An Introduction to Nonparametric Hierarchical Bayesian Modelling with a Focus on Multi-Agent Learning

Volker Tresp[1] and Kai Yu[2]

[1] Siemens AG, 81730 München, Germany
Volker.Tresp@siemens.com,
[2] Siemens AG, 81730 München, Germany
Kai.Yu@siemens.com

Abstract. In this chapter, we address the situation where agents need to learn from one another by exchanging learned knowledge. We employ hierarchical Bayesian modelling, which provides a powerful and principled solution. We point out some shortcomings of parametric hierarchical Bayesian modelling and thus focus on a nonparametric approach. Nonparametric hierarchical Bayesian modelling has its roots in Bayesian statistics and, in the form of Dirichlet process mixture modelling, was recently introduced into the machine learning community. In this chapter, we hope to provide an accessible introduction to this particular branch of statistics. We present the standard sampling-based learning algorithms and introduce a particular EM learning approach that leads to efficient and plausible solutions. We illustrate the effectiveness of our approach in context of a recommendation engine where our approach allows the principled combination of content-based and collaborative filtering.

1 Introduction

There are many occasions where agents should "learn" from one another. As an example, the effectiveness of a treatment for a cardiac disease is a function of the severity of the disease and patient characteristics but might also vary from hospital to hospital (due to hidden factors such as varying patient population, staff training, local expertise, ...). Thus models that predict the outcomes for different hospitals should be quite similar but will also be different to some degree. Despite the differences in the models it would be advantageous if various models could benefit from each other's learned knowledge, in particular in the case that there is only a small data set available for each hospital. A similar situation arises in the design of recommendation engines that predict the interests of users in various items. Essentially each user is an individual and one should learn a personal model for each user. On the other hand if few training data points for the active user are available one would like to benefit from the recommendations of like-minded users, as in collaborative filtering. In machine learning, the scenarios described are known as transfer learning or meta learning.

In the Bayesian literature this framework falls into hierarchical Bayesian (HB) modelling. The basic idea in HB modelling is that information between different models can be exchanged via common hyperparameters. In this chapter, we provide an introduction to HB modelling. We emphasize that, in our view, HB by itself is useful but also severely limited since it is inflexible in the representation of the "learned prior". Additional flexibility is obtained by a process called Dirichlet enhancement in which the prior distribution is specified in terms of a highly flexible multinomial distribution with a Dirichlet prior. Of particular interest is the limit that the number of states in the multinomial becomes infinite in which case we obtain a Dirichlet process and our hierarchical model becomes a Dirichlet process mixture model.[3] Dirichlet process mixture models originated in Bayesian statistics [11] [1] and recently found growing interest in the machine learning community, in particular in the context of infinite mixture models. A particular advantage of Dirichlet process mixture models is that the number of components required for achieving a good overall model is automatically determined by the algorithm. In the problem setting described in this chapter this feature is of minor interest in comparison to the benefits achieved by the transfer of learned knowledge via HB modelling. We describe the standard sampling approach for inference in Dirichlet process mixture models and also introduce a particular expectation maximization (EM) solution that is powerful and efficient in the frameworks addressed in this chapter.

The chapter is organized as follows. In the following section we provide an intuitive motivation for nonparametric HB modelling and present the first algorithmic solution to the problem. In Section 3 we introduce HB modelling more systematically and discuss some of its shortcomings. In Section 4 we introduce the process of a Dirichlet enhancement, which is a first step towards nonparametric HB modelling. The finite-dimensional approach presented in Section 4 is not of great practical interest by itself but provides the basis for the infinite-dimensional nonparametric HB models described in Section 5. We discuss stochastic sampling and EM as approaches towards parameter inference. In Section 6 we illustrate the effectiveness of our approach using the example of a recommendation engine where our approach allows the principled combination of content-based filtering and collaborative filtering. In Section 7 we discuss related work, in particular recent work on infinite models. In Section 8 we provide conclusions.

2 Intuitive Introduction

2.1 Bayesian Modelling

We will develop the ideas based on two-class classification although the same concepts are valid for general probabilistic models, e.g., for regression and density estimation. Readers who want to fresh up on Bayesian statistics may consult

[3] Dirichlet process mixture models are also known as mixtures of Dirichlet processes (MDPs).

the excellent tutorial [15]. Let $P(Y = y|x, \theta)$ denote the probability that Y assumes the state $y \in \{0, 1\}$ given features x and given a parameter set $\theta = \{\theta_j\}_j$. In a Bayesian setting one defines an a priori distribution $P(\theta|h_{prior})$ with hyperparameters $h = h_{prior}$. Both prior distribution and hyperparameters specify one's prior belief. The prior belief is typically rather unspecific or non-informative and thus the prior distribution should place nonzero probability on all reasonable model parameters.

As example, in Figure 1A the prior distribution might be specified as a Gaussian distribution with

$$P(\theta|h_{prior}) = \mathcal{N}(\theta|\mu_{prior}, \Sigma_{prior})$$

with $h_{prior} = \{\mu_{prior}, \Sigma_{prior}\}$.

Bayesian learning means updating the parameter distribution based on available training data. Given a data set with N_D data points $D = \{(x_n, y_n)\}_{n=1}^{N_D}$ one can calculate the posterior parameter density using Bayes formula as

$$P(\theta|D, h_{prior}) = \frac{1}{P(D)} P(D|\theta) P(\theta|h_{prior})$$

where, in our classification example, assuming exchangeability,

$$P(D|\theta) = \prod_{n=1}^{N_D} P(y_n|x_n, \theta).$$

Note that in this chapter we do not treat the inputs x probabilistically and focus on the modelling of the condition probability distribution $y|x$.

For classifying a new pattern we obtain the predictive distribution

$$P(Y = y|x, D, h_{prior}) = \int P(Y = y|x, \theta) P(\theta|D, h_{prior}) d\theta.$$

If additional data points become available, the posterior parameter distribution $P(\theta|D, h_{prior})$ now assumes the role of the new "learned prior" , i.e., the available knowledge prior to the arrival of the additional data. In the case that the prior distribution is conjugate to the likelihood function, we obtain

$$P(\theta|D, h_{prior}) = P(\theta|h_{post}),$$

i.e., the posterior parameter distribution has the functional form of the prior distribution but with new hyperparameters h_{post}. Returning to our example, we would expect that

$$P(\theta|h_{post}) = \mathcal{N}(\theta|\mu_{post}, \Sigma_{post})$$

with $h_{post} = \{\mu_{post}, \Sigma_{post}\}$ and where $\lim_{N_D \to \infty} \det \Sigma_{post} = 0$, i.e., the posterior distribution become increasingly concentrated (Figure 1B) with an increasing number of data points and asymptotically is locally peaked at the maximum likelihood solution

$$\theta^{ML} := \arg\max_{\theta} P(D|\theta).$$

2.2 Hierarchical Bayesian Modelling

Now assume, that we have obtained M data sets $\{D_j\}_{j=1}^M$ for related but not identical settings and we have trained M *different models* with parameters $\{\theta_j\}_{j=1}^M$ on those data sets. For the sake of argument let's assume that each data set is sufficiently large such that $P(\theta_j|D_j, h_{prior})$ is sharply peaked at the maximum likelihood (ML) estimate θ_j^{ML}. Let $\{\theta_k^{ML}\}_{k=1}^M$ denote the maximum likelihood estimates for the M models. Recall that since the models were trained on different data sets generated from different settings, the maximum likelihood parameter values are not identical. Figure 1C illustrates the set of maximum likelihood parameter estimates. Now, if a new model concerns a related problem, then it makes sense to select new hyperparameters h_{hb} such that $P(\theta|h_{hb})$ approximates the empirical distribution given by the maximum likelihood parameter estimates instead of using the original uninformed prior $P(\theta|h_{prior})$. In this way the new model can inherit knowledge acquired not only from its own data set but also from the other models.

Returning to our example, we would expect that for a new setting with a new model with parameters θ_{M+1}

$$P(\theta_{M+1}|\{D_j\}_{j=1}^M) \approx P(\theta_{M+1}|h_{hb}) \tag{1}$$

where, in the example, $P(\theta_{M+1}|h_{hb}) = \mathcal{N}(\theta_{M+1}|\mu_{hb}, \Sigma_{hb})$, with $h_{hb} = \{\mu_{hb}, \Sigma_{hb}\}$ and where now in the non-degenerate case

$$\lim_{M\to\infty} \det \Sigma_{hb} > 0$$

and the entries of Σ_{hb} converge to fixed typically nonzero values (Figure 1C). What we have just described is the basis for hierarchical Bayesian modelling that we will introduce more formally in Section 3.

2.3 Nonparametric Hierarchical Bayesian Modelling

In more cases than not, the empirical distribution of the maximum likelihood parameters $\{\theta_k^{ML}\}_{k=1}^M$ will not fall into the class of distributions that can be described by $P(\theta|h)$ for any h. If the assumed noninformative prior is too inflexible to truthfully model the learned prior, then this is a severe limitation of the classical HB approach. See for example Figure 1D. Thus we might prefer a nonparametric approximation in the form of the empirical nonparametric distribution of the maximum likelihood parameters

$$P(\theta_{M+1}|\{D_j\}_{j=1}^M) \approx \frac{1}{M}\sum_{k=1}^{M} \delta_{\theta_k^{ML}},$$

where $\delta_{\theta_k^{ML}}$ is a distribution concentrated at a single point θ_k^{ML}.

Now if we receive the data set D_{M+1} for the new setting, we predict

$$P(Y_{M+1} = y|x, D_{M+1}) \approx \frac{1}{C}\int P(Y_{M+1} = y|x, \theta)P(D_{M+1}|\theta)\sum_{j=1}^{M} \delta_{\theta_j^{ML}}\, d\theta$$

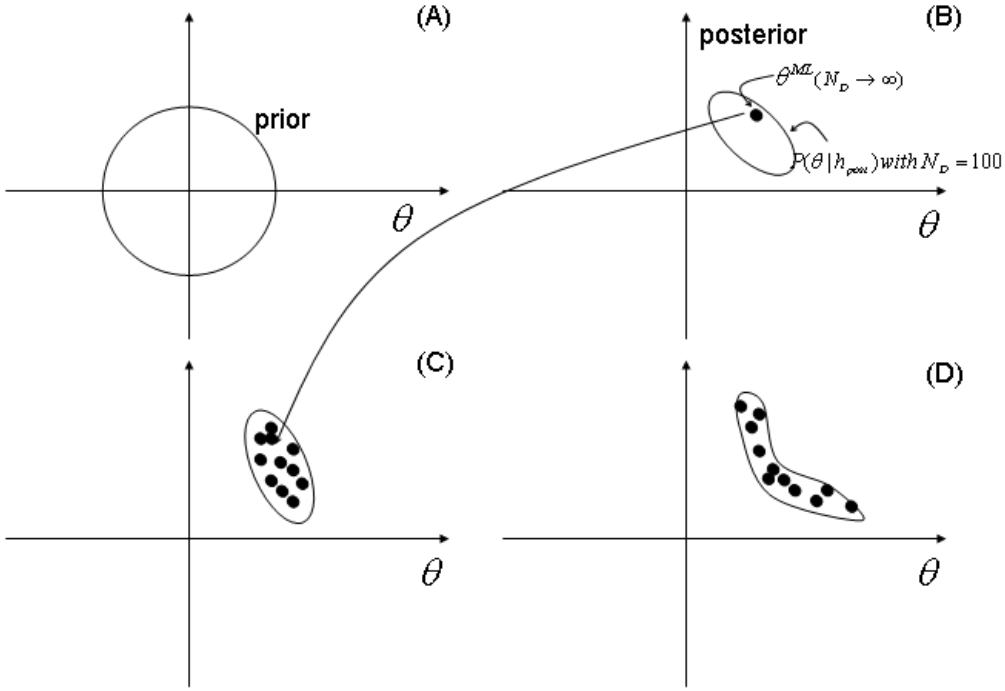

Fig. 1. A: The circle indicates the standard deviation of the prior Gaussian distribution with mean zero representing $P(\theta|h_{prior}) = \mathcal{N}(\theta|\mu_{prior}, \Sigma_{prior})$. B: The posterior parameter distribution $P(\theta|h_{post}) = \mathcal{N}(\theta|\mu_{post}, \Sigma_{post})$ with lets say $N_D = 100$ data points; shape and location of the Gaussian have changed. With $N_D \to \infty$, $P(\theta|h_{post})$ is concentrated at the maximum likelihood estimate θ^{ML}. C: Set of maximum likelihood estimates $\{\theta_j^{ML}\}_{j=1}^M$ and approximation $P(\theta_{M+1}|\{D_j\}_{j=1}^M) \approx \mathcal{N}(\theta|\mu_{hb}, \Sigma_{hb})$. The implicit assumption in HB modelling is that this distribution can be approximated by a member of the family of distributions assumed for the prior, i.e., in this example a Gaussian distribution. D: Here is an example where this distribution cannot be approximated by a Gaussian distribution. Thus, nonparametric HB with $P(\theta_{M+1}|\{D_j\}_{j=1}^M) \approx \frac{1}{M}\sum_{k=1}^M \delta_{\theta_k^{ML}}$ is more appropriate.

$$= \frac{1}{C} \sum_{j=1}^{M} P(D_{M+1}|\theta_j^{ML}) P(Y = y|x, \theta_j^{ML}) \quad (2)$$

where $C = \sum_{j=1}^{M} P(D_{M+1}|\theta_j^{ML})$ normalizes the distribution. Here and in the following, capital C stands for an appropriate normalization constant. Note that the result (Equation 2) is very intuitive. To make a prediction for setting $M+1$ for input x, each model $1, \ldots, M$ makes a prediction using its maximum likelihood parameter estimate and this prediction is then weighted with the probability that this model explains the data points D_{M+1} of the setting of interest. This means that initially, with only few data points available for setting $M+1$, the predictions of all previous models are essentially averaged. With more data points available for setting $M+1$, models that agree well with the data D_{M+1} obtain a higher weight.

3 Hierarchical Bayesian Modelling

In this and the following sections we will introduce HB modelling and non-parametric Bayesian modelling more formally. We start with HB. Recall that in Section 2.2 we essentially learned new hyperparameters h_{hb} to communicate learned knowledge. This is exactly the basis for the knowledge transfer via common hyperparameters in the framework of HB modelling. The joint probabilistic HB model is written as (Figure 2A)

$$P(h) \prod_{j=1}^{M} P(D_j|\theta_j) P(\theta_j|h). \quad (3)$$

The hyperparameters h —now considered to be random variables with prior distribution $P(h)$— are common to all models whereas each model has its own parameters $\{\theta_j\}_{j=1}^{M}$. Given the hyperparameters, the models are exchangeable, which means that the probabilistic model is invariant to a permutation (re-indexing) of the models.[4]

Now, for a model $M+1$ that did not yet receive any data points, we obtain as a full Bayesian version of Equation 1

$$P(\theta_{M+1}|\{D_j\}_{j=1}^{M}) \propto \int \left[P(\theta_{M+1}|h) P(h) \prod_{j=1}^{M} \int P(\theta_j|h) P(D_j|\theta_j) \, d\theta_j \right] dh. \quad (4)$$

In all but the simplest cases, the inference based on the HB model in Equation 4 does not lead to closed-form solutions and one typically relies on Markov Chain

[4] In contrast to the HB modelling assumption if we would assume that the models are all identical, then all data points are exchangeable and the probabilistic model is $P(h)P(\theta|h) \prod_{j=1}^{M} P(D_j|\theta)$, which would lead to one global model. The other extreme is that all models are independent $\prod_{j=1}^{M} P(h_j)P(D_j|\theta)P(\theta_j|h_j)$, which would result in M independent models.

Monte Carlo (MCMC) approximations. We do not want to get deeper into the issues of learning parametric HB models since we already concluded that the conventional HB approach is too limited for many applications. Readers more interested in the basics of HB modelling may consult [12].

4 Dirichlet Enhanced Hierarchical Bayesian Modelling

4.1 The Basic Idea

To alleviate the problem of HB we have to specify a parameterization of the prior parameter distribution that on the one hand can represent the assumed noninformative prior knowledge but also is flexible enough to be able to appropriately represent the "learned" prior to be communicated to a new model. The concept we are applying here is sometimes referred to as Dirichlet enhancement [10] and the basic idea is to replace the parametric prior distribution by a finite or infinite multinomial distribution with a Dirichlet prior. The essential features are that, first, the multinomial distribution by itself poses no constraint on the distributions that can be represented and that, second, the noninformative prior knowledge can be encoded in the form of the base distribution of the Dirichlet (which we will introduce further down). In this section we will consider the case that the model parameters can only assume values out of a given finite set of size K. The finite case is mathematically considerably easier and already introduces the main features of Dirichlet enhanced HB modelling. From an application point of view the case that $K \to \infty$ is of greater importance and will be discussed on the the following section.

To represent the model parameters we introduce a random variable Θ_j for each model j that can be in states $\theta_1, \ldots, \theta_K$. We further assume that a particular state is chosen by a multinomial distribution such that, for all j, $P(\Theta_j = \theta_k|g) = g_k$ with $g_k > 0$ and $\sum_{k=1}^{K} g_k = 1$ such that the probabilities $g_k, k = 1, \ldots, K$ play the role of the hyperparameters (previously the h)(Figure 2B). We specify our prior belief in terms of the conjugate prior that in this case is a Dirichlet distribution, i.e.,

$$P(g) = \mathrm{Dir}(g|\tau\alpha_1, \ldots, \tau\alpha_K) = \frac{1}{C} \prod_{k=1}^{K} g_k^{\tau\alpha_k - 1}$$

where $g = \{g_i\}_{i=1}^{K}$, $\alpha = \{\alpha_i\}_{i=1}^{K}$, $\alpha_k \geq 0$, $\sum_{k=1}^{K} \alpha_k = 1$ and with precision parameter $\tau > 0$. A description of the properties of a multinomial model with a Dirichlet prior including most equations used in this section can be found in the already mentioned tutorial [15]. A sample of a Dirichlet distribution is a probability distribution and the precision parameter τ corresponds to an equivalent sample size or weight. We can integrate out g and have $P(\Theta_j = \theta_j) = \alpha_j, j = 1, \ldots, K$.[5] Thus we can specify our non-informative prior belief by defining the $\alpha_j, j = 1, \ldots, K$

[5] Incidentally, the most likely configuration is also $g = \alpha$.

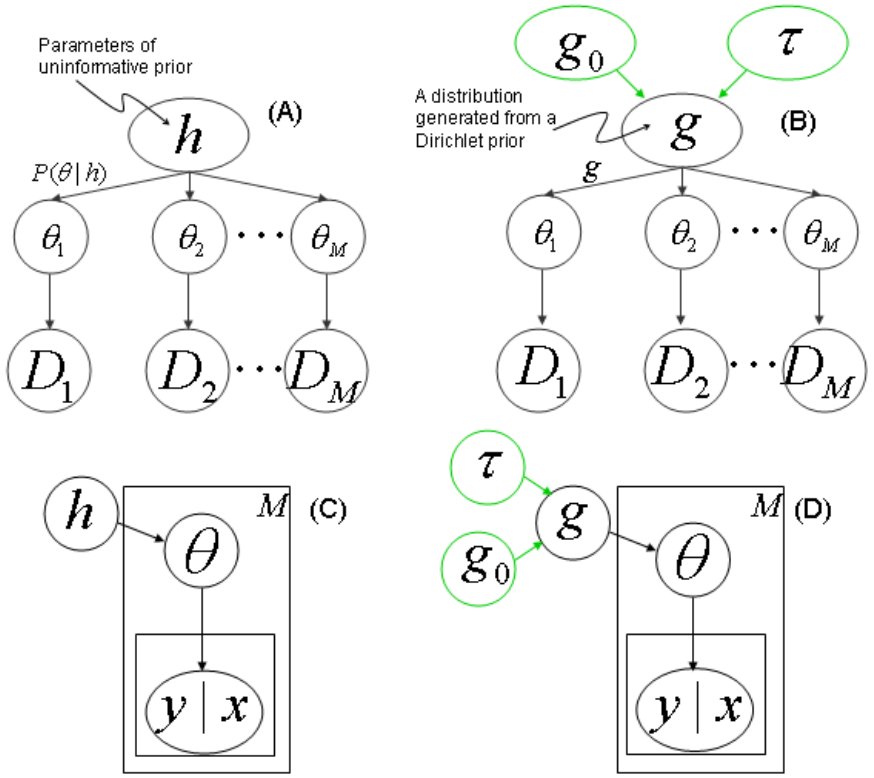

Fig. 2. A: A HB model. B: A Dirichlet enhancement HB model. C: A plate model for HB. The large plate indicates that M samples from $P(\theta|h)$ are generated; the smaller plate indicates that, repeatedly, data points are generated for each θ. D: A plate model for the Dirichlet enhanced HB. In B and D the finite dimensional hyperparameters h are replaced by the distribution g. In the finite-dimensional case, g is finite-dimensional and is generated from a Dirichlet distribution. In the infinite-dimensional case, g is infinite-dimensional and is generated from a Dirichlet process. We also indicate that, in the latter case, the prior distribution for g is defined using a base distribution G_0 with density g_0 and concentration parameter τ (see Section 5)

.

and the $\theta_j, j = 1, \ldots, K$ appropriately. The solution used in the following is to randomly select θ_j from $P(\Theta_j)$ and set $\alpha_j = 1/K, j = 1, \ldots, K$ (Figure 3 (top)). This is quite similar to the implementation of the non-informative prior belief in the infinite model of Section 5 where $K \to \infty$.

The joint distribution of the Dirichlet enhanced model is now (compare Equation 3 and Figure 2)

$$P(g) \prod_{j=1}^{M} P(D_j|\Theta_j = \theta_j) P(\Theta_j = \theta_j|g). \tag{5}$$

4.2 Sampling from a Dirichlet Model

First, we consider the simpler model $P(g)P(\Theta|g)$ consisting of a Dirichlet prior for g and a multinomial likelihood. We assume that for a fixed (but potentially unknown) g, N repeated samples of Θ are drawn. These samples form the set D_θ. Let's assume that in D_θ we have N_k instances of θ_k with $N = \sum_{k=1}^{K} N_k$.

Since the Dirichlet distribution is conjugate to the multinomial distribution, we obtain for the posterior distributions for g also a Dirichlet distribution with

$$P(g|D_\theta) = \mathrm{Dir}(g|\tau\alpha_1 + N_1, \ldots, \tau\alpha_K + N_k).$$

A nice property is that one can integrate out g to obtain the posterior predictive density [15]

$$P(\Theta = \theta_k|D_\theta) = \frac{\tau\alpha_k + N_k}{\tau + N}. \tag{6}$$

Equations 6 says that we can conveniently calculate the predictive distribution without the need for the explicit estimation of g. This is of great importance in the next section in the context of Dirichlet processes where g is infinite dimensional and could not explicitly be represented. According to Equation 6, a state becomes more likely if it has previously been observed with high frequency.

Note that Equation 6 also specifies how a new sample can be generated given previously *generated* samples D_θ. This sampling procedure generates data points from a fixed (but potentially unknown) g generated by the Dirichlet prior. Asymptotically, g can be inferred from the samples by noting that $g_k = \lim_{N\to\infty} N_k/N$.

The generation of samples according to Equation 6 is called a Pólya urn sampling process or a Chinese restaurant sampling process (for a recent discussion, see [27]). The essential feature is that if a state is sampled in the past, the probability that the same state is selected in the future is increased. This might be compared to a "Chinese restaurant" where customers select with higher probability a table that is already occupied by customers, or the Pólya urn where, if one draws a ball with a certain color, more than one ball with the same color is replaced and thus the probability of picking the same color in the future is increased.

From Equation 6 it is clear that if the precision parameter τ is large, many samples are generated independently from the base distribution α but if τ is

small, the first few samples quickly dominate the sampling procedure and the subsequently generated samples are quite clustered (see Figure 3).

4.3 Gibbs Sampling for Dirichlet Enhanced HB

We now return to the Dirichlet enhanced HB model from Equation 5 where for each setting j we have access to the data sets D_j with likelihood functions $P(D_j|\Theta_j = \theta_k)$. We will discuss two approaches for parameter inference in the HB setting. In this subsection we introduce Gibbs sampling, which is particularly attractive if K is large. Readers, not familiar with Gibbs sampling should consult [13]. The second approach is an EM solution, which is quite effective for smaller K and will be discussed in the next subsection.

Based on Equation 5 we can derive the conditional distribution of a variable of interest, say Θ_j, given samples from the remaining variables and given the data sets as

$$P(\Theta_j = \theta_k|\{\Theta_l\}_{l\neq j}, \{D_l\}_{l=1}^M) = \frac{1}{C}P(D_j|\Theta_j = \theta_k)\ P(\Theta_j = \theta_k|\{\Theta_l\}_{l\neq j})$$

$$= \frac{1}{C}(\tau\alpha_k + N_k)P(D_j|\Theta_j = \theta_k) \tag{7}$$

where we have N_k assignments of $\Theta_l = \theta_k$ in the remaining variables with $l \neq j$ and $\sum_k N_k = M - 1$. Note that we have integrated out g as in Subsection 4.2.

Thus a sample θ_k for setting j becomes more likely, if θ_k explains the D_j-th data set well and if either it is favored by the prior distribution (large α_k) or if θ_k is a sample already selected by the other models (large N_k). This latter property, that samples for different models influence each other, results in a sharing of information between the different models, as intended in HB modelling.

Note that the representation is upper limited by $\min(M, K)$, thus Gibbs sampling is particularly interesting for large K, i.e., if $K >> M$.

4.4 EM for Dirichlet Enhanced HB

We now discuss the EM solution to learning in Dirichlet enhanced HB. Here, we treat $\{\Theta_j\}_{j=1}^M$ as unknown variables, that we integrate out, and the goal is to find the MAP estimate of g that is defined as

$$g^{(MAP)} := \arg\max_g P(g) \prod_{j=1}^{M} \sum_{k=1}^{K} P(\Theta_j = \theta_k|g)P(D_j|\Theta_j = \theta_k).$$

The EM algorithm iterates for $t = 0, 1, 2, \ldots$ the E-step and the M-step. At iteration t, the E-step estimates [15], for $k = 1, \ldots, M, m = 1, \ldots, K$

$$\hat{P}^{(t)}(\Theta_k = \theta_m|D_k) = \frac{\hat{P}^{(t)}(D_k|\Theta_k = \theta_m)\hat{P}^{(t)}(\Theta_k = \theta_m)}{\sum_{l=1}^{K} \hat{P}^{(t)}(D_k|\Theta_k = \theta_l)\hat{P}^{(t)}(\Theta_k = \theta_l)} \tag{8}$$

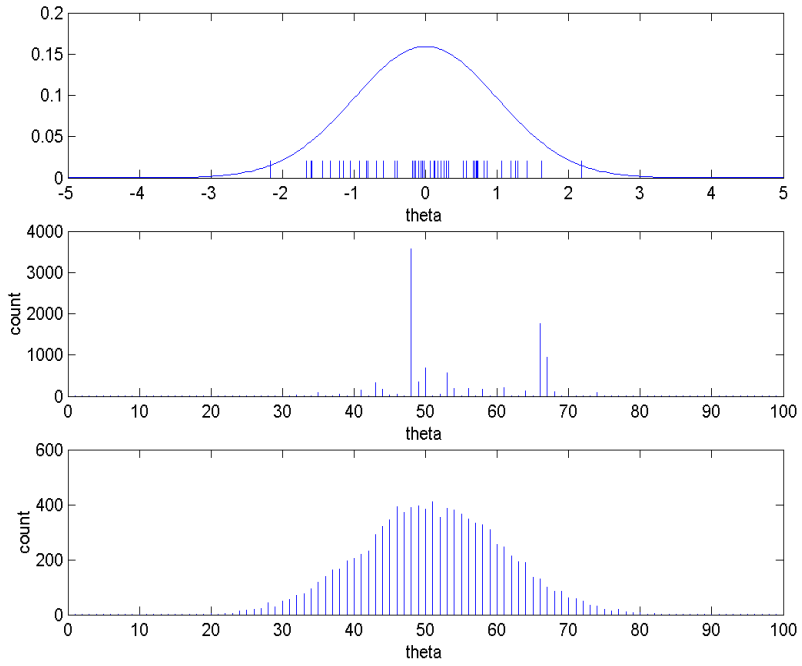

Fig. 3. Top: Subjective non-informative prior (Gaussian) and samples generated from this prior. These samples can be used for Dirichlet enhancement. Center: Samples from a distribution that was generated by a Dirichlet distribution with a Gaussian base distribution with precision $\tau = 10$. Clustering is quite apparent. Although the positions of the samples represent the base distribution, the counts are neither uniform nor follow the base distribution. Counts reflect the Pólya urn process (Section 4.2) or, equivalently, the stick breaking process (Section 5.1). Thus, that the ragged structure is *not* a result of a finite sample size —100000 samples were drawn— but is an inherent property of a distribution generated by a Dirichlet distribution, resp. Dirichlet process. Bottom: Same, but with $\tau = 10000$. With a large precision parameter, samples are drawn predominantly independently from the base distribution. Again, 100000 samples were drawn.

The M-step updates for $k = 1, \ldots, M, m = 1, \ldots, K$

$$\hat{P}^{(t+1)}(\Theta_k = \theta_m) = \hat{g}_m^{t+1}$$

with

$$\hat{g}_m^{t+1} = \frac{1}{\tau + M}\left(\tau\alpha_m + \sum_{j=1}^{M} \hat{P}^{(t)}(\Theta_j = \theta_m | D_j)\right).$$

After convergence, the prediction of an active model $a \in 1, \ldots, M$ becomes

$$P(Y_a = y | x, \{D_j\}_{j=1}^{M}) \approx$$

$$\frac{\sum_{k=1}^{K} \hat{P}(\Theta_a = \theta_k) P(D_a | \Theta_a = \theta_k) P(Y_a = y | x, \Theta_a = \theta_k)}{\sum_{k=1}^{K} \hat{P}(\Theta_a = \theta_k) P(D_a | \Theta_a = \theta_k)}. \tag{9}$$

Note that this solution is similar to the heuristically motivated solution of Equation 2 in the sense that predictions of the models are weighted by the probability with which those models explain the data set of the active model. The differences are that first, we have an additional weighting constant $\hat{P}(\Theta_a = \theta_k)$ that evaluates the overall relevance of a model and second, we assumed here that the parameters were generated rather unspecifically from the base distribution whereas in the heuristic solution they correspond to maximum likelihood estimates.

5 Hierarchical Bayesian Modelling with Infinite Models

Dirichlet enhancement has practical relevance only if we let $K \to \infty$, which is the case we consider in this section. The reason is that, with finite K and by simply sampling from the prior distribution as described in Section 4.1, it is unlikely that parameters leading to good models will be included.

The transition $K \to \infty$ leads us to nonparametric HB, where, as in the finite-dimensional case, the $\theta_k, k = 1, \ldots$, are sampled randomly from the base distribution. In this context we need to first introduce some properties of the Dirichlet process, which is a generalization of the Dirichlet distribution to infinite dimensions.

5.1 Dirichlet Process

The Dirichlet Process (DP) is of central importance in nonparametric Bayesian modelling. A formal definition can be found in the appendix. A DP is written as $\mathrm{DP}(G_0, \tau)$ where G_0 is the base distribution with probability density g_0 that corresponds to the α_j in the finite-dimensional case; $\tau \geq 0$ is the concentration parameter. Please, compare this definition to the definition of a Dirichlet distribution in Section 4.1. [6] As in the case of the Dirichlet distribution we can

[6] In the literature one often finds the notation α_0 for the concentration parameter.

use the Pólya urn representation to sample from a distribution generated by a Dirichlet process. Given previous samples $\{\theta_l\}_{l=1}^{j-1}$ generated from a distribution generated from a DP with base distribution g_0 and precision τ, the $j-th$ sample is generated from the probability density

$$P(\theta_j|\{\theta_l\}_{l=1}^{j-1}) = \frac{\tau g_0(\theta_j) + \sum_{k=1}^{j-1} \delta_{\theta_k}}{\tau + j - 1}. \tag{10}$$

Note that this formula is a direct generalization of the finite-dimensional case, Equation 6. Samples are generated with probability proportional to τ from the base distribution and with increasing probability proportional to $j-1$ from an already existing sample. Thus, for small τ we observe the same clustering effect as in the finite dimensional case (Figure 3). A mathematical treatment of nonparametric Bayesian modelling and the Dirichlet processes can be found in [14].

Equation 10 specifies how samples are generated from a distribution that is a sample from a DP. It is also possible to generate directly a sample from such a distribution by using the so-called stick breaking process (for a definition consult [26] or [27]) according to which this distribution can be written as an infinite sum of weighted delta functions placed at samples randomly selected from the base distribution,

$$g = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}. \tag{11}$$

The $\beta_k \geq 0$ and with $\sum_k \beta_k = 1$ only depend on τ and are generated by the stick breaking process, which is based on a sequence of independent beta random variables. Note that even if the base distribution G_0 is smooth, a sample distribution is discrete in nature.

5.2 Nonparametric Bayesian Modelling for Dirichlet Enhanced HB

In the context of an infinite model, i.e. a DP, the HB model of Equation 5 is called a Dirichlet process mixture model. The conditional probability distribution required for Gibbs sampling becomes [10] [32]

$$P(\theta_j|\{\theta_l\}_{l\neq j}, D_j) = \frac{1}{C}\left(\tau g_0(\theta_j) + \sum_{l:l\neq j} \delta_{\theta_l}\right) P(D_j|\theta_j)$$

$$= \frac{1}{C}\left(\tau \tilde{P}(D_j)\tilde{P}(\theta_j|D_j) + \sum_{l:l\neq j} \delta_{\theta_l} P(D_j|\theta_l)\right)$$

where

$$\tilde{P}(D_j) := \int P(D_j|\theta) g_0(\theta) d\theta, \quad \tilde{P}(\theta_j|D_j) := P(D_j|\theta_j) g_0(\theta_j)/\tilde{P}(D_j),$$

and where $\{\theta_l\}_{l \neq j}$ are the values of the remaining models. Note that this is a direct generalization of Equations 7. With probability proportional to $\tau \tilde{P}(D_j)$ a sample is generated from $\tilde{P}(\theta_j|D_j)$ and with probability proportional to $P(D_j|\theta_l)$ we take an existing sample θ_l. Note that our notation hides the fact that several θ_l might be identical, increasing the selection probability accordingly. This parameter clustering is particularly strong if τ is small in which case the number of distinct parameters is typically much smaller than M. Note also that, despite the fact that we are considering infinite models, computational load per round and memory requirements only grow proportional to M. This semi-automated determination of the number of distinct models is an important feature and was the focus of some recent work (see Section 7) but is not of central interest in the HB framework presented here.

The presented Gibbs sampling approach was introduced by Escobar [9]. Since in Gibbs sampling only one parameter is re-sampled at a time, the clustering of the parameters makes it difficult for the sampling procedure to modify parameter values. In the appendix we describe a mixture of models approach introduces by MacEachern [19] that turns out to be equivalent to the presented model. Gibbs sampling based on that model exhibits much better mixing properties. The blocked Gibbs sampler that is based on a finite stick-breaking prior provides another computational attractive sampling procedure [18]. A comprehensive overview of sampling techniques for Dirichlet process mixture models can be found in [21].

5.3 Variational EM

In a nonparametric setting our EM equations from Subsection 4.4 cannot directly be applied since a distribution generated by a Dirichlet process is infinite-dimensional. In [29] the authors discuss a one-step EM solution. Here, we discuss an EM solution that can be derived from a variational approximation that approximates probability densities of the E-step in Equation 8 by a simpler approximating density [30]. We propose a sum of weighted delta functions defined at the maximum likelihood estimates of the models, i.e.,

$$\hat{P}(\theta_j|D_j) \approx q_j(\theta_j) = \sum_{k=1}^{M} \xi_{j,k} \delta_{\theta_k^{ML}} \quad j = 1, \ldots, M \tag{12}$$

where $\xi_{j,k}$ are the variational parameters with $\xi_{j,k} \geq 0$ and $\sum_{k=1}^{M} \xi_{j,k} = 1$. In each variational E-step, the variational parameters are adapted such that KL-divergence between the variational approximation and $\hat{P}^{(t)}(\theta_j|D_j)$ is minimized.

As a generalization to the finite-dimensional case we propose as update equations for $t = 1, 2, \ldots$:

$$\xi_{j,k}^t = \frac{P(D_j|\theta_k^{ML})\hat{P}^{(t)}(\theta_k^{ML})}{\sum_{k=1}^{M} P(D_j|\theta_k^{ML})\hat{P}^{(t)}(\theta_k^{ML})} \quad j = 1, \ldots, M \quad k = 1, \ldots, M. \tag{13}$$

The M-step updates

$$\hat{P}^{(t+1)}(\theta_k^{ML}) = \frac{1}{\tau + M}\left(\tau g_0(\theta_k^{ML}) + \sum_{l=1}^{M} \xi_l \delta_{\theta_l^{ML}}\right) \quad k = 1, \ldots, M$$

with $\xi_l = \sum_{j=1}^{M} \xi_{j,l}$.

Note that the EM iterations are quite simple since many terms, such as $P(D_j|\theta_k^{ML})$, don't change in the iterations. Also note the similarity to the finite-dimensional case in Section 4.4.

Now, the prediction of an active model $a \in 1, \ldots, M$ becomes

$$P(Y_a = y|x, \{D_j\}_{j=1}^{M}) \approx$$

$$\frac{\tau \tilde{P}(D_a)\tilde{P}(Y_a = y|x, D_a) + \sum_{k=1}^{M} \xi_k P(D_a|\theta_k^{ML}) P(Y_a = y|x, \theta_k^{ML})}{\tau \tilde{P}(D_a) + \sum_{k=1}^{M} \xi_k P(D_a|\theta_k^{ML})} \tag{14}$$

where we use

$$\tilde{P}(D_a) := \int g_0(\theta) P(D_a|\theta)\, d\theta$$

$$\tilde{P}(Y_a = y|x, D_a) := \frac{1}{\tilde{P}(D_a)} \int g_0(\theta) P(D_a|\theta) P(Y_a = y|x, \theta)\, d\theta.$$

Note the great similarity of this prediction equation to the prediction equation for the finite dimensional case (Equation 9) and the heuristically defined solution of Equation 2: the second term in the numerator in Equation 14 contains the model predictions using maximum likelihood parameter estimates, weighted by the probability that models agrees with the data set of the active model $P(D_a|\theta_k^{ML})$. Here, additional relevance weights ξ_k are included, which represent the overall relevance of the models. If we look at Equation 13, it becomes clear that the contribution of the j-th setting to the relevance weight ξ_k is essentially determined by the term $P(D_j|\theta_k^{ML})$ which means that a setting j which has received a small number of data points contributes to all ξ_k, whereas a setting j which receives a large number of data points will mostly contribute to ξ_j. In our experiments we found that the weight of a model prediction in Equation 14 is mostly determined by the term $P(D_a|\theta_k^{ML})$ and that the ξ_k are more or less of the same magnitude and thus have only a minor influence. Thus in many applications one might refrain from the fitting of the variational parameters $\xi_{j,k}$ and use Equation 14 with $\xi_k = 1, k = 1, \ldots, M$.

The first term in the numerator of Equation 14 puts additional weight on the prediction of the active model. In particular, it consists of the Bayesian prediction of the active model a based on its own data $\tilde{P}(Y_a = y|x, D_a)$ weighted by τ and the evidence of the data of the active model $\tilde{P}(D_a)$. The latter term evaluates the correctness of the prior modelling assumption.

Equation 14 is equivalent to the Bayesian prediction of the active model if we use a prior proportional to

$$\tau g_0(\theta) + \sum_{k=1}^{M} \xi_k \delta_{\theta_k^{ML}}$$

which illustrates the similarity of this solution to the heuristically defined solution of of Equation 2, in particular with $\tau \to 0$. With $\tau \to \infty$ the prediction of the active model is simply the prediction of the active model trained on its own data, i.e., all models are independent and only rely on their own data. With a finite τ we obtain the HB solution.

If compared to the Gibbs sampling approach, the variational EM solution has several important advantages. Each model can be trained independently from the other models just based on its own data set. Thus the solution is easy to train, modular, and efficient and an additional model can easily be incorporated. From a theoretical perspective, Gibbs sampling might be more appealing but one should note its typically slow convergence and the slow mixing of the sampling process in practice.

An advantage of the sampling approach is that it leads to an automated clustering of the models, a feature that is not achieved in the variational EM solution. On the other hand, if such a model clustering is of prime importance, one can achieve it, for example, by a corresponding postprocessing step.

The variational approximation of Equation 12 uses maximum likelihood parameter estimates. This has the advantage that asymptotically, with an increasingly large number of data points for the active model, the overall prediction converges to the prediction of the active model. The same feature can be achieved if the variational approximation is based on the maximum a posteriori (MAP) parameter estimates of the models. The MAP estimate is more appropriate if only few training data points are available. Alternatively, one could select for the variational approximation sets of samples obtained from the posterior parameter distributions for each model.

6 A Recommendation Engine

In this section we provide a summary of the application of nonparametric hierarchical Bayesian modelling to information filtering. A more detailed description can be found in [30].

Information filtering denotes a family of techniques that try to understand people's information needs, and then help them find the right information items while filtering out undesired ones. In a very wide range of applications, such as spam email filtering, news filtering, recommender systems for products (e.g., books, movies, CDs), and web navigation, information filtering is playing an increasingly important role. Content-based filtering (CBF) and collaborative filtering (CF) represent the two major information filtering technologies.

CBF has its root in the concept of *relevance feedback* in the information retrieval literature (e.g., Rocchio's algorithm [25]). It explores the similarity of contents between information items (e.g., articles, paintings, music), to infer which of the yet unseen items might be of interest for the active user, based on some annotated examples previously given by the user. In contrast, collaborative filtering methods typically accumulate a database of item ratings—explicitly or implicitly—cast by a large set of users. The prediction of ratings for the

active user is solely based on the ratings provided by all other users, under the assumption that like-minded users share similar information needs. The method does not rely on a description of the item's content.

One major difficulty in designing CBF systems lies in extracting content features that are sufficiently indicative. There is often a large gap between low-level content features (visual, auditory, or others) and high-level user interests (like or dislike a painting or a CD). In some other circumstances, the features are not available at all.

On the other hand, pure CF only relies on user preferences, without incorporating the actual content of items. CF often suffers from the extreme sparsity of the available data set, in the sense that users typically rate only very few items, thus making it difficult to compare the interests of two users. Furthermore, pure CF can not handle items for which no user has previously given a rating. Such cases are easily handled in CBF systems, which can make predictions based on the content of the new item.

We combine CF and CBF under the framework of nonparametric hierarchical Bayesian modelling which leads to a model that combines the advantages of both approaches. Essentially a CBF model is formed for every user and the predictions are combined using the nonparametric HB approach using variational EM as described in Section 5.3.

In our application, we focus on a survey of 642 paintings of 30 artists. A web-based online survey is built to gather user ratings. In the survey, each user gave ratings, i.e., "like", "dislike", or "not sure", to a randomly selected set of paintings. Finally we got a total of $N = 190$ users' ratings. On average, each of them had rated 89 paintings.

For each painting, we calculate the *color histogram* (216-dim.), the *correlagram* (256-dim.), the *first and second color moments* (9-dim.) and the *pyramid wavelet texture* (10-dim.) to form a 491-dimensional feature vector.

We will examine the performance of various algorithms in terms of their accuracy in predicting users' interests in paintings. We used as our base user models a probabilistic version of the support vector machine (SVM) [22] with Gaussian kernels. *Hybrid filtering 1* implements the nonparametric HB approach using variational EM as described in Section 5.3; *Hybrid filtering 2* is identical, except that we set $\tau = 0$; for *SVM Content-Based filtering* (CBF) we use $\tau \to \infty$ and obtain independent user models; *Collaborative filtering* (CF) combines a society of advisory users' preferences to predict an active user's preferences. The combination is weighted by the *Pearson correlation* between the active user and the other advisory users' preferences. The algorithm applied here is described in [7];

These algorithms are evaluated using two metrics. One is *Top-L accuracy*, i.e., the proportion of truly liked paintings among L top ranked paintings. Since normal users only care about the quality of the first set of returned items, this quantity reflects the *subjective* quality of an information filter system. Secondly, we evaluated the ROC (receiver operating characteristics) curve, which plots *sensitivity* versus 1-*specificity*. Sensitivity is defined as the probability that a

good painting is recommended by the system; and specificity is the probability that a disliked painting is rejected by the system. By changing the cut point (e.g., return top 10 or 20 paintings), a curve can be plotted. ROC curve is insensitive to the prior distribution of liked (or disliked) paintings. The area under the curve, called *ROC sensitivity*, measures the *objective* quality of the ranking. A higher ROC sensitivity indicates a better ranking.

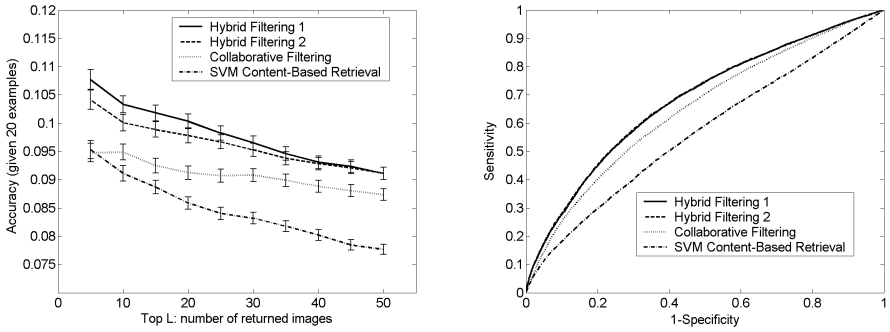

Fig. 4. Left: Top-L accuracy. Right: ROC curves. From [30].

In the application it was not required that a user rates all of the 642 paintings in the survey; thus for each user we only partially know the "ground truth" of preferences. As a result, the true top-L accuracy cannot be computed. We thus adopt as accuracy measure the fraction of *known* liked paintings in the top ranked L paintings. The quantity is smaller than true accuracy because *unknown* liked paintings are missing in the measurement. However, in our survey, the presentation of paintings to users is completely random, thus the distributions of rated/unrated paintings in both unranked and ranked lists are also random. This randomness dose not change the relative values of the compared methods. Thus in the evaluation of the experiment it still makes sense to use the adopted accuracy measurement to compare the three retrieval methods. The ROC curves are insensitive to this problem.

In our experiments, we used a 10-fold cross validation scheme, in which we pick up each fold as a set of active users and treat the rest as users in the data base. We fix the number of given examples for each active user to be 20 (10 positive and 10 negative), and predict the user's interests in the remaining paintings. For each active user, recommendations for 10 different paintings are calculated. Finally, the overall average performances and error bars are computed. 4 shows the results. Both Top-L accuracy and ROC curve clearly indicate that the two hybrid algorithms outperform CF and CBF. We found that the extracted painting features are poor indicators of human interests, which is the reason for the bad performance of CBF. The ROC curves of the two hybrid filtering algorithms

are essentially overlapping. However, Top-L accuracy suggests that hybrid filtering 1 is slightly better.

7 Related Work in Machine Learning

Dirichlet process mixture models were introduced into machine learning by Neal [20] [21] who used them to realize infinite mixture models. Dirichlet processes were applied to realize infinite mixtures of Gaussians [23], infinite mixtures of Gaussian experts [24] and infinite hidden Markov models [3]. These models are also based on nonparametric HB modelling but the application focus is different: In these papers, there are no repeated measurements for a given model (i.e., in the plate model of Figure 2D, $N = 1$) and the focus is on model-based soft clustering using an infinite mixture approach and on the realization of an infinite mixture of experts solution. An inherent advantage of Dirichlet process mixture modelling is that the number of clusters does not need to be specified in advance but is automatically determined via the sampling process. A small precision parameter τ leads to few clusters whereas a large precision parameter leads to many clusters. Thus in those applications a sensible tuning of τ is required. In those papers the sampling procedure described in the appendix is used. A hierarchical Dirichlet process model was recently introduced to model hierarchical unsupervised structures [27]. Mathematically demanding variational mean-field approximations were applied to Dirichlet processes in [6] and [31]. Some of the work on the development of self-organizing maps for the clustering of probabilistic models can also be related to nonparametric HB modelling [17].

Examples of the application of HB to machine learning are probabilistic clustering [8], the finite-dimensional HB approach by [4] [5] who used HB in the context of a model for latent semantic analysis and information retrieval and the application of neural networks models to HB [16] [2].

8 Conclusions

Nonparametric hierarchical Bayesian modelling is a powerful and flexible approach for multi-agent learning if agents need to share learned knowledge. We introduced the basic background and the common inference approach via Gibbs sampling. We described a variational EM solution that leads to excellent results in a multi-agent learning framework. The main advantages of the EM solution are its modularity, low computational complexity, intuitive plausibility and good performance. Many variants of nonparametric hierarchical Bayesian modelling have been used in the literature with various combinations of model specific parameters, shared parameters and Dirichlet enhanced distributions and with varying levels of hierarchies (see, for example, [28] and [27]). Thus nonparametric hierarchical Bayesian modelling is quite flexible and might find an increasing number of applications in multi-agent learning.

9 Appendix

9.1 Definition of a Dirichlet Process

The theorem asserts the existence of a Dirichlet process and also serves as a definition [14]. Let $(\mathbb{R}, \mathcal{B})$ be the real line with the Borel σ-algebra $\mathcal{B}$ and let $M(\mathbb{R})$ bet the set of probability measures on $\mathbb{R}$, equipped with the σ-algebra $\mathcal{B}_M$.

Theorem 1 *Let α be a finite measure on $(\mathbb{R}, \mathcal{B})$. Then there exists a unique probability measure D_α on $M(\mathbb{R})$ called the Dirichlet process with parameters α satisfying:*

For every partition $B_1, B_2, ..., B_k$ of $\mathbb{R}$ by Borel sets $(P(B_1), P(B_2), \ldots, P(B_k))$ is $Dir(\alpha(B_1), \alpha((B_2), \ldots, \alpha((B_k))$

9.2 Equivalence of Dirichlet Enhanced HB to Mixture Models

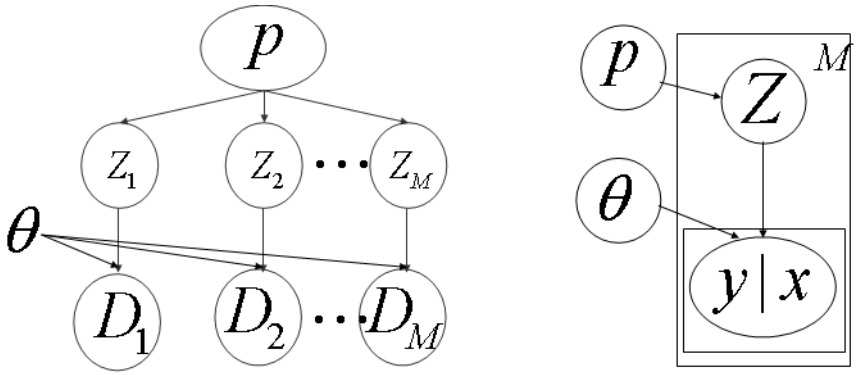

Fig. 5. Left: The mixture model. Right: The plate model. Note that in contrast to the HB model, all parameters are global parameters.

Finite Mixture Models: In Section 4 we had concluded that the prior distribution in HB must be made flexible and we introduced the process of Dirichlet enhancement. Thus we obtained a highly flexible prior distribution that permitted the sharing of knowledge between models. An alternative solution is the mixture of models approach presented here (see Figure 5).

The predictive model of the mixture model is

$$P(Y = y|x) = \sum_{k=1}^{K} P(Z = k)P(Y = y|x, k)$$

where Z is a latent variable with states $1, \ldots, K$. It is now uncertain which model generated the data for the active setting such that

$$P(D_a, Y_a = y|x) = \sum_{k=1}^{K} P(Z = k)P(D_a|k)P(Y_a = y|x, k).$$

To classify a new pattern, we thus obtain

$$P(Y_a = y|x, D_a) = \frac{\sum_{k=1}^{K} P(Z = k)P(D_a|k)P(Y_a = y|x, k)}{\sum_{k=1}^{K} P(Z = k)P(D_a|k)}.$$

Please, note the similarity of this equation to Equation 9 that deals with the finite-dimensional Dirichlet enhanced case.

It now turns out that there is an exact equivalence with the finite-dimensional Dirichlet enhanced model in Section 4 if:

- the likelihood models for HB and the mixture approach are identical

$$P(Y_a = y|x, k) = P(Y_a = y|x, \theta_k),$$

- the same parameter vectors $\{\theta_k\}_{k=1}^{K}$ are selected,
- the prior for Z is a multinomial,

$$P(Z = k) = p_k$$

- which is generated from a Dirichlet distribution with

$$p_1, \ldots, p_K \sim \text{Dir}(\tau\alpha_1, \ldots, \tau\alpha_K).$$

Details can be found in [21] and the graphical model and plate model are shown in Figure 5.

Infinite Mixture Models: It turns out that the equivalence also holds if we let $K \to \infty$ in which case we obtain an infinite mixture model, which is equivalent to a Dirichlet process mixture model (Section 5), if we chose as prior parameter distribution

$$P(\theta_k) = g_0(\theta_k) \quad \forall k,$$

and with

$$p_1, \ldots, p_K \sim \text{Dir}(\tau/K, \ldots, \tau/K).$$

Stochastic sampling based on this model can be implemented as follows [21]: One first updates the latent variables $\{Z_j\}_{j=1}^{M}$. Let consider the update of Z_j, which denotes the latent variable which is associated with the j-th model (Figure 5). As in nonparametric HB, a new sample depends on the states of the latent variables of the remaining variables which might also be clustered. Let N_k be the number of variables in the set $\{Z_l\}_{l=1}^{M}$, which are in state k, *without counting the state of* Z_j, i.e., $\sum_k N_k = M - 1$.

Then for all states with $N_k > 0$

$$P(Z_j = k | \{Z_l\}_{l \neq j}, D_j, \theta) \propto N_k P(D_j | \theta_k).$$

A new state is generated with probability

$$P(Z_j \neq k \text{ for all } k \neq j | \{Z_l\}_{l \neq j}, D_i, \theta) \propto \frac{1}{C} \tau \tilde{P}(D_j)$$

with $\tilde{P}(D_j) := \int g_0(\theta) P(D_j|\theta)\, d\theta$. In the first case, the j-th model inherits the parameters of the models assigned to state k and in the latter case, a new θ is drawn from $P(\theta|D_j)$.

Typically after one update of all latent variables, the model parameters are all updated. E.g., for all models in state k, a new θ_k is drawn from

$$\frac{1}{C} g_0(\theta_k) \prod_{\{j: Z_j = k\}} P(D_j|\theta_k).$$

The advantage of this sampling scheme is that at each round all parameters are re-sampled and typically assume new values whereas in the sampling schemes described in Section 5.2 it is rather unlikely that clustered parameters will assume new values since only one parameter is re-estimated at a time.

Neal [21] discusses additional advanced sampling techniques.

References

1. Antoniak, C. E.: Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. Annals of Statistics **2** (1974) 1152-1174
2. Bakker, B., Heskes, T.: Task Clustering and Gating for Bayesian Multitask Learning. Journal of Machine Learning Research, **4** (2003)
3. Beal, M. J., Ghahramani, Z., Rasmussen, C. E.: The Infinite Hidden Markov Model. Advances in Neural Information Processing Systems **14** (2002)
4. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3** (2003)
5. Blei, D. M., Jordan, M. I., Ng, A. Y. : Hierarchical Bayesian Modelling for Applications in Information Retrieval. Bayesian Statistics **7**. Oxford University Press (2003)
6. Blei, D. M., Jordan, M. I.: Variational methods for the Dirichlet process. To appear in Proceedings of the 21st International Conference on Machine Learning (2004)
7. Breese, J. S, Heckerman, D., Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (1998)
8. Cadez, I., Smyth, P.: Probabilistic Clustering using Hierarchical Models. TR No 99-16, Dept. of Information and Computer Science. University of California, Irvine (1999)
9. Escobar, M. D.: Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means. Unpublished PhD dissertation, Yale University (1988)

10. Escobar, M. D., West, M.: Computing Bayesian Nonparametric Hierarchical Models. Practical Nonparametric and Semiparametric Bayesian Statistics, D. Dey, P. Müller, D. Sinha (eds.), Springer (1998)
11. Ferguson, T. S.: A Bayesian Analysis of some Nonparametric Problems. Annals of Statistics **1** (1973) 209-230
12. Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B.: Bayesian Data Analysis. CRC press (2003)
13. Gilks, W. R., Richardson, S., Spiegelhalter, D. J.: Markov Chain Monte Carlo in Practice. CRC press (1995)
14. Gosh, J. K, Ramamoorthi, R. V.: Bayesian Nonparametrics. Springer Series in Statistics (2002)
15. Heckerman, D.: A Tutorial on Learning with Bayesian Networks. Technical report MSR-TR-95-06 of Microsoft Research (1995)
16. Heskes, T.: Empirical Bayes for Learning to Learn. In Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2000) 367–374
17. Holmen, J., Tresp, V., Simula, O.: A Self-Organizing Map for Clustering Probabilistic Models. Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN'99) **2** (1999)
18. Ishwaran, H., James, L. F. : Gibbs Sampling Methods for Stick-Breaking Priors. Journal of the American Statistical Association, Vol. 96, No. 453 (2001)
19. MacEachern, S. M.: Estimating Normal Means with Conjugate Style Dirichlet Process Prior. Technical report No. 487, Department of Statistics, The Ohio State University (1992)
20. Neal, R, M.: Bayesian Mixture Modeling by Monte Carlo Simulation. Technical Teport No. DCR-TR-91-2, Department of Computer Science, University of Toronto (1991)
21. Neal, R, M.: Markov Chain Sampling Methdos for Dirichlet Process Mixture Models. Technical report No. 9815, Department of Statistics, University of Toronto (1998)
22. Platt, J. C.: Probabilities for SV machines. In Advances in Large Margin Classifiers. MIT Press (1999)
23. Rasmussen, C. E.: The Infinite Gaussian Mixture Model. Advances in Neural Information Processing Systems **12** (2000)
24. Rasmussen, C. E., Ghahramani, Z.: Infinite Mixtures of Gaussian Process Experts. Advances in Neural Information Processing Systems **14** (2002)
25. Rocchio, J. J.: Relevance Feedback in Information Retrieval. The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice Hall (1971)
26. Sethuraman, J.: A Constructive definition of Dirichlet Priors. Statistica Sinica **4** (1994)
27. Teh, Y. W. , Jordan, M. I., Beal, M. J., Blei, D. M.: Hierarchical Dirichlet Proceses. Technical Report 653, UC Berkeley Statistics (2004)
28. Tomlinson, G., Escobar, M.: Analysis of Densities. Talk given at the Joint Statistical Meeting (2003)
29. Yu, K., Schwaighofer, A., Tresp, V., Ma, W.-Y., Zhang, H.: Collaborative Ensemble Learning: Combining Collaborative and Content-Based Information Filtering via Hierarchical Bayes. Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI) **19** (2003)
30. Yu, K., Tresp, V., Yu S.: A Nonparametric Bayesian Framework for Information Filtering. To appear in the proceedings of the 27th Annual International ACM SIGIR Conference (2004)

31. Yu, K., Yu S., Tresp, V. : Dirichlet Enhanced Latent Semantic Analysis. Technical Report (2004)
32. West, M., Müller, P., Escobar, M. D.: Hierarchical Priors and Mixture Models, with Application in Regression and Density Estimation. Aspects of Uncertainty: A Tribute to D. V. Lindley, A.F.M. Smith and P. Freeman, (eds.), Wiley New York (1994) 363–386