# HIERARCHICAL BAYESIAN LEARNING FOR ELECTRICAL TRANSIENT CLASSIFICATION

*Matthieu Sanquer* [⋆†] *Florent Chatelain*[⋆] *Mabrouka El-Guedri*[†] *Nadine Martin*[⋆]

[⋆] University of Grenoble, GIPSA-lab, 961 rue de la Houille Blanche,
BP 46, 38402, St Martin D'Hères, France
[†] Electricité de France – R&D, 4 Quai Watier, 78401 Chatou, France

**ABSTRACT**

This paper addresses the problem of the supervised signal classification, by using a hierarchical Bayesian method. Each signal is characterized by a set of parameters, the features, which are estimated from a set of learning signals. Moreover, these parameters are distributed according to a class-specific posterior distribution which allows one to capture the variability of the features within the same class. Within the hierarchical Bayesian framework, the feature extraction step and the learning step can be performed jointly. Unfortunately, the estimation of the class-specific distribution parameters requires the computation of intractable multi-dimensional integrals. Then a Markov-chain Monte Carlo (MCMC) algorithm is used to sample the posterior distributions of the features over all the training signals of each class. An application to electrical transient classification for non-intrusive load monitoring is introduced. Simulations over real-world electrical transients signals are driven and show the capacity of the proposed methodology to discriminate two classes of transients.

***Index Terms***— Hierarchical Bayesian model, MCMC methods, supervised classification, curve fitting, smooth transition regression model, non intrusive appliance load monitoring.

## 1. INTRODUCTION

Bayesian inference is an usual method in a learning framework [1]. This allows classification algorithms to be derived, based on the posterior distributions of some features that characterize a class. These posterior distributions are inferred on some learning signals. Finally, the classification task is usually performed for a given signal by choosing the class that maximizes the posterior probability of the features that have been extracted.

In standard classification methods, feature extraction and learning are performed separately. This approach is adequate when the feature extraction step is straightforward. However, in some cases, the feature extraction requires the use of computational estimation methods as, for instance, Markov chain Monte Carlo (MCMC) sampling methods in a hierarchical Bayesian approach. Under such circumstances, it is interesting to perform jointly the feature extraction and the learning tasks [2]. Furthermore, such a strategy offers the possibility to include some prior knowledge at different levels of abstraction in the hierarchical model. Then, feature variability within a class is taken into account in order to support classification.

The detection and the classification of electrical transients are useful in the context of non intrusive load monitoring (NILM). Machine learning methods have received a great attention to tackle NILM [3, 4]. The so called *microscopic* methods focus on the analysis of the waveform of the electrical transients. In fact, some seminal works- [5] have shown that, when an appliance is turned on, it generates an electrical transient which is characteristic to the kind of the appliance. Consequently, some methods have already been investigated to detect and classify these appliances by fitting the transients that appear in the load curves according to some deterministic pattern [6]. However, these deterministic rules lead to some specific and manual feature extraction methods which are difficult to be generalized to real-world applications.

The main contribution of the present work is to study a general fully probabilistic approach to model and to classify the different electrical appliance transients. Thus, it benefits from the many advantages such as flexibility, confidence values, robustness... offered by probabilistic pattern recognition methods. The considered work extends the smooth transition regression modeling proposed in [7, 8] to the supervised classification problem. In [7, 8], a hierarchical Bayesian model was introduced to fit an unique transient with a view of achieving a sparse representation. In a machine learning framework, it seems now quite natural to account for some *overhypotheses* [9] on the feature variability. In our hierarchical Bayesian framework, these overhypotheses reduces to some hyperpriors common to all the signals of the same class. As a consequence, the feature extraction specific for each signal of a given class, and, the learning over all different electrical transient classes are performed jointly.

This paper is outlined as follows. The hierarchical Bayesian learning method based on the smooth transition regression parameters is given in the second section. The MCMC algorithm derived to infer the signal features and to learn the class-specific parameters is presented in section III. Some simulations conducted on real-world electrical transients are reported in section IV. Finally, some concluding remarks are exposed in the last section.

## 2. HIERARCHICAL BAYESIAN LEARNING

In a supervised classification context, the set of the training samples is denoted as $\mathcal{X}$. In our application, each sample $\mathbf{x} \in \mathcal{X}$ stands for the time series associated with a training signal. The class-label of $c(\mathbf{x}) \in \mathcal{C} = [1, \dots, N_C]$ is known for each example $\mathbf{x}$, and $N_C$ denotes the number of different classes. For all $c \in \mathcal{C}$, the subset $\mathcal{X}_c = \{\mathbf{x} \in \mathcal{X} \text{ such that } c(\mathbf{x}) = c\}$ contains all the samples belonging to class labeled as $c$, whereas the number of samples in this class is denoted as $N_c = \mathrm{card}(\mathcal{X}_c)$. Finally, $\mathbf{x}_{i,c}$ stands for the $i^{th}$ sample in the class $c$.

The classification problem formulated in a Bayesian framework reduces to the computation of the posterior distributions $f(c|\mathbf{x}, \mathcal{X}_c)$ for each $c \in \mathcal{C}$ and for any given signal that does not belong to the training set : $\mathbf{x} \notin \mathcal{X}_c$. Assuming a zero-one loss function, the signal $\mathbf{x}$ is classified according to the following maximum a posteriori decision rule:

$$c(\mathbf{x}) = \underset{c \in \mathcal{C}}{\mathrm{argmax}} f(c|\mathbf{x}, \mathcal{X}_c). \tag{1}$$

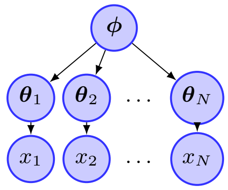



**Fig. 1**: Directed acyclic graph of the parameters and hyperparameters of the hierarchical Bayesian model - $\phi$ are the hyperparameters - $\theta_i$ are the features/parameters of the signal $x_i$

In the context of a model-based classification, a set of parameters (or features) $\boldsymbol{\theta}$ is extracted from an input signal $\mathbf{x}$, using a parametric model. This model is defined by its likelihood function $f(\mathbf{x}|\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, where $\Theta$ denotes the feature space.

### 2.1. Hierarchical Bayesian model

We will assume now that the signals belong to a given class $c$. For the sake of simplicity, the $c$ subscript is omitted in the following. The training signal $\mathbf{x}_i$ of the class $c$ is modeled according to the following hierarchical model:

$$\mathbf{x}_i \sim f(\mathbf{x}|\boldsymbol{\theta}_i), \tag{2}$$
$$\boldsymbol{\theta}_i \sim f(\boldsymbol{\theta}|\boldsymbol{\phi}), \tag{3}$$
$$\boldsymbol{\phi} \sim f(\boldsymbol{\phi}|c), \tag{4}$$

where $\boldsymbol{\phi}$ is set of hyperparameters which define the distribution of the parameters $\boldsymbol{\theta}$ within a class. These expressions show that, first, each training signal $\mathbf{x}_i$ is governed by a specific vector of features $\boldsymbol{\theta}_i$. This favors the ability of the model to account for the variability between the features of the same class. Secondly, all the feature vectors $\boldsymbol{\theta}_i$ associated with all the training signals are governed by the same class-specific distribution. The hyperparameters $\boldsymbol{\phi}$ of this distribution depends only on the considered class $c$. This ensures the ability of the model to categorize the different features in the same class. Last, the distribution $f(\boldsymbol{\phi}|c)$ corresponds to the prior on these class-specific parameters. These hypotheses introduced in the hierarchical Bayesian model are set up at different levels of abstraction. This is in agreement with the notion of overhypothesis, as introduced, for instance, in [9]. These relations are depicted on the directed acyclic graph of the parameters and hyperparameters in Figure 1. Each stage of the underlying hierarchical Bayesian method is related to one of the standard steps of supervised learning:

1. Estimate $\boldsymbol{\theta}_i$ from $\mathbf{x}_i$ is the feature extraction stage,
2. Estimate $\boldsymbol{\phi}$ from $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$ is the learning stage.

It is of note that these two steps are performed jointly within the hierarchical Bayesian framework. This is an appealing characteristic, these two steps being performed successively in a standard classification framework. Indeed, the estimation of the features of each training signal is improved if the feature distribution of the signals of the same class is taken into account.

In order to derive the classification rule, one needs to learn the parameters $\boldsymbol{\phi}$ from the training examples $\mathcal{X}$. The following expression of the joint posterior of the class-specific parameters $\boldsymbol{\phi}$ and the signal-specific parameters $\boldsymbol{\theta}_i$ is derived, assuming that the training examples are independent conditionally to $\boldsymbol{\phi}$

$$f(\boldsymbol{\phi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N|\mathcal{X}) \propto f(\boldsymbol{\phi}) \prod_{i=1}^{N} f(\mathbf{x}_i|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i|\boldsymbol{\phi}). \tag{5}$$

Finally, the following expression of $f(\boldsymbol{\phi}|\mathcal{X})$ is deduced from (5)

$$f(\boldsymbol{\phi}|\mathcal{X}) \propto \quad f(\boldsymbol{\phi}) \textstyle\prod_{i=1}^{N} \int f(\mathbf{x}_i|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i|\boldsymbol{\phi}) d\boldsymbol{\theta}_i. \tag{6}$$

Note that (5) is the most useful equation in our hierarchical Bayesian method as the analytical integration over the parameter space in (6) is not tractable in the general case.

### 2.2. Smooth transition regression model for transient modeling

For the processing of the electrical transients of our database, a specific model has been introduced in [7, 8]. We briefly recall in this section the basics of this model. The signal is modeled as a sequence of $K$ constant consumption level connected by a sequence of $K-1$ transition functions centered around the time instants $\boldsymbol{\tau}_i = [\tau_{1,i}, \ldots, \tau_{K-1,i}]^T$. Then, for all $j = 1, \ldots, n_i$

$$\mathbf{x}_i[j] = \sum_{k=1}^{K} [\pi_{k-1,i}(t_j) - \pi_{k,i}(t_j)]\beta_{k,i} + \epsilon_i[j], \tag{7}$$

where $n_i$ is the length of the time series, $\epsilon_i$ is an i.i.d centered Gaussian noise vector with a variance $\sigma_i^2$, $\boldsymbol{\beta}_i = [\beta_{1,i}, \ldots, \beta_{K,i}]^T$ is a the vector of the active power consumption levels, and $\pi_{0,i}, \ldots, \pi_{K,i}$ is the set of the transition functions. These transition functions are chosen among the family of stretched exponential functions

$$\pi_{k,i}(t) = \begin{cases} 1 - \exp\left[\left(\frac{t-\tau_{k,i}}{\lambda_{k,i}}\right)^{\exp(\alpha_{k,i})}\right] & t > \tau_{k,i} \\ 0 & t < \tau_{k,i} \end{cases}. \tag{8}$$

Each transition function $\pi_{k,i}$ is parameterized according to a location parameter $\tau_{k,i}$, a scale parameter $\lambda_{k,i}$ and a shape parameter $\alpha_{k,i}$. The number of components $K$ is supposed to be known for each class of signal.

The linear relation between $\mathbf{x}_i$ and $\boldsymbol{\beta}_i$ is captured by the matrix $Z_i$, the entries of which depend on the transition function parameters

$$\mathbf{x}_i = Z_i \boldsymbol{\beta}_i + \epsilon_i. \tag{9}$$

The likelihood function for this model reads

$$f(\mathbf{x}_i|\theta_i) \propto \frac{1}{(\sigma_i^2)^{\frac{n_i}{2}}} \exp\left(\frac{1}{2\sigma_i^2}(\mathbf{x}_i - Z_i\boldsymbol{\beta}_i)^T(\mathbf{x}_i - Z_i\boldsymbol{\beta}_i)\right), \tag{10}$$

where the parameter vector is $\boldsymbol{\theta}_i = \{\sigma_i^2, \boldsymbol{\beta}_i, \boldsymbol{\eta}_i\}$, with $\boldsymbol{\eta}_i$ being the set of the transition function parameters $\boldsymbol{\eta}_i = (\tau_{k,i}, \lambda_{k,i}, \alpha_{k,i})_{k=1,\ldots,K-1}$

### 2.3. Prior distribution

In this work, we assume the classical hypothesis which leads to the naive Bayes classifier [1] (i.e conditional independence of the features given the class), which yields the following expression of the class-specific distribution over the parameter space

$$f(\boldsymbol{\theta}_i|\boldsymbol{\phi}) = f(\sigma_i^2|\rho_\sigma) f(\boldsymbol{\beta}_i) \prod_{k=1}^{K-1} f(\lambda_{k,i}|\rho_{\lambda k}) f(\alpha_{k,i}|\mu_{\alpha k}, \sigma^2_{\alpha k}),$$

where $\boldsymbol{\phi} = \{\rho_\sigma, \boldsymbol{\rho_\lambda}, \boldsymbol{\mu_\alpha}, \boldsymbol{\sigma^2_\alpha}\}$ is the hyperparameter vector, whereas $\boldsymbol{\rho_\lambda} = (\rho_{\lambda k})_{k=1,\ldots,K-1}$, $\boldsymbol{\mu_\alpha} = (\mu_{\alpha k})_{k=1,\ldots,K-1}$, and $\boldsymbol{\sigma^2_\alpha} = (\sigma^2_{\alpha k})_{k=1,\ldots,K-1}$ are the parameters of the hyperpriors.

Conjugate inverse-gamma prior and g-prior are chosen for the variance of the observation noise $\sigma_i^2$ and the coefficients $\boldsymbol{\beta}_i$ respectively

$$\sigma_i^2|\rho_\sigma \sim \mathcal{IG}(1, \rho_\sigma), \tag{11}$$
$$\boldsymbol{\beta}_i|\delta^2 \sim \mathcal{N}(\mathbf{0}, \sigma^2\delta^2({Z_i}^T Z_i)^{-1}), \tag{12}$$

with $\mathcal{IG}$ being the inverse-gamma distribution, $\mathcal{N}$ being the normal distribution and $\delta^2$ being fixed to $\delta^2 = 50$.

As the dependence of the likelihood with respect to the shape and scale parameters of the transitions functions is non standard, conjugate priors cannot be selected. Since no prior information is available on these parameters, vague priors defined on a support in agreement with the parameter spaces are considered. This leads to the following gamma and normal priors for the scale and shape parameters respectively

$$\lambda_{k,i}|(\nu_{\lambda k}, \rho_{\lambda_k}) \sim \mathcal{G}(\nu_{\lambda k}, \rho_{\lambda_k}), \quad (13)$$
$$\alpha_{k,i}|(\mu_{\alpha k}, \sigma^2_{\alpha k}) \sim \mathcal{N}(\mu_{\alpha k}, \sigma^2_{\alpha k}). \quad (14)$$

with $\mathcal{G}$ being the gamma distribution with the $\nu_{\lambda k}$ being fixed to the deterministic value $\nu_{\lambda k} = 1$ for all $k = 1, \ldots, K-1$. The set of hyperparameter $\phi$ represents the distribution of parameters within a class. The prior distributions of these hyperparameters are called hyperpriors.

### 2.4. Hyperprior distribution

Since no information on these hyperparameters is available a priori, non informative, or sufficiently vague, priors are chosen

$$f(\mu_{\alpha k}, \sigma^2_{\alpha k}, \rho_{\lambda k}) \propto \frac{1}{\sigma^2_{\alpha k}} \frac{1}{\rho_{\lambda k}} \mathbb{I}_{\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}^+}(\mu_{\alpha k}, \sigma^2_{\alpha k}, \rho_{\lambda k}) \quad (15)$$
$$f(\rho_\sigma) \propto \frac{1}{\rho_\sigma} \mathbb{I}_{\mathbb{R}^+}(\rho_\sigma), \quad (16)$$

with $\mathbb{I}_A$ being the indicator function for the set $A$.

### 2.5. Marginalized posterior

The full joint posterior of class-specific parameters $\phi$ and parameters of each signal $\theta_i$ is deduced from the likelihood, the prior distributions and the hyperprior distribution using eq. (5). Then, the parameters $\boldsymbol{\beta}_i, \sigma_i^2$ and hyperparameters $\rho_{\lambda k}, \mu_{\alpha k}, \sigma^2_{\alpha k}$ are integrated out. Then the marginalize posterior distribution is :

$$f(\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \rho_\sigma|\mathcal{X}) \propto \quad (17)$$
$$\frac{f(\rho_\sigma)\prod_{k=1}^{K-1} f((\lambda_{ki})_{i=1,\ldots,N}) f((\alpha_{ki})_{i=1,\ldots,N})}{\prod_{i=1}^{N}\left(\rho_\sigma + \frac{1}{2}\left(x_i^T x_i - \frac{\delta^2}{1+\delta^2} x_i^T {Z_i}^T ({Z_i}^T Z_i)^{-1} Z_i x_i\right)\right)^{\frac{n_i}{2}+1}}$$

with

$$f((\lambda_{ki})_{i=1,\ldots,N}) \propto \left(\sum_{i=1}^{N} \lambda_{ki}\right)^{-N\nu_{\lambda k}} \prod_{i=1}^{N} \lambda_{ki}^{-(\nu_{\lambda k}-1)}, \quad (18)$$
$$f((\alpha_{ki})_{i=1,\ldots,N}) \propto \left(\sum_{i=1}^{N}\left(\alpha_{ki} - \frac{1}{N}\sum_{j=1}^{N}\alpha_{kj}\right)^2\right)^{-\frac{N}{2}}. \quad (19)$$

However analytical estimation of the remaining parameters and hyperparameters is not tractable. In this case, we sample the posterior distribution (5) by using an MCMC algorithm [10].

## 3. MCMC ALGORITHM

The standard Metropolis-Hastings algorithm is used to sample the parameters of the transitions $\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\alpha}$ for each training signal from marginal posterior distribution (17). That is, at the iteration $t$, one of the transitions $k \in [1, K-1]$ from the $i$th signal is selected and a proposal $(\tilde{\tau_k}, \tilde{\lambda_k}, \tilde{\alpha_k})$ is drawn from a proposition distribution $q(\tau, \lambda, \alpha)$. The proposal is accepted with the probability

$$P = \min\left(1, \frac{f(\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\alpha}}|\mathcal{X}) q(\tau^{(t)}, \lambda^{(t)}, \alpha^{(t)})}{f(\boldsymbol{\tau}^{(t)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\alpha}^{(t)}|\mathcal{X}) q(\tilde{\tau}, \tilde{\lambda}, \tilde{\alpha})}\right) \quad (20)$$

The reader is invited to read [8] for a full description of the proposition distribution $q(\tau, \lambda, \alpha)$.

An acceptation-rejection move of the transitions configuration is attempted iteratively for each training signal. A sketch of the overall MCMC algorithm is

- *for each training signal :* $i = 1, \ldots, N$
  - *select one of the transitions* $k \in [1, K-1]$
  - *draw a sample of* $\tilde{\tau_k}, \tilde{\lambda_k}, \tilde{\alpha_k}$ *according to the proposition distribution* $q(\tau, \lambda, \alpha)$.
  - *accept the proposal with the probability* $P$ *defined in* (20)
  - *sample* $\sigma_i^2$ *according to his marginal posterior distribution* $\sigma_i^2|\boldsymbol{\eta}_i, \rho_\sigma \sim \mathcal{IG}\left(\frac{n_i}{2}+1, \rho_\sigma + \frac{1}{2}\left(x_i^T x_i - \frac{\delta^2}{1+\delta^2}\right)\right)$
- *sample* $\rho_\sigma$ *according to his marginal posterior distribution* $\rho_\sigma \sim \mathcal{G}\left(N, \left[\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right]^{-1}\right)$

This algorithm generates a Markov chain of parameters $(\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ asymptotically distributed according to their marginalized posterior $f(\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\alpha}|\mathcal{X})$ (5). It is of note that, even if the hyperparameters $\boldsymbol{\phi}$ have been integrated out, it is straightforward to derive their conditional posterior from the full joint posterior (5) thanks to the choice of conjugate hyperpriors. Then, the hyperparameters $\boldsymbol{\phi}$ can be sampled from their conditional distribution in an additional Gibbs move within the MCMC algorithm. Furthermore this distribution, denoted for each $c \in \mathcal{C}$ as

$$\boldsymbol{\phi}|c \sim f_l(\boldsymbol{\phi}|\mathcal{X}_c), \quad (21)$$

is now of particular interest since it summerized the information provided by all the training set of signals for a given class $c$. This is the result of the learning step for our hierarchical Bayesian learning model.

### 3.1. Bayes factor computation

To evaluate the capacity of our method to discriminate efficiently two classes of transient, we estimate the Bayes factor $b_{12}$ [11] of class $c_1$ against class $c_2$ which is defined for a signal $x$ as

$$b_{12} = \frac{f(c_1|x)}{f(c_2|x)} = \frac{\int\int f(\mathbf{x}|\theta) f(\boldsymbol{\theta}|\boldsymbol{\phi}) f_l(\boldsymbol{\phi}|\mathcal{X}_{c_1}) d\boldsymbol{\theta} d\boldsymbol{\phi}}{\int\int f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{\phi}) f_l(\boldsymbol{\phi}|\mathcal{X}_{c_2}) d\boldsymbol{\theta} d\boldsymbol{\phi}}. \quad (22)$$

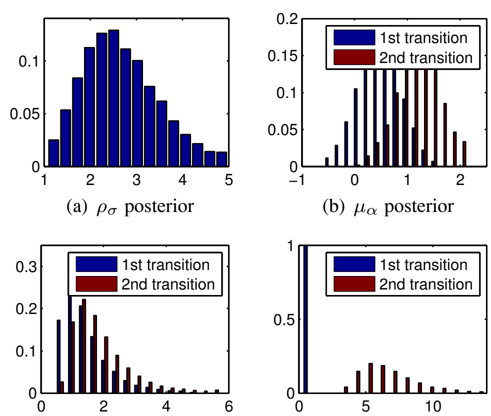
Since the integrals over the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are not tractable, we use Monte-Carlo integration to compute the Bayes factor [12]. More precisely, we sample the posterior distribution of the model (7) over a joint space created by a class indicator $c \in \mathcal{C}$, the parameters and the hyperparameters of the hierarchical Bayesian model. Then the MCMC estimate of the Bayes factor $\hat{b}_{12}$ is obtained as the ratio of the number of occurences of $c = c_1$ against the number of occurences of $c = c_2$ within the samples of the class indicator $c$. Note that for each test signal $x$ the parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{\phi}$ are re-sampled together with the class indicator $c$ of the test signal. The hyperparameters $\boldsymbol{\phi}$ are sampled according to the distribution $f_l(\boldsymbol{\phi}|\mathcal{X}_c)$ learned previously, whereas the parameters $\boldsymbol{\theta}$ are drawn according to their conditionnal distribution $f(\boldsymbol{\theta}|\boldsymbol{\phi})$. The full description of the MCMC algorithm used for the sampling of $c$ is out of the scope of this communication (see [12] for more details).

## 4. RESULTS

The hierarchical Bayesian learning framework introduced in this work has been used in the context of nonintrusive load monitoring [5, 4, 3], more precisely for the identification of electrical transient produced by one appliance being turned on [6]. This issue is quite challenging and, to our knowledge, their is no standard method or public dataset of such transient signals so that we can give comparative results to assess the performance of the method introduced in this work. Moreover, using the smooth regression model to extract some features from the signal yields a trans-dimensional space: each class is not described with the same number of features. Unfortunately, standard supervised learning methods [13] can not deal directly with such kind of a feature space.

The method has been tested over a database of real-world electrical transient, provided by EDF R&D. The transients are generated by $N_C = 2$ classes of appliances, $c_1 =$ "vacuum cleaner" and $c_2 =$"refrigerator", the number of signals being $N_{c_1} = 18$ and $N_{c_2} = 36$ for each class respectively. The database has been split in half to form a learning set and a test set of signals. The number of component $K$ is fixed for each class, using prior work [7, 8]: $K_1 = 3$ for the class $c_1$ and $K_2 = 4$ for the class $c_2$. The MCMC algorithm has been applied to generate samples from the joint posterior distribution of parameters and hyperparameters in order to perform the feature extraction and the learning of each class. The first $2 \times 10^3$ iterations which corresponds to the burn-in period have been thrown away. The next $3 \times 10^3$ iterations are used to estimate the posterior distribution $f_l(\phi|\mathcal{X}_c)$ of the hyperparameters over the training signals. These posteriors are depicted in Figure 2 for the class $c_1$. Similar results are obtained for the class $c_2$ but are not displayed for brevity reasons. It is interesting to note that these distributions are quite regular although they summarize all the training set of signals.

Within each class, two different signals with their curve fitting estimates derived from the smooth transition model are also depicted in Figures 3 and 4 for classes $c_1$ and $c_2$ respectively. This figures show that the feature parameters $\boldsymbol{\theta}$ can take quite different values from one signal to another within a same class. This emphasizes their variability within the class and thus the interest of the proposed hierarchical Bayesian approach where the learning is performed on the hyperparameters $\boldsymbol{\phi}$ rather than directly on the parameters $\boldsymbol{\theta}$.

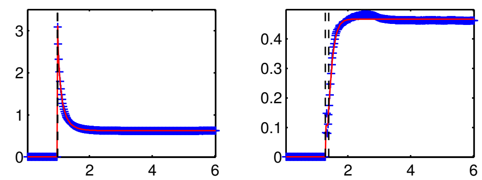



(a) $\rho_\sigma$ posterior (b) $\mu_\alpha$ posterior

(c) $\sigma_\alpha^2$ posterior (d) $\rho_\lambda$ posterior

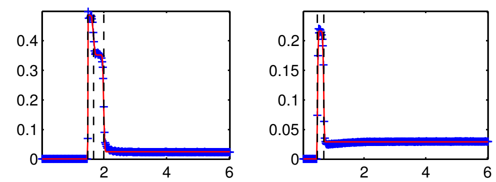
**Fig. 2**: Hyperparameter learned distribution $f_l(\boldsymbol{\phi}|\mathcal{X}_{c_1})$ for $c_1$

Finally, to assess the ability of this method to discriminate the two classes of transients, the MCMC estimate of the Bayes factor $\hat{b}_{12}$ has been computed for each test signal. Table 1 reports these estimates for the 8 test signals of the class $c_1$, numbered from 1 to 8, and for the 8 first test signals of the class $c_2$. The value $\hat{b}_{12}$ for the test signals of the class $c_2$ which are not shown in table 1 is 0. It means that, the convergence of the MCMC sampler being reached, the label parameter $c$ takes only the value $c_2$. Within the $N_C = 2$ classes, the two signals depicted in Figures 3 ($c_1$ class) and 4 ($c_2$ class) correspond to the greater and the lesser value of the Bayes factor estimates, whose values are indicated in the corresponding subcaption. The $8^{th}$ signal of class $c_2$, shown in figure 4(b) is classified in the wrong class, $\hat{b}_{12}$ being greater than one. The $6^{th}$ signal of class $c_1$, shown in figure 3(b) is well classified though $\hat{b}_{12}$ is barely greater than one. Nevertheless, the results suggest that the proposed model is discriminant with respect to this two classes. Indeed, except for the two signals mentioned, every test signals are classified with a substantial evidence in favor of the correct model ($B > 3.2$ according to Kass's interpretation of the Bayes factor [11]).



(a) $2^{nd}$ test signal of class $c_1$ $\hat{b}_{12} = 58$ (b) $6^{th}$ test signal of class $c_1$ $\hat{b}_{12} = 1.6$

**Fig. 3**: Two examples of the active power (kW) versus time (s) for vacuum cleaner transient (blue) with their smooth transition regression fit (red) and the position of the transitions (black)



(a) $1^{st}$ test signal of class $c_2$ $\hat{b}_{12} = 0$ (b) $8^{th}$ test signal of class $c_2$ $\hat{b}_{12} = 5.5$

**Fig. 4**: Two examples of the active power (kW) versus time (s) for refrigerator transient (blue) with their smooth transition regression fit (red) and the position of the transitions (black)

| signal | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $c_1$ | 30 | **58** | 14 | 13 | 37 | **1.6** | 5.0 | 14 |
| $c_2$ | **0** | 0 | 0 | 0 | 0 | 0 | 0 | **5.5** |

**Table 1**: MCMC estimates of the Bayes factor $\hat{b}_{12}$ for each test signal (columns) of the two classes $c_1$ and $c_2$ (rows). The bold values correspond to the test signals shown in Figures 3 and 4.

## 5. CONCLUSION

A hierarchical Bayesian method has been introduced to perform jointly the feature extraction and the class learning over a database of electrical transients. With this methodology, the variability of the signal features within a class is summarized by the posterior distribution thanks to a few number of hyperparameters. The results obtained over a few number of test signals suggest that this method is able to separate efficiently some appliance classes without overlearning the training set of signals. However, to be fully convincing, the evaluation of the classification performance has now to be conducted over larger data sets, and over other classes of transients. The extension of the method to detect and classify multiple transients in a single signal is also under investigation.

## 6. REFERENCES


[1] Pedro Domingos and Michael Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss,” *Machine Learning*, vol. 29, pp. 103137, 1997.

[2] Manuel Davy, Christian Doncarli, and Jean-Yves Tourneret, “Classification of Chirp Signals Using Hierarchical Bayesian Learning and MCMC Methods,” *IEEE Trans. On Signal Processing*, vol. 50, no. 2, pp. 377–388, 2002.

[3] J. Liang, S. Ng, G. Kendall, and J. Cheng, “Load signature study - part i: Basic concept, structure, and methodology,” *Power Delivery, IEEE Transactions on*, vol. 25, no. 2, pp. 551 –560, april 2010.

[4] M. Zeifman and K. Roth, “Nonintrusive appliance load monitoring: Review and outlook,” *IEEE Trans. on Consumer Electronics*, vol. 57, no. 1, pp. 76 –84, february 2011.

[5] George W. Hart, “Nonintrusive Appliance Load Monitoring,” *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[6] Steven Leeb, *A conjoint pattern recognition approach to nonintrusive load monitoring*, Ph.D. thesis, Massachussets Institute of Technology, 1993.

[7] Matthieu Sanquer, Florent Chatelain, Mabrouka El Guedri, and Nadine Martin, “A reversible jump mcmc algorithm for bayesian curve fitting using smooth transition regression models.,” in *Proceedings of ICASSP 2011*, Prague, Czech Republic, May 2011, pp. 3960–3963.

[8] Matthieu Sanquer, Florent Chatelain, Mabrouka El Guedri, and Nadine Martin, “A smooth transition model for multipleregime time series,” *IEEE Trans. on Signal Processing*, in press.

[9] Charles Kemp, Amy Pefors, and Joshua B. Tenenbaum, “Learning overhypotheses with hierarchical bayesian models,” *Developmental Science*, vol. 10, no. 3, pp. 307–321, 2007.

[10] Christian Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer-Verlag, 2nd edition, 2001.

[11] Robert E. Kass and Adrian E. Raftery, “Bayes factors,” *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 791, 1995.

[12] Cong Han and Bradley P Carlin, “Markov chain monte carlo methods for computing bayes factors,” *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1122–1132, 2001.

[13] C.M. Bishop, *Pattern recognition and machine learning*, Information Science and Statistics. Springer-Verlag, 2nd edition, 2006.