Sequential Bayesian updating for big data

Zita Oravecz
The Pennsylvania State University
Matt Huentelman
Translational Genomics Research Institute
Joachim Vandekerckhove
University of California, Irvine

Abstract

The velocity, volume, and variety of big data present both challenges and opportunities for cognitive science. We introduce sequential Bayesian updating as a tool to mine these three core properties. In the Bayesian approach, we summarize the current state of knowledge regarding parameters in terms of their posterior distributions, and use these as prior distributions when new data become available. Crucially, we construct posterior distributions in such a way that we avoid having to repeat computing the likelihood of old data as new data become available, allowing the propagation of information without great computational demand. As a result, these Bayesian methods allow continuous inference on voluminous information streams in a timely manner. We illustrate the advantages of sequential Bayesian updating with data from the *MindCrowd project*, in which crowd-sourced data are used to study Alzheimer's Dementia. We fit an extended LATER (Linear Approach to Threshold with Ergodic Rate) model to reaction time data from the project in order to separate two distinct aspects of cognitive functioning: speed of information accumulation and caution.

Introduction

The big data era offers multiple sources of data, with measurements that contain a variety of information in large volumes. For example, neuroimaging data from a participant might be complemented with a battery of personality tests and a set of cognitive-behavioral data. At the same time, with brain imaging equipment more widely accessible the number of participants is unlikely to remain limited to a handful per study. These advancements allow us to investigate cognitive

phenomena from various angles, and the synthesis of these perspectives requires highly complex models. Cognitive science is slated to update its set of methods to foster a more sophisticated, systematic study of human cognition.

Cognitive science has traditionally relied on explicative models to summarize observed data. Even simple cognitive measurement models (such as, e.g., process dissociation) are non-linear and can capture complex processes in more interesting terms than additive elements of true score and noise (such as in regression). With more complex cognitive process models (e.g., Ratcliff, 1978) we can study underlying mechanisms in meaningful terms and extract important facets of cognitive functioning from raw behavioral data.

In this chapter, we focus on methods that can be used for predominantly *model-driven* statistical inference. Here we use model-driven as a distinctive term, to separate these methods from the largely data-driven ones, such as those in machine learning (for Bayesian methods in machine learning see Zhu, Chen, & Hu, 2014). In practice, the specifics of a research question, together with relevant domain knowledge, will inform the choice of methods. Our purpose is not to advocate one set of methods over the other, but to offer by example an insight into what model-driven methods can achieve. In particular, we will focus on how Bayesian methods can be employed to perform model-driven inference for big data in an efficient and statistically coherent manner.

The primary reasoning behind considering model-based inference lies in the fact that big data is often voluminous in both length (number of units, e.g., people) and width (number of variables, e.g., cognitive measures). While increases in computing power can help data-driven exploration, this doubly exponential problem of 'thick' datasets often calls for domain-specific expertise. As a start, simple data curation can help to select variables that matter. These chosen variables can then be combined into a coherent narrative (in the form of a mathematical model), which opens up new ways of understanding the complex problem of human cognition.

First we will review why classical statistical methods are often unsuited for big data purposes. The reason is largely a lack of flexibility in existing methods, but also the assumptions that are typically made for mathematical convenience, and the particular way of drawing inference from data. Then we will elaborate on how Bayesian methods, by contrast, form a principled framework for interpreting parameter estimates and making predictions. A particular problem with Bayesian methods, however, is that they can be extremely demanding in terms of computational load, so one focus of this chapter is on how to reconcile these issues with big data problems. Finally, an example application will focus on a crowd-sourced data set as part of a research project on Alzheimer's Dementia.

Two schools of statistical inference

Broadly speaking, there exist two schools of thought in contemporary statistics. In psychological and cognitive science, the *frequentist* (or *classical*) school maintains a dominant position. Here, we will argue that the *Bayesian* school (see, e.g., Gelman, Carlin, Stern, & Rubin, 2013;

Kruschke, 2014; Lee & Wagenmakers, 2013), which is rising in popularity, holds particular promise for the future of big data.

The most fundamental difference between the frequentist and the Bayesian schools lies in the use and interpretation of *uncertainty*—possibly the most central issue in statistical inference. In classical statistics (null hypothesis significance testing/NHST, α and p-values, and confidence intervals), the thought process of inference starts with an existing hypothesis—usually, the null hypothesis $\mathcal{H}_0$. The classical reasoning goes: "assuming that the null hypothesis is true, how surprising are the data I have observed?" The word "surprising" in this context has a very specific meaning. It means "the probability of a set of observations that is *at least as extreme* as the real data." In the case of a common t-test where the null hypothesis is that a difference truly is zero but the observation is t_d, the surprise is given by the probability of observing a t statistic that is at least as far away from zero as t_d (i.e., larger than t_d if t_d was positive, and smaller if it was negative). If this probability is small, then the data are considered to be very surprising, or unlikely, "under the null," and the null hypothesis is rejected in favor of the alternative hypothesis $\mathcal{H}_A$. This conditional probability of certain constellations of data given a specific model ($\mathcal{H}_0$) is commonly known as the p-value.

One common counterargument to this line of reasoning is that *just because the data are unlikely under $\mathcal{H}_0$ does not imply that they are likely under any other hypothesis*—it is possible for data to simply be unlikely under all hypotheses that are being considered. This argument is somewhat counterintuitive because it is tempting to think that the probabilities under consideration should sum up to one. A counterexample is easy to construct. Consider a fictional person K who plays the lottery:

Premise— either K is a person who is not Bertrand Russell ($\mathcal{H}_0$) or K is Bertrand Russell ($\mathcal{H}_A$)
Premise— if K is a person who is not Bertrand Russell (i.e., if $\mathcal{H}_0$ is true), the probability p of K winning the lottery is very small: $p(\text{win} \mid \mathcal{H}_0) < \alpha$
Premise— K wins the lottery (an event with $p < \alpha$ has occurred)
Conclusion— therefore, $\mathcal{H}_0$ is false, and K is Bertrand Russell $\lightning$

The absurdity is obvious in this example: conclusions from this method of reasoning are entirely determined by which hypothesis was arbitrarily chosen as the null, and clearly the probabilities $p(e|\mathcal{H}_0)$ and $p(e|\mathcal{H}_A)$ do not necessarily add up to one.[1] For more discussion on the peculiarities of the p-values see for example Wagenmakers (2007).

In a more fundamental sense, adherents of the two frameworks think about data and parameters in rather different ways. The classical framework considers the data to be random: the current data to be analyzed is just one possible instance of thousands of hypothetical data sets—a population that is assumed to exist and that we could observe if we could re-run the study or exper-

[1] If our example seems far-fetched, consider that the existence of a counterexample means one of two things. Either (a) p-values are *never* a logically valid method of inference, or (b) p-values are *sometimes* a logically valid method of inference, but there exist necessary boundary conditions on the use of p-values that must be tested whenever p-values are applied. No such boundary conditions are known to the authors.

iment with the *exact* same settings. The observed data are then interpreted against the backdrop of this population of hypothetical data in order to determine how surprising the outcome was. The inferred hypothesis itself does not bear any probabilistic meaning: in the classical sense parameters and hypotheses are fixed, meaning that there exists a "true" parameter value, an exact value for a parameter that is waiting to be found. The only probabilistic statements made are about data: how likely were the data, and if we collect more data and compute confidence intervals, what are the probabilistic properties of our conclusions?[2] It is tempting to invert the probabilistic statement and make it about the underlying truth rather than about the data (e.g., "what is the probability that $\mathcal{H}_A$ is true," or "what is the probability that the results are due to chance," or "what is the probability these results will reappear in a replication?"), however, such statements can only be evaluated with the use of Bayes' rule (see below).

Big data applications in some sense preempt thoughts of hypothetical data sets—we have a large amount of data at hand and the size of the sample often approaches that of the population. Therefore in these settings it is more coherent to assume that the data are fixed and we compute the probability distributions of parameter values based on the information contained by all data available at present.

Moreover, a common goal in big data analysis is to make prediction about future trends. Frequentist inference can only assign probabilities to random events and to long-run frequencies, and is not equipped to make statements that are conditioned on past data. In fact, by relying on frequentist inference "one would not be allowed to model business forecast, industrial processes, demographics patterns, or for that matter real-life sample surveys, all of which involve uncertainties that cannot simply represented by physical randomization" (Gelman, 2006, p. 149). To summarize, with Bayesian modeling uncertainty can be directly addressed in terms of probability statements. To further illustrate the advantages of Bayesian modeling, we first review some of its basic principles.

Principles of Bayesian statistics

Bayesian methods are used to update current knowledge as information (data) comes in. The core of Bayesian statistical inference is the posterior distribution of the parameters, which contains the most up-to-date information about models and parameters. The posterior is proportional to the product of the likelihood and a prior distribution. The latter allows us to introduce information into current inference based on past data. The likelihood describes the assumed data generating mechanism. Formally, by using Bayes' rule of conditional probability we can estimate the probability distribution of parameters given the data:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}, \tag{1}$$

[2] These long-run guarantees of classical methods have issues in their own right, which we will not discuss here. More on problematic interpretation of confidence intervals can be found in Hoekstra, Morey, Rouder, and Wagenmakers (2014).

where $\boldsymbol{\theta}$ stands for the vector of all parameters in the model and $\mathcal{D}$ denotes the data. The left hand side is referred to as the *posterior distribution*. $p(\mathcal{D}|\boldsymbol{\theta})$ is the *likelihood* of the data $\mathcal{D}$ given $\boldsymbol{\theta}$. The second factor $p(\boldsymbol{\theta})$ in the numerator is the *prior distribution* on $\boldsymbol{\theta}$, which incorporates prior information on the parameter of interest and formalizes the current state of our knowledge of the parameters (before having seen the *current* data, but after having seen all *past* data). The denominator, $p(\mathcal{D})$, is the probability of the data averaged over all models under consideration. It does not depend on the model parameters and serves as a normalization constant in the equation above. The posterior distribution can often be obtained using only the repeated application of Bayes' rule (Eq. 1) and the law of total probability:

$$p(a) = \int_B p(a \mid b)p(b)db, \tag{2}$$

where B is the domain of the random variable b. For example, Equation 2 can be used to obtain $p(\mathcal{D}) = \int_{\boldsymbol{\Theta}} p(\mathcal{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$.

That wretched prior[3]

The most frequent criticism on Bayesian statistics involves the necessity of specifying a prior distribution on the parameters of interest, even in cases when one has no idea which values are likely to occur (e.g., Trafimow & Marks, 2015). A reply to this criticism is that the need for a specified prior distribution is not a weakness of Bayesian statistics but a necessary condition for principled statistical inference. Alternatively, solutions in terms of uninformative prior distributions have been offered (e.g., Jaynes, 2003).

Interestingly, however, from the perspective of big data, prior specification is a blessing rather than a curse: Through specifying informative prior distributions based on past data (or, crucially, *previous, smaller subsets of a large dataset*), the data at hand (or other parts of some large dataset) can be analyzed without having to re-fit the model for past data, while retaining the information from past data in the newly derived parameter estimates. A worked-out example of this principle appears at the end of this section, but Figure 1 shows a graphical example of how the conditional posterior distribution of a certain parameter (that is, the posterior distribution of that parameter *conditional on all the other parameters*) updates and becomes more informative as data are added. At the outset, we have next to no knowledge of the parameter, as expressed in the flat prior distribution. The prior becomes more peaked when more information becomes available, and with each update the parameter estimate is less noisy (i.e., has lower posterior standard deviation). With informative priors, convergence can be fast even if only a handful of new data points are added at a time.

[3]This expression is due to Lindley (2004).

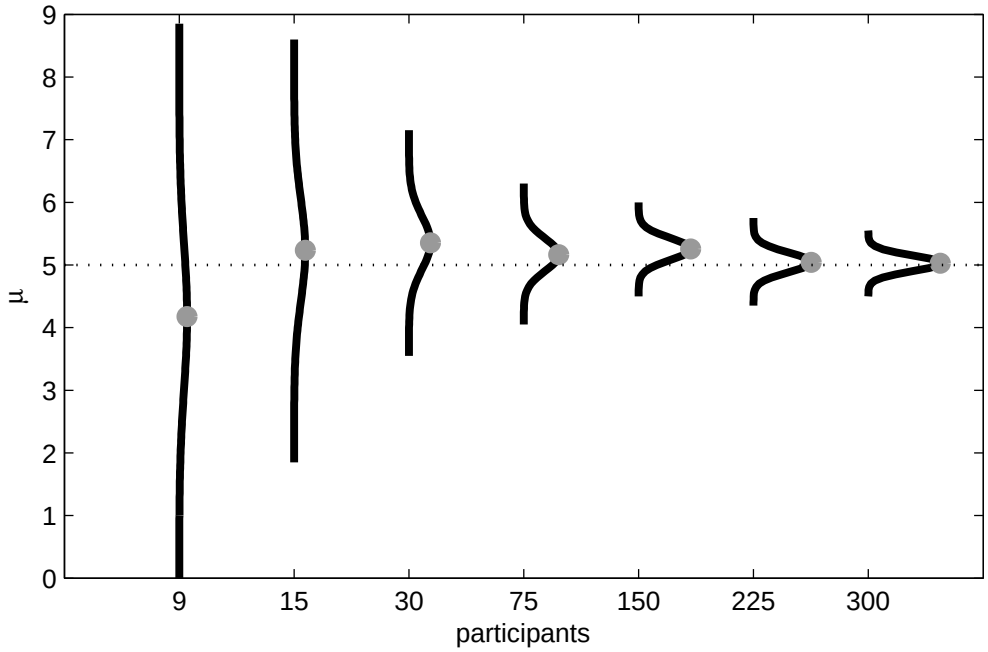

Figure 1. : Sequential updating of the conditional posterior distribution of a parameter μ. The parameter μ was simulated to be 5, and the probability density function of the parameter given all the available data is updated with some number of participants at a time (the total number is given on the horizontal axis). The distibution concentrates around the true value as N increases.

Obtaining the posterior

Statistical inference in the Bayesian framework typically focuses on the full posterior distribution. When models are simple (e.g., linear models), the analytical form of the posterior can be derived and posterior statistics can be calculated directly. Most often however, posteriors are far too complex to obtain through straightforward derivation. In these cases approximate Bayesian inference can be applied. We can divide these into two categories: *structural* and *stochastic* approaches. Structural approaches (e.g., variational Bayes; Fox & Roberts, 2012) aim to find an analytical proxy (variational distribution) of the model parameters that are maximally similar to the posterior—as defined by some closeness/divergence criterion—but have a simpler form. Posterior statistics are then based on this proxy. Once this is derived and tested for a specific model, inference can be carried out very efficiently (e.g., Ostwald, Kirilina, Starke, & Blankenburg, 2014). However, finding a proxy posterior distribution for new models can be a labor of some tedium. On the other hand, stochastic (sampling-based) techniques are implemented in ready-to-use generic inference engines such as WinBUGS ("Bayesian inference Using Gibbs Sampling"; Lunn, Thomas, Best, & Spiegelhalter, 2000), JAGS ("Just Another Gibbs Sampler"; Plummer, 2003), and, more

recently, Stan (Stan Development Team, 2013). Moreover, they provide an asymptotically exact representation of the posterior via Markov chain Monte Carlo (MCMC) sampling schemes. While the computational cost of sampling may be prohibitive when considering the large volumes of data in big data applications, the readiness of these methods to fit a wide range of models make them an appealing alternative and call for the development of techniques to overcome the computational hurdles. Later in this chapter we will describe how sequential Bayesian updating can be a useful technique to consider.

Another quantity that is important for statistical inference in the Bayesian framework is the *Bayes factor*, which is used to compare two models against each other. The computational details to obtain the Bayes factor can be found in the literature (e.g., Vandekerckhove, Matzke, & Wagenmakers, in press; Verdinelli & Wasserman, 1995), but for our purposes it suffices to know that the Bayes factor expresses the degree to which the available evidence should sway our beliefs from one model to another. A Bayes factor of one indicates no change in belief, whereas a Bayes factor of ten for model A over B indicates that we should be ten times more confident in A over B after seeing the data than we were before.

Sequential updating with Bayesian methods

A canonical example in statistical inference is that of "the Lady Tasting Tea" (Lindley, 1993; Fisher, 1935). In an account by Clarke (1991), Ronald Fisher was once visited upon by his colleague, a Dr. Muriel, who during the course of a party reprimanded Fisher for pouring tea into a cup first, and milk second. She claimed to be able to discern the difference and to prefer the reverse order. Fisher, exclaiming that "surely, it makes no difference," proceeded to set up a blind tasting experiment with four pairs of cups. Dr. Muriel correctly identified her preferred cup each time.

The pivotal quantity in this simple example is the rate π of correct identifications. We are interested in the posterior distribution of the parameter π given the data. Call that distribution $p(\pi \mid C, N)$, where C is the number of correct judgments out of N trials. By Bayes' theorem,

$$p(\pi \mid C, N) = \frac{P(C, N \mid \pi)\, p(\pi)}{P(C, N)}.$$

In this case, the *likelihood*, or the probability of observing the data takes the form of a binomial distribution, and is

$$P(C, N \mid \pi) = \binom{N}{C} \pi^C (1 - \pi)^{N-C}.$$

The marginal likelihood of the data, also known as the *evidence*, is

$$P(C, N) = \int_0^1 P(C, N \mid \pi)\, p(\pi) d\pi.$$

Finally, the *prior* can be set to a Beta distribution with shape parameters α and β:

$$p(\pi) = \text{Beta}\,(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}.$$

The mean of this prior distribution is $\frac{\alpha}{\alpha+\beta}$. In order to allow all possible values of rate π to be a-priori equally likely, set $\alpha = \beta = 1$, implying a prior mean of 0.5.

These elements can be combined to compute the posterior distribution of π given the data. To simplify this calculation, isolate all factors that contain the parameter π and collect the rest in a scale factor S that is independent of rate π:

$$\begin{aligned} p\,(\pi \mid C, N) &= \frac{\left[\binom{N}{C}\pi^{C}\,(1-\pi)^{N-C}\right]\left[\frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}\right]}{\int_0^1 P\,(C, N \mid \pi)\,p(\pi)d\pi} \\ &= S\pi^{C}\,(1-\pi)^{N-C}\,\pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &= S\pi^{C+\alpha-1}(1-\pi)^{N-C+\beta-1}. \end{aligned}$$

Now use the knowledge that the posterior distribution must be a *proper distribution* (i.e., it must integrate to 1), so that S can be determined as that unique value that ensures propriety. We exploit the similarity to the binomial distribution to obtain:

$$\begin{aligned} p\,(\pi \mid C, N) &= \binom{N+\alpha+\beta-2}{C+\alpha} \times \\ &\quad \pi^{C+\alpha-1}(1-\pi)^{N-C+\beta-1}, \end{aligned} \tag{3}$$

which corresponds to a Beta distribution with updated parameters: $\text{Beta}(\alpha + C, \beta + N - C)$, and with posterior mean $\frac{\alpha+C}{\alpha+\beta+N}$.[4] Note that if we choose the "flat prior" parameters $\alpha = \beta = 1$, then the posterior reduces to the likelihood.

More interestingly, however, "today's posterior is tomorrow's prior" (Lindley, 1972, p. 2). Suppose that we observe new data from a second round of tastings, with some sample size N' and C' correct identifications. We can then combine this new information using the posterior as a new prior, using the exact same methods:

$$\begin{aligned} P\left(\pi \mid C, N, C', N'\right) &= \binom{N+N'+\alpha+\beta-2}{C+C'+\alpha-1} \times \\ &\quad \pi^{C+C'+\alpha-1}(1-\pi)^{N+N'-(C+C')+\beta-1}, \end{aligned}$$

which corresponds to a Beta distribution with updated parameters: $\text{Beta}(\alpha+C+C', \beta+N+N'-C-C')$, and with posterior mean $\frac{\alpha+C+C'}{\alpha+\beta+N+N'}$.

[4]The *variance* of the Beta distribution is defined as: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, which becomes $\frac{(\alpha+C)(\beta+N-C)}{(\alpha+\beta+N)^2(\alpha+\beta+N+1)}$. The posterior uncertainty regarding the parameter is hence a strictly decreasing function of the added sample size N.

Crucially, note the similarity of this equation to Equation 3: this function is exactly what would have been obtained if $C + C'$ correct judgments had been seen in $N + N'$ trials done *all at once*: The prior distribution of π is updated by the data $(C + C', N + N')$ as if there had only ever been one round of tastings. The Bayesian method of sequential updating is coherent in this sense: datasets can be partitioned into smaller parts and yet contribute to the posterior distribution with equal validity.

We also note here that sequential updating does not always lead to an analytically tractable solution. The example above has the special property that the prior distribution of the parameter of interest (the Beta prior for the rate parameter π) is of the same distributional form as the posterior distribution. This property is called *conjugacy*; Information from the data enters into the Beta distribution by changing the *parameters* of the Beta prior, but not its parametric form. Many simple and moderately complex problems can be described in terms of conjugate priors and likelihoods. For models where the conjugacy property is not met, nonparametric techniques have to be applied to summarize information in the posterior distribution. Our example application will have conjugate properties, and we provide further information on nonparametric modeling in the Discussion section.

Advantages of sequential analysis in big data applications

The method of sequential Bayesian updating (SBU) can address the two computational hurdles of big data: volume and velocity. The combination of these solutions with the possibility of fitting cognitive models that can exploit the variety in big data through flexible modeling makes SBU a useful tool for research problems in cognitive science.

SBU is not the only way to deal with computational challenges in Bayesian inference, and we mention some techniques based on parallelization in the Discussion section. In choosing SBU our focus is more dominantly on the time-varying aspect of data size and on online inference: data are assumed to accumulate over time and—presumably—sharpen the inference. In SBU all data batches but the first are analyzed using informative priors, which should speed up convergence relative to the parallel techniques.

As described above, the procedure of SBU is to summarize ones current state of knowledge regarding parameters in terms of their posterior distributions, and use these as prior distributions when new data become available. Crucially, we construct posterior distributions in such a way that we avoid having to repeat computing the likelihood of old data as new data become available. We can address the three main properties of big data as follows:

Volume One can think of big data simply as large data set that it is infeasible to analyze at once on the available hardware. Through SBU, one can partition a large data set into smaller, more manageable batches, perform model fitting on those sequentially, using each batch's posterior distribution as a prior for the next batch. This procedure avoids having to store large data sets in

memory at any given time.

Velocity Bayesian decision rules are by default sequential in nature, which makes them suitable for handling big data streams. Unlike the frequentist paradigm, Bayesian methods allow for inferences and decisions to be made at any arbitrary point in the data stream, without loss of consistency. Information about past data is kept online by means of the posterior distributions of the model parameters that sufficiently summarize the data generation process. The likelihood only needs to be calculated for the new data point to update the model parameters' posteriors. We will focus on cases where data are streaming continuously and a relatively complex model is fit to the data. These principles scale seamlessly and can be applied where a large volume of data is analyzed with complex models.

Variety Big data means a lot of information coming in from different sources. One needs complex models to combine different sources of information (see van der Linden, 2007, for a general method for combining information across sources). For example, often not only neuroimaging data are collected, but several behavioral measures are available (e.g., the Human Connectome Project). In such a case, one could combine a neural model describing fMRI data with a cognitive model describing behavioral data (see Turner, Forstmann, Wagenmakers, Sederberg, & Steyvers, 2013, for an application in cognitive neuroscience). Off-the-shelf software packages are not ready to make inference with novel complex models, while Bayesian tools provide us with a possibility to fit practically any model regardless of complexity.

Application: MindCrowd—Crowdsourcing in the service of understanding Alzheimer's Dementia

MindCrowd

MindCrowd (TGen and The University of Arizona; `www.mindcrowd.org`) is a large-scale research project that uses web-based crowdsourcing to study Alzheimer's Dementia (AD). The focus is on the assessment of cognition in a large cohort of healthy adults of all ages. The project is in its first phase, where web-based memory testing is conducted through two tasks: an attention task resulting in simple reaction times (of five trials) and a pair-associated learning task with three stages of recall. Moreover, a set of covariates are collected including age, gender, marital status, education, whether the participant or a family member have been diagnosed with AD, and more. The goal is to collect data from one million people and select various profiles. Then in a second phase more intensive cognitive testing will be carried out complimented by DNA sampling and additional demographic questions. MindCrowd was launched in April of 2013 and has recruited over 40,000 test takers who have completed both tasks and answered at least 80% of the demographic questions. The analyses presented here are based on 22,246 participants whose data were available at the time of writing. The goal of the MindCrowd project is to collect data for one million participants. With

sequential Bayesian updating, inference regarding substantively interesting parameters can be kept up-to-date in a continuous fashion, adding only newly arriving data to a prior that is itself based on previous data. This means, for example, that when the last responses (the ones closer to the one million mark) arrive, computing the posterior distribution will be fast.

Modeling simple reaction time with the LATER model

Data collected through the MindCrowd website provides us with several opportunities for cognitive modeling. We will focus here on the attention task of the MindCrowd project: a vigilance task in which participants are asked to respond as fast as they can to an appearing stimulus, with a randomized interstimulus interval. The stimulus was a fuchsia disk and participants were instructed to hit enter/return as soon as they saw it appear on their screen. At the time of writing the task is still available on the MindCrowd website.

We will apply an hierarchical extension of a widely used process model for RT called the *Linear Approach to Threshold with Ergodic Rate* model (LATER; Reddi & Carpenter, 2000; Ratcliff, Carpenter, & Reddi, 2001). The LATER model is one of a large class of sequential-sampling models, in which it is assumed that during the course of a trial, information is accumulated sequentially until a criterial amount of information is reached, upon which a response is executed. In the LATER model, the accumulation process is assumed to be linear, approaching a fixed threshold, with a rate that is random from trial to trial. A graphical illustration of the model is shown in Figure 2.

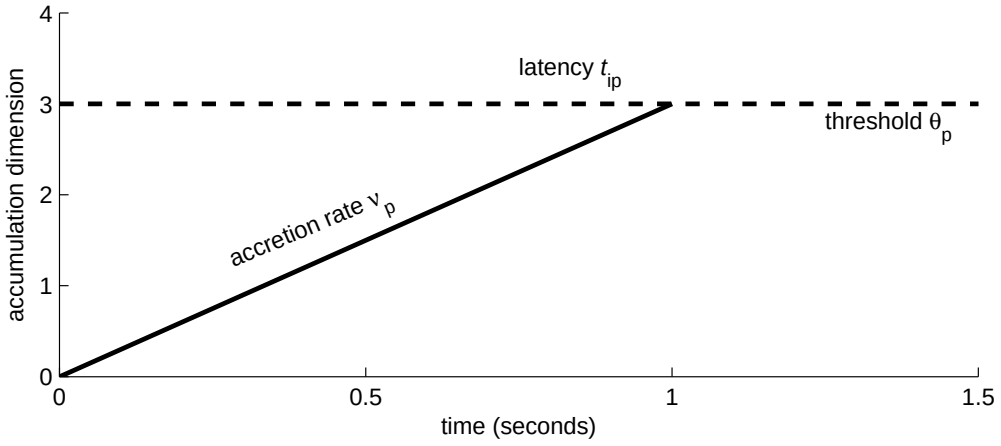

Figure 2. : Graphical illustration of the LATER model.

The LATER model describes the latency distributions of observed RTs by characterizing the decision-making process in terms of two cognitive variables, namely (1) person-specific *caution* θ_p, or the amount of information needed by p to respond (the "threshold"), and (2) the average *rate of information accumulation* ν_p (the "accretion rate"). In taking this approach, we are fitting a

probabilistic model to the observed behavioral data. We think of this probabilistic abstraction as the *generative model*, and it characterizes our assumptions regarding the process by which the data come about. More specifically, at each trial i, a single, trial-specific, realization of the accretion rate, denoted as z_{pi}, is generated according to a unit-variance Gaussian distribution:

$$z_{pi} \sim N(\nu_p, 1), \tag{4}$$

where $\sim$ is the common notation used to indicate that the variable on the left side is a draw from the distribution on the right side. The predicted response time at trial i is then $t_{pi} = \frac{\theta_p}{z_{pi}}$; that is, the person-specific caution θ_p divided by the person-specific rate at the i^{th} trial: z_{pi}.

Rearranging this expression yields that $z_{pi} = \frac{\theta_p}{t_{pi}}$, which by Equation 4 follows a Gaussian distribution with mean ν_p and variance 1. It further follows that

$$\frac{z_{pi}}{\theta_p} = \frac{1}{t_{pi}} \sim N\left(\frac{\nu_p}{\theta_p}, \frac{1}{\theta_p^2}\right). \tag{5}$$

where ν_p remains the accretion rate parameter for person p, capturing their information processing speed, and θ_p is the threshold parameter implying their caution in responding.

In what follows, we will apply a regression structure to the accretion rate parameter in order to quantify between-person differences in speed of information processing (see, e.g., Vandekerckhove, Tuerlinckx, & Lee, 2011, on hierarchical Bayesian approaches to cognitive models). To the best of our knowledge this is the first application of a hierarchical Bayesian LATER model. The caution parameter, θ, is positive by definition—and is closely related to the precision of the accretion distribution—so we choose a gamma distribution on the square of caution, θ_p^2 to inherit the conjugacy of that distribution:

$$\theta_p^2 \sim \Gamma(s_\theta, r_\theta), \tag{6}$$

with shape s_θ and rate r_θ the parameters of the gamma distribution on θ_p^2. Furthermore, assume that C covariates are measured and x_{pc} denotes the score of person p on covariate c ($c = 1, \ldots, C$). All person-specific covariate scores are collected into a vector of length $C + 1$, denoted as $\mathbf{x}_p = (1, x_{p1}, x_{p2}, \ldots, x_{pC})^{\mathrm{T}}$. The assumed distribution of the accumulation rate parameter ν_p is then:

$$\nu_p \sim N(\mathbf{x}_p\boldsymbol{\beta}, \sigma^2) \tag{7}$$

Finally, we need to choose priors for the remaining parameters of interest. The regression terms follow a multivariate normal distribution, specified as: $\boldsymbol{\beta} \sim MVN\left(M_\beta, \Sigma_\beta\right)$, with M_β a vector of zeros, Σ_β a covariance matrix with 0.1 on the diagonal and 0 elsewhere. We specify gamma prior on the inverse of the residual variance (i.e., on the precision): $\sigma^{-2} \sim \Gamma(s_\sigma, r_\sigma)$, where $s_\sigma = r_\sigma = 0.01$. Fitting the specified model through sequential Bayesian updating is described in the next section.

Study design

We analyzed the reaction time data of $N = 21{,}947$ participants (each providing at most 5 valid trials) from the MindCrowd project. While the original sample size was slightly larger (22,246) we

discarded data from participants whose covariate information was missing. We also omitted reaction times above 5 seconds or below 180ms, which are unrealistic for a simple vigilance task. As part of the project several covariates are collected. From this pool, we chose the following variables for inclusion in our analysis: age, gender, and whether the participant or a family member[5] had been diagnosed with AD. Our interest is in the effect of the presence of AD on the speed of information processing, and its possible interaction with age. The hierarchical LATER model we construct for this purpose is very similar to a classical regression model, with the main difference being that the predicted distribution of the data is not a normal distribution, but rather the distribution of RTs as predicted by a LATER model. The 'target' of the regression analysis is therefore not the mean of a normal distribution but the accretion rate of a LATER process. For illustration, we write out the mean of the person-specific information accumulation rates (ν_p) from Eq. 7 as a function of age, sex, AD diagnosis and the interaction of age and AD diagnosis, and the corresponding regression terms:

$$\mathbf{x}_p\boldsymbol{\beta} \quad = \quad \beta_0 + \beta_1 \mathrm{AGE}_p + \beta_2 \mathrm{SEX}_p + \beta_3 \mathrm{ALZ}_p + \beta_4 \mathrm{AGE}_p \mathrm{ALZ}_p. \tag{8}$$

The key regression equation (the mean in Eq. 7; worked out in Equation 8), together with Equations 5, 6, and 7 completes our model.

For carrying out the analysis, we specify prior distributions on the parameters in Equations 6, 7, and 8 (i.e., for the regression terms $\boldsymbol{\beta}$, for the inverse of the residual variance σ^{-2}, and for the person-specific caution θ_p). The parametric forms of these priors (namely, the multivariate normal distribution and the gamma distribution) are chosen to be conjugate with the Gaussian likelihood of the data. The sequential Bayesian updating then proceeds as follows: As described above, we specify standard non-informative prior distributions for the first batch of data. We then obtain posterior samples from JAGS. Once JAGS returns the results, we summarize these samples in terms of the conditional posterior distributions of the parameters of interest. More specifically, for the regression terms, we calculate the mean vector and the covariance matrix of the multivariate normal distribution based on the posterior samples. The mean vector expresses our best current state of knowledge on the regression terms, the variances on the diagonal quantify the uncertainty in these, and the covariances in the off-diagonal positions capture possible trade-offs due to correlation in the covariates. These posterior summary statistics sufficiently summarize our knowledge on the parameter given the data, up to a small computational error due to deriving these posterior summarizes through sampling with JAGS, instead of deriving them analytically. The same principle applies for the residual precision parameter, σ^{-2}, in terms of the shape parameters (s_σ, r_σ) of its Gamma distribution. Finally, we plug these estimated distributions in as priors for the next batch of data.

In the current analysis we use exclusively conjugate priors (i.e., where the parametric form of the prior on the parameter combined with the likelihood of the model results in a conditional posterior distribution of the same parametric form but with updated parameters based on the

[5]The phrasing of the item was: "Have you, a sibling, or one of your parents been diagnosed with Alzheimer's disease? Yes, No, NA." The variable one only took two values in the current dataset: 1—a first degree family member has AD (including respondent, around 4000 respondents); 0—there is no first degree relative with AD in the family.

data). However, not all models can be formulated by relying only on conjugate priors. In these cases, conjugacy can be forced with the use of non-parametric methods, but this is beyond the scope of the current chapter (but see the Discussion section for further guidelines).

The analyses presented here were run on a desktop computer with a 3.40GHz CPU and 16GB RAM. While in principle in this phase of the project (with $N = 21947$) we could have analysed the entire data set on this machine, for the purposes of demonstration we divided the full dataset into 40 batches. In a later phase of the MindCrowd project the sample size will increase substantially, to an expected one million participants, in which case—due to the desktop computer's RAM limitations—batch processing will be required rather than optional.

We implemented the model in JAGS using a homegrown MATLAB interface.[6] The analysis took approximately 10 minutes to run. From the first batch of data, parameters were estimated by running 5 chains with 1,500 iterations each, discarding the first 1,000 samples as burnin.[7]

From the second batch until the last, we ran 5 chains with 800 iterations each, from which 500 were discarded as burnin. The reason why we choose a shorter adaptation for the second batch was that the algorithm was now better "informed" by the prior distributions of the parameters inferred from the first batch, so that we expect faster convergence to the highest posterior density area. The final sample size was 1,500 samples. Convergence of the 5 chains was tested by the $\hat{R}$ statistics. $\hat{R}$ was lower than 1.01 for all parameters (with the standard criterion being $\hat{R} < 1.1$).

Results from the hierarchical Bayesian LATER model

Figure 3 shows the evolution of the distribution of β_4 as more data are introduced. The results of our regression-style analysis are displayed in Table 1. Parameters β_1 and β_2 show posterior distributions that are clearly far away from zero, indicating high confidence in the existence of an effect. β_1 is negative, indicating that speed of information processing decreases with advancing age. β_2 is positive, indicating an advantage for men over women in terms of speed of information processing. Parameters β_3 and β_4, however, do not show clear effects. In both cases, the posterior mean is close to zero. In the case of β_3 – the predictor on whether the participant or a family member has been diagnosed with AD – the value 0 is included in its 95% credibility interval (i.e., the interval between the 2.5 and 97.5 percentiles), and the Bayes factor indicates weak evidence for

[6] All scripts are available from the following https://git.psu.edu/zzo1/ChapterSBU. MindCrowd's data are proprietary.

[7] These burnin samples serve two purposes. First, when a model is initialized, JAGS enters an adaptive mode during which the sampling algorithms modifies its behaviour for increased efficiency. These changes in the algorithm violate the *detailed balance* requirement of Markov chains, so that there is no guarantee that the so generated samples converge to the desired stationary distribution. Second, to ensure that the samplers are exploring the posterior parameter space sufficiently, the sampling algorithm is restarted several times with dispersed starting values and it is checked whether all these solution converge into the same area (as opposed to being stuck in a local optimum, e.g.). Posterior inference should be based on samples that form a Markov chain and are converged into the same area and have "forgotten" their initial values. In the current analysis the samplers are run independently 5 times (i.e., we run 5 chains). The independence of these MCMC chains implies that they can be run in parallel, which we do.

Table 1:: Summary of the regression weights where response speed was modeled with the LATER model and the information accumulation rate were regressed on age, gender, AD in the family, the interaction of age and AD in the family.

	Predictor	Mean	SD	95% CrI	BF_{ALT}
β_0	Intercept	5.6280	0.0268	(5.575 , 5.6800)	$\gg 10^{10}$
β_1	Age	−0.7878	0.0196	(−0.8261 , −0.7492)	$\gg 10^{10}$
β_2	Gender	0.6185	0.0368	(0.5471 , 0.6891)	$\gg 10^{10}$
β_3	AD	−0.0704	0.0672	(−0.2011 , 0.0560)	0.37
β_4	Age×AD	0.0273	0.0602	(−0.0912 , 0.1478)	0.21

Note: Mean and SD are posterior mean and standard deviation. CrI stands for 'credibility interval'. BF_{ALT} is the Savage-Dickey approximation (Verdinelli & Wasserman, 1995) to the Bayes factor in favor of the (alternative) hypothesis that $\beta \neq 0$.

the null hypothesis (i.e., no effect, that is $\beta_3 = 0$). More precisely, the Bayes factor is 2.7 (1/0.37) in favor of the null hypothesis. Similarly, there is no evidence that information accumulation changes in relation to the interaction of age and the presence of AD – in fact, the Bayes factor for β_4 shows moderate evidence in favor of the null hypothesis of no effect ($1/0.21 = 4.76$).

Especially in the case of big data, it is important to draw a distinction between *statistical significance*—the ability of the data to help us distinguish effects from non-effects—and *practical significance*—the degree to which an extant effect influences people. In the current data set, the difference in (mean) predicted RT between a male participant (group mean accretion rate $\bar{\nu}_m$) and a female participant (group mean accretion rate $\bar{\nu}_f$) is approximately $\bar{\theta}\left(\frac{1}{\bar{\nu}_f} - \frac{1}{\bar{\nu}_m}\right)$, which with our results works out to about 10ms. Hence, while the difference between these two groups is *detectable* (the Bayes factor against the null is more than 1000:1), it is small enough that any daily-life consequences are difficult to imagine.

To summarize, our cognitive model allows us to cast light on the information processing system that is assumed to underlie the simple RT measures. The process model identifies a parameter of interest—in this case, a rate of information accumulation—and inferences can then be drawn in terms of this parameter. Caution in the responding is factored into the inference, treated as a nuisance variable, and separated from the accumulation rate.

Combining cognitive models

The MindCrowd website currently tests volunteers not only on the vigilance task, but also on a paired-associate learning (PAL) task. Cognitive models exist to model underlying processes

in these decisions as well (e.g., multinomial models for measuring storage and retrieval processes; Rouder & Batchelder, 1998). In the hierarchical Bayesian modeling framework, we could combine data from these two tasks together by specifying a joint hyperprior distribution of the parameters of the model for PAL and the model for the RTs (e.g., Pe, Vandekerckhove, & Kuppens, 2013; Vandekerckhove, 2014). Combining these joint modeling techniques that were originally developed in psychometrics (e.g. van der Linden, 2007) with Bayesian modeling can offer a flexible unified framework for drawing inference from data that would classically be analyzed separately, thereby partially addressing the "variety" aspect of big data challenges.

Discussion

In this chapter, we discussed one way in which Bayesian methods can contribute to the challenges introduced by big data. A core aspect of Bayesian inference—the *sequential updating* that is at the heart of the Bayesian paradigm—allows researchers to partition large data sets so that they become more manageable under hardware constraints. We have focused on one specific method for exploiting the sequential updating property, namely using conjugate priors, which lead to closed-form posterior distributions that can be characterized with only a few sufficient statistics, and in turn serve as priors for future data. This particular method is limited because it requires conjugacy of the focal parameters. However, we were able to apply it to a nontrivial cognitive model (the hierarchical LATER model) and draw interesting process-level conclusions. For more complex models, priors and posteriors could be expressed in nonparametric ways (Gershman & Blei, 2012). This method solves the need for conjugacy, but will itself introduce new computational challenges. The sequential updating method is computationally efficient because it collapses posterior samples into sufficient statistics, but also because the informative priors that are generated from the first batches of data speed up convergence of later batches.

Our method has also assumed a certain *stationarity of data*; That is, it was assumed that as the data came in, the *true* parameters of the model did not change. However, there are many real-world scenarios—ranging from negotiation theory, learning psychology, and EEG analysis, over epidemiology, ecology, and climate change, to industrial process control, fraud detection, and stock market predictions—where the stationarity assumption would clearly be violated and the academic interest would be in *change point detection* (e.g., Adams & MacKay, 2007). Within our current approach, a change point detection model would require that the parameters relevant to the regime-switches are explicitly included, and posteriors over these parameters can be updated as data becomes available.

Moving beyond sequential updating, there exist other methods for obtaining samples of a posterior distribution using large data sets. For example, the Consensus Monte Carlo Algorithm (Scott, Blocker, & Bonassi, 2013) or the Embarrassingly Parallel, Asymptotically Exact MCMC algorithm (Neiswanger, Wang, & Xing, 2014) both rely on distributing the computational load across a larger hardware infrastructure and reducing the total "wall time" required for an analysis. The method we present here has the advantage of not requiring a large dedicated computation

infrastructure and can be run on a regular desktop computer, with the size of the data affecting only the computation time.

All of these methods rely on Bayesian inference. As we have argued extensively, we believe that Bayesian methods are not only useful and feasible in a big data context, but are in fact superior from a philosophical point of view. Classical inference is well known to generate bias against the null hypothesis, and this bias increases with increasing data size. Recent attempts to reform statistical practice in the psychological sciences (Cumming, 2014) shift the focus of statistical analysis to parameter estimation, but with this there remain several major issues. First, the estimation framework is still based in classical statistics and does not take into account the prior distribution of parameters of interest. Second, it is not clear if inference is possible at all in this framework, and "dichotomous thinking" is discouraged entirely (though it is tempting to wrongly interpret confidence intervals as posterior distributions, and to decide that an effect is present if the interval does not contain zero). These recent recommendations seem to us to throw the dichotomous baby away with the NHST bathwater, while a Bayesian approach (as we and many others have demonstrated) is logically consistent, does allow for inferential statements, and allows one to collect evidence in favor of a null hypothesis. Especially in the case of big data, these are highly desirable qualities that are not shared by classical methods, and we recommend Bayesian inference as a default method.

Acknowledgements

JV was supported by grant #48192 from the John Templeton Foundation and by NSF grant #1230118 from the Methods, Measurements, and Statistics panel.

References

Adams, R. P., & MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Clarke, C. (1991). Invited commentary on r. a. fisher. *American Journal of Epidemiology*, *134*(12), 1371-1374. Retrieved from `http://aje.oxfordjournals.org/content/134/12/1371.short`

Cumming, G. (2014). The new statistics why and how. *Psychological science*, *25*(1), 7-29.

Fisher, R. A. (1935). The design of experiments.

Fox, C., & Roberts, S. (2012). A tutorial on variational Bayes. *Artificial Intelligence Review*, *38*, 85–95.

Gelman, A. (2006). The boxer, the wrestler, and the coin flip: a paradox of robust bayesian inference and belief functions. *American Statistician*, *60*, 146–150.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian Data Analysis (third ed.)*. Boca Raton, FL.: Chapman & Hall/CRC.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*, 1–12.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychological Bulletin and Review*.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Kruschke, J. K. (2014). *Doing Bayesian data analysis : A tutorial with R, JAGS and Stan (second ed.)*. London: Academic Press / Elsevier.

Lee, M. D., & Wagenmakers, E. (2013). *Bayesian cognitive modeling*. New York: Cambridge.

Lindley, D. (1972). *Bayesian statistics: A review*. SIAM.

Lindley, D. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*(1), 22–25.

Lindley, D. (2004). That wretched prior. *Significance*, *1*(2), 85–87.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Neiswanger, W., Wang, C., & Xing, E. A. (2014). Asymptotically exact, embarrassingly parallel MCMC. *http://arxiv.org/pdf/1311.4780v2.pdf* , *1311.4780*.

Ostwald, D., Kirilina, E., Starke, L., & Blankenburg, F. (2014). A tutorial on variational bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, *60*, 1-19.

Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion*, *13*(4), 739.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003)* (pp. 20–22).

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., Carpenter, R. H. S., & Reddi, B. A. J. (2001). Putting noise into neurophysiological models of simple decision making. *Nature Neuroscience*, *6*, 336–337.

Reddi, B. A., & Carpenter, R. H. S. (2000). The influence of urgency on decision time. *Nature Neuroscience*, *3*, 827–830.

Rouder, J. N., & Batchelder, W. H. (1998). Multinomial models for measuring storage and retrieval processes in paired associate learning. In C. E. Dowling, F. S. Roberts, & P. Theuns (Eds.), *Recent progress in mathematical psychology* (pp. 195–226). New York: Psychology Press.

Scott, S. L., Blocker, A. W., & Bonassi, F. V. (2013). Bayes and Big Data: The Consensus Monte Carlo algorithm. In *Paper presented at the 2013 EFab@Bayes 250 Workshop*.

Stan Development Team. (2013). *Stan: A C++ Library for Probability and Sampling, Version 1.3*. Retrieved from `http://mc-stan.org/`

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1-2.

Turner, B. M., Forstmann, B. U., Wagenmakers, S. D., E.-J.and Brown, Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (in press). *Model comparison and the principle of parsimony*. Oxford: Oxford University Press.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*, 44–62.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, *90*(430), 614–618.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Zhu, J., Chen, J., & Hu, W. (2014). Big learning with Bayesian methods. *http://arxiv.org/pdf/1411.6370.pdf* , *1411.6370v1*.

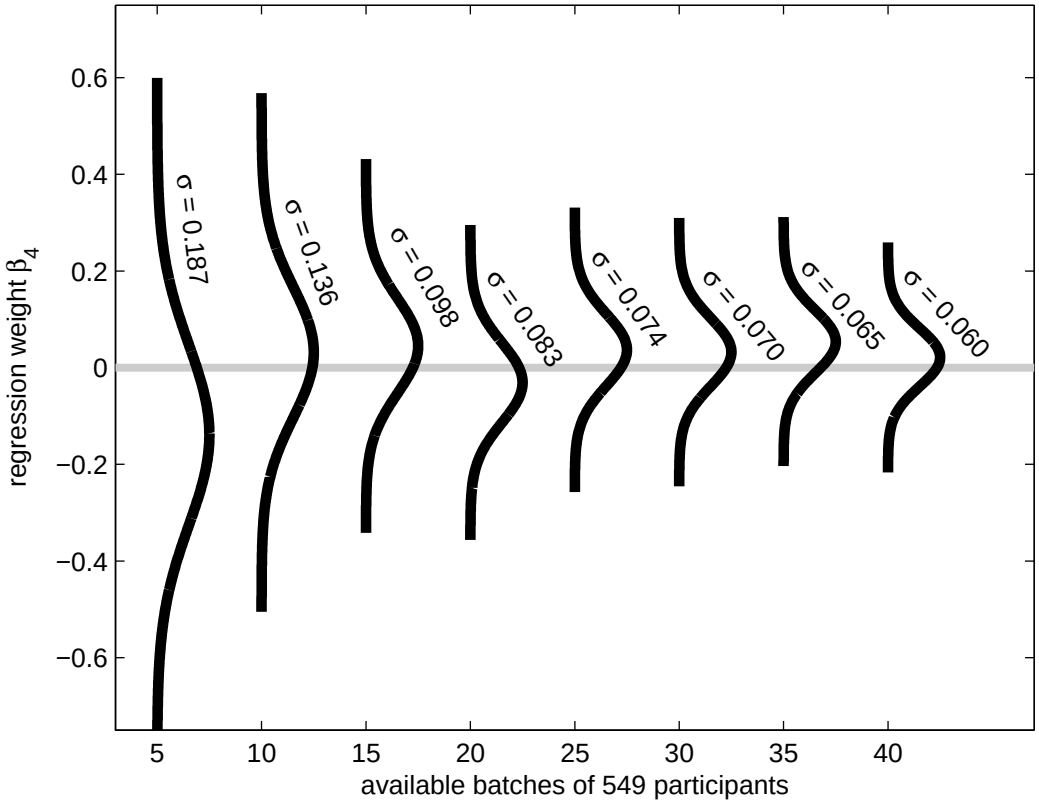

Figure 3. : Sequence of conditional posterior distributions for the regression coefficient parameter β_4—the weight of the AD-by-age interaction regressed on the speed of information processing parameter. As each batch of participants is added to the analysis, the our knowledge of β_4 is updated and the posterior standard deviation decreases while the posterior mean converges to a stable value (in this case, near 0).