Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification

Thomas V. Wiecki, Jeffrey Poland, Michael Frank

July 10, 2014

1 Appendix

The following serves as a reference for the mathematical details of the methods motivated above.

1.1 Parameters used in simulation study

The below table contains the group means of the parameters used to create subjects of two groups. Each individual subject was created by adding normally distributed noise of $\sigma = .1$ to the group mean.

Parameter	Group 1	Group 2
non-decision time	.3	.25
drift-rate	1	1.2
threshold	2	2.2

1.2 Drift-Diffusion Model

Mathematically, the DDM is defined by a stochastic differential equation called the Wiener process with drift:

$$dW \sim \mathcal{N}(v, \sigma^2) \tag{1}$$

where v represents the drift-rate and σ the variance. As we often only observe the response times of subjects we are interested in the wiener first passage time (wfpt) – the time it takes W to cross one of two boundaries. Assuming two absorbing boundaries of this process and through some fairly sophisticated math (see e.g. Smith, 2000) it is possible to analytically derive the time this process will first pass one of the two boundaries (i.e. the wiener first passage time; wfpt). This probability distribution[1] then serves as the likelihood function for the DDM.

1.3 Bayesian Inference

1.3.1 Hierarchical Bayesian modeling

Bayesian methods require specification of a generative process in form of a likelihood function that produced the observed data x given some parameters θ. By specifying our prior belief we can use Bayes formula to invert the generative model and make inference on the probability of parameters θ:

$$P(\theta|x) = \frac{P(x|\theta) * P(\theta)}{P(x)} \tag{2}$$

[1] the wfpt will not be a distribution rather than a single value because of the stochasticity of the wiener process

Where $P(x|\theta)$ is the likelihood and $P(\theta)$ is the prior probability. Computation of the marginal likelihood $P(x)$ requires integration (or summation in the discrete case) over the complete parameter space Θ:

$$P(x) = \int_{\Theta} P(x, \theta) \, \mathrm{d}\theta \tag{3}$$

Note that in most scenarios this integral is analytically intractable. Sampling methods like Markov-Chain Monte Carlo (MCMC) (Gamerman and Lopes, 2006) circumvent this problem by providing a way to produce samples from the posterior distribution. These methods have been used with great success in many different scenarios (Gelman et al., 2003) and will be discussed in more detail below.

A hierarchical model has a particular benefit to cognitive modeling where data is often scarce. We can construct a hierarchical model to more adequately capture the likely similarity structure of our data. As above, observed data points of each subject $x_{i,j}$ (where $i = 1, \ldots, S_j$ data points per subject and $j = 1, \ldots, N$ for N subjects) are distributed according to some likelihood function $f|\theta$. We now assume that individual subject parameters θ_j are normal distributed around a group mean with a specific group variance ($\lambda = (\mu, \sigma)$ with hyperprior G_0) resulting in the following generative description:

$$\mu, \sigma \sim G_0() \tag{4}$$
$$\theta_j \sim \mathcal{N}(\mu, \sigma^2) \tag{5}$$
$$x_{i,j} \sim f(\theta_j) \tag{6}$$

See figure 1 for the corresponding graphical model description.

Another way to look at this hierarchical model is to consider that our fixed prior on θ from formula (2) is actually a random variable (in our case a normal distribution) parameterized by λ which leads to the following posterior formulation:

$$P(\theta, \lambda|x) = \frac{P(x|\theta) * P(\theta|\lambda) * P(\lambda)}{P(x)} \tag{7}$$

Note that we can factorize $P(x|\theta)$ and $P(\theta|\lambda)$ due to their conditional independence. This formulation also makes apparent that the posterior contains estimation of the individual subject parameters θ_j and group parameters λ.

1.3.2 Empirical Bayesian Approximation

Empirical Bayes can be regarded as an approximation of equation (7). To derive this approximation consider $P(\theta|x)$ which we can calculate by integrating over $P(\lambda)$:

$$P(\theta|x) = \frac{P(x|\theta)}{P(x)} \int P(\theta|\lambda) P(\lambda) \, \mathrm{d}\lambda \tag{8}$$

Now, if the true distribution $P(\theta|\lambda)$ is sharply peaked, the integral can be replaced with the point estimate of its peak $\lambda^\star$:

$$P(\theta|x) \simeq \frac{P(x|\theta) P(\theta|\lambda^\star)}{P(x|\lambda^\star)} \tag{9}$$

Note, however, that $\lambda^\star$ depends itself on $P(\theta|x)$. One algorithm to solve this interdependence is Expectation Maximization (EM) (Dempster et al., 1977). EM is an iterative algorithm that alternates between computing the expectation of $P(\theta|x)$ (this can be easily done by Laplace Approximation

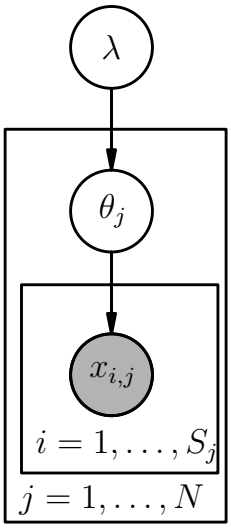

Figure 1: Graphical notation of a hierarchical model. Circles represent continuous random variables. Arrows connecting circles specify conditional dependence between random variables. Shaded circles represent observed data. Finally, plates around graphical nodes mean that multiple identical, independent distributed random variables exist.

(Azevedo-filho and Shachter, 1994)) and then maximizing the prior point estimate $\lambda^\star$ based on the current values obtained by the expectation step. This updated point estimate is then used in turn to recompute the expectation. The algorithm is run until convergence or some other criterion in reached. This approach is used for example by Huys et al. (2012) to fit their reinforcement learning models.

1.3.3 Markov-Chain Monte-Carlo

As mentioned above, the posterior is often intractable to compute analytically. While Empirical Bayes provides a useful approximation, an alternative approach is to estimate the full posterior by drawing samples from it. One way to achieve this is to construct a Markov-Chain that has the same equilibrium distribution as the posterior (Gamerman and Lopes, 2006). Algorithms of this class are called Markov-Chain Monte Carlo (MCMC) samplers.

One common and widely applicable algorithm is Metropolis-Hastings (Chib and Greenberg, 1995; Andrieu et al., 2003). Assume we wanted to generate samples θ from the posterior $p(\theta|x)$. In general, we can not sample from $p(\theta|x)$ directly. Metropolis-Hastings instead generates samples θ^t from a proposal distribution $q(\theta^t|\theta^{t-1})$ where the next position θ^t only depends on the previous position at θ^{t-1} (i.e. the Markov-property). For simplicity we will assume that this proposal distribution is symmetrical; i.e. $q(\theta^t|\theta^{t-1}) = q(\theta^{t-1}|\theta^t)$. A common choice for the proposal distribution is the Normal distribution, formally:

$$\theta^t \sim \mathcal{N}(\theta^{t-1}, \sigma^2) \tag{10}$$

The proposed jump to θ^t is then accepted with probability α:

$$\alpha = \min(1, \frac{p(\theta^t|x)}{p(\theta^{t-1}|x)}) \tag{11}$$

In other words, the probability of accepting a jump depends on the probability ratio of the proposed jump position θ^t to the previous position θ^{t-1}. Critically, in this probability ratio, the intractable integral in the denominator (i.e. $p(x) = \int p(x, \theta)\, \mathrm{d}\theta$) cancels out. This can be seen by applying Bayes formula (2):

$$\frac{p(\theta^t|x)}{p(\theta^{t-1}|x)} = \frac{\frac{p(x|\theta^t)p(\theta^t)}{p(x)}}{\frac{p(x|\theta^{t-1})p(\theta^{t-1})}{p(x)}} = \frac{p(x|\theta^t)p(\theta^t)}{p(x|\theta^{t-1})p(\theta^{t-1})} \tag{12}$$

Thus, to calculate the probability of accepting a jump we only have to evaluate the likelihood and prior, *not* the intractable posterior.

Note that θ^0 has to be initialized at some position and can not directly be sampled from the posterior. From this initial position, the Markov chain will explore other parts of the parameter space and only gradually approach the posterior region. The first samples generated are thus not from the true posterior and are often discarded as "burn-in". Note moreover that once the algorithm reaches a region of high probability it will continue to explore lower probability regions in the posterior, albeit with lower frequency. This random-walk behavior is due to the probability ratio α which allows Metropolis-Hastings to also sometimes accept jumps from a high probability position to a low probability position.

Another common algorithm is Gibbs sampling that iteratively updates each individual random variable conditional on the other random variables set to their last sampled value (e.g Frey and Jojic, 2005). Starting at some configuration θ^0, the algorithm makes T iterations over each random variable θ_i. At each iteration t each random variable is sampled conditional on the current $(t-1)$ value of all other random variables that it depends on:

$$\theta_i^t \sim p(\theta_i^{(t)}|\theta_{i\neq j}^{(t-1)}) \tag{13}$$

Critically, $\theta_{i\neq j}^{(t-1)}$ are treated as constant. The sampled value of $\theta_i^{(t)}$ will then be treated as fixed while sampling the other random variables.

Note that while Gibbs sampling never rejects a sample (which often leads to faster convergence and better mixing), in contrast to Metropolis-Hastings, it does require sampling from the conditional distribution which is not always tractable.

1.4 Likelihood free methods

Several likelihood-free methods have emerged in the past (for a review, see Turner and Van Zandt (2012)). Instead of an analytical solution of the likelihood function, these methods require a sampling process that can simulate a set of data points from a generative model for each θ. We will call the simulated data y and the observed data x. Approximate Bayesian Computation (ABC) relies on a distance measure $\rho(x, y)$ that compares how similar the simulated data y is to the observed data x (commonly, this distance measure relies on summary statistics). We can then use the Metropolis-Hastings algorithm introduced before and change the acceptance ration α (11) to use $\rho(x, y)$ instead of a likelihood function.

$$\alpha = \begin{cases} \min(1, \frac{p(\theta^t)}{p(\theta^{t-1})}) & \text{if } \rho(x, y) \leq \epsilon_0 \\ 0 & \text{if } \rho(x, y) \geq \epsilon_0 \end{cases} \tag{14}$$

where ϵ_0 is an acceptance threshold. Large ϵ_0 will result in higher proposal acceptance probability but a worse estimation of the posterior while small ϵ_0 will lead to better posterior estimation but

slower convergence.

An alternative approach to ABC is to construct a synthetic likelihood function based on summary statistics (Wood, 2010). Specifically, we sample N_r multiple data sets $y_{1,\ldots,N_r}$ from the generative process. We then compute summary statistics $s_{1,\ldots,N_r}$ for each simulated data set[2]. Based on these summary statistics we then construct the synthetic likelihood function to evaluate θ (see figure 2 for an illustration):

$$p(x|\theta) \simeq \mathcal{N}(S(x); \mu_\theta, \Sigma_\theta) \tag{15}$$

This synthetic likelihood function based on summary statistics can then be used as a drop-in replacement for e.g. the Metropolis-Hastings algorithm outlined above.

1.5 Model Comparison

Computational models often allow formulation of several plausible accounts of cognitive behavior. One way to differentiate between these various plausible hypotheses as expressed by alternative models is model comparison: which of several alternative models provides the best explanation of the data? In the following we review various methods and metrics to compare hierarchical models. The most critical property for model comparison is that model complexity gets penalized because more complex models have greater degrees of freedom and could thus overfit data. Several model comparison measures have been devised.

1.5.1 Deviance Information Criterion

The Deviance Information Criterion (DIC) is a measure which trades off model complexity and model fit (Spiegelhalter et al., 2002). Several similar measures exist such as Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). However, both these measures use the number of parameters as a proxy for model complexity. While a reasonable approximation to the complexity of non-hierarchical models, the relationship between model parameters (some of which are latent) and complexity in hierarchical models is more intricate. The DIC measure instead infers the number of parameters from the posterior. The DIC is computed as follows:

$$\text{DIC} = \bar{D} + pD \tag{16}$$

where

$$pD = \bar{D} - \hat{D} \tag{17}$$

$\bar{D}$ is the posterior mean of the deviance (i.e. $-2 * \log(\text{likelihood})$) and $\hat{D}$ is a point estimate of the deviance obtained by substituting in the posterior means. Loosely, $\bar{D}$ represents how well the model fits the data on average while $\hat{D}$ captures the deviance at the best fitting parameter combination. pD then acts as a measure related to the posterior variability and used as a proxy for the effective number of parameters. Complex models with many parameters will tend to have higher posterior variability and thus result in increased pD penalization.

Note that the only parameters that affect $\hat{D}$ directly in our hierarchical model (equation 7) are the subject parameters θ_i. Thus, DIC estimates model fit based on how well individual subjects explain the observed data.

[2]The summary statistics must (i) be sufficient and (ii) normally distributed

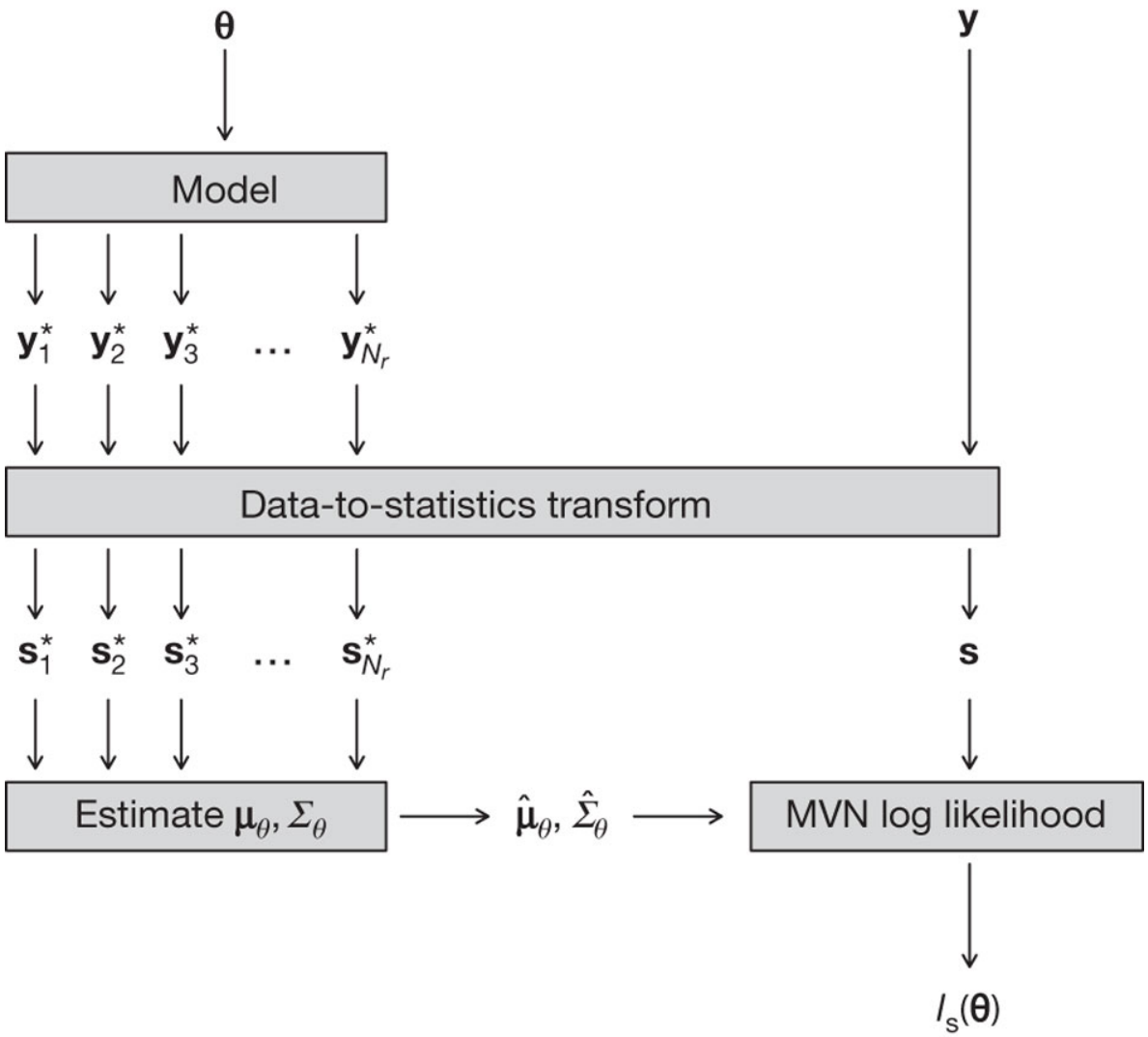

Figure 2: Construction of a synthetic likelihood. To evaluate parameter vector θ, N_r data sets $y_{1,\ldots,N_r}$ are sampled from the generative model. On each sampled data set summary statistics $s_{1,\ldots,N_r}$ are computed. Based on these summary statistics a multivariate normally distribution is approximated with mean μ_θ and covariance matrix Σ_θ. The likelihood is approximated by evaluating summary statistics of the actual data on the log normal distribution with the estimated μ_θ and Σ_θ. Reproduced from (Wood, 2010).

1.5.2 BIC

The Bayesian Information Criterion (BIC) is defined as follows:

$$\text{BIC} = -2 * \log p(x|\hat{\theta}^{ML}) + k * \log(n) \tag{18}$$

where k is the number of free parameters, n is the number of data points, x is the observed data and $\log p(x|k)$ is the likelihood of the parameters given the data (Schwarz, 1978).

While BIC can not directly be applied to hierarchical models (as outlined above), it is possible to integrate out individual subject parameters (e.g. Huys et al., 2012):

$$\log p(x|\hat{\theta}^{ML}) = \sum_i \log \int p(x_i|h)p(h|\hat{\theta}^{ML})\,\mathrm{d}h \tag{19}$$

where x_i is the data belonging to the ith subject. The resulting score is called integrated BIC.

Since the subject parameters are integrated out, integrated BIC estimates how well the group parameters are able to explain the observed data.

1.5.3 Bayes Factor

Another measure to compare two models is the Bayes Factor (BF) (Kass and Raftery, 1995). It is defined as the ratio between the marginal model probabilities of the two models:

$$BF = \frac{p(x|M_1)}{p(x|M_2)} = \frac{\int p(\theta_1|M_1)p(x|\theta_1, M_1)\,\mathrm{d}\theta_1}{\int p(\theta_2|M_2)p(x|\theta_2, M_2)\,\mathrm{d}\theta_2} \tag{20}$$

The magnitude of this ratio informs the degree one should belief in one model compared to the other.

As BF integrates out subject *and* group parameters this model comparison measure should be used when different classes of models are to be compared in their capacity to explain observed data.

1.6 Mixture Models

1.6.1 Gaussian Mixture Models

Mixture models infer k number of clusters in a data set. The assumption of normally distributed clusters leads to a Gaussian Mixture Model (GMM) with a probability density function as follows:

$$p(x|\pi, \mu_{1,\ldots,K}, \sigma_{1,\ldots,K}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k^2) \tag{21}$$

Each observed data point x_i can be created by drawing a sample from the normal distribution selected by the unobserved indicator variable z_i which itself is distributed according to a multinomial distribution π:

$$\mu_k, \sigma_k \sim \mathrm{G}_0() \tag{22}$$

$$z_i \sim \pi \tag{23}$$

$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2) \tag{24}$$

where the base measure G_0 defines the prior for μ_k and σ_k. To simplify the inference it is often advisable to use a conjugate prior for these paramters. For example, the normal distribution is the conjugate prior for a normal distribution with known variance:

$$\mu_k \sim \mathcal{N}(\mu_0, \sigma_0) \tag{25}$$

In a similar fashion, we can assign the mixture weights a symmetric Dirichlet prior:

$$\pi \sim \mathrm{Dir}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}) \tag{26}$$

Note that the GMM assumes a mixture distribution on the level of the observed data x_i. However, in our relevant case of a multi-level hierarchical model we need to place the mixture at the level of the latent subject parameters instead of the observed data. As before, we use the subject index $j = 1, \ldots, N$.

$$\mu_k, \sigma_k \sim \mathrm{G}_0() \tag{27}$$
$$\pi \sim \mathrm{Dir}(\alpha) \tag{28}$$
$$z_j \sim \mathrm{Categorical}(\pi) \tag{29}$$
$$\theta_j \sim \mathcal{N}(\mu_{z_j}, \sigma^2_{z_j}) \tag{30}$$
$$x_{i,j} \sim \mathrm{f}(\theta_j) \tag{31}$$

Where f denotes the likelihood function.

Interestingly, the famous K-Means clustering algorithm is identical to a Gaussian Mixture Model (GMM) in the limit $\sigma^2 \to 0$ (Kulis et al., 2012). K-Means is an expectation maximization (EM) algorithm that alternates between an expectation step during which data points are assigned to their nearest cluster centroids and a maximization step during which new cluster centroids are estimated. This algorithm is repeated until convergence is reached (i.e. no points are reassigned to new clusters).

1.6.2 Dirichlet Process Gaussian Mixture Models

Dirichlet processes Gaussian mixture models (DPGMMs) belong to the class of Bayesian non-parametrics (Antoniak, 1974). They can be viewed as a variant of GMMs with the critical difference that they assume an infinite number of potential mixture components (see Gershman and Blei (2012) for a review). Such mixture models can infer sub-groups when the data is heterogeneous as is generally the case in patient populations. While the mindset describing these methods was their application towards the SSM, their applicability is much more general than that. For example, the case-studies described above which used, among others, RL models to identify differences between HC and psychiatric patients could easily be embedded into this hierarchical Bayesian mixture model framework we outlined here. Such a combined model would estimate model parameters and identify subgroups simultaneously. There are multiple benefits to such an approach. First, computational models fitted via hierarchical Bayesian estimation provide a tool to accurately describe the neurocognitive functional profile of individuals. Second, the mixture model approach is ideally suited to deal with the heterogeneity in patients but also healthy controls (Fair et al., 2012). Third, by testing psychiatric patients with a range of diagnoses (as opposed to most previous research studies that only compare patients with a single diagnosis, e.g. SZ, to controls) we might be able to identify shared pathogenic cascades as suggested by Buckholtz and Meyer-Lindenberg (2012).

$$p(x|\pi, \mu_{1,\ldots,\infty}, \sigma_{1,\ldots,\infty}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k^2) \tag{32}$$

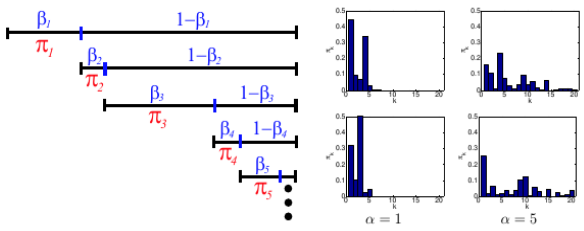

Figure 3: Left: Stick-breaking process. At each iteration (starting from the top) a π is broken off with relative length $\sim \text{Beta}(1, \alpha)$. Right: Histogram over different realizations of the stick-breaking process. As can be seen, higher values of hyperprior α lead to a more spread out distribution. Taken from Eric Sudderth's PhD thesis.

As above, we specify our generative mixture model:

$$\mu_k, \sigma_k \sim \text{G}_0() \tag{33}$$
$$z_i \sim \text{Categorical}(\pi) \tag{34}$$
$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i}) \tag{35}$$

with the critical difference of replacing the hyperprior π with the *stick breaking process* (Sethuraman, 1991):

$$\pi \sim \text{StickBreaking}(\alpha) \tag{36}$$

The stick-breaking process is a realization of a Dirichlet process (DP). Specifically, $\pi = \{\pi_k\}_{k=1}^{\infty}$ is an infinite sequence of mixture weights derived from the following process:

$$\beta_k \sim \text{Beta}(1, \alpha) \tag{37}$$
$$\pi_k \sim \beta_k * \prod_{l=1}^{k-1}(1 - \beta_l) \tag{38}$$

with $\alpha > 0$. See figure 3 for a visual explanation.

The Chinese Restaurant Process (CRP) – named after the apparent infinite seating capacity in Chinese restaurants – allows for a more succinct model formulation. Consider that customers z_i are coming into the restaurant and are seated at table k with probability:

$$p(z_i = k | z_{1,\ldots,n-1}, \alpha, K) = \frac{n_k + \alpha/K}{n - 1 + \alpha}$$

where $k = 1 \ldots K$ is the table and n_k is the number of customers already sitting at table k (see figure 4 for an illustration). It can be seen that in the limit as $K \to \infty$ this expression becomes:

$$p(z_i = k | z_{1,\ldots,n-1}, \alpha) = \frac{n_k}{n - 1 + \alpha}$$

Thus, as customers are social, the probability of seating customer z_i to table k is proportional the number of customers already sitting at that table. This desirable clustering property is also known as

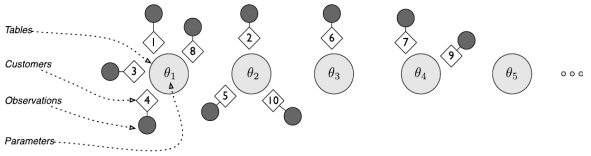

Figure 4: Illustration of the Chinese Restaurant Process. Customers are seated at tables with parameters θ. The more customers are already seated at a table, the higher the probability that future customers are seated at the same table (i.e. clustering property). Taken from Gershman and Blei (2012).

the "rich get richer".

Note that for an individual empty table k at which no customer has been seated (i.e. $n_k = 0$) the probability of seating a new customer to that table goes to 0 in the limit as $K \to \infty$. However, at the same time the number of empty tables approaches infinity. Consider that we have so far seated L customers to tables and the set $\mathbf{Q}$ contains all empty tables such that there are $|\mathbf{Q}| = K - L$ empty tables in the restaurant. The probability of seating a customer z_i at an empty table becomes:

$$p(z_i \in \mathbf{Q} | \mathbf{z}_{1,\dots,n-1}, \alpha) = \frac{\alpha}{n - 1 + \alpha}$$

As can be seen, the probability of starting a new table is proportional to the concentration parameter α. Intuitively, large values of the dispersion parameter α lead to more clusters being used.

Thus, while the Stick-Breaking process sampled mixture weights from which we had to infer cluster assignments, the CRP allows for direct sampling of cluster assignments. The resulting model can then be written as:

$$\mu_k, \sigma_k \sim \mathrm{G}_0() \tag{39}$$
$$z_{1,\dots,N} \sim \mathrm{CRP}(\alpha) \tag{40}$$
$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i}) \tag{41}$$

Finally, in a hierarchical group model we would need to place the infinite mixture on the subject level rather than the observed data level:

$$\mu_k, \sigma_k \sim \mathrm{G}_0() \tag{42}$$
$$z_j \sim \mathrm{CRP}(\alpha) \tag{43}$$
$$\theta_j \sim \mathcal{N}(\mu_{z_j}, \sigma^2_{z_j}) \tag{44}$$
$$x_{i,j} \sim \mathrm{F}(\theta_j) \tag{45}$$

See figure 5 for a graphical model description.

Note that while the potential number of clusters is infinite, any realization of this process will always lead to a finite number of clusters as we always have finite amounts of data. However, this method allows the addition (or subtraction) of new clusters as new data becomes available.

References

C Andrieu, N De Freitas, A Doucet, and MI Jordan. An introduction to MCMC for machine learning. *Machine learning*, 2003. URL `http://www.springerlink.com/index/xh62794161k70540.pdf`.

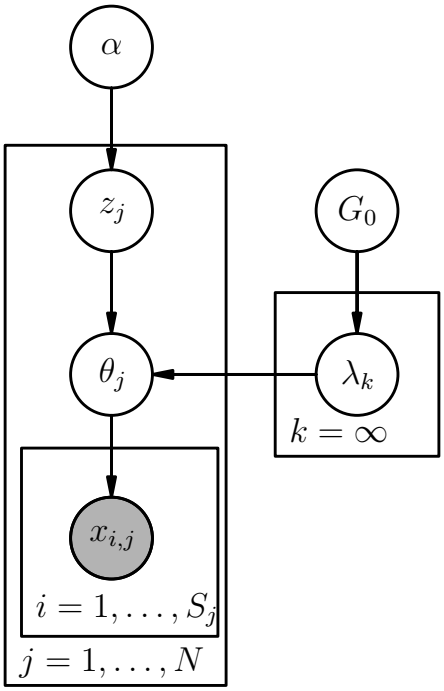

Figure 5: Graphical model representation of the hierarchical Dirichlet process mixture model. Group parameters $\lambda_k = (\mu_k, \sigma_k)$. See text for details.

Adriano Azevedo-filho and Ross D Shachter. Laplace s Method Approximations for Probabilistic Inference in Belief Networks with Continuous Variables. pages 28–36, 1994.

S Chib and E Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 1995. URL http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1995.10476177.

AP Dempster, NM Laird, and DB Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. ...*, 39(1):1–38, 1977. URL http://www.jstor.org/stable/10.2307/2984875.

BJ Frey and N Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *Pattern Analysis and Machine Intelligence, ...*, 2005. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1471706.

D Gamerman and HF Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. 2006. URL http://books.google.com/books?hl=en&lr=&id=yPvECi_L3bwC&oi=fnd&pg=PR13&dq=gamerman+bayesian&ots=Nis

A Gelman, JB Carlin, HS Stern, and DB Rubin. *Bayesian data analysis*. 2003. URL http://books.google.com/books?hl=en&lr=&id=TNYhnkXQSjAC&oi=fnd&pg=PP1&dq=Gelman+Carlin+Stern+04&ots

Samuel J. Gershman and David M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, February 2012. ISSN 00222496. doi: 10.1016/j.jmp.2011.08.004. URL http://linkinghub.elsevier.com/retrieve/pii/S002224961100071X.

Quentin J. M. Huys, Neir Eshel, Elizabeth O'Nions, Luke Sheridan, Peter Dayan, and Jonathan P. Roiser. Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees. *PLoS Computational Biology*, 8(3):e1002410, March 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002410. URL http://dx.plos.org/10.1371/journal.pcbi.1002410.

Robert E. Kass and Adrian E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995. ISSN 0162-1459. doi: 10.1080/01621459.1995.10476572. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572.

Brian Kulis, Michael I Jordan, Jordan Eecs, and Berkeley Edu. Revisiting k-means : New Algorithms via Bayesian Nonparametrics. 2012.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, March 1978. ISSN 2168-8966. URL http://projecteuclid.org/euclid.aos/1176344136.

J Sethuraman. A constructive definition of Dirichlet priors. 1991. URL http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA238689.

Pl Smith. Stochastic Dynamic Models of Response Time and Accuracy: A Foundational Primer. *Journal of mathematical psychology*, 44(3):408–463, September 2000. ISSN 0022-2496. doi: 10.1006/jmps.1999.1260. URL http://www.ncbi.nlm.nih.gov/pubmed/10973778.

DJ Spiegelhalter, NG Best, and Bradley P. Carlin. Bayesian measures of model complexity and fit. *Journal of the Royal ...*, 2002. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353/full.

Brandon M. Turner and Trisha Van Zandt. A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85, April 2012. ISSN 00222496. doi: 10.1016/j.jmp.2012.02.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0022249612000272.

Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–4, August 2010. ISSN 1476-4687. doi: 10.1038/nature09319. URL `http://www.ncbi.nlm.nih.gov/pubmed/20703226`.