what sorts of skewness parameters might be reasonable? This can be circular, and having a quick prior to begin with certainly helps out. There has to be a first analysis for there to be a second analysis. On the other hand, (1) we do not want to call this a benchmark analysis; (2) in the next novel situation, we would rather not have to wait for another in depth series of enlightening analyses from Rubio, Steel and colleagues over a period of at least 15 years before we can start doing analyses, and (3) after we've developed some expertise in the model, we would prefer to use an informative prior.

To build experience with the skew models, we looked at eleven diverse variables separately for males and females from the Australian Institute of Sport (AIS) data (Cook and Weisberg 2009). There are $n = 100$ females and $n = 102$ males in the data set. Variables are Sex, Ht (Height in cm); Wt (Weight in kg); LBM (Lean body mass); RCC (Red cell count); WCC (White cell count); Hc (Hematocrit); Hg (Hemoglobin); Ferr (Plasma ferritin concentration); BMI (Body mass index = weight/height^2); SSF (Sum of skin folds); %Bfat (% body fat). Also included is the sport the person plays, which we do not use here.

We fit a skew-normal and a skew-$t(4)$ model using the epsilon-skew parameterization and the AG-beta$(1, 1)$ prior to the 11 variables, separately for males and females using BEAST (Drummond et al. 2012). Point estimates for the epsilon-skew parameter (taking values from -1 to 1) are plotted in Figure 1 for the normal model on the x-axis against the corresponding estimate from the $t(4)$ model on the y-axis.

Table 1 gives summary statistics for all 11 variables × 2 sexes = 22 analyses for the normal and the $t(4)$ analyses. Variable/sex pairs are ordered by the posterior mean of the skewness parameter in the normal model.

In Figure 1, the normal model estimates are generally closer to 1 than the $t(4)$ model estimates. Posterior standard deviations (PSD) range from .06 to .23 for the normal and from .06 to .30 for the $t(4)$ and are large enough that if error bars $\pm 2 *$PSD parallel to the appropriate axis were added to the means, the error bars would cross the $x = y$

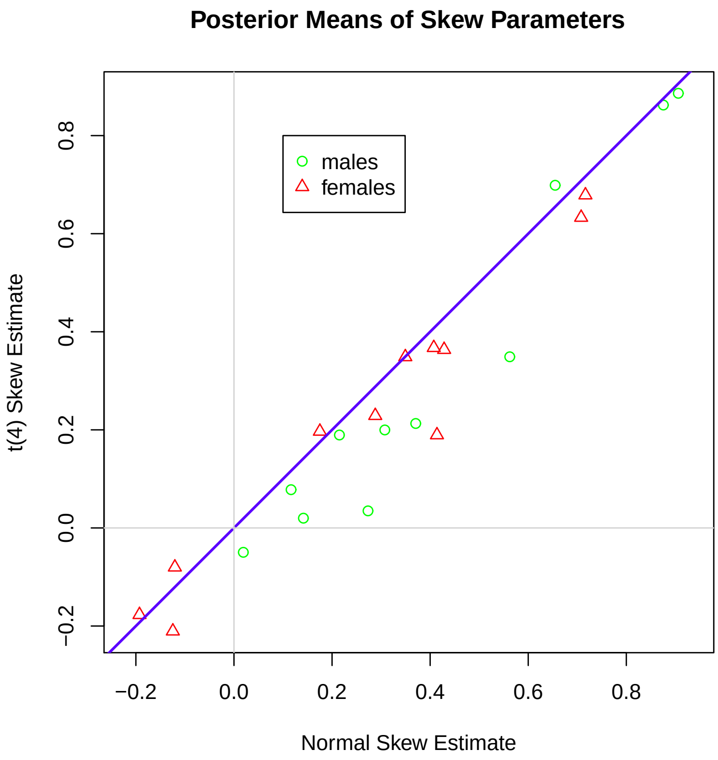

Figure 1: Normal (x) and $t(4)$ (y) point estimates for the epsilon-skew parameter for 22 data sets, 11 variables for males and females.

		Normal				$t(4)$			
				HPD region				HPD region	
Variable	sex	mean	SD	lower	upper	mean	SD	lower	upper
Ht	female	-0.19	0.11	-0.41	0.03	-0.18	0.13	-0.45	0.07
Wt	female	-0.12	0.13	-0.36	0.15	-0.21	0.15	-0.48	0.10
LBM	female	-0.12	0.12	-0.35	0.12	-0.08	0.13	-0.34	0.16
Ht	male	0.02	0.15	-0.28	0.32	-0.05	0.18	-0.40	0.29
LBM	male	0.12	0.13	-0.15	0.37	0.08	0.18	-0.25	0.42
RCC	male	0.14	0.10	-0.05	0.34	0.02	0.12	-0.21	0.27
Hg	female	0.18	0.24	-0.27	0.61	0.20	0.30	-0.35	0.70
Wt	male	0.21	0.14	-0.06	0.48	0.19	0.18	-0.19	0.50
Hc	male	0.27	0.12	0.06	0.52	0.03	0.13	-0.23	0.28
BMI	female	0.29	0.13	0.04	0.53	0.23	0.14	-0.04	0.50
Hg	male	0.31	0.11	0.08	0.51	0.20	0.15	-0.10	0.47
WCC	female	0.35	0.12	0.11	0.59	0.35	0.14	0.07	0.61
WCC	male	0.37	0.15	0.08	0.69	0.21	0.14	-0.07	0.46
Hc	female	0.41	0.21	-0.02	0.79	0.37	0.23	-0.12	0.74
%Bfat	female	0.41	0.22	-0.02	0.79	0.19	0.28	-0.25	0.75
RCC	female	0.43	0.14	0.16	0.70	0.36	0.15	0.07	0.64
BMI	male	0.56	0.15	0.27	0.86	0.35	0.14	0.07	0.63
Ferr	male	0.65	0.10	0.46	0.82	0.70	0.11	0.49	0.88
Ferr	female	0.71	0.11	0.48	0.90	0.63	0.13	0.38	0.88
SSF	female	0.72	0.13	0.45	0.94	0.68	0.16	0.32	0.93
SSF	male	0.88	0.07	0.75	1.00	0.86	0.07	0.71	0.98
%Bfat	male	0.91	0.06	0.79	1.00	0.89	0.06	0.77	1.00

Table 1: Posterior means, posterior standard deviations (SD) and 95% Highest Posterior Density (HPD) regions for the epsilon-skew parameter. Rows are sorted by the posterior mean of the epsilon skew parameter.

line in all 22×2 cases. We presume the normal model estimates are greater than the $t(4)$ because the long tails of the $t(4)$ downweight outlying observations that contribute to the normal model estimating greater skewness.

Using a skew-t model with unknown degrees of freedom such as in Fernández and Steel (1998), we are curious how strongly correlated the skewness parameter and the t degrees of freedom parameter would be in the posterior. Relatedly, we are curious how outliers affect the posteriors of the skewness and the degrees of freedom parameters in the general skew-t model. A data analytic issue is whether QQ plots, histograms, or kernel density plots or something else entirely would best help the data analyst quickly identify both that data is skewed and that the skew normal or skew t would be a good choice for analyzing the data. We had assumed that QQ plots would be quite useful in this regard, but limited experience with the AIS data, and RS14's Figure 3 made us look at histograms and density plots, and we found histograms/density plots to be easier to interpret than QQ plots. It's possible that QQ plots will become more informative with some study.

The variables Ferr, BMI, SSF, and %BFat have rather large estimated skewness parameters to the point where for males, SSF and %BFat are better described as half-normal. Briefly inspecting QQ plots (not shown), histograms (also not shown) and density plots (figure 2) of these variables by sex, the data illustrate skewness and several high outliers. Figure 2 shows kernel density estimates for Ferr, BMI, SSF and %BFAT variables for both sexes. The variance differences between the sexes and the common scaling of the x-axis make it hard to visualize both densities. The data have skewness, right outliers, possible multi-modality or even an outlying cluster, and shoulders. Further experience, such as (i) drawing random samples from skew-normals and skew-t's and plotting, (ii) trying a variety of other models such as Fernández and Steel (1998), will perhaps tell us whether skew-normals or skew-ts are the best we could do to model this data.

Looking at Figure 1 and (a) taking the point estimates at face value and (b) pretending these variables are representative, one might argue for a prior for the skew parameter that was uniform on $[-.2, 1]$. This ignores the risk that occurs if the next analyses actually needed a skew parameter outside this range versus the minimal loss from extending the uniform density full width to $[-1, 1]$, which we prefer, and which is RS14's figure 2(b). Slightly more informative, one might ask for a prior that had mode at .3 and covered full range, perhaps a beta$(13, 7)$ scaled to $[-1, 1]$ which gives little prior mass to values below $-.2$.

5 Discussion

We don't dispute the need for skewed data models. We're not so sanguine about the need to explore and expand on Jeffreys priors, though many of our friends and colleagues do spend time on this. But given that we didn't yet investigate goodness of fit for the AIS data, we're left wondering if it would be easier to fit a Dirichlet process mixture (DPM) model to skewed data. Issues and topics surrounding use of the DPM and its generalizations for skewed data include (i) choice of mixture components; (ii) whether

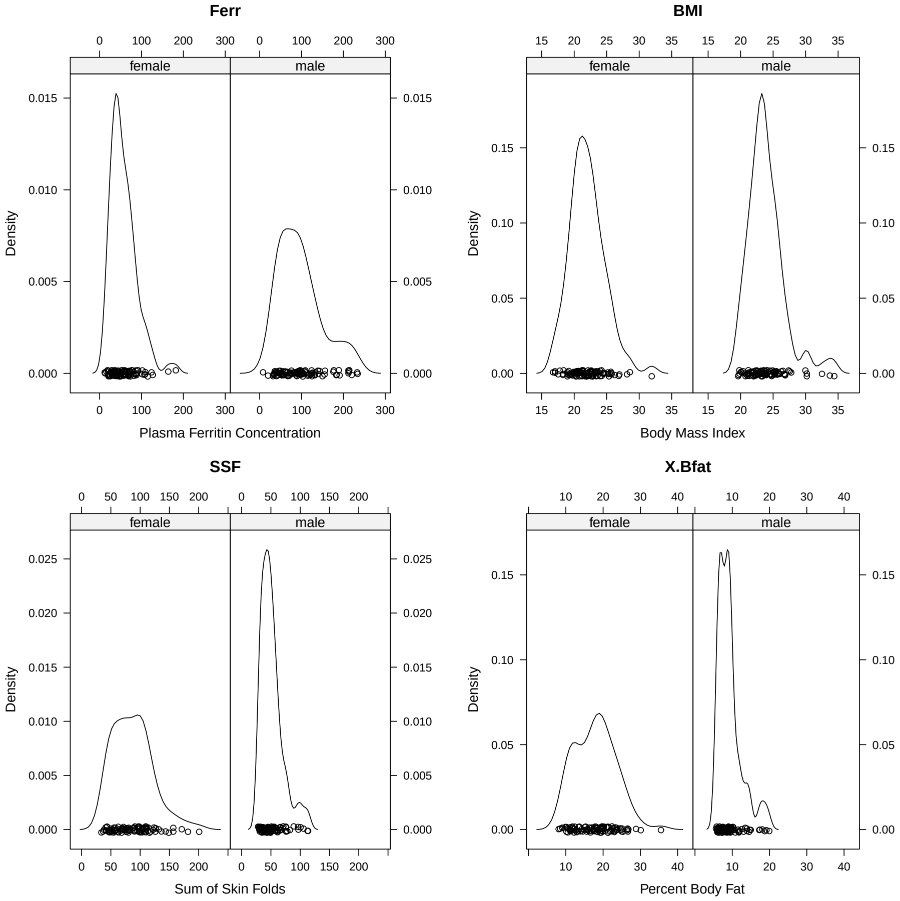

Figure 2: Density plots by sex for Ferr, BMI, SSF, and %BFAT

to soften up the data by initially taking a log or other transformation; (iii) development of measures to assess the posterior distribution of the skewness, particularly when there isn't an overt skewness parameter in the model; (iv) whether there is a variant of the normal mixture model DPM that could do a better job of fitting skewed data (i.e. large probability components need to be to the left of small probability components); and (v) how to think about priors so as to set sensible priors for the DPM model. For skewed variables, would there be any benefit to fitting DPM models that had skew-normal components rather than the usual normal components? Would taking a log and using normal mixture components be sufficient? Finally, prior specification is an important topic that we do spend a fair amount of time on; setting priors for the DPM in different settings is high on our target list of interesting, entertaining, enjoyable and important problems.

References

Cook, R. D. and Weisberg, S. (2009). *Applied Regression Including Computing and Graphics*. Wiley. 32

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). "Bayesian phylogenetics with BEAUti and the BEAST 1.7." *Molecular Biology and Evolution*, 29(8): 1969–1973. 32

Eaton, M. L. and Sudderth, W. D. (1998). "A new predictive distribution for normal multivariate linear models." *Sankhyā: The Indian Journal of Statistics, Series A*, 60: 363–382. 29

Edwards, W., Lindman, H., and Savage, L. J. (1963). "Bayesian statistical inference for psychological research." *Psychological Review*, 70(3): 193–242. 30

Fernández, C. and Steel, M. F. (1998). "On Bayesian modeling of fat tails and skewness." *Journal of the American Statistical Association*, 93(441): 359–371. 35

Geisser, S. (1965). "Bayesian estimation in multivariate analysis." *The Annals of Mathematical Statistics*, 36(1): 150–159. 29

Geyer, C. J. (1992). "Practical Markov chain Monte Carlo." *Statistical Science*, 7(4): 473–483. 30

Hobert, J. P. and Casella, G. (1996). "The effect of improper priors on Gibbs sampling in hierarchical linear mixed models." *Journal of the American Statistical Association*, 91(436): 1461–1473. 30

Ibrahim, J. G. and Laud, P. W. (1991). "On Bayesian analysis of generalized linear models using Jeffreys's prior." *Journal of the American Statistical Association*, 86(416): 981–986. 29

Ioannidis, J. P. (2005). "Why most published research findings are false." *PLoS Medicine*, 2(8): e124. 31

Jeffreys, H. (1998). *The Theory of Probability, third edition*. Oxford University Press. 29

Johnson, V. E. (2013). "Revised standards for statistical evidence." *Proceedings of the National Academy of Sciences*, 110(48): 19313–19317. 31

Keyes, T. K. and Levy, M. S. (1996). "Goodness of prediction fit for multivariate linear models." *Journal of the American Statistical Association*, 91(433): 191–197. 29

Liang, L.-J. and Weiss, R. E. (2007). "A Hierarchical Semiparametric Regression Model for Combining HIV-1 Phylogenetic Analyses Using Iterative Reweighting Algorithms." *Biometrics*, 63(3): 733–741. 32

Liang, L.-J., Weiss, R. E., Redelings, B., and Suchard, M. A. (2009). "Improving phylogenetic analyses by incorporating additional information from genetic sequence databases." *Bioinformatics*, 25(19): 2530–2536. 31, 32

Rubio, F. J. and Steel, M. F. J. (2014). "Inference in Two-Piece Location-Scale Models with Jeffreys Priors." *Bayesian Analysis*, 9: 1–22. 30

Stein, C. (1956). "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 197–206. 29

Tiao, G. C. and Tan, W. (1965). "Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance-components." *Biometrika*, 52: 37–53. 29

Acknowledgments

REW was supported by the UCLA CHIPTS and UCLA CFAR funded by National Institutes of Health (NIH) grants 5P30MH058107-15 and AI28697–CORE H. MAS was supported by NIH grants R01 HG006139 and R01 AI107034 and National Science Foundation grants DMS126153 and IIS1251151.